ALEXANDRU IOAN CUZA
UNIVERSITY of IAŞI

Proceedings of the Eighth
Global WordNet Conference

Editors:

Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, Piek Vossen

Bucharest, Romania, January 27-30, 2016

ISBN 978-973-0-20728-6

Preface

This eighth meeting of the international Wordnet community coincides with the 15th anniversary of the Global WordNet Association and the 30th anniversary of the Princeton WordNet. We are delighted to welcome old and new colleagues from many countries and four continents who construct wordnets, ontologies and related tools, as well as colleagues who apply such resources in a wide range of Natural Language Applications or pursue research in lexical semantics.

The number of wordnets has risen to over 150 and includes – besides all the major world languages – many less-studied languages such as Albanian and Nepali. Wordnets have become a principal tool in computational linguistics and NLP, and *wordnet*, *SemCor* and *synset* have entered the language as common nouns. Coming together and sharing some of the results of our work is an important part of the larger collaborative effort to better understand both universal and particular properties of human languages.

Many people have donated their time and effort to make this meeting possible: the review committee, the local organizers and their helpers (Eric Curea, Maria Mitrofan, Elena Irimia), our sponsors (PIM, QATAR Airways, Oxford University Press), EasyChair and our host, the Romanian Academy. Above all, thanks go to you, the contributors, for traveling to Bucharest to present your work, listen and discuss.

January, 2016 — Christiane Fellbaum, Corina Forăscu,

Bucharest — Verginica Mititelu, Piek Vossen

Table of contents

Program Committee

Eneko Agirre, University of the Basque Country, Spain

Eduard Barbu, Translated.net, Italy

Francis Bond, Nanyang Technological University, Singapore

Sonja Bosch, University of South Africa, South Africa

Alexandru Ceaușu, Euroscript Luxembourg S.à r.l., Luxembourg

Dan Cristea, Alexandru Ioan Cuza University of Iași, Romania

Agata Cybulska, VU University Amsterdam / Oracle Corporation, the Netherlands

Tsvetana Dimitrova, Institute for Bulgarian Language, Bulgaria

Marieke van Erp, VU University Amsterdam, the Netherlands

Christiane Fellbaum, Princeton University, USA

Darja Fiser, University of Ljubljana, Slovenia

Antske Fokkens, VU University Amsterdam, the Netherlands

Corina Forăscu, Alexandru Ioan Cuza University of Iași & RACAI, Romania

Ales Horak, Masaryk University, Czech Republic

Florentina Hristea, University of Bucharest, Romania

Shu-Kai Hsieh, Graduate Institute of Linguistics at National Taiwan University, Taiwan

Radu Ion, Microsoft, Ireland

Hitoshi Isahara, Toyohashi University of Technology, Japan

Ruben Izquierdo Bevia, VU University Amsterdam, the Netherlands

Kaarel Kaljurand, Nuance Communications, Austria

Kyoko Kanzaki, Toyohashi University of Technology, Japan

Svetla Koeva, Institute for Bulgarian Language, Bulgaria

Cvetana Krstev, University of Belgrade, Serbia

Margit Langemets, Institute of the Estonian Language, Estonia

Bernardo Magnini, Fondazione Bruno Kessler, Italy

Verginica Mititelu, RACAI, Romania

Sanni Nimb, Society for Danish Language and Literature, Denmark

Kemal Oflazer, Carnegie Mellon University, Qatar

Heili Orav, University of Tartu, Estonia

Karel Pala, Masaryk University, Czech Republic

Adam Pease, IPsoft, USA

Bolette Pedersen, University of Copenhagen, Denmark

Ted Pedersen, University of Minnesota, USA

Maciej Piasecki, Wroclaw University of Technology, Poland

Alexandre Rademaker, IBM Research, FGV/EMAp Brazil

German Rigau, University of the Basque Country, Spain

Horacio Rodriguez, Universitat Politecnica de Catalunya, Spain

Shikhar Kr. Sarma, Gauhati University, India

Roxane Segers, VU University Amsterdam, the Netherlands

Virach Sornlertlamvanich, SIIT, Thammasart University, Thailand

Dan Ştefănescu, Vantage Labs, USA

Dan Tufiș RACAI, Romania

Gloria Vasquez, Lleida University, Spain

Zygmunt Vetulani, Adam Mickiewicz University, Poland

Piek Vossen, VU University Amsterdam, the Netherlands

Additional Reviewers

Anna Feltracco

Filip Ilievski

Vojtech Kovar

Egoitz Laparra

Simone Magnolini

Zuzana Neverilova

Adam Rambousek

Conference Chairs

Christiane Fellbaum, Princeton University, USA

Piek Vossen, VU University Amsterdam, the Netherlands

Organising Committee

Verginica Mititelu, RACAI, Romania

Corina Forăscu, Alexandru Ioan Cuza University of Iași & RACAI, Romania

The Awful German Language: How to cope with the Semantics of Nominal Compounds in GermaNet and in Natural Language Processing

Erhard Hinrichs
University of Tübingen
Tübingen, Germany
erhard.hinrichs@uni-tuebingen.de

Abstract

The title for my presentation borrows from Mark Twain's well-known 1880 essay "The Awful German Language", where Twain cites pervasive nominal compounding in German as one of the pieces of evidence for the "awfulness" of the language. Two much cited examples of noun compounds that are included in the Duden dictionary of German are *Kraftfahrzeughaftpflichtversicherung* (‘motor car liability insurance’) and *Donaudampfschifffahrtsgesellschaft* (‘Danube steamboat shipping company’). Any dictionary of German, including the German word net GermaNet, has to offer an account of such compound words. Currently, GermaNet contains more than 55,000 nominal compounds. As the coverage of nouns in GermaNet is extended, new noun entries are almost always compounds.

In this talk I will present an account of how to model nominal compounds in GermaNet with particular focus on the semantic relations that hold between the constituents of a compound, e.g., the WHOLE-PART relation in the case of *Roboterarm* ('robot arm') or the LOCATION relation in the case of *Berghütte* ('mountain hut'). This account, developed jointly with Reinhild Barkey, Corina Dima, Verena Henrich, Christina Hoppermann, and Heike Telljohann, borrows heavily from previous research on semantic relations in theoretical linguistics, psycholinguistics, and computational linguistics.

The second part of the talk will focus on using the semantic modelling of nominal compounds in a word net for the automatic classification of semantic relations for (novel) compound words. Here, I will present the results of recent collaborative work with Corina Dima and Daniil Sorokin, using machine learning techniques such as support vector machines as well as deep neural network classifiers and a variety of publicly available word-embeddings, which have been developed in the framework of distributional semantics.

Adverbs in the Sanskrit Wordnet

Tanuja P. Ajotikar
Dept. of South Asian Studies
Harvard University, Cambridge, MA
The Sanskrit Library
tanuja@sanskritlibrary.org

Malhar Kulkarni
Indian Institute of Technology Mumbai
Powai, Mumbai, India
malharku@gmail.com

Abstract

The wordnet contains part-of-speech categories such as noun, verb, adjective and adverb. In Sanskrit, there is no formal distinction among nouns, adjectives and adverbs. This poses the question, is an adverb a separate category in Sanskrit? If not, then how do we accommodate it in a lexical resource? To investigate the issue, we attempt to study the complex nature of adverbs in Sanskrit and the policies adopted by Sanskrit lexicographers that would guide us in storing them in the Sanskrit wordnet.

1 Introduction

An adverb is an open-class lexical category that modifies the meaning of verbs, adjectives (including numbers) and other adverbs, but not nouns.[1] It can also modify a phrase or a clause. The category of adverb indicates: (a) manner, (b) time, (c) place, (d) cause, and (e) answers to the questions how, where, when and how much.

Fellbaum (1998, p. 61) describes adverbs as a heterogeneous group in which not only adverbs derived from adjectives are included but also phrases used adverbially. Some of these phrases are included in WordNet. These phrases are mainly frozen phrases that are used widely.

In this paper we discuss those adverbs which modify verbs, and how modern Sanskrit lexicography deals with them. Kulkarni et al. (2011) briefly discussed the issues regarding adverbs in the Sanskrit wordnet. We focused primarily on how modern Sanskrit lexicographers have dealt with them. The study of their methodology can guide us in forming a policy for representing adverbs in the Sanskrit wordnet.

[1] http://www.odlt.org

2 Adverbs in Sanskrit

The Sanskrit grammatical tradition does not divide words into many categories. It divides words into two divisions: words that take nominal affixes and words that take verbal affixes. The words in the second division are verbs. Those in the first division are nouns, adjectives, adverbs, particles, etc., i.e., non-verbs. This is because unlike languages like English, Sanskrit does not have distinct forms for each part of speech. One cannot categorize a word merely by looking at its form. This is why there is not a formal category for adjective or adverb in traditional Sanskrit grammar. There is no equivalent term in Sanskrit for adjective or adverb in the modern sense (See Joshi (1967), Gombrich (1979)). Sanskrit can be analyzed under word classes other than noun and verb. Bhat (1991) observes that adjectives in Sankrit form a sub-group of nouns. Likewise, adverbs, except indiclinables, form a subgroup of nouns. Attempts were first made in the 19th century to describe Sanskrit using various word classes. Monier-Williams (1846), Wilson (1841), Speijer (1886), Whitney (1879) and Macdonell (1927) discuss adverbs in Sanskrit.[2] A summary of the description of adverbs given by these scholars is as follows:

- The non-derived words listed by traditional grammar and termed 'indeclinable' are used as adverbs, e.g., *uccaiḥ* 'high,' *nīcaiḥ* 'below,' *ārāt* 'distant,' etc.

- Compounds, like *avyayībhāva*, are used as adverbs, e.g., *yathāśakti* 'according to power or ability.'[3] Some of the *bahuvrīhi*

[2] We refer to these works because Macdonell, Wilson and Monier-Williams compiled bilingual dictionaries. We refer to their works to study how far they follow their description in their dictionaries.

[3] In the sentence *yathāśakti dātavyam* 'you may give according to your ability,' the compound *yathāśakti* modifies the action. Hence, it is an adverb.

compounds are also used as adverbs, e.g., *keśākeśi* 'hair to hair' (i.e., head to head).[4]

- Words formed by adding certain affixes, such as *śas*, *dhā*, etc., are used as adverbs. The affix *śas* is added after a nominal base or a number word in the sense of *vīpsā* 'repetition.' Words like *śataśaḥ* 'hundred times' are formed by adding this affix. The affix *dhā* is added after a number word in the sense of *vidhā* 'division or part.' Words like *dvidhā* 'twofold' or *tridhā* 'threefold' are formed by adding this affix. Words formed by adding certain affixes after a nominal base are considered indeclinable by the traditional grammarians.

- The accusative, instrumental, ablative and locative cases of a noun or an adjective are used as adverbs, e.g., *mandam* 'slowly,' *vegena* 'hastily,' *javāt* 'speedily,' *sannidhau* 'near.'

This summary shows that we can classify adverbs in Sanskrit in three main groups: words that are unanalyzable in parts, such as a base and an affix; words that formed by secondary derivation, such as adding an affix or forming a compound; and words that have an adverbial sense but belong to a class of words which are not adverbs, for example, the accusative or instrumental case of any noun or adjective. A morphological analysis of these words would categorize them under nouns because they are formed by adding the same affixes that are added after a noun, even though their function differs. In other words, qualifying a verb or an adjective in Sanskrit does not require the use of a distinct morphological form. The difficulty in dealing with adverbs in Sanskrit arises only if we have a form-based idea of word classes. It becomes lexically opaque to judge a category simply by looking at the form. The adverb is a functional category in Sanskrit, not formal one. Hence, adverbs pose a problem in Sanskrit lexicography because they lack a distinguishing form and they are functional.

[4]In the sentence *te keśākeśi yuddhyante* 'they battled hair to hair', the compound *keśākeśi* also modifies the action so it is an adverb.

2.1 The importance of part-of-speech categories in lexical entries

The nature of adverbs in Sanskrit is complex, so it is a matter of discussion what the exact relationship is between a part-of-speech category and a dictionary. Lexemes do not occur in isolation. They form part of a phrase or sentence. In this way, the role of a lexicon is to structure sentences. Lexemes form an important part, as they determine the syntactic structure of sentences. Each and every lexeme plays a certain role in a sentence. The morphological and syntactic behavior of a lexeme determines its class. This class is designated as a part-of-speech category. It is also called a word class, lexical class or lexical category. Noun, verb, adjective and adverb are major word classes. Thus, a lexicon, which is an inventory of lexemes, contains these major word classes to denote the morphological and syntactic behavior of the lexemes listed in it. The morphological and syntactic behavior of a language decides what kind of information a lexicon should contain.

In Sanskrit, where there is no formal distinction between adverb and noun (with the exception of indeclinables), the following question arises: Should an adverb be a separate category in a Sanskrit lexicon? It would be interesting to study the policy adopted in the available lexical resources of Sanskrit, which range from 1819 C.E. to 1981 C.E, to answer this question. The examples below were given by Gombrich (1979):

- *atra* 'here'
- *ciram* 'for a long time'
- *javena* 'speedily'
- *tūsṇīm* 'silently'
- *vividhaprakāram* 'variedly'
- *śīghram* 'quickly'

Gombrich observes that the first, second and fourth examples are found in the traditional grammar. However, the rest of the adverbs are not recognized as such. His article is important because he has thoroughly discussed the position of traditional Sanskrit grammarians on adverbs, and given an historical account of the concept of adverb. He points out that words that function as adverbs are not grammatically analyzed; instead, they are simply listed by traditional grammarians. There is

no process of deriving adverbs from adjectives. Hence, *ciram*, *cirāt*, *cirasya* 'for a long time,'[5] which might be derived from the same word, are listed separately. Their status is independent. This forms a base for entering these words in a lexicon as separate lexemes.

2.2 Adverbs in the list above and the treatment they receive in dictionaries

We consulted eighteen dictionaries of Sanskrit to study the treatment given to the above-mentioned adverbs. Two of these eighteen dictionaries are monolingual and the rest are bilingual. Among those bilingual dictionaries, (Goldstücker (1856) and Ghatge (1981)) are not complete. These eighteen dictionaries are listed chronologically below:

- Radhakanatdeva, (Monolingual), 1819–1858.
- Wilson H. H., Sanskrit–English, 1832.
- Yates W., Sanskrit–English, 1846.
- Bopp F., Sanskrit–French, 1847.
- Böhtlingk, O. and Roth R., Sanskrit–German, 1855–1875.
- Goldstükar T., Sanskrit–English, 1856.
- Benfey, T., Sanskrit–English, 1866.
- Burnouf É., Sanskrit–French, 1866.
- Böhtlingk, O., Sanskrit–German, 1879-1889.
- Monier-Williams M., Sanskrit–English, 1872.
- Bhattacharya T., (Monolingual), 1873.
- Cappeller, C., Sanskrit–German, 1887.
- Apte V. S., Sanskrit-English, 1890.
- Cappeller, C., Sanskrit–English, 1891.
- Macdonell A. A. Sanskrit–English 1893
- Monier-Williams M., Leumann, and Cappeller, Sanskrit–English, 1899.
- Stchoupak, N., Nitti, L. and Renou L., Sanskrit–French, 1932.
- Ghatge, A. M., Sanskrit–English (Encyclopedic dictionary on historical principles), 1981.

Let us analyze how the above-listed adverbs are treated in these Sanskrit dictionaries.

2.2.1 *atra*

Atra, which means 'here,' is an indeclinable according to the traditional Sanskrit grammarians, whereas its treatment in dictionaries varies. It is derived from the pronoun *etad* 'this' by adding the affix *tral*. It is termed indeclinable by the rule *taddhitaścāsarvavibhakatiḥ* A.1.1.38.[6] There are more such words formed by adding the affix *tral*, such as, *tatra* 'there,' *kutra* 'where,' etc. We will discuss only *atra* in detail in this paper.

Derivation of *atra*

etad tral
a tra (*etad* is replaced by *a*)
atra

All the lexicographers treat it as an adverb except Monier-Williams (1872), Monier-Williams, Leumann, and Cappeller (1899), Apte (1890) and Goldstücker (1856). These lexicographers consider it indeclinable, as does Radhakantadeva (1819–1858) and Bhattacharya (1873–1884). Cappeller (1887) does not assign any category to it, but describes it morphologically. We can observe that the lexicographers who use the term indeclinable as a part-of-speech category follow traditional grammar. Other lexicographers, though aware of this analysis do not follow the traditional grammar.

2.2.2 *tūṣṇīm*

The traditional Sanskrit grammarians list words which are non-derivable. That list gets the status of indeclinable. The word under discussion is a member of this list. *Tūṣṇīm*, which means 'silently,' is categorized as an indeclinable. Radhakantadeva (1819–1858), Wilson (1832), Monier-Williams (1872), Monier-Williams, Leumann, and Cappeller (1899), Bhattacharya (1873–1884) and Apte (1890) follow the tradition and indicate its category as indeclinable. The rest of the lexicographers assign it to the category of adverb. Here also we can observe that Radhakantadeva (1819–1858) and Bhattacharya (1873–1884) are consistent in following the traditional grammar. Those lexicographers who label it an adverb are

[5] These forms resemble the accusative singular, ablative singular and genitive singular, respectively, of a nominal base which ends in short *a*.

[6] This is a rule in Pāṇini's *Aṣṭādhyāyī*. It assigns the term *avyaya* 'indeclinable' to those words which end in the affixes termed *taddhita*, and are not used in all cases.

also consistent in analyzing indeclinables listed by the traditional grammarians as adverbs.

2.2.3 *ciram*

Ciram means 'for a long time.' It can be analyzed as the accusative case of *cira*. The traditional grammarians of Sanskrit treat it as an indeclinable, as they include it in the list of nonderivable words. They do not analyze it as a nominal form, even though lexicographers vary in their analysis. Macdonell (1893), Yates (1846), Bopp (1847), Cappeller (1887), Cappeller (1891) assign an adverb category to it. Wilson (1832), Monier-Williams (1872) and Monier-Williams, Leumann, and Cappeller (1899) treat it as an indeclinable. Apte (1890), Böhtlingk and Roth (1855–1875), Benfey (1866) and Burnouf (1866) describe its adverbial role, but do not assign an adverb category to it.

Macdonell (1893), Böhtlingk (1879–1889), Monier-Williams (1872), Monier-Williams, Leumann, and Cappeller (1899), Benfey (1866) and Burnouf (1866) list it under *cira*. Thus, they assume that all forms of *cira* are derivable–forms such as *ciram* (formally identical to the accusative singular of a nominal base which ends in short *a*); *cireṇa* (formally identical to the instrumental singular of a nominal base which ends in short *a*); *cirāya* (formally identical to the dative singular of a nominal base which ends in short *a*); *cirāt* (formally identical to the ablative singular of a nominal base which ends in short *a*); and *cirasya* (formally identical to the genitive singular of a nominal base which ends in short *a*). These are given separately by Radhakantadeva (1819–1858) and Bhattacharya (1873–1884), who treat these forms as indeclinable. This evidence is sufficient to say that *ciram*, *cireṇa* and *cirāya*, *cirāt*, *cirasya* are different words according to them–not declensions of *cira*, which is contrary to the western lexicographers' treatment. Thus, western lexicographers do not follow the traditional grammar in this case. Radhakantadeva (1819–1858) and Bhattacharya (1873–1884) follow the tradition and maintain their independent status.

2.2.4 *javena*

This is the instrumental singular of *java* 'speed.' None of the lexica records this form as an adverb, but its ablative form is assigned an adverb category by Cappeller (1887). Böhtlingk (1879–1889) notes its ablative form, and gives its meaning as *eiligst* (haste), *alsbald* (soon). Stchoupak, Nitti, and Renou (1932) note its accusative and ablative forms and give its meaning as *rapidement*, *vivement* (quickly, sharply). They do not assign any category to it. But the meanings given certainly reflect its adverbial use. The instrumental case of *java* 'speed' does not occur in dictionaries and hence is not recognized as an adverb. Accordingly, words like *raṃhasā, vegena, vegāt* 'speedily' should be recognized as adverbs since they are instrumental and ablative singular forms of *raṃhas* and *vega* 'speed' respectively. However, these also do not occur in dictionaries.

2.2.5 *vividhaprakāram*

The word *vividhaprakāram* 'variedly' is not found in any of the dictionaries. It is the accusative singular form of *vividhaprakāra* which is a *karmadhāraya* (endocentric) compound.

2.2.6 *śīghram*

The word *śīghram* 'quickly' is the nominative and accusative singular form of *śīghra* 'quick.' In the present context it is the accusative singular form. All the lexicographers consider it an adverb, except for Monier-Williams (1872), Monier-Williams, Leumann, and Cappeller (1899) and Apte (1890) who consider it an indeclinable. Stchoupak, Nitti, and Renou (1932) do not consider *śīghra* an indeclinable or an adverb but rather an adjective. Burnouf (1866) mentions its gender and accusative form, but does not assign any category. Yates (1846) mentions its neuter gender by giving the nominative form, as well as assigns an adverb category to it. All of these lexicographers have analyzed it as derived from *śīghra* which is an adjective. Monier-Williams (1872) and Monier-Williams, Leumann, and Cappeller (1899) do not use the adjective category. Instead, they use the abbreviation *mfn* (masculine, feminine and neuter) to show that the word is used in all genders. Wilson (1832) and Cappeller (1887) record *śīghra* as a neuter word; thus, they consider it a noun. Radhakantadeva (1819–1858) and Bhattacharya (1873–1884) list *śīghra* and indicate its gender as neuter. Then they mention its adjectival use through the term *tadvati tri* (i.e., having that (speed)). It can be inferred that they consider *śīghra* a noun since they note its gender, but do not mention its adverbial use. All of the lexicographers, except for Radhakantadeva (1819–1858) and Bhattacharya (1873–1884), take into consid-

eration the adverbial *śīghram*, but do not consider it an independent lexeme.

2.2.7 *yathāśakti*

The word *yathāśakti* 'according to one's power or ability' is an *avyayībhāva* compound. Radhakantadeva (1819–1858), Bhattacharya (1873–1884), Monier-Williams (1872), Monier-Williams, Leumann, and Cappeller (1899) and Apte (1890) give its category as indeclinable following the traditional analysis. Benfey (1866), Bopp (1847), Macdonell (1893) do not list this word, even though other *avyayībhāva* compounds are assigned to the adverb category.

3 Observations on the basis of the previous section

This investigation gives rise to certain observations. We may say that *tūṣnīm*, *atra* and *yathāśakti* are formal adverbs.

Ciram can be derived from *cira*, but its other forms like *cireṇa*, *cirāya*, *cirāt*, *cirasya* are also used as adverbs. So whether to analyze it formally or functionally is a matter of debate. Radhakantadeva (1819–1858) and Bhattacharya (1873–1884) treat all these forms as synonyms on the basis of the *Amarakośa* (a 6th century A.D. Sanskrit thesaurus), and do not mention them under one lexeme, i.e., *cira*. Hence, we may say that it is also a formal adverb on the basis of the monolingual dictionaries.

Śīghram is also treated as a form of *śīghra*, which is an adjective according to western lexicographers. Hence, we may say that it is an adverbial not an adverb, whereas Radhakantadeva (1819–1858) and Bhattacharya (1873–1884) treat it as a noun. They also take into consideration its use as an adjective. If we follow modern western lexicographers, then *śīghram* is an adverbial. If we follow monolingual dictionaries, then it is neither an adverb nor an adverbial. In this way, it is difficult to decide the exact criterion by which to label its category.

Javena is an adverbial. None of the lexica assign it to the category of adverb. Cappeller (1887), it should be noted, cites its adverbial use in the ablative case. Interestingly, Bhattacharya (1873–1884) cites an example under *java* where it occurs in the instrumental case, but he is silent about its part-of-speech category. The one example given by Gombrich that is not found in any of these dictionaries is *vividhaprakāram*.

Table 1: **The number of completed synsets for each part-of-speech category in Sanskrit wordnet**

Nouns	27563
Verbs	1247
Adjectives	4031
Adverbs	264
Total	33117

On the basis of this investigation, we may say that there is no single policy adopted by modern Sanskrit lexicographers to record adverbs. Even after this investigation, doubts regarding the category of certain forms remain.

4 Adverbs in Sanskrit wordnet

These lexica are in print form and written purely from the point of view of human use. Hence, a single entry contains a lot of information. Multiple functions of a word can be listed under one entry. But when a lexical resource is built for machines, then this strategy cannot be adopted. Multiple functions of a word are stored separately. In other words, there is more than one entry for the same word based on its meanings and functions, whatever information is necessary to make it explicit for a machine.

The Sanskrit wordnet is being developed by following the expansion approach, and its source is the Hindi wordnet. It is a well known fact that Sanskrit is a morphologically rich language. So a proper policy should be adopted for part-of-speech categories that take into account their nature. A long and rich tradition of Sanskrit grammar guides us in this regard. Following the tradition, we accept the verbal roots given in the list of verbal roots known as the *dhātupāṭha* after removing their metalinguistic features. For nouns, we enter the nominative singular form, and we enter the base forms of adjectives.

Given the discussion above, should the Sanskrit worndet have a separate category called 'indeclinable' which links to the relevant synsets in the Hindi wordnet, or should it just retain the category of adverb? A wordnet recognizes a separate category for function words even though none are actually included in it. Indeclinables in Sanskrit consist of function words as well as content words. Hence it is difficult to adopt the category 'indeclin-

able' in the Sanskrit wordnet, which may harm the basic principle of a wordnet. To avoid this, we retain the adverb category. Thus, we follow western lexicographers who assign the adverb category to those words which are indeclinables and which can be termed *formal* adverbs. These words appear without any change in the Sanskrit wordnet, e.g., *atra* 'here,' *iha* 'here,' etc. They appear in the same synset (id 2647).[7] The compound *yathāśakti* is also entered without any change.[8]

The issue of adverbials remains to be solved. How do we store the oblique cases of nouns or adjectives that are used as adverbs? If they are stored in their base forms, their role as an adverb is restricted. Not all of the forms are used as adverbs. The Sanskrit wordnet resolves this issue by storing the declined forms. For example, *śīghram*, *śīghreṇa*, *javena*, *javāt* appear in one synset (id 1922).[9] At the same time, there is a separate entry (id 5118) for *śīghra*.[10] In this way, we may say that the Sanskrit wordnet stores adverbials. We do not claim that this phenomenon is recognized for the first time in the history of Sanskrit lexicography. It is implicit by its representation in the dictionaries. We make it explicit for computational processing so that it will be helpful for an automatic parser of Sanskrit. Such a parser would benefit from a lexical resource that contains both adverbs and adverbials.

5 Adverbs in the Hindi and Sanskrit wordnets

The discussion in the previous sections focuses on adverbs as a part-of-speech category. In this section, we address two issues regarding the linking of synsets of adverbs.

1. It is difficult to link a synset in the source language if it uses an adverb to express what the target language conveys by using pre-verbs that are bound morphemes.

2. According to the policy of the expansion approach, we cannot link a synset whose part-of-speech category in the source language differs from that in the target language. For example, if the source language uses a noun or an adjective, and the target language uses an adverb to convey the same lexical concept, then we cannot link these synsets.

These are cases of language divergence that become apparent when Sanskrit is analyzed in comparison to other languages. Let us take an example for each of the two above–metioned issues.

5.1 Adverbs in Hindi and preverbs in Sanskrit

Hindi Synset id 10819
Gloss: *lauṭakara phira apane sthāna para* 'Returning to his own place again.'
Example: *Mohana kala hi videśa se vāpasa āyā* 'Mohana came back yesterday from abroad.'
Synset: *vāpasa vāpisa* 'back'

Sanskrit uses the preverb and verb combination to convey the meaning 'back.' It does not use an independent word. The preverb *prati* is used with verbs of motion. We cannot store preverbs separately in synsets because they are bound morphemes. So the synset in the Hindi wordnet is not linkable to the Sanskrit wordnet. This aspect of preverbs that conveys adverbial sense becomes apparent when Sanskrit is analyzed in the context of another language, i.e., Hindi.

5.2 Cross part-of-speech category

Hindi Synset id 11374
Gloss: *āṃkhoṃ ke sāmanevālā* 'the one who is in front of eyes.'
Example: *śikṣaka ne chātroṃ ko pratyakṣa ghaṭanā para ādhārita nibaṃdha likhane ko kahā.* 'The teacher asked students to write an essay based on an actual incident.'
Synset: *pratyakṣa sākṣāt anvakṣa aparokṣa samakṣa nayanagochara* 'evident.'

The Sanskrit word *pratyakṣa*, which is an *avyayībhāva* compound, is not an adjective in the sense of 'evident' but an adverb. When this word was borrowed in Hindi, its category changed. So the synset in Hindi is not linkable to the Sanskrit wordnet under the adjective category. Cross part-of-speech category linkage would be a solution for this problem.

6 Adverbs and their relations

There are two kinds of relations, 'derived from' and 'modifies verb,' for adverbs in the Hindi wordnet, and so also in the Sanskrit wordnet. Both of

[7] The source synset in Hindi is *yahāṃ isa jagaha itaḥ ita iha ihāṃ ihavāṃ ĩṃghe īhāṃ yahāṃ*

[8] The source synset in Hindi is id 9882 *yathāśakti, yathāsambhava, bhaarasaka, yathāsādhya, kṣamatānusāra, yathākṣama* 'according to one's power or ability.'

[9] The linked Hindi synset contains more than 30 words such as *jhaṭpaṭ, caṭpaṭ*, etc.

[10] The linked Hindi synset is *tīvra, druta, teja, etc.*

these relations cross the part-of-speech category. The first relation is between a noun and an adverb or between an adjective and an adverb, and the second relation is between a verb and an adverb. The adverbials, such as *vegena*, are easy to link by this relation. In this case, *vega* 'speed' is a noun which is linkable to *vegena* with the relation 'derived from.' The non-derived adverbs such as *uccaiḥ* 'high,' *nīcaiḥ* 'below,' and *śanaiḥ* 'slowly' cannot be linked with any other noun or adjective because they are frozen forms. These non-derived adverbs may not present a complex situation, as there is only one form. The complexities arise with words like *cira* 'for a long time.' If adverbs such as *ciram, cirasya,* etc. are considered as derived from *cira,* then there should be a separate synset in the adjective category. It is hard to form such a separate synset because it is not used as an adjective. If these adverbs are considered non-derived, then they cannot be linked to any other synset with the relation 'derived from.'

The compound *yathāśakti*, for example, is derived from *yathā* and *śakti*. Should it be linked to both of these words? Currently, it is linked only to *śakti*. Thus, it is a matter of concern whether compounds should be linked to one or more of their components. In this way, there is a need for more analysis regarding the relations of adverbs in Sanskrit.

7 Conclusion

From the above discussion, it is clear that adverbs in Sanskrit are formal as well as functional, and that they have not received any uniform treatment in the hands of lexicographers. Formal adverbs are easy to store under the adverb category in the Sanskrit wordnet. The real challenge is with the nominal forms, adverbially used. It is the Sanskrit wordnet's contribution to lexicalize the adverbials, especially the declined forms of nouns and adjectives. The real challenge is to collect all of the possible cases. Currently, the Sanskrit wordnet stores those cases that are available in the lexical sources it uses.

The case of adverbs in Sanskrit reveals the complexity of their nature. Clearly, a lexicon developed for a machine use will need to adopt strategies suitable for its system.

8 Acknowledgement

We thank Mr. Peter Voorlas for his valuable help in editing this paper.

References

Apte, Vaman Shivram. (1890). *The Practical Sanskrit-English Dictionary*. Delhi: Motilal Banarasidas.

Benfey, Theodore (1866). *A Sanskrit-English Dictionary*. London: Longmans, Green, and Co.

Bhat, D. N. S. (1991). *An Introduction to Indian Grammars: Part Three:Adjectives. A report being submitted to The University Grants commission*. .Three.

Bhattacharya, Taranatha Tarkavacaspati (1873–1884). *Vācaspatya Bṛhatsaṃskṛtābhidhāna*. Calcutta: Kavya Prakash Press.

Böhtlingk, Otto von (1879–1889). *Sanskrit Wörtebuch. in Kürzer Fassung*. St. Petersburg: Kaiserlichen Akademie der Wissenschaften.

Böhtlingk, Otto von and Rudolph von Roth (1855–1875). *Sanskrit Wörtebuch. in Kürzer Fassung*. St. Petersburg: Buchdruckerei Der kaiderlichen Akadamie Der Wissenschften.

Bopp, Francisco (1847). *Glossarium Sanscritum. omnes radices et vocabula usitatissima explicantur et cum vocabulis graecis, latinis, germanicis, lithuanicis, slavicis, celticis comparantur*. Berlin: Dümmler.

Burnouf, Émile (1866). *Dictionnaire classique Sanscrit-Francais. oú sont coordonnés, revisés et complétés les travaux de Wilson, Bopp, Westergaard, Johnson etc. et contenant le devanagari, sa transcription européene, l'interprétation, les racines et de nombreux rapprochements philologiques, publié sous les auspices de M. Rouland, ministre de l'instruction publique*. Paris: Adrien-Maisonneuve.

Cappeller, Carl (1887). *Sanskrit-Wörterbuch. nach den Petersburger Wörterbüchern bearbeitet*. Strassburg: Karl J. Trübner.

– (1891). *A Sanskrit-English Dictionary. Based upon the St. Petersburg Lexicons*. Strassburg: Karl J. Trübner.

Fellbaum, Christianne. (1998). *Wordnet: an electronic lexical database*. Cambridge: MIT Press.

Ghatge, A. M., ed. (1981). *An encyclopedic dictionary of Sanskrit on historical principles*. Vol. 2. Poona: Deccan College Post Graduate and Research Institute.

Goldstücker, Theodor (1856). *A Dictionary in Sanskrit and English. Extended and improved from the second edition of the dictionary of Professor H. H. Wilson, with his sanction and concurrence, together with a supplement, grammatical appendices and an index serving as an English-Sanskrit vocabulary*. Berlin: A. Asher and Co.

Gombrich, Richard (1979). "'He cooks softly': Adverbs in Sanskrit. In Honor of Thomas Burrow". In: *Bulletin of the School of Oriental and African Studies, University of London* 42 no. 2, pp. 244–256.

Joshi, Shivaram Dattatreya (1967). "Adjectives and Substantives as a Single Class in the 'Parts of Speech'". In: *Journal of University of Poona Humanities Section*, pp. 19–30.

Kulkarni, Malhar et al. (2011). "Adverbs in Sanskrit Wordnet". In: *Icon 2011*. URL: http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/02_iwn_Adverbs%20in%20SWN.pdf.

Macdonell, Arthur Anthony (1893). *A Sanskrit-English dictionary. being a practical handbook with transliteration, accentuation, and etymological analysis throughout*. London: Longmans, Green, and Co.

– (1927). *Sanskrit grammar for students*. third. Oxford: Oxford University Press.

Monier-Williams, Monier (1846). *An elementary grammar of Sanskrit language. partly in roman character*. London: W. H. Allen.

– (1872). *A Sanskrit-English Dictionary. etymologically and philologically arranged with special reference to Greek, Latin, Gothic, German, Anglo-Saxon, and other cognate Indo-European languages*. London: The Clarendon Press.

Monier-Williams, Monier, Ernst Leumann, and Carl Cappeller (1899). *A Sanskrit-English Dictionary. Etymologically and philologically arranged with special reference to cognate Indo-European languages new edition, greatly enlarged and improved*. Oxford: The Clarendon Press.

Radhakantadeva (1819–1858). *Śabdakalpaduma*. 1st ed. Varanasi: Chaukhamba Sanskrit Series.

Speijer, J. S. (1886). *Sanskrit Syntax. with an intrduciton of H. Kern*. first. New Delhi: Motilal Banarasidas Publishers.

Stchoupak, Nadine, Luigia Nitti, and Louis Renou (1932). *Dictionnaire Sanskrit-Français*. Paris: Librairie d'Amérique et d'Orient, Adrien Maisonneuve.

Whitney, W. D. (1879). *Sanskrit Grammar*. 1st ed. Cambridge: Harvard University Press.

Wilson, Horace Hayman (1832). *A dictionary in Sanscrit and English. translated, amended, and enlarged from an original compilation, prepared by learned natives for the College of Fort William*. Calcutta: Printed at the Education press.

– (1841). *Anintroduction to the grammar of the Sanskrit language. for the use of early students*. London: J. Mandon and co.

Yates, William (1846). *A dictionary in Sanscrit and English. designed for the use of private students and of Indian colleges and schools*. Calcutta: Baptist Mission Press.

WSD in monolingual dictionaries for Russian WordNet

Daniil Alexeyevsky
Higher School of Economics, National research University, Moscow, Russia
dalexeyevsky@hse.ru

Anastasiya V. Temchenko
Moscow, Russia
avtemko@gmail.com

Abstract

Russian Language is currently poorly supported with WordNet-like resources. One of the new efforts for building Russian WordNet involves mining the monolingual dictionaries. While most steps of the building process are straightforward, word sense disambiguation (WSD) is a source of problems. Due to limited word context specific WSD mechanism is required for each kind of relations mined. This paper describes the WSD method used for mining hypernym relations. First part of the paper explains the main reasons for choosing monolingual dictionaries as the primary source of information for Russian language WordNet and states some problems faced during the information extraction. The second part defines algorithm used to extract hyponym-hypernym pair. The third part describes the algorithm used for WSD

1 Introduction

After the development of Princeton WordNet (Fellbaum, 2012), two main approaches were widely exploited to create WordNet for any given language: dictionary-based concept (Brazilian Portuguese WordNet, Dias-da-Silva *et al.*, 2002) and translation-based approach (see for example, Turkish WordNet, Bilgin *et al.*, 2004). The last one assumes that there is a correlation between synset and hyponym hierarchy in different languages, even in the languages that come from distant families. Bilgin *et al.* employ bilingual dictionaries for building the Turkish WordNet using existing WordNets.

Multilingual resources represent the next stage in WordNet history. EuroWordNet, described by Vossen (1998), was build for Dutch, Italian, Spanish, German, French, Czech, Estonian and English languages. Tufis *et al.* (2004) explain the methods used to create BalkaNet for Bulgarian, Greek, Romanian, Serbian and Turkish languages. These projects developed monolingual WordNets for a group of languages and aligned them to the structure of Princeton WordNet by the means of Inter-Lingual-Index.

Several attempts were made to create Russian WordNet. Azarova *et al.* (2002) attempted to create Russian WordNet from scratch using merge approach: first the authors created the core of the Base Concepts by combining the most frequent Russian words and so-called "core of the national mental lexicon", extracted from the Russian Word Association Thesaurus, and then proceeded with linking the structure of RussNet to EuroWordNet. The result, according to project's site[1], contains more than 5500 synsets, which are not published for general use. Group of Balkova *et al.* (2004) started a large project based on bilingual and monolingual dictionaries and manual lexicographer work. As for 2004, the project is reported to have nearly 145 000 synsets (Balkova *et al.* 2004), but no website is available (Loukachevitch and Dobrov, 2014). Gelfenbeyn *et al.* (2003) used direct machine translation without any manual interference or proofreading to create a resource for Russian WordNet[2]. Project RuThes by Loukachevitch and Dobrov (2014), which differs in structure from the canonical Princeton WordNet, is a linguistically motivated ontology and contains 158 000 words and 53 500 concepts at the moment of writing. YARN (Yet Another RussNet) project, described by Ustalov (2014), is based on the crowdsourcing approach towards creating WordNet-like machine readable open online thesaurus and contains at the time of writing more than 46 500

[1] http://project.phil.spbgu.ru/RussNet/, last update June 14, 2005

[2] Available for download at http://www.wordnet.ru

synsets and more than 119 500 words, but lacks any type of relation between synsets.

This paper describes one step of semi-automated effort towards building Russian WordNet. The work is based on the hypothesis that existing monolingual dictionaries are the most reliable resource for creating the core of Russian WordNet. Due to absence of open machine-readable dictionaries (MRD) for Russian Language the work involves shallow sectioning of a non machine-readable dictionary (non-MRD). This paper focuses on automatic extraction of hypernyms from Russian dictionary over a limited number of article types. Experts then evaluate the results manually.

1.1 Parsing the Dictionary

As far as our knowledge extends, there is no Russian monolingual dictionary that was designed and structured according to machine-readable dictionary (MRD) principles and is also available for public use.

There exist two Russian Government Standards that specify structure for machine readable thesauri (Standard, 2008), but they are not widely obeyed.

Some printed monolingual dictionaries are available in form of scanned and proof-read texts or online resources. For example, http://dic.academic.ru/ offers online access to 5 monolingual Russian dictionaries and more than 100 theme-specific encyclopedias. Each dictionary article is presented as one unparsed text entry. Resource http://www.lingvoda.ru/dictionaries/, supported by ABBYY, publishes user-created dictionaries in Dictionary Specification Language (DSL) format. DSL purpose is to describe how the article is displayed. DSL operates in terms *of italic, sub-article, reference-to-article* and contains no instrument to specify type of relations. This seems to be closest to MRD among available resources. Fully automated information extraction is out of the question in this case. When using non-MRD we have faced with number of problems that should be addressed before any future processing can be started:

1. Words and word senses at the article head are not marked by unique numeric identifiers.
2. Words used in article definitions are not disambiguated, so creating a link from a word in a definition to article defining the word sense is not trivial task.
3. Many contractions and special symbols are used.
4. Circular references exist; this is expected for synonyms and base lexicon, but uncalled for in sister terms, hypernyms, and pairs of articles with more complex relations.
5. The lexicon used in definitions is nearly equal to or larger than the lexicon of the dictionary.

In general, ordinary monolingual dictionaries, compiled by lexicographers, were not intended for future automated parsing and analysis. As stated in Ide and Véronis (1994), when converting typeset dictionaries to more suitable format researchers are forced to deal with:

1. Difficulties when converting from the original format, that often requires development of complex dedicated grammar, as previously showed by Neff and Boguraev (1989).
2. Inconsistencies and variations in definition format and meta-text;
3. Partiality of information, since some critical information in definitions is considered common knowledge and is omitted.

Research by Ide and Véronis (1994) gives us hope that using monolingual dictionaries is the best source of lexical information for WordNet. First they show that one dictionary may lack significant amount of relevant hypernym links (around 50-70%). Next they collect hypernym links from merged set of dictionaries and in the resulting set of hypernym links only 5% are missing or inconsistent as compared with expert created ontology.

Their work is partly based on work by Hearst (1998) who introduced patterns for parsing definitions in traditional monolingual dictionaries.

One notable work for word sense disambiguation using text definitions from articles was performed by Lesk (1986). The approach is based on intersecting set of words in word context with set of words in different definitions of the word being disambiguated. The approach was further extended by Navigli (2009) to use corpus bootstrapping to compensate for restricted context in dictionary articles.

In this paper we propose yet another extension of Lesk's algorithm based on semantic similarity databases.

2 Building the Russian WordNet

Specific aim of this work is to create a bulk of noun synsets and hypernym relations between them for further manual filtering and editing. To simplify the task we assume that every word sense defined in a dictionary represents a unique synset. Furthermore we only consider one kind of word definitions: such definitions that start with nominative case noun phrase. E. g.: *rus. ВЕНТИЛЯ́ЦИЯ: Процесс воздухообмена в лёгких. eng.'VENTILATION: Process of gas exchange in lungs'*. We adhere to hypothesis that in this kind of definitions top noun in the NP is hypernym. In order to build a relation between word sense and its hypernym we need to decide which sense of hypernym word is used in the definition. This step is the focus of this work.

2.1 The Dictionary

The work is based on the Big Russian Explanatory Dictionary (BRED) by Kuznetsov S.A. (2008). The dictionary has rich structure and includes morphological, word derivation, grammatical, phonetic, etymological information, three-level sense hierarchy, usage examples and quotes from classical literature and proverbs. The electronic version of the dictionary is produced by OCR and proofreading with very high quality (less than 1 error in 1000 words overall). The version also has sectioning markup of lower quality, with FPR in range 1~10 in 1000 tag uses for the section tags of our interest.

We developed specific preprocessor for the dictionary that extracts word, its definition and usage examples (if any) from each article. We call every such triplet word sense, and give it unique numeric ID. A article can have reference to derived word or synonym instead of text definition. Type of the reference is not annotated in the dictionary. We preserve such references in a special slot of word sense. The preprocessor produces a CSV table with senses.

2.2 Hypernym candidates

Given a word sense *W* we produce a list of all candidate hypernym senses.

Ideally under our assumption the first nominative case noun in *W*'s definition is a hypernym. However, due to variance in article definition styles and imperfect morphological disambiguation used, some words before the actual hypernym are erroneously considered candidate hypernym. To mitigate this we consider each of the first three nominative nouns candidate hypernyms. For each such noun we add each of its senses as candidate hypernym senses.

If sense *W* is defined by reference rather than by textual definition, we add both every sense of referenced word and each of its candidate hypernym senses to the list of candidate hypernym senses of *W*.

2.3 Disambiguation pipeline

We have developed a pipeline for massively testing different disambiguation setups. The pipeline is preceded by obtaining common data: word lemmas, morphological information, word frequency.

For the pipeline we broke down the task of disambiguation into steps. For each step we presented several alternative implementations. These are:

1. Represent candidate hyponym-hypernym sense pair as a Cartesian product of list of words in hyponym sense and list of words in hypernym sense, repeats retained.
2. Calculate numerical metric of words similarity. This is the point we strive to improve. As a baseline we used: random number, inverse dictionary definition number; classic Lesk algorithm. We also introduce several new metrics described below.
3. Apply compensation function for word frequency. We assume that coincidence of frequent words in to definitions gives us much less information about their relatedness than coincidence of infrequent words. We try the following compensation functions: no compensation, divide by logarithm of word frequency, divide by word frequency.
4. Apply non-parametric normalization function to similarity measure. Some of the metrics produce values with very large variance. This leads to situations where one matching pair of words outweighs a lot of outright mismatching pairs. To mitigate this we attempted to apply these functions to reduce variance: linear (no normalization), logarithm, Gaussian, and logistic curve.
5. Apply adjustment function to prioritize the first noun in each definition. While extracting candidate hypernyms the algorithm retained up to three candidate nouns in each article. Our hypothesis states that the first one is most likely the hypernym. We apply penalty to the metric depending

on candidate hypernym position within hyponym definition. We tested the following penalties: no penalty, divide by word number, divide by exponent of word number.

6. Aggregate weights of individual pairs of words. We test two aggregation functions: average weight and sum of best N weights. In the last case we repeat the sequence of weights if there were less than N pairs. We also tested the following values of N: 2, 4, 8, 16, 32.

Finally, the algorithm returns candidate hypernym with the highest score.

2.4 Testing setup

For testing the algorithms we selected words in several domains for manual markup. We determined domain as a connected component in a graph of word senses and hypernyms produced by one of the algorithms. Each annotator was given the task to disambiguate every sense for every word in such domain. Given a triplet an annotator assigns either no hypernyms or one hypernym; in exceptional cases assigning two hypernyms for a sense is allowed.

One domain with 175 senses defining 90 nouns and noun phrases was given to two annotators to estimate inter-annotator agreement. Both annotators assigned 145 hypernyms within the set. Of those only 93 matched, resulting in 64% inter-annotator agreement.

The 93 identically assigned hyponym-hypernym pairs were used as a core dataset for testing results. Additional 300 word senses were marked up to verify the results on larger datasets. The algorithms described were tested on both of the datasets.

2.5 Our Approach to Disambiguation

In this section we describe various alternatives to metric function on step 2 of the pipeline.

One known problem with Lesk algorithm is that it uses only word co-occurrence when calculating overlap rate (Basile *et al.*, 2004) and does not extract information from synonyms or inflected words. In our test it worked surprisingly well on the dictionary corpus, finding twice as many correct hypernym senses as the random baseline. We strive to improve that result for dictionary definition texts.

Russian language has rich word derivation through variation of word suffixes. The first obvious enhancement to Lesk algorithm to account for this is to assign similarity scores to words based on length of common prefix. In the results we refer to this metric as advanced Lesk.

Another approach to enhance Lesk algorithm is to detect cases where two different words are semantically related. To this end we picked up a database of word associations Serelex (Panchenko *et al*, 2013). It assigns a score on a 0 to infinity scale to a pair of noun lemmas roughly describing their semantic similarity. As a possible way to score words that are not nouns in Serelex we truncate a few characters off the ends of both words and search for the best pair matching the prefixes in Serelex. (See prefix "serelex" in Table 1).

We tested several hypotheses on how these two metrics can be used to improve the resulting performance. The tests were: to use only Lesk; to use only Serelex; to use Serelex where possible and fallback to advanced Lesk for cases where no answer was available; and to sum the results of Serelex and Lesk. Since Serelex has a specific distribution of scores we adjusted the advanced Lesk score to produce similar distribution.

For each estimator we performed full search through available variations on steps 3-6 of the pipeline and selected the best on the core set and estimated again on the larger dataset.

Test results are given in the Table 1:

Algorithm	CoreSet	LargeSet
random	30.8%	23.9%
first sense	38.7%	37.7%
naive Lesk	51.6%	**41.3%**
serelex	49.5%	38.0%
advanced Lesk	53.8%	33.3%
serelex with adjusted Lesk fallback	52.7%	36.3%
serelex + adjusted Lesk	52.7%	38.3%
prefix serelex	**53.8%**	38.0%

Table 1. Precision of different WSD algorithms.

3 Discussion

The low resulting quality of disambiguation seems to be a result of several factors: overall difficulty of the task (inter-annotator agreement is 64%), quality of input dictionaries, quality of used similarity database. We also seem to have missed some important linguistic or systemic features of text as well. Notably, the algorithms presented are still generically-applicable and do not use hypernym information.

Despite the low precision in determining the exact hypernyms, the pipeline produces thematically related chains of words. Examples of

chains, extracted by *prefix Serelex* algorithm are given below with English translation and comparison to Princeton WordNet (here ">>" symbolises *IS_A* relation):

- *rus. спираль >> кривая >> линия* eng.'spiral >> curve >> *line'* compared to *PWN spiral >> curve, curved shape >> line >> shape >> attribute >> abstraction >> entity*
- *rus. передняя >> комната >> помещение* eng. 'anteroom >> room >> premises' compared to PWN *room >> room >> area >> structure >> artifact >> whole >> object >> physical entity >> entity*
- *rus. рост >> высота >> расстояние* eng. 'stature, height >> height >> distance' compared to PWN *stature, height >> bodily property >> property >> attribute >> abstraction >> entity*

Dictionary parsing quality appears to be crucial for the current work, and the dictionary we selected provides us with a huge set of difficulties: abbreviations; alternating language in sense definitions; not all head words are lemmas (e.g. plural for nouns that have singular); poor quality of sectioning in OCR. Sectioning within BRED presents a large problem due to underspecified vaguely nested nature of sections. Properly digitized openly published Russian dictionary is really wished for.

Another problem with the dictionary is presence of nearly-identical definitions for the same term. Due to restricted context in dictionary in some cases it is difficult even for a human annotator to guess correctly whether a given pair of definitions describes the same concepts or two very distinct ones. This is especially true with abstract terms like *time (rus.: время)*, but physical entities like *field (rus.: поле)* also present such troubles.

One further step to building the Russian WordNet is to differentiate hypernyms from synonyms and co-hyponyms. Currently we hope to achieve this through classification of definitions and developing morphosyntactic templates to match different relation types within them. This is out of the scope of the current article though.

4 Conclusion

In this work we present a new pipeline for disambiguating and testing disambiguation frameworks for building WordNet relations from raw dictionary data in Russian language[3].

We described new algorithm for hypernym disambiguation which performs somewhat better than baseline in cases where annotators agree. The possibility for better disambiguation of specific relation types within dictionaries to be still open.

The resulting network, though noisy, is very suitable for rapid manual filtering.

References

Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., and Oparin, I. 2002. *Russnet: Building a lexical database for the russian language.* In Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas: 60-64.

Basile, P., Caputo, A., and Semeraro, G. 2014. *An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model*. In Proceedings of COLING: 1591-1600.

Bilgin, O., Çetinoğlu, Ö., and Oflazer, K. 2004. *Building a wordnet for Turkish*.Romanian Journal of Information Science and Technology, 7(1-2):163-172.

Balkova, V., Sukhonogov, A., and Yablonsky, S. 2004. *Russian wordnet. From UML-notation to Internet/Intranet Database Implementation.* In Proceedings of the Second Global Wordnet Conference.

Dias-da-Silva, B. C., de Oliveira, M. F., and de Moraes, H. R. 2002. *Groundwork for the development of the Brazilian Portuguese Wordnet.* In Advances in natural language processing:189-196.

Fellbaum, C. 2012. *WordNet.* The Encyclopedia of Applied Linguistics.

Gelfenbeyn, I., Goncharuk, A., Lehelt, V., Lipatov, A. and Shilo, V. 2003. *Automatic translation of WordNet semantic network to Russian language*. In Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003.

Hearst, M. A. 1998. *Automated discovery of WordNet relations.* WordNet: an electronic lexical database: 131-153.

Ide, N., Véronis, J. 1994. *Machine Readable Dictionaries: What have we learned, where do we*

[3] Available at http://bitbucket.org/dendik/yarn-pipeline

go. In Proceedings of the International Workshop on the Future of Lexical Research, Beijing, China: 137-146.

Ide, N., Véronis, J. 1993. *Refining taxonomies extracted from machine-readable dictionaries*. In Hockey, S., Ide, N. Research in Humanities Computing 2, Oxford University Press.

Kuznetsov S.A. Кузнецов, С. А. 2008. *Новейший большой толковый словарь русского языка*. СПб.: РИПОЛ-Норинт.

Lesk, M. 1986. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.* In Proceedings of the 5th annual international conference on Systems documentation: 24-26

Loukachevitch, N., Dobrov, B. 2014. *RuThes linguistic ontology vs. Russian Wordnets.*GWC 2014: Proceedings of the 7th Global Wordnet Conference: 154–162.

Navigli, R. (2009, March). *Using cycles and quasi-cycles to disambiguate dictionary glosses.* In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: 594-602.

Neff, M. S., and Boguraev, B. K. 1989. *Dictionaries, dictionary grammars and dictionary entry parsing.* In Proceedings of the 27th annual meeting on Association for Computational Linguistics: 91-101.

Panchenko, A., Romanov, P., Morozova, O., Naets, H., Philippovich, A., Romanov, A., and Fairon, C. 2013. *Serelex: Search and visualization of semantically related words.* In Advances in Information Retrieval: 837-840.

Standard, G. O. S. T. 2008. Standard 7.0.47-2008, *Format for representation on machine-readable media of information retrieval languages vocabularies and terminological data*.

Tufis, D., Cristea, D., Stamou, S. 2004. *BalkaNet: Aims, Methods, Results and Perspectives.* A General Overview In: D. Tufiş (ed): Special Issue on BalkaNet. Romanian Journal on Science and Technology of Information.

Ustalov, D. 2014. *Enhancing Russian Wordnets Using the Force of the Crowd.* In Analysis of Images, Social Networks and Texts. Third International Conference, AIST 2014. Springer International Publishing: 257-264.

Vossen, P. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Network*. Dordrecht.

Playing Alias - efficiency for *wordnet(s)*

Sven Aller
University of Tartu
sven.aller@ut.ee

Heili Orav
University of Tartu
heili.orav@ut.ee

Kadri Vare
University of Tartu
kadri.vare@ut.ee

Sirli Zupping
University of Tartu
sirli.zupping@ut.ee

Abstract

This paper describes an electronic variant of popular word game Alias where people have to guess words according to their associations via synonyms, opposites, hyperonyms etc. Lexical data comes from the Estonian Wordnet. The computer game Alias which draws information from Estonian Wordnet is useful at least for two reasons: it creates an opportunity to learn language through play, and it helps to evaluate and improve the quality of Estonian Wordnet.

1 Introduction

WordNet[1] is one of the most well-known lexico-semantic resources which is not used simply as a thesaurus for linguistic knowledge but also for language technology applications of language technology. Tony Veale has said that "WordNet … has found myriad applications in the field of natural language processing[2]" (i.e word sense disambiguation, ontologies, wordnets for opinion mining or sentiment analysis etc).

Estonian Wordnet (EstWN)[3] has grown quite large in size and our team is consistently working on the wordnet quality improvement. Since it is fairly complicated to revise concepts and their semantic relations manually (even one-by-one), automatic or semi-automatic ways for checking and discovering errors are preferred. For checking the consistency of EstWN different test patterns (Lohk 2015), also word frequency lists and corpora were used. One of the possibilities is to use gamification in language learning, namely a word explanation game called Alias. The Estonian computer game Alias[4] uses nouns, verbs, adjectives and adverbs present in EstWN[5]. In this paper we describe firstly how Alias is compiled and secondly, how it helps to improve the quality of EstWN. Although the data for learning language is quite useful and interesting, it is not the primary focus of this paper.

2 Estonian Wordnet

When setting up the Estonian WordNet we followed the principles of Princeton WordNet and EuroWordnet[6]. EstWN was built as a part of the EWN project (EuroWordNet-2 from the beginning of January 1998) and thus used the extension method as a starting point. It means that Base Concepts from English were translated into Estonian as a first basis for a monolingual extension. The extensions have been compiled manually from Estonian monolingual dictionaries and other monolingual resources (like frequency lists from Corpora of Written Estonian[7]).

EstWN includes nouns, verbs, adjectives and adverbs; as well as a set of multiword units. The database currently (September 2015; version 72) contains approximately 75 000 concepts (within more than 95 000 words) which are connected with approx 210 000 semantic relations and work is still in progress.

3 Design of the computer game Alias

Based on Princeton WordNet a game for word sense labeling has been created (Venhuizen et al 2013)[8]. Since obtaining gold standard data for

[1] http://wordnet.princeton.edu
[2] http://www.odcsss.ie/node/39
[3] http://www.cl.ut.ee/ressursid/teksaurus/
[4] http://keeleressursid.ee/alias/
[5] http://www.cl.ut.ee/ressursid/teksaurus/
[6] http://www.illc.uva.nl/EuroWordNet/
[7] http://www.cl.ut.ee/korpused/
[8] http://wordrobe.housing.rug.nl/Wordrobe

word sense disambiguation is costly, they are using gamification for collecting semantically annotated data. Another game that uses Princeton WordNet is an on-line questions game Piclick[9]. This is an implementation of twenty questions game, where one person thinks of a concept while the other asks him a series of yes/no questions and attempts to guess what his partner thinks of (Rzeniewicz and Szymanski, 2013).

One of the computer games which uses concepts and relations between these concepts is called word explanation game Alias, where the goal is to explain words to one's partner using different hints. These hints are typically definitions, synonyms, antonyms, hyperonyms and hyponyms etc, which are mostly present in wordnet making it suitable knowledge base for Alias' game engine.

Alias as a computer game is designed to be used by non-experts, non-linguists, and for players to play for fun. One of the main crowdsourcing platform is Amazon's Mechanical Turk, where workers get paid. In Alias game it assumed that contributors are awarded with entertainment and players are challenged to win more points than the computer.

The computer chooses a random word and shows different hints which are supposed to help a player guess the right words. For each word up to 12 randomly chosen hints are given. Hints are given to a player in sequence. If the player does not guess the word by the last hint, the point will be given to the computer.

Alias is written in PHP and it is web-based. Considering the game's architecture the EstWN database is somewhat modified – Alias uses only these synsets which have at least three hints to show (synonyms or other semantic relations), which in turn means, that at least three hints for a player are assured.

3.1 Different levels of Alias

The EstWN contains of words, which have very different usage frequencies and it can be quite complicated to guess the words, which are rarely used (mostly adverbs, i.e *criss-cross*) or domain-specific (i.e grammatical categories in linguistics, *ablative case*) for example. For this reason words for Alias game are selected in comparison of the word frequency lists from the Corpus of Written Estonian[10] and only these words from the synsets that belong to the frequency list are selected for playing. Following Table 1 shows the numbers of words per word classes of different levels in Alias game. Words are selected as follows: words from EstWN which are also in the list of most frequent words, this means that conjunctives and pronouns are left out from the frequent words, since they do not exist in EstWN. Also, only one member of the synset is taken from the frequent words list, for example if both synset members are in the frequency list ('kid' and 'child') then only the first is chosen.

Table 1. Numbers of words of different levels in Alias game

	Beginner (selected from 1000 frequent words)	**Intermediate (selected from 5000 frequent words)**	**Expert (selected from 10000 frequent words)**
Nouns	333	1654	2863
Verbs	161	583	883
Adjectives	56	315	528
Adverbs	99	251	384
All	649	2803	4658

Based on that information there are three different levels: beginner level contains of 649 words (selected from 1000 frequent), intermediate level contains of 2803 words (selected from 5000 frequent) and expert level of 4658 words (selected from 10 000 frequent). Homonyms are connected, the word *bank*, for example, displays hints from the meanings of both institution and natural object.

3.2 Questions for Alias

There are 55 different types of semantic relations present on Alias game (as it is in EstWN). In addition also definitions and example-sentences are used. Every type of semantic relation is related to a certain sentence template, which is presented to a player. The sentences should be simple in the sense that an average user is supposed to under-

[9] https://kask.eti.pg.gda.pl/pinqee/game

[10] http://www.cl.ut.ee/ressursid/sagedused/ (only in Estonian)

stand the questions that present different semantic relations.

Here are presented some of the sentence templates which Alias uses for questions:

- antonym – It's opposite for ___ (for example "*It's opposite for a man*")
- fuzzynym – It's somehow related to _____ (for example "*It's somehow related to the word elegance*")

Similarly to original board game Alias the computer game also asks words in dictionary form – nouns in nominative and verbs in infinitive form.

Estonian language is rich in compound words and in EstWN many hyponyms contain of their hyperonym as the second part of the compound word.

1. For example: one type of *kaabu* 'hat' is *vilt+kaabu* 'trilby hat'

If the compound word consists of the word that is currently guessed, the similar stems of the words are removed (see example 2). The same rule applies also in the original board game. Since Estonian is rich in cases, persons and in inflectional system, then it is quite complicated to find the word with the similar stem. The morphological analyzer [11] is used to compare the lemmas in hint to the lemma of the asked word. If they match, then the similar stem is replaced with a gap.

2. For example:

Question:

See on teatud liiki õunapuu.
This has a type of appletree.
is replaced
See on teatud tüüpi õuna_____
This has a type of apple_____

Answer: Puu (*Tree*)

Question:

You can use this word like that:
Bring back my pony to me
is replaced with
Bring ____ my pony to me

Answer: Back

4 Some statistics from play log

Since the December 2014 Alias is played 664 times. During these games, 2571 words have been asked, it means that average 3,87 words per game are guessed. As the Table 2 shows, the correctly guessed words percentage differed largely across different semantic relations and definitions or examples used.

All the semantic relations present in EstWN are also used in Alias. Of course there are some relations in EstWN, which are not so frequent – role_instrument or has_mero_member for example, which means that they are also asked less frequently during the game. Table 2 states that the top-guessed relation is role_instrument even though it occurred only 5 times, so we can say that it is not statistically so important as definitions and antonym relation for example.

Groups (as group_role, group_xpos, group_holo, group_involved, group_derive) are connected in table because they share the same sentence template for hints. These sentence templates will be changed in the next version of the game.

5 Discussion

George Miller, as a psycholinguist was interested in how the human semantic memory is organized (Miller 1998), which type of relations are most typical between words and concepts.

In addition to (psycho)linguistic tests, some conclusions/inferences can be drawn using log files of game Alias as well. Results give us feedback which relations are clear, which are too fuzzy or too general or just too strange. For example: *migration* involved_location *residence, abode*. Piek Vossen's (2002) test for location_involved relation is:

(A/an) X is the place where the Y happens.

So, it is obvious that relation between *migration* and *residence* needs to be corrected in EstWN.

As you can see from the Table 2, there is a slight difference between guessing hints containing of hyperonyms (7.2%) and hyponyms (9.1%), the latter shows slightly better results. Hyperonyms might be too general, they might have multiple hyponyms, for example 'to run – to move'. While giving a hyponym as hint, for example 'to run – to sprint', opens the meaning of the word more precisely.

Since fuzzynym-hints do not appear to be very useful for players (only 7.1%), we can assume, that the connections and associations presented by fuzzynyms are too vague. Some of the fuzzynynms can be assigned to a more specific semantic relation, for example 'doctor' and 'stetoscope' or 'postman' and 'postbag' which denote something that belongs to some certain pro-

[11] http://www.filosoft.ee/html_morf_et/

fession. But, as we could see from the play logs, there are many fuzzynyms completely distant, for example ‘presentation’ and ‘evolution’, ‘painting’ and ‘education’ etc.

From the player’s perspective the definitions (21.3%) and examples (18.2%) are one of the most successful hint for guessing the right word. In many cases we can see from logs that various hints with semantic relations do not help the player, but definition and explanation – also even if they are the first hints – are very informative. This means that as a concept based database EstWN needs to have clear definitions and good examples to open the meanings of concepts.

The meaning of the word is quite well guessed while hints present synonyms (here Variants, 14.5% right answers) or antonyms (33.7%) and near antonyms (9.0%) or near synonyms (9.4%). It is intuitively simpler to guess for example the word ‘kiss’ by its synonym ‘buss’ than its hyperonym ‘touch’ or verb ‘to buy’ by its antonym ‘to sell’ than its hyperonym ‘to acquire’.

Hints that contain of functional relations (i.e role, meronymy) are usually very clear to a player, of course these indicate to concrete objects. The role-relation can connect both nouns to nouns and nouns to verbs. For example the verb ‘to run’ has been guessed by its role_agent ‘runner’ but not by its hyperonym ‘to move’.

The logs from beginner and even intermediate level can indicate to problems of the main vocabulary, for example for a question: this is near synonym for the word ‘swamp bridge’ the correct answer should be ‘road’. Of course this near synonym link is not correct and should be revised also in EstWN.

In many aspects this game reflects that the associations of words/concepts are free and arbitrary in human minds. For example, illegible (sloppy, quickly written) handwriting can remind us the doctors’ style of handwriting. But still it is possible – if considered carefully and thoroughly – find a certain system, which is similar to the one Georg Miller started to create a model of the human mental lexicon. In „On wordnets and relations“ (Piasecki et al 2013) is mentioned that forming a synset (in the sense of wordnet) is a quite difficult task and has been largely left to the intuition of people who build wordnets. Game gives us a chance to check how similar the compilers intuition is to a player’s intuition.

6 Conclusion

The play logs contain of valuable information for a lexicographer and using this for improvement of EstWN is quite a new approach. The EstWN has benefited from the Alias game in many ways. Firstly it was possible to determine completely false synsets and/or the non-suitable semantic relations. Secondly it was possible to correct some of the semantic relations. Thirdly some of the definitions were improved and made more precise. The correction work has grown more systematic, since more log files have become available. As an addition to revising and correcting synsets and their relations it was interesting to observe which hints were more informative to players than the others. It gives us good feedback if there is any semantic relation too general, too narrow or just too vague.

Not less important is the value to Alias game and it working principles. If studying the logs more thoroughly it is possible to improve the quality of Alias, for example how to choose concepts, how to sort, choose, form and present hints etc. This game is adjustable for every language which has their own wordnet.

Researchers of Polish Wordnet (Maziarz et al 2013) have said that “Synonymy is intended as the cornerstone of a wordnet, hypernymy – its backbone, meronymy – its essential glue”. After analyzed the log files of Alias-game we can say that traditional definitions and antonyms are clearer to a player with no linguistic background.

References

Fellbaum, Ch., 2010. WordNet, in: Poli, R., Healy, M., Kameas, A. (Eds.), Theory and Applications of Ontology: Computer Applications. Springer Netherlands, pp. 231–243.

Lohk, A. 2015. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Thesis on Informatics and System Engineering C105. Press of Tallinn University of Technology.

Maziarz, M., Piasecki, M.; Szpakowicz, S. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. – Language Resources and Evaluation 47 (3), 769–796.

Miller, G. A. 1998. Nouns in WordNet. – WordNet. An Electronic Lexical Database. Ed. Christiane Fellbaum. Cambridge, MA: The MIT Press, 23–46.

Piasecki, M.; Szpakowicz, S.; Fellbaum, Ch.; Pedersen, B.S. 2013. On wordnets and relations. – Language Resources and Evaluation 47 (3), 757–767.

Rzeniewicz, J.; Szymański, J. 2013. Bringing Common Sense To Wordnet With A Word Game. *Computational Collective Intelligence. Technologies and Applications* (2013): 296-305. Web. 14 Sept. 2015.

Venhuizen, N.J., Basile, V., Evang, K. and Bos, J. 2013. Gamification for Word Sense Labeling. In Katrin Erk and Alexander Koller (eds.), *Proceedings of the 10th International Conference on Computational Semantics (IWCS'13) -- Short Papers*, 397-403, Potsdam, Germany.

Vossen, P. 2002. EuroWordNet General Document. Version 3. Final. July 1, 2002. http://www.vossen.info/docs/2002/EWNGeneral.pdf (27.10.2015).

Table 2. Results of playing by different relations

Relation	Occurence	Right cases	Right cases (%)	Wrong cases
role_instrument	5	3	60.0%	2
role_agent	17	7	41.2%	10
antonym	86	29	33.7%	57
causes	18	6	33.3%	12
has_holo_madeof	23	6	26.1%	17
DEFINITION	1390	296	21.3%	1094
is_caused_by	31	6	19.4%	25
EXAMPLE	1136	207	18.2%	929
group_role	41	6	14.6%	35
VARIANTS	1597	232	14.5%	1365
has_mero_member	7	1	14.3%	6
has_mero_madeof	7	1	14.3%	6
has_meronym	26	3	11.5%	23
has_mero_part	36	4	11.1%	32
has_holo_member	18	2	11.1%	16
group_involved	42	4	9.5%	38
near_synonym	577	54	9.4%	523
has_hyponym	2123	194	9.1%	1929
near_antonym	200	18	9.0%	182
group_holo	60	5	8.3%	55
has_mero_location	12	1	8.3%	11
role_location	13	1	7.7%	12
has_hyperonym	994	72	7.2%	922
has_xpos_hyponym	152	11	7.2%	141

fuzzynym	622	44	7.1%	578
group_xpos	313	19	6.1%	294
state_of	84	4	4.8%	80
be_in_state	45	1	2.2%	44
is_subevent_of	4	0	0.0%	4
has_mero_portion	2	0	0.0%	2
has_holo_portion	2	0	0.0%	2
role_target_direction	1	0	0.0%	1
has_subevent	1	0	0.0%	1
role_manner	1	0	0.0%	1
has_holo_location	0	0	0.0%	0
belongs_to_class	0	0	0.0%	0
group_derive	0	0	0.0%	0
role_source_direction	0	0	0.0%	0
has_instance	0	0	0.0%	0
role_direction	0	0	0.0%	0

Detecting Most Frequent Sense using Word Embeddings and BabelNet

Harpreet Singh Arora
Computer Science and Engineering, Academy of Technology, Hooghly, India
harpreet.singharora@aot.edu.in

Sudha Bhingardive
Department of Computer Science and Engineering, IIT Bombay, India
sudha@cse.iitb.ac.in

Pushpak Bhattacharyya
Department of Computer Science and Engineering, IIT Bombay, India
pb@cse.iitb.ac.in

Abstract

Since the inception of the SENSEVAL evaluation exercises there has been a great deal of recent research into Word Sense Disambiguation (WSD). Over the years, various supervised, unsupervised and knowledge based WSD systems have been proposed. Beating the first sense heuristics is a challenging task for these systems. In this paper, we present our work on Most Frequent Sense (MFS) detection using Word Embeddings and BabelNet features. The semantic features from BabelNet *viz.*, synsets, gloss, relations, *etc.* are used for generating sense embeddings. We compare word embedding of a word with its sense embeddings to obtain the MFS with the highest similarity. The MFS is detected for six languages *viz.*, English, Spanish, Russian, German, French and Italian. However, this approach can be applied to any language provided that word embeddings are available for that language.

1 Introduction

Word Sense Disambiguation or WSD refers to the task of computationally identifying the sense of a word in a given context. It is one of the oldest and toughest problems in the area of Natural Language Processing (NLP). WSD is considered to be an AI-complete problem (Navigli et al., 2009) *i.e.*, it is one of the hardest problems in the field of Artificial Intelligence. Various approaches for word sense disambiguation have been explored in recent years. Two of the widely used approaches for WSD are – disambiguation using the annotated training data called as supervised WSD and disambiguation without the annotated training data called as unsupervised WSD.

MFS is considered to be a very powerful heuristics for word sense disambiguation. With sophisticated methods, it is difficult to outperform this baseline. The MFS baseline is created with the help of a sense annotated corpus wherein the frequencies of individual senses are learnt. It is found that, only 5 out of 26 WSD systems submitted to SENSEVAL-3, were able to beat this baseline. The success of the MFS baseline is mainly due to the frequency distribution of senses, with the shape of the sense rank versus frequency graph being a Zipfian curve. Unsupervised approaches were found very difficult to beat the MFS baseline, while supervised approaches generally perform better than the MFS baseline.

In our paper, we have extended the work done by Bhingardive et al. (2015). They have used word embeddings along with features from WordNet for the detection of MFS. We used word embeddings and features from BabelNet for detecting MFS. Our approach works for all part-of-speech (POS) categories and is currently implemented for six different languages *viz.*, English, Spanish, Russian, German, French and Italian. This approach can be easily extended to other languages if word embeddings for the specific language are available.

The paper is organized as follows: Section 2 briefs the related work. Section 3 explains BabelNet. Our approach is given in section 4. Experiments are presented in section 5 followed by conclusion.

2 Related Work

McCarthy et al. (2007) proposed an unsupervised approach for finding the predominant sense using an automatic thesaurus. They used WordNet similarity for identifying the predominant sense. This approach outperforms the SemCor baseline for words with SemCor frequency below five.
Bhingardive et al. (2015) compared the word embedding of a word with all its sense embedding

to obtain the predominant sense with the highest similarity. They created sense embeddings using various features of WordNet.
Preiss et al. (2009) refine the most frequent sense baseline for word sense disambiguation using a number of novel word sense disambiguation techniques.

3 BabelNet

BabelNet (Navigli et al., 2012) is a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a semantic network. It connects concepts and named entities in a very large network of semantic relations, made up of more than 13 million entries, called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages.
BabelNet v3.0 covers 271 languages and is obtained from the automatic integration of:

- WordNet[1] - a popular computational lexicon of English.
- Open Multilingual WordNet[2] - a collection of WordNets available in different languages.
- Wikipedia[3] - the largest collaborative multi-lingual Web encyclopedia.
- OmegaWiki[4] - a large collaborative multi-lingual dictionary.
- Wiktionary[5] - a collaborative project to produce a free-content multilingual dictionary.
- Wikidata[6] - a free knowledge base that can be read and edited by humans and machines alike.

BabelNet provides API for Java, Python, PHP, Javascript, Ruby and SPARQL.

4 Our Approach

We propose an approach for detecting the MFS which is an extension of the work done by Bhingardive et al. (2015). Our approach follows an iterative procedure to detect the MFS of any word given its POS and language. It works for six different languages *viz.*, English, Spanish, Russian, German, French and Italian. We used BabelNet as a lexical resource, as it contains additional information as compared to WordNet. This approach uses pre-trained Google Word Embeddings[7] for English language, and for all other languages Polyglot[8] Word Embeddings are used.

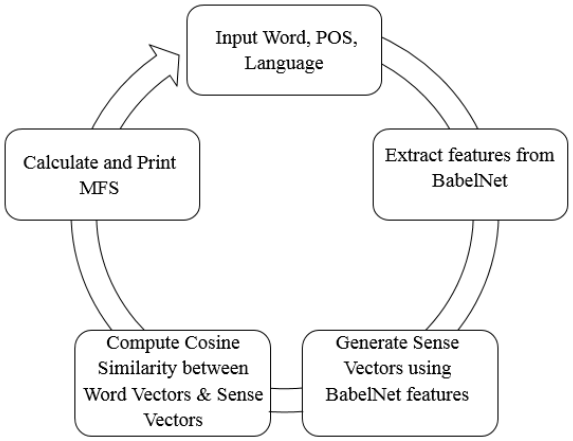


Figure 1. Steps followed by our approach

The steps followed by our approach as shown in figure 1 are as follows -

1. The system takes a word, POS and language code as an input**.**
2. For every sense of a word, features such as synset members, gloss, hypernym, *etc*. are extracted from BabelNet.
3. Sense embeddings or sense vectors are calculated by using this feature set.
4. Cosine similarity is computed between word vector (word embedding) of an input word and its sense vectors.
5. Sense vector which has maximum cosine similarity with the input word vector is treated as the MFS for that word.

4.1 Calculating Sense Vectors

4.1.1 Creation of BOW

Bag of Words (BOW): Bag of words for each sense of a word are created by extracting context words from each individual feature from BabelNet. BOWs obtained for each feature are, BOWS for synset members (S), BOWG for

[1] http://wordnet.princeton.edu/
[2] http://compling.hss.ntu.edu.sg/omw/
[3] http://www.wikipedia.org/
[4] http://www.omegawiki.org/
[5] http://www.wiktionary.org/
[6] https://www.wikidata.org/
[7] https://code.google.com/p/word2vec/
[8] http://polyglot.readthedocs.org/en/latest/Embeddings.html

content words in the gloss (G), BOWHS for synset members of the hypernym synset (HS), BOWHG for content words in the gloss of hypernym synsets (HG).

Word Embeddings: Word embedding or word vector is a low dimensional real valued vector which captures semantic and syntactic features of a word.

Sense Embeddings: Sense embedding or sense vector is similar to word embedding which is also a low dimensional real valued vector. It is created by taking average of word embeddings of each word in the BOW.

4.1.2 Filtering BOW

Filtering of BOWs are done to reduce the noise. The following procedure is used to filter BOWs:

1. Words for which word embeddings are not available are excluded from BOW.
2. From this BOW, the most relevant words are picked using following steps:
    a. Select a word from BOW
    b. The cosine similarity of that word with each of the remaining words in the BOW is computed.
    c. If the average cosine similarity lies between the threshold values 0.35 and 0.4, then we keep the word in the BOW else it is discarded. It is found that values above 0.4 were discarding many useful words while the values below 0.35 were accepting irrelevant words resulting in increasing the noise. Hence, the threshold range of 0.35 - 0.4 was chosen by performing several experiments.

For example, consider the input as -

Word: *cricket*
POS: *NOUN*
Language code: *EN*

Let $BOWG_1$ be the BOW of a gloss feature for the sport sense (S_1) of a word *cricket*.

$BOWG_1$ = {*Cricket is a bat and ball game played between two teams of 11 players each on a field at the center of which is a rectangular 22-yard long pitch*}

After removing stop words and words for which word embeddings are not available, we get the updated $BOWG_1$ as**,**

$BOWG_1$ = {bat ball game played two teams}

Now, the cosine similarity of each word in $BOWG_1$ with other words in $BOWG_1$ is computed to get the most relevant words which can represent the sense S_1. For instance, for a word *game*, the average cosine similarity was found to be 0.38 which falls in the selected threshold. Hence, the word *game* is not filtered from the $BOWG_1$. Table 1 shows how the word *game* is selected based on the average cosine similarity score.

Word	Gloss Members	Cosine Similarity
game	played	0.50
game	ball	0.49
game	bat	0.30
game	two	0.17
game	teams	0.44

Table 1: Cosine similarity scores of a word *game*

Average Cosine Score (*game*) =
(0.51 + 0.49 + 0.30 + 0.17 + 0.44)/5 = **0.38**
Similar process is carried out for each word of BOW.

4.2 Detecting MFS

In our approach we are detecting MFS in an iterative fashion. In each iteration we are checking which type of BOWs (BOWS, BOWG, BOWHS, and BOWHG) are sufficient to detect the MFS. This can be observed in figure 2.

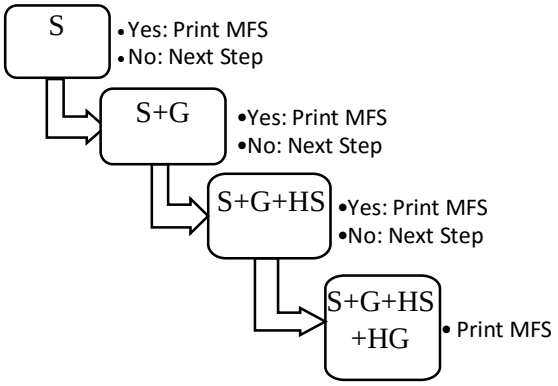


Figure 2: Iterative process of detecting MFS

In figure 2, we can see how BOWs are used to create sense vectors in an iterative fashion to get

the MFS. If synset members (S) are sufficient to get the MFS then our algorithm prints the MFS and stops, otherwise other BOWs of various features like gloss (G), synset members of the hypernym synsets (HS), content words in the gloss of the hypernym synsets (HG) are used iteratively to get the MFS. The algorithm is as follows:

1. For each sense *i* of a word:
   a. VEC(i) = Create_sense_vector ($BOWS_i$)
   b. VEC(W) = word vector of the input word
   c. SCORE(i) = cosine_similarity (VEC(i), VEC(W))
2. Arrange this SCORE in descending order according to the similarity score.
3. If (SCORE(0) – SCORE(1)) > threshold:
   a. MFS = Sense(SCORE(0))
   b. Print MFS
   c. End
4. Else:
   a. Run Steps 1 to 3 for (BOWSi + BOWGi)
5. If (SCORE(0) – SCORE(1)) > threshold:
   a. Run Steps 1 to 4 for (BOWSi + BOWGi + BOWHSi)
6. Else:
   a. Print MFS
   b. End
7. If (SCORE(0) – SCORE(1)) > threshold:
   a. Run Steps 1 to 4 for (BOWSi + BOWGi + BOWHSi + BOWHGi)
   b. Print MFS
   c. End
8. Else:
   a. Print MFS
   b. End

Where,

- VEC(i) denotes sense vector of an input word.
- SCORE (v1, v2) is cosine similarity between word vector v1 and sense vector v2.
- SENSE (SCORE(i)) is the sense corresponding to SCORE(i).
- Ambiguity is resolved by comparing the score of most similar sense and second most similar sense, obtained after Step 2. Step 3 checks if the difference between their score is above threshold →0.02 (This threshold was chosen after conducting various experiments with other threshold figures. The average difference between two most similar senses was found to be 0.02). There is a net speed-up in the procedure, as the computation time is significantly abridged as compared to Bhingardive et al. (2015). As we are using an iterative procedure for detecting the MFS, our approach, most of the times gives a better result as compared to Bhingardive et al. (2015) which we have manually verified.

5 Experiment and Results

We used pre-trained Google's word vectors as word embedding for English language, for all other languages Polyglot's word embeddings are used. Due to lack of availability of gold data, we could not compare our results with MFS results obtained from BabelNet. Upon considering Princeton WordNet as gold data, we cannot equate our results with it because they might be semantically similar but not syntactically.

6 Conclusion

We have proposed an approach for detecting the most frequent sense for a word using BabelNet as a lexical resource. BabelNet is preferred as a resource since it incorporates data not only from Princeton WordNet but also from sources. Hence the volume of ambiguity is reduced by a significant proportion. Our approached follows an iterative procedure until a suitable context is found to detect the MFS of a word. It is currently working for English, Russian, Italian, French, German, and Spanish languages. However, it can be easily ported across multiple languages. An API is developed for detecting MFS using BabelNet which can be publically made available in future.

References

Calvo Hiram and Alexander Gelbukh. 2014. Finding the Most Frequent Sense of a Word by the Length of Its Definition. Human-Inspired Computing and its Applications. Springer International Publishing.

Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll. 2007. Unsupervised Acquisition of Predominant Word Senses. Computational Linguistics, 33 (4) pp 553-590.

George A. Miller. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Judita Preiss. 2009. Refining the most frequent sense baseline. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future

Directions. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41.2: 10.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193 (2012): 217-250

Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised Most Frequent Sense Detection using Word Embeddings. North American Chapter of the Association for Computational Linguistics (NAACL), Denver, Colorado.

Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary

Martin Benjamin
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
martin.benjamin@epfl.ch

Abstract

The data compiled through many Wordnet projects can be a rich source of seed information for a multilingual dictionary. However, the original Princeton WordNet was not intended as a dictionary *per se*, and spawning other languages from it introduces inherent ambiguity that confounds precise inter-lingual linking. This paper discusses a new presentation of existing Wordnet data that displays joints (distance between predicted links) and substitution (degree of equivalence between confirmed pairs) as a two-tiered horizontal ontology. Improvements to make Wordnet data function as lexicography include term-specific English definitions where the topical synset glosses are inadequate, validation of mappings between each member of an English synset and each member of the synsets from other languages, removal of erroneous translation terms, creation of own-language definitions for the many languages where those are absent, and validation of predicted links between non-English pairs. The paper describes the current state and future directions of a system to crowdsource human review and expansion of Wordnet data, using gamification to build consensus validated, dictionary caliber data for languages now in the Global WordNet as well as new languages that do not have formal Wordnet projects of their own.

1. Introduction

When viewed from the perspective of creating a concept-based multilingual dictionary, the Global WordNet (GWN) is filled with both treasure and risk. The Kamusi Project has imported the freely available data from the Open Multilingual Wordnet (OMW) as seed for further dictionary development. In doing so, we have encountered issues with current Wordnet implementations[1] that we hope to contribute toward resolving. Section 2 describes the work we have done to make existing OMW data available in a format that might add value for the public over previous distributions. Section 3 discusses problems encountered with using Wordnet data as the basis for detailed lexicography. Section 4 details the systems we are implementing to (1) offer improved data for current Wordnets and to (2) use as a basis for building parallel data for many more languages.

2. Converting synsets to concept-specific lemmas.

In structuring a multilingual dictionary, Kamusi has determined that each concept/spelling pair within a language should be a distinct node; "light" (not heavy) is different from "light" (not dark) is different from "light" (not serious). This arrangement is compatible overall with the Princeton WordNet (PWN), which separates each sense it has identified for a given English spelling. However, PWN clusters other terms with the same general meaning in the same "synset", such as {*cloth, fabric, material, textile*}, so part of the conversion of PWN to the Kamusi structure is to make each member a separate node, each linked as a synonym to all others, while retaining for each the Wordnet working definition.

Wordnets for different languages are matched to PWN by synset (Bond and Foster 2013). PWN's own search engine shows the terms in the OMW that correspond to a synset, marked by language, with no further navigation possible between languages (see figure 1). The OMW search interface better shows the different synsets that are linked to the English concept (see figure 2), and also allows users to seek synsets in a second language that match through English to a search term in a first. For Kamusi, by contrast, the matrix of relationships between the individual terms within Wordnet synsets is the multilingual problematic. With English concepts and translation equivalents granted a debatable

[1] This paper uses "Wordnet" as a collective noun to signify the web of projects that adopt the synset and ontological approach, and that largely adhere to the same concept set, while also referring to individual Wordnets that exist for specific languages.

assumption of validity, Kamusi has now linked the individual terms in the synsets in each language independently, with the matches inferred through English shown as second degree. In the example of "light" (not dark) in figure 3, the concept as defined in English links to two different nodes in Catalan, "brillant" and "illuminós", and two nodes in Spanish, "claro" and "luminoso". These particular senses of "claro" and "luminoso" in turn link individually to "brillant" and "illuminós", and all five of the preceding terms have independently negotiable relationships with Japanese "明るい" and "明らか", Croatian "svjetleći" and "svijetao", and onward through the languages available in OMW.

When new terms are matched to the concept in Kamusi for non-Wordnet languages, for example a Quechua equivalent matched to Spanish, links are formed, with degree of separation indicated, to all of the existing terms within the multilingual relation set.

The data from OMW includes 117,659 synsets from PWN, matched to varying amounts among 26 languages and two variants (for Chinese and Norwegian), resulting in approximately 1.2 million individual nodes. Some large relation sets include 150 or more terms as equivalents among languages, which can produce upwards of 11,000 individual links; while server resources have not been expended to tally the total links in the data, at least ten million term pairs have been mapped.

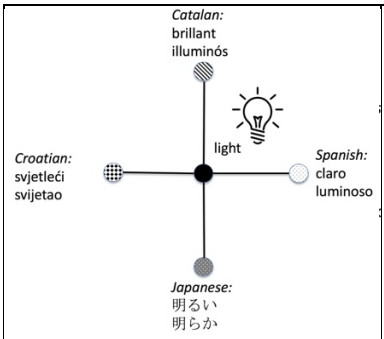


Figure 1: Terms linked to English synset (PWN method)

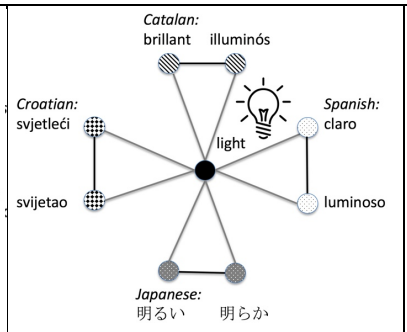


Figure 2: Multilingual synsets linked to English synset (OMW method)

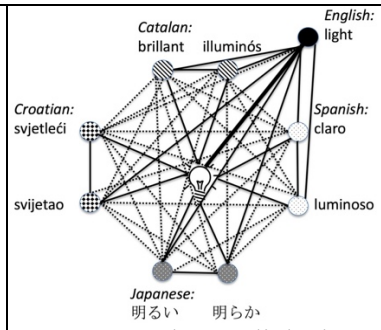


Figure 3: Each term linked to concept and each other, with joints (distance) and substitutes (type of equivalence) tracked (Kamusi method)

3. Problems with Wordnet as lexicography

Having thus worked at length with the data in OMW, we have encountered a number of limitations that bear mentioning and further work.

It is important to acknowledge that Wordnet was never intended to be a definitive dictionary, for English or any other language. The intent of the word list was to provide data for non linguistic research, initially in psychology (Miller et al 1990, Miller and Fellbaum 2007). It is thus not a criticism to state that it does not fulfill a role it was not designed for. However, in the absence of a better large and well organized set of freely available terms and definitions, it has taken on the de facto role of a universal lexicon, linked not only across languages but also across numerous projects related to computational linguistics. We suggest that Wordnet can be retrofitted for incorporation within a more lexicographically oriented resource, without losing its strong bonds across languages and projects.

The first problem is that many of the English definitions in the PWN data are inadequate, some to the point of error. Many of the definitions were written by the founder of the project, who was not a lexicographer and was faced with the immense task of producing good-enough ways of understanding tens of thousands of terms. The data is thus peppered with definitions such as "elevator car: *where passengers ride up and down*"; the sense is clear to a knowledgeable speaker, but would not suffice for a credible dictionary. Sometimes the definition is a problem for one member of a synset, either because the terms do not have identical meanings (e.g., verb "eat, feed: *take in food; used of animals only*" is valid for "feed" but not for "eat") or because that term forms the nub of the explanation used to define the group (e.g., verb "visit, call, call in: *pay a brief visit*" functions for "call" and "call in", but

is a tautology for "visit"). Some definitions are simply wrong; a law practice, as a lexicalizable multiword expression, is not "*the practice of law*", but a business through which lawyers conduct their profession.

The consequence of a wrong definition is that the errors propagate through reproductions, projects, and languages. Fixing mistakes is thus an opaque journey through long-completed Wordnet projects that are unlikely to be reopened, in languages that can only be corrected by their speaker communities if they are alerted to the issues and provided with the tools to make the necessary changes. All three languages that attempt an equivalent for "law practice" completely miss the true English sense (perhaps the other 25 groups were too stymied by the tautology to attempt a translation), so Finnish, Thai, and Spanish parties must somehow be alerted that the PWN definition has been modified, and given the platform to review and revise the term in their language. Further, the original PWN definition must be maintained with an indication that it had been deprecated, so projects like BabelNet[2] and VisuWords[3] that link to or build upon it (Navigli and Ponzetto 2010) can see the adjustments flagged, and update themselves accordingly. Unfortunately, numerous websites have replicated the existing PWN data in apparently static form (e.g., vocabulary.com[4]), so the current data will live in many places forever.

The second problem is that many errors exist in the equivalents that other languages map to English. For example, the French word "lumière", always a noun, translates to a few senses of English "light", mostly in regard to things that shine and figuratively in respect to illuminating knowledge. As rendered in the WOLF French Wordnet, however, "lumière" is mapped to 45 senses of "light", as a noun, verb, or adjective, with meanings such as "insubstantial", "less than the full amount", and "alight from (a horse)". Of similar concern, "light" as visible radiation is mapped to 24 different terms in Polish, and the synset with "illuminate" is given 20 equivalents in both Indonesian and Malaysian. While most languages have a lively list of expressions for some common concepts such as "goodbye", large sets of synonyms for most concepts indicate an overly broad brush in the Wordnet compilation. In the Polish example, the purported synonyms include a range of things related to brightness, such as "zaćmienie", which is an eclipse. As with poor English definitions, poor translations and clustering are unlikely to be fixed because their compilation projects have expired with no system in place for updating data.

These issues point to a third problem, a conceptual limitation that our concept-specific rearrangement of the data described above in section 2 seeks to address. A strength of Wordnet, and indeed its main organizing principle, is the highly detailed ontologies through which concepts are related (Vossen et al 1998, Vossen 1998)), such as hyponymy (this is a type of that) and meronymy (this is a part of that), e.g. a ship is a type of vessel and a deck is a part of a ship (Fellbaum 1998). These precise vertical ontologies are not matched, however, with a method for understanding horizontal distinctions within a synset (Derwojedowa et al 2008). Every term within a synset is defined as "this" same thing, e.g. E={approximate, estimate, gauge, guess, judge}, "judge tentatively or form an estimate of (quantities or time)," is all one notion.[5] Moreover, every term in every synset linked from every other language in GWN is bequeathed with the same meaning, in this example including 6 terms in Croatian, 11 in Japanese including orthographic variations, 20 in Arabic, 22 in Indonesian, and 24 in Malaysian; any term in {ثمّن, قضائ يا حكم, ر أيا كان, ثمن, قوارب, عليحكم, قوم, حزر, خَمَّنَ, خمن, فصل, تبأر, قدر, قيّم, حكم, قدّر, ظن, ماشى عسعة عين, قلس, اسدتذتچ} is equivalent to any term in {見立てる, 見積る, 予算+する, 目算, 積もる, 目算+する, 見積もる, 予算, 積る, 推算, 推算+する}. Where the English synset elides the large difference between guessing and gauging, the multilingual composite compounds the weakness of the assumption of strict equivalence. The Arabic terms do not all share a meaning with each other, nor are all the Japanese terms internal synonyms, leaving no way to determine whether اسدتذتچ is a viable translation for 積もる.[6] Any term produced by a

[2] http://babelnet.org/synset?word=bn:00050277n&details=1&orig=law%20practice&lang=EN

[3] http://visuwords.com/law%20practice

[4] http://www.vocabulary.com/dictionary/law%20practice

[5] http://wordnet-rdf.princeton.edu/wn31/200674352-v

[6] To evaluate these two blindly-chosen terms, bilingual informants translated both synsets, yielding information similar to what the processes in section 4 are designed to elicit. The Arabic term is substantially more definitive ("concluded") than the Japanese ("pile up like discussions during an absence"). {1. ثمّن, evaluated; 2. عليحكم, judged; 3. قوارب, compared; 4. ثمن, price; 5. ر أيا كان, had an idea about; 6. قضائ يا حكم, verdict; 7. قيّم, evaluated; 8. قدر, considered; 9. تبأر, focused; 10. فصل, separated; 11. خمن,

contributor in one language has a 1/E chance of being a direct translation of one of the English synset members, so any two cross-language terms in GWN have a $1/E^2$ chance of corresponding via the English intermediary with each other; in the example, E=5, any thoughtfully-produced term has a 20% of matching a specific term pertaining to assessing amounts, and any two non-English terms have a 4% chance of having been selected as best equivalents of the same English term. Linking the terms computationally is a prodigious shortcut to find likely pairs, but it is not lexicography.

If, however, we see the synset as a grouping of things that share a topical relationship rather than a strict meaning, we can resolve the problem by adding levels of detail similar to the vertical Wordnet ontologies. Kamusi splits the topical lumping of synonymy into what what can be seen as a two-tier horizontal ontology, joints and substitutes, that extends the conceptualization of a multilingual lexicon from a grid (Fellbaum and Vossen 2007) to a matrix.

1. "Joints" is the relationship that shows that terms have been linked transitively as synonyms (synset members) or translations. Joints are evaluated numerically by the degree of separation between links that have, in principle, some element of human confirmation.[7] A first generation joint indicates that two terms have been manually paired, a second generation joint links though one pivot term, third generation has two intermediary terms, etc. With data from GWN, the presumption of manual linking is cloudy; all members of an English synset have been manually linked to each other, all members of internal synsets for most other languages have been manually linked unless the Wordnet was assembled computationally, and most other-language synsets have been manually linked to the English synset, but that does not mean that اسدتذتج or 積もるhave been manually linked to "guess" or "gauge". In the current import, joints within a language are all shown as first generation (to be re-filtered as "synonyms" in due course), and joints between each term in an English synset and each member of a linked synset are also shown as first generation, i.e., اسدتذتج is said to be a first generation joint with both guess and gauge, as is 積もる, with the Arabic and Japanese terms therefore set as second generation. A future method to validate joints is described below in section 4.8.

2. "Substitutes" speaks to the degree of equivalence between terms. Whether in-language synonyms or cross-language translations, terms are either "parallel" or "similar", with the additional possibility that a translation is an "explanatory phrase" invented in one language to fill a lexical gap for a concept that is indigenous to another (Benjamin 2014b). Pending programming will provide fields on Kamusi similar to those for definitions. These fields provide space for the differences between "similar" substitutes to be elaborated, such as the distinction between "arm" in English that is the body part from the shoulder to the wrist versus "mkono" in Swahili that extends from the shoulder to the fingertips. Substitution relationships can in principle be followed across joint relationships, so that the degree of equivalence can be tracked along with the degree of separation, a task for future coding. For the data imported from OMW, all substitution relations have been set initially to "parallel", putting aside judgments about equivalence for a more distant future.

A fourth limitation with using Wordnet as a dictionary end-product is that it is incomplete in some essential ways. Wordnet cannot be faulted for not including every sense of every English term, much less every term from other languages, as that was never its mission. However, terms or senses that are not in Wordnet, such as "light" as a traffic signal, or "lightsaber", should be included – or at least includable – in a dictionary that

---

guessed; 12. خَمَّنَ, quantified; 13. حزر, guessed; 14. قوم, measured; 15. اسدتذتج , concluded; 16. قلس , measured; 17. ماشى عسعة ع ين, set capacity of; 18. ظن, doubted; 19. قدّر, evaluated; 20. حكم put to trial};{1. 見立てる to judge or diagnose [kanji for see and stand up] (make a visual estimation such as a physical exam, or take measurements for clothing); 2. 見積る, 3. 見積もる to estimate [kanji for see and stack] (predict price and time for a job); 4. 予算+する, 5. 予算 to estimate or budget [kanji for calculate and beforehand] (calculate anticipated expenses); 6. 目算, 7. 目算+する to estimate [kanji for calculate and look] (an inexact number such as ml in a cup or remaining moves in Go); 8. 積もる, 9. 積る to estimate [kanji for stack] (uncountable things such as snow or emotions); 10. 推算, 11. 推算+する estimation [kanji for calculate and guess] (less-knowable or unknowable things such as a coin flip, the size of a crowd, or evaluation of a crime scene)}.

[7] This assumption does not necessarily hold, as some Wordnets are built using automatic generation techniques (Atserias et al 1997, de Melo and Weikum 2008, Oliver 2014). The tendency for error in computationally-derived datasets is amply displayed WOLF French Wordnet (Wordnet Libre du Français) (Sagot and Fišer 2012, http://alpage.inria.fr/~sagot/wolf-en.html)

aspires toward a thorough representation of a language. If a concept is missing in PWN, moreover, it stands little chance of appearing in other language Wordnets, and conversely there is no chance for a concept indigenous to another language to join the global Wordnet concept set. Within the scope of the Wordnet vision, relationships that have not been found by Wordnet editors cannot be forged by readers, such as proposing that "boat" and "ship" be joined in a synset. Further, the lack of own-language definitions in most languages leaves the impression that the meaning of each term can be encapsulated in the English definition of the corresponding synset, to the extent that the attributed definition for "zaćmienie" is, exactly and erroneously, "electromagnetic radiation that can produce a visual sensation". Finally, and again because it is out of scope, Wordnet does not include a great deal of information that is relevant for dictionary or data purposes, such as word forms (Spanish "invitado" does not indicate an association with "invitada", "invitados", and "invitadas").

A final limitation with Wordnet is that projects for many languages have licenses that restrict the use of the data, if the data can be located at all. For example, the Romanian Wordnet is distributed with a "no derivatives" license. This means that the data cannot be imported into the multilingual structure described above, because linking Romanian to Slovenian would be a derivative product. Nor could the data be expanded, with Romanian definitions or with information such as the female form "invitátă" corresponding to the given masculine "invitát". Furthermore, the Romanian data has a "no redistribution" restriction, so its use in a project that makes its data shareable or downloadable seems proscribed. GermaNet is even more restrictive, only allowing the data to be used for internal research within an institution. The openness or lack thereof of Wordnets is indicated at http://globalwordnet.org/wordnets-in-the-world. Bringing restricted Wordnets into a dictionary project does not offer new technical challenges, but is only possible if the creators choose to amend their licenses.

4. Tools and techniques for adding and improving Wordnet data

Wordnet's popularity stems in part from its openness to the mash-ups others create from the core PWN data. In that spirit, Kamusi has developed tools that will transform the open Wordnet data into data that is appropriate for dictionaries and additional technological applications, using automated procedures as a starting point for human lexicographic review (Pianta, Bentivogli, and Girardi 2002). At the same time, these tools are designed to keep the data in synch with existing Wordnet instances, in such a way that transformations generated by Kamusi can be reincorporated in PWN or other language projects when and if their maintainers desire.

The primary new tools developed by Kamusi that can transform Wordnet data are a set of crowd-sourcing applications that include games embedded within Facebook and (still in alpha development) on mobile devices (Benjamin 2014a, Benjamin 2015). These games ask players to answer targeted questions about their language, for which they receive various rewards when their answers adhere to the consensus. The games build data progressively, such that a definition that has been approved for English can be shown to people producing equivalents or definitions for other languages.

These systems can transform Wordnet seed data into dictionary data, in several ways:

1. Each English definition will be reviewed as it pertains to the individual members of a synset, and improved when the participants find it appropriate. Players are shown the existing Wordnet "working definition", and given the opportunity to either suggest their own definition, vote for the Wordnet definition, or vote for a contribution from another player. Once a definition passes the consensus threshold, it is published to Kamusi and used for subsequent game modes. If the Wordnet definition has been replaced, it is shown on Kamusi as deprecated.
2. Definitions in their own languages for terms from other Wordnets will be generated using the same procedure. This feature will be introduced after players have had the chance to validate existing translations against a critical mass of finalized English definitions, e.g. a new English definition for "law practice" will first be given to Spanish speakers to verify or replace the current matched Spanish term, and only afterwards will the approved Spanish term be advanced to the definition game.
3. Existing translations of PWN will be validated term by term. For example, Polish players will assuredly approve "światło" for the sense of visible light, but reject "zaćmienie". This mode has not been developed at time of writing, the

need only becoming evident through examination of the data imported in mid 2015, but is anticipated for quick completion. Terms that are evicted from a defined synset, like "zaćmienie", will be moved through a sequence of games to produce definitions, translations, and sense matches.

4. Concepts from PWN that are not already matched in other Wordnet languages will be elicited. For example, the Arabic WordNet has only 10,000 synsets, so more than 100,000 concepts remain untouched. In the game, players are shown a defined English term and asked to provide an equivalent term in their language. Terms that pass the consensus threshold are added to Kamusi, while non-winning terms are passed to another mode to see whether they are synonyms for the concept.

5. Languages that do not have existing Wordnet projects will be opened to their speakers, using the improved English definition set and the game modes described above. Because the elicitation list used in the games is inherently linked to Wordnet, Wordnets for these other languages will be created as a default outcome. This opens GWN to languages that do not have formal organizations to take on the trouble of creating a Wordnet project, including building tools from scratch (e.g. Wijesiri et al 2014), but do have passionate speakers who will contribute through crowd methods.

6. Languages that have existing but restricted Wordnet projects, like German, will be opened for their speakers to start from scratch. This is a phenomenal waste of time and energy, if one can speak frankly in an academic paper, but, barring changes in license restrictions, may be the fastest way to acquire reliable data that can be used in an open resource.

7. One already-developed game calls on players to judge whether usages gleaned from Twitter or more formal corpora (currently configured for Wikipedia and the Helsinki Corpus of Swahili, but the technique can be applied more widely) are good examples to illustrate a particular sense. Most Wordnets lack usage examples, so this game can fill that gap for many languages. Future game modes will elicit additional lexical and ontological information, some of which falls within the scope of what is sought within Wordnets.

8. A future game mode, which will be activated after languages have sufficient numbers of defined entries, will ask users to confirm joints established through English for their language pairs. For example, "światło" and "lumière" will be shown with their respective own-language definitions, and a registered Polish/ French speaker will vote whether the two concepts match. This game can only be played after sufficient data for the concerned languages has been gathered in the English-confirmation mode described above in paragraph 4.3. The result will be validated aligned Wordnets for numerous language pairs.

9. Work on other tracks within Kamusi will introduce many terms and senses that are not part of PWN or other Wordnets. These concepts will be made available to language teams, and some could form part of an extended multilingual Wordnet desiderata.

5. Conclusions

This paper has discussed two difficulties with using Global Wordnet as the source for a formal multilingual dictionary. First, Wordnet does not do things it was not intended to do, but that are needed for lexicography, such as differentiation of terms grouped topically in synsets and matching those concept distinctions across languages. Second, some of the things it does do bear improvement, either in quantity (completion of the full PWN set of synsets in other languages, production of own-language definitions), quality, or access. Fortunately, the open approach with which Wordnet was designed makes it possible to retrofit the data with English definitions that may be more sensible than those initially drafted, and with revised equivalents in other languages when necessary, without severing the bonds that have already been built across languages and projects. The broad inter-lingual predictions made possible by GWN have been refined by charting the joints between members of a topical group, and will further show the degree to which confirmed pairs can substitute for each other. The work will not be easy, involving recruiting many crowd members from many languages, as well as oversight from authoritative arbiters. However, many of the tools have already been developed, and are being rolled out gradually as Kamusi musters the resources to foster speaker communities and manage the incoming data flow. As time goes on, the data produced by various Wordnet projects will lie at the core of a more comprehensive multilingual dictionary, and the data from the dictionary project will be available for the further refinement of existing and future Wordnets.

References

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. 1997. Combining multiple methods for the automatic construction of multi-lingualWordNets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, pages 327–338.

Martin Benjamin. 2014a. Collaboration in the Production of a Massively Multilingual Lexicon. In *LREC 2014 Conference Proceedings*. Reykjavik (Iceland).

Martin Benjamin. 2014b. Elephant Beer and Shinto Gates: Managing Similar Concepts in a Multilingual Database. In *Proceedings of the 7th International Global WordNet Conference*. Tartu (Estonia)

Martin Benjamin. 2015. Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. In *AsiaLex 2015 Conference Proceedings*, Hong Kong.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, Concepts and Relations in the Construction of Polish WordNet. In *Proceedings of the Global WordNet Conference 2008*, Szeged (Hungary), pages 167–68.

Christiane Fellbaum. 1998, ed.. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Christiane Fellbaum and Piek Vossen. 2007. Connecting the Universal to the Specific: Towards the Global Grid. In *Proceedings of the First International Workshop on Intercultural Communication*, Kyoto (Japan).

Gerard de Melo and Gerhard Weikum. 2008. On the utility of automatically generated wordnets. In *Proceedings of 4th Global WordNet Conference, GWC 2008*, Szeged, Hungary. University of Szeged. Pages 147–161.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller and Christiane Fellbaum. 2007. WordNet then and now. In *Language Resources and Evaluation*, 41:209–214.

Roberto Navigli and Simone Paolo Ponzetto. 2010. *BabelNet: building a very large multilingual semantic network*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 216–225, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858704.

Antoni Oliver. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. In *Proceedings of the 7th International Global WordNet Conference*, Tartu (Estonia), pages 7-15.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st Int'l Conference on Global WordNet*, Mysore (India), 293-302.

Benoît Sagot and Darja Fišer. 2012. Automatic Extension of WOLF. In *Proceedings of the 6th International Global Wordnet Conference*, Matsue, (Japan)

Piek Vossen, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. *The EuroWordNet Base Concepts and Top Ontology*. Deliverable D017 D, 34:D036.

Piek Vossen. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32(2-3):73–89.

Indeewari Wijesiri et al. 2014. Building a WordNet for Sinhala. In *Proceedings of the 7th International Global WordNet Conference*. Tartu (Estonia)

Ancient Greek WordNet meets the Dynamic Lexicon: the example of the fragments of the Greek Historians

Monica Berti
Gregory R. Crane
Tariq Yousef
Institute of Computer Science,
University of Leipzig
Leipzig, Germany
{name.surname}@uni-leipzig.de

Yuri Bizzoni
Federico Boschetti
Riccardo Del Gratta
CNR-ILC "A. Zampolli",
Via Moruzzi 1
Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

The Ancient Greek WordNet (AGWN) and the Dynamic Lexicon (DL) are multilingual resources to study the lexicon of Ancient Greek texts and their translations. Both AGWN and DL are works in progress that need accuracy improvement and manual validation. After a detailed description of the current state of each work, this paper illustrates a methodology to cross AGWN and DL data, in order to mutually score the items of each resource according to the evidence provided by the other resource. The training data is based on the corpus of the Digital Fragmenta Historicorum Graecorum (DFHG), which includes ancient Greek texts with Latin translations.

1 Introduction

The Ancient Greek WordNet (AGWN) and the Dynamic Lexicon (DL), which will be illustrated in detail in the next sections (see sections 2 and 4), are complementary resources to study the Ancient Greek lexicon. AGWN is based on the paradigmatic axis provided by bilingual dictionaries, while DL is based on the syntagmatic axis provided by historical and literary texts aligned to their scholarly translations. Both of them have been created automatically and they need to be corrected and extended. In this specific case the data is taken from the Digital Fragmenta Historicorum Graecorum (DFHG), which is a corpus of quotations and text reuses of ancient Greek lost historians and their Latin translations provided by the editor Karl Müller (Berti et al., 2014 2015; Yousef, 2015)[1]. This corpus is part of LOFTS (Leipzig Open Fragmentary Texts Series) at the Humboldt Chair of Digital Humanities at the University of Leipzig. We have been using this collection because it is big enough to include many different sources preserving information about Greek historians. Instead of working with extant authors, the DFHG allows us to focus on specific topics related to ancient Greek lost historiography and on the language of text reuse within this domain. The working hypothesis is that the evidence provided by Dynamic Lexicon Greek - Latin pairs is relevant to score the Greek word - conceptual node (synset) associations in the Ancient Greek WordNet and, on the other hand, that the evidence provided by AGWN Greek word - Latin translations is relevant to score the DL Greek - Latin pairs.

2 Ancient Greek WordNet

The creation of the Ancient Greek WordNet has been outlined in (Bizzoni et al., 2014). It is based on digitized Greek-English bilingual dictionaries (in particular the Liddell-Scott-Jones and the Middle Liddell provided by the Perseus Project[2]): first, Greek-English pairs (Greek words and English translations) are extracted from the dictionaries; then, the English word is projected onto the Princeton WordNet (PWN) (Fellbaum, 1998). If the English word is in PWN, then its synsets are assigned to the Greek word; the same goes for its lexical relations with other lemmas. Thus AGWN is created "bootstrapping" data from different datasets. As a bootstrapped process, its result is quite inaccurate. For example, induced polysemy (from English) maps the Greek verb ἔχω *-échō-* over 170 English words (including "cut", "make", "brake" ...). On the contrary, when the English word is not in PWN, the Greek word of the pair is excluded from AGWN, thus strongly reducing the coverage of AGWN for the entire Greek lexicon to c.a 30%.

[1] http://opengreekandlatin.github.io/dfhg-dev/

[2] http://www.perseus.tufts.edu

Currently, AGWN is linked not only to PWN, but also to other WordNets, in particular to the Latin WordNet (LWN) (Minozzi, 2009) and to the Italian WordNet (IWN) (Roventini et al., 2003). The way these WordNets are interconnected follows the guidelines illustrated in (Vossen, 1998; Rodríguez et al., 2008), by using English as the bridge language. As a consequence, Greek and Latin and/or Greek and Italian are linked through the common sense(s) in English.

3 The conceptual structure of Ancient Greek WordNet

Sharing a unique conceptual network among different languages is a good solution when the civilizations expressed by those languages are very similar, due to the effects of the globalization. In this case, only few conceptual nodes must be inserted when a concept is lexicalized in the source language but not in the target language, and few nodes must be deactivated when a concept is only lexicalized in the target language, but not in the source language.

On the contrary, when the civilizations expressed by the source and the target languages are highly dissimilar, the conceptual network needs to be heavily restructured.

As illustrated in the introduction, the conceptual network of AGWN is originally based on PWN, but the glosses of the synsets and the semantic relations can be modified through a web interface.[3]

4 Dynamic Lexicon

The Dynamic Lexicon is an increasing multilingual resource constituted by bilingual dictionaries (Greek/English, Latin/English, Greek/Latin), which have been created through the direct automated alignment of original texts with their translations or through a triangulation with a bridge language.

The first version of the DL[4] is a National Endowment for the Humanities (NEH)[5] co-funded project developed at Tufts University (Medford, MA) by the Perseus Project, whereas the second version is under development at the University of Leipzig by the Open Philology. Project[6]

[3] http://www.languagelibrary.eu/new_ewnui

[4] http://nlp.perseus.tufts.edu/lexicon

[5] http://www.neh.gov/about

[6] http://www.dh.uni-leipzig.de

5 Bilingual Dictionary Extraction

This section investigates a simple and effective method for automatic extraction of a bilingual lexicon (Ancient Greek/Latin) from the available aligned bilingual texts (Greek/English and Latin/English) in the Perseus Digital Library using English as a bridge language.

The data comes from the corpus of the DFHG and consists of 163 parallel documents aligned at a word level (104 Ancient Greek/English files and 59 Latin/English). The Greek-English dataset consists approximately of *210K* sentence pairs with *4,32M* Greek words, whereas the Latin-English dataset consists approximately of *123K* sentence pairs with *2,33M* Latin words. The parallel texts are aligned on a sentence level using Moore's Bilingual Sentence Aligner (Moore, 2002), which aligns the sentences with a very high precision (one-to-one alignment).[7] Then the GIZA++ toolkit[8] is used to align the sentence pairs at the level of individual words. Table 1 introduces statistics about the DFHG parallel corpus, while Figure 1 displays the used workflow. Note that the number of words in Table 1 is the total number of words in the documents, whereas the aligned pairs are the number of aligned words in the documents. Some words are not aligned at all, therefore the number of aligned words is smaller than the total number of words.

	Ancient Greek	Latin
Files	104	59
Sentences	$210K$	$132K$
Words	$4,32M$	$2,33M$
Aligned words	$3,34M$	$1,71M$
Distinct words	$872K$	$575K$

Table 1: Size of the corpora.

5.1 Preprocessing

The data sets provided by the workflow in Figure 1 are available in XML format. Each document is identified (through an *id*) in the Perseus Digital Library and consists of sentences in the orig-

[7] Sentences have been segmented using punctuation marks excluding commas.

[8] GIZA++ is an extension of the program GIZA which was developed by the Statistical Machine Translation team at the Center for Language and Speech Processing at Johns-Hopkins University.

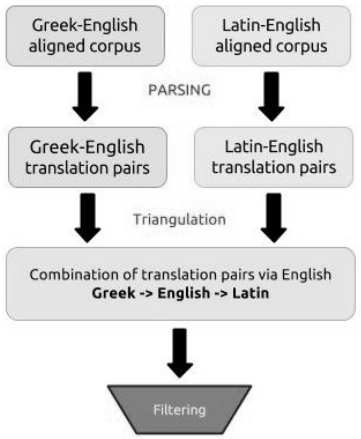


Figure 1: Explanation of the method

inal language (Ancient Greek or Latin) and their translation in English, as reported in Figure 2 (A). Each Latin or Greek word is aligned to one word in the English text (one-to-one Alignment), but in some cases a word in the original language could be aligned to many words (one-to-many / many-to-one) or not aligned at all, cf. Figure 2 (B).

Lemmatization of English translations will produce better results, because that will reduce the number of translation candidates as we can see in this example: The Greek word λέγειν *-légein-* is translated with ("say", "speak", "tell", "speaking", "said", "saying", "mention", "says", "spoke"). Many of the translation candidates share the same lemma (*say* for "said", "saying", "says"), (*speak*, "speaking", "spoken"). Before the lemmatization there were 9 translation candidates and after the lemmatization there are only four candidates, showing therefore the change of frequencies.

Table 2 shows how the lemmatization process recalculates the frequencies and percentages of each single translation.

5.2 Triangulation

Triangulation is based on the assumption that two expressions are likely to be translations if they are translations of the same word in a third language. We will use triangulation to extract the Greek-Latin pairs via English. In order to do that, we query our datasets to get the Greek and Latin words that share the same English translation along with their frequencies, see Figure 3.

The English word *ship* is associated to the Greek word ναῦς *-naûs-* (54.8%), to ναός *-naós-* (21.5%) and so on; the same English word *ship* is associated to the Latin word *navis* (65.3%), to *no* (23.8%), and so on.

Lemma	Freq.	%	Word	Freq.	%
say	719	46.8	say	551	36
			said	89	6
			saying	54	3.5
			says	25	1.5
speak	621	40.6	speak	492	32
			speaking	110	7
			spoke	19	1.2
tell	149	9.7	tell	149	9.7
mention	45	2.9	mention	45	2.9

Table 2: Lemmas and words:frequencies and percentages

The extracted pairs via triangulation are the correct association {ναῦς, *navis*} and the wrong associations {ναῦς, *no*} (*ship-to swim*), {ναός, *navis*} (*temple-ship*), {ναός, *no*} (*temple-to swim*). These pairs don't have the same level of relatedness, therefore we have to filter the results to keep only strong related pairs, as exposed in Section 5.3.

5.3 Translation-Pairs filtering

The translation pairs are not completely correct, because there are still some translation errors. In order to eliminate incorrect pairs, we will use a similarity metric to measure the similarity or the relatedness between every Greek-Latin pairs. The Jaccard coefficient (Jaccard, 1901) measures the similarity between finite sample sets (in our case two sets), and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \quad (1)$$

A and B in equation 1 are two vectors of translation probabilities (Greek-English, Latin-English). For example, the relatedness[9] between the Greek word πόλις and the Latin word *civitas* is reported in Figure 4.

We have to determine a threshold to classify the translation pairs as accepted or not accepted. High threshold yields high accuracy lexicon but with less number of entries, whereas low threshold produce more translation pairs with lower accuracy. The accuracy of the method depends on two factors:

[9]In the calculation we use the fact that *city* and *state* are shared English translation between πόλις *-pólis-* and *civitas*

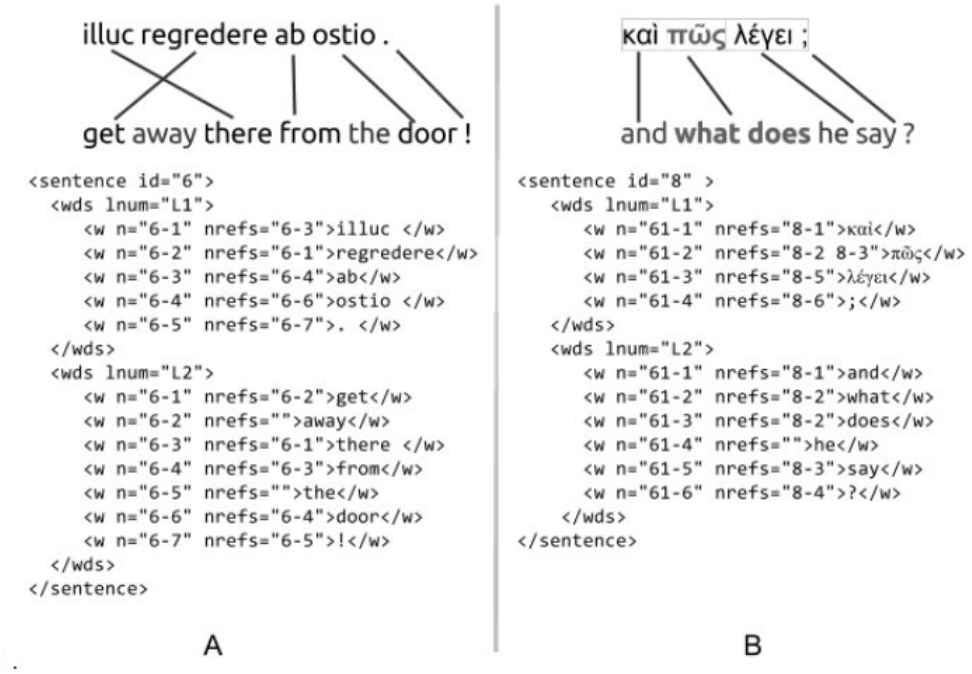


Figure 2: The aligned sentences in XML format

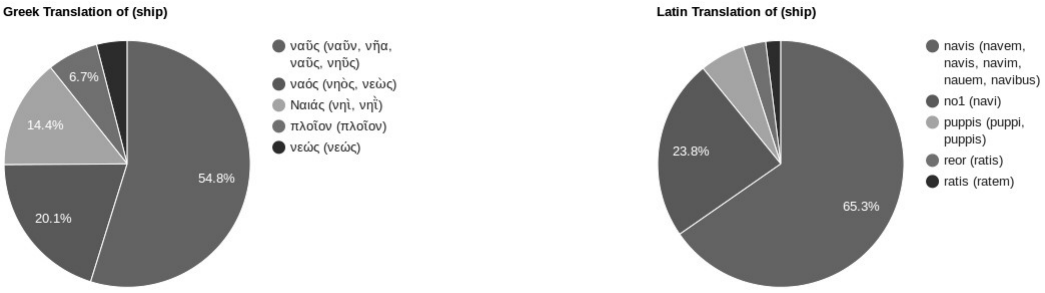


Figure 3: An example of triangulation

(πόλις civitas) = (72.9 + 19.5 + 74 +18.7) = 185.1
(πόλις civitas) = (100 + 100)=200
J(πόλις , civitas) = 185.1/200= 92.55 %

civitas	city	72.9%
civitas	state	19.5%
civitas	citizenship	2.9%
civitas	citizen	2.6%
civitas	country	2.1%

πόλις	city	74 %
πόλις	state	18.7 %
πόλις	athens	3 %
πόλις	town	3 %
πόλις	of	1.3 %

Figure 4: Use of Jaccard algorithm for aligning πόλις to *civitas*

The size of aligned-parallel corpora plays important role to improve the accuracy of the produced lexicon: bigger corpora produce better translation probability distribution and more translation candidates which yield a more accurate lexicon. In addition to that bigger corpora cover more words

The quality of the aligner used to align the parallel corpora: manually aligned corpora yield more accurate results, whereas automatic alignment tools produce some noisy translations; in our case GIZA++ has been used to align the parallel corpora.

6 Evaluating and extending the AGWN through evidence provided by the Dynamic Lexicon and vice versa

Students and scholars that evaluate and extend the AGWN synset items need to compare online dictionaries and other lexical resources. The DL can provide evidence for this purpose, especially to discover relevant missing correspondences. An example should clarify.

In AGWN we can find the association *minister* (eng) / *minister* (lat) / διάκτορος *-diáktoros-* (grc), but not *minister* (eng) / *minister* (lat) / διάκονος *-diákonos-* (grc), which is instead provided by the DL. If we consult the bilingual dictionary Liddell-Scott-Jones, we find out that διάκτορς "taken as minister, =διάκονος". The automatic parser used to bootstrap AGWN from bilingual dictionaries has not processed this information, so the DL provides a hint for the integration of this missed item in the correct synset of AGWN.

Complementary, the DL is missing the triplet *minister* (eng) / *minister* (lat) / διάκτορος (grc), which would be a relevant translation, even if not attested by the aligned bilingual texts of the training corpus. Moreover, AGWN can be used to add scoring criteria to the DL system, by tuning the results with a further piece of evidence, which reinforces the Jaccard score.

For example, the score of the correct association {ναῦς, *navis*}, discussed in Section 5.2 is reinforced, due to its presence in AGWN, whereas the scores of the wrong associations {ναῦς, *no*}, {ναός, *navis*} and {ναός, *no*} are weakened, due to their absence in AGWN.

7 Future work

The next step is the creation of a gold standard both for AGWN and for DL, in order to quantify the gain in terms of precision and recall that we can obtain by crossing AGWN and DL data.

8 Conclusion

In conclusion, we think that the paradigmatic approach, by extraction of bilingual pairs from dictionaries, and the syntagmatic approach, by extraction of bilingual pairs from aligned texts, are complementary for the study of Ancient Greek semantics and that they can be integrated, in order to mutually improve the performances of both of them.

References

Monica Berti, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova, and Gregory Crane. 2014-2015. The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors. *Journal of the Text Encoding Initiative*, 8.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The Making of Ancient Greek WordNet. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Stefano Minozzi. 2009. The Latin WordNet Project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*, pages 707–716.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Piek Vossen, and Christiane Fellbaum. 2008. Arabic Wordnet: Current State and Future Extensions. In *Proceedings of the Fourth International Global WordNet - Conference – GWC 2008*, pages 387–406, January.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Christian Girardi, Bernardo Magnini, Rita Marinelli, and Antonio Zampolli. 2003. ItalWordNet: building a large semantic database for the automatic treatment of italian. *Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI*, 2:745–791.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Tariq Yousef. 2015. Word Alignment and Named-Entity Recognition applied to Greek Text Reuse, school = Alexander von Humboldt Lehrstuhl für Digital Humanities, Universität Leipzig. Master's thesis.

IndoWordNet::Similarity
Computing Semantic Similarity and Relatedness using IndoWordNet

Sudha Bhingardive, Hanumant Redkar, Prateek Sappadla, Dhirendra Singh, Pushpak Bhattacharyya
Center for Indian Language Technologies,
Indian Institute of Technology Bombay, India
{bhingardivesudha, hanumantredkar, prateek2693, dhiru.research, pushpakbh}@gmail.com

Abstract

Semantic similarity and relatedness measures play an important role in natural language processing applications. In this paper, we present the IndoWordNet::Similarity tool and interface, designed for computing the semantic similarity and relatedness between two words in IndoWordNet. A java based tool and a web interface has been developed to compute this semantic similarity and relatedness. Also, Java API has been developed for this purpose. This tool, web interface and the API are made available for the research purpose.

1 Introduction

The Semantic Similarity is defined as a concept whereby a set of words are assigned a metric based on the likeliness of the semantic content. It is easy for humans with their cognitive abilities to judge the semantic similarity between two given words or concepts. For example, a human can quite easily say that the words *apple* and *mango* are more similar than the words *apple* and *car*. There is some understanding of how humans are able to perform this task of assigning similarities. However, measuring similarity computationally is a challenging task and attracts a considerable amount of research interest over the years. Another term very closely related to similarity is Semantic Relatedness. For example, *money* and *bank* would seem to be more closely related than *money* and *cash*. In past, various measures of similarity and relatedness have been proposed. These measures are developed based on the lexical structure of the WordNet, statistical information derived from the corpora or a combination of both. These measures are now widely used in various natural language processing applications such as Word Sense Disambiguation, Information Retrieval, Information Extraction, Question Answering, *etc.*

We have developed IndoWordNet::Similarity tool, interface and API for computing the semantic similarity or relatedness for the Indian Languages using IndoWordNet.

The paper is organized as follows. Section 2 describes the IndoWordNet. Semantic similarity and relatedness measures are discussed in section 3. Section 4 details the IndoWordNet::Similarity. Related work is presented in section 5. Section 6 concludes the paper and points to the future work.

2 IndoWordNet

WordNet[1] is a lexical resource composed of synsets and semantic relations. Synset is a set of synonyms representing distinct concept. Synsets are linked with basic semantic relations like hypernymy, hyponymy, meronymy, holonymy, troponymy, *etc*. and lexical relations like antonymy, gradation, *etc.* IndoWordNet (Bhattacharyya, 2010) is the multilingual WordNet for Indian languages. It includes eighteen Indian languages *viz.*, *Assamese*, *Bengali*, *Bodo*, *Gujarati*, *Hindi*, *Kannada*, *Kashmiri*, *Konkani*, *Malayalam*, *Manipuri*, *Marathi*, *Nepali*, *Odia*, *Punjabi*, *Sanskrit*, *Tamil*, *Telugu*, *Urdu*, *etc*. Initially, Hindi WordNet[2] was created manually taking reference from Princeton WordNet. Similarly, other Indian language Word-

[1] http://wordnet.princeton.edu/
[2] http://www.cfilt.iitb.ac.in/indowordnet/

Nets were created from Hindi WordNet using expansion approach and following the three principles of synset creation. In this paper, we present the IndoWordNet::Similarity tool, interface and API, which help in computing similarity and relatedness of words / concepts in Indian language WordNets.

3 Overview of Semantic Similarity and Relatedness Measures

Over the years, various semantic similarity and relatedness measures have been proposed. These measures are classified based on the path length, information content and the gloss overlap. Some of them are described below.

3.1 Path Length Based Measure

These measures are based on the length of the path linking two synsets and the position of synset the WordNet taxonomy.

3.1.1 Shortest Path Length Measure

This is the most intuitive way of measuring the similarity between two synsets. It calculates the semantic similarity between a pair of synsets depending on the number of links existing between them in the WordNet taxonomy. The shorter the length of the path between them, the more related they are. The inverse relation between the length of the path and similarity can be characterized as follows:

$$sim_{path} = \frac{1}{shortest\_path\_length(S_1,S_2)}$$

$$sim_{path} = 2 * D - shortest\_path\_length(S_1,S_2)$$

Where, S_1 *and* S_2 are synsets and *D* is the maximum depth of the taxonomy.

3.1.2 Leacock and Chodorow's Measure

This measure proposed by Leacock and Chodorow's (1998) computes the length of the shortest path between two synsets and scales it by the depth *D* of the *IS-A* hierarchy.

$$sim_{lch} = -\log(\frac{shortest\_path\_length(S_1,S_2)}{2 * D})$$

Where, S_1 and S_2 are the synsets and *D* represents the maximum depth of the taxonomy.

3.1.3 Wu and Palmer Measures

This measure proposed by Wu & Palmer (1994) calculates the similarity by considering the depths of the two synsets, along with the depth of the lowest common subsumer (LCS). The formula is given as,

$$sim_{wup} = \frac{2 * depth(LCS(S_1,S_2))}{depth(S_1) + depth(S_2)}$$

Where, S_1 and S_2 are the synsets and $LCS(S_1,S_2)$ represents the lowest common subsumer of S_1 and S_2.

3.2 Information Content Based Measure

These measures are based on the information content of the synsets. Information content of a synset measures the specificity or the generality of that synset, *i.e.* how specific to a topic the synset is.

3.2.1 Resnik's Measure

Resnik (1995) defines the semantic similarity of two synsets as the amount of information they share in common. It is given as,

$$sim_{resnik} = IC(LCS(S_1,S_2))$$

This measure depends completely upon the information content of the lowest common subsumer of the two synsets whose relatedness we wish to measure.

3.2.2 Jiang and Conrath's Measure

A measure introduced by Jiang and Conrath (1997) addresses the limitations of the Resnik measure. It incorporates the information content of the two synsets, along with that of their lowest common subsumer. This measure is given by the formula:

$$distance_{jcn}(S_1,S_2) = IC(S_1) + IC(S_2) - (2\,IC(LCS(S_1,S_2)))$$

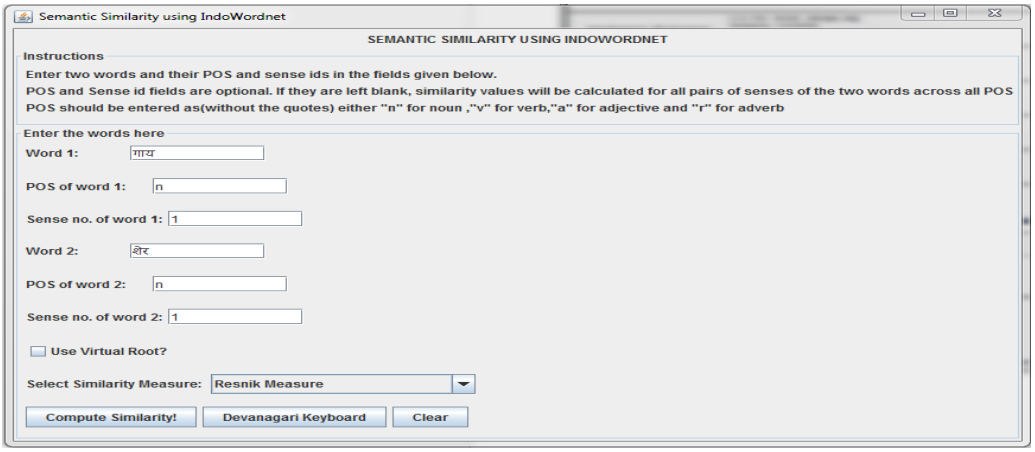


Figure 1: IndoWordNet::Similarity Tool

Where, *IC* determines the information content of a synset and *LCS* determines the lowest common subsuming concept of two given concepts.

3.3 Gloss Overlap Measures

Lesk (1986) defines the relatedness in terms of dictionary definition overlap of given synsets. Further, the extended Lesk measure (Banerjee and Pedersen, 2003) computes the relatedness between two synsets by considering their own gloss as well as by considering the gloss of their related synsets.

4 IndoWordNet::Similarity

We have developed IndoWordNet::Similarity tool, web based interface and API to measure the semantic similarity and relatedness for a pair of words / synsets in the IndoWordNet.

4.1 IndoWordNet::Similarity Tool

The IndoWordNet::Similarity[3] tool is implemented using Java. The user interface layout and its features are given below.

4.1.1 User Interface Layout

The main window of the tool is as shown in Figure 1. In order to use this tool, user needs to provide the following inputs:

- User can enter the pair of words for which similarity to be computed.
- User can specify the part-of-speech and the sense number for the given two words for calculating the similarity. If user doesn't provide these details then the tool computes the similarity between all possible pair of senses of the two input words over all parts-of-speech.
- Drop-box is provided for selecting the type of similarity measure.
- Check-box is provided for virtual root option.

Depending on the user query the similarity is calculated and displayed in an output window.

4.1.2 Features

- This is system independent portable standalone Java Application.
- Option such as part-of-speech and sense-id are optional.
- If user doesn't provide part-of-speech and sense-id option, then similarity is calculated for all possible pair of senses of the given words.
- If the virtual root node option is enabled then one hypothetical root is created which connects all roots of the taxonomy. This allows similarity values to be calculated between any pair of nouns or verbs.

4.2 IndoWordNet::Similarity API

IndoWordNet::Similarity Application Programming Interface (API) has been developed using Java which provides functions to compute the semantic similarity and relatedness using various measures. API provides three types of functions for each measure.

1. A function which takes only two words as

[3] http://www.cfilt.iitb.ac.in/iwnsimilarity

parameters and returns the similarity score between all possible senses of the two words.

2. A function which takes two words along with part-of-speech, sense-id and returns the similarity score between the particular senses as specified by the user.
3. A function which takes only two words as parameters and returns the maximum similarity between two words among all possible sense pairs. Some of the API functions are mentioned below:

API Function	Computes
public SimilarityValue[] getPathSimilarity(String word1, String pos1, int sid1, String word2, String pos2, int sid2, boolean use_virtual_root)	Path Similarity
public SimilarityValue[] getPathSimilarity(String word1,String word2,boolean use_virtual_root)	Path Similarity
public SimilarityValue getMaxPathSimilarity(String word1, String word2, boolean use_virtual_root)	Maximum Path Similarity

Table 1. Important functions of IndoWordNet::Similarity API

4.3 IndoWordNet::Similarity Web Interface

IndoWordNet::Similarity Web Interface has been developed using Php and MySql which provides a simple interface to compute the semantic similarity and relatedness using various measures. Figure 2 shows the IndoWordNet::Similarity web interface.

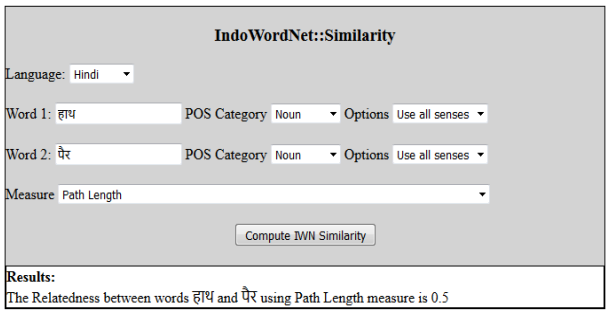


Figure 2. IndoWordNet::Similarity Web Interface

5 Related Work

WordNet::Similarity[4] (Pedersen *et. al*. 2004) is freely available software for measuring the semantic similarity and relatedness for English WordNet. This application uses an open source Perl module for measuring the semantic distance between words. It provides various semantic similarity and relatedness measures using WordNets. Given two synsets, it returns numeric score showing their degree of similarity or relatedness according to the various measures that all rely on WordNet in different ways. It also provides support for estimating the information content values from untagged corpora, including plain text, the Penn Treebank, or the British National Corpus[5].

WS4J[6] (WordNet Similarity for Java) provides a pure Java API for several published semantic similarity and relatedness algorithms. WordNet Similarity is also integrated in NLTK tool[7]. However, the need to make entirely different application for IndoWordNet lies in its multilingual nature which supports 19 Indian language WordNets. Hence, we developed the IndoWordNet::Similarity tool, web interface and API for calculating the similarity and relatedness.

6 Conclusion

We have developed the IndoWordNet::Similarity tool, web interface for computing the semantic similarity and relatedness measures for the IndoWordNet. Also, a java API has also been developed for accessing the similarity measures. The tool and the API can be used in various NLP areas such as Word Sense Disambiguation, Information Retrieval, Information Extraction, Question Answering, *etc*. In future, the other measures of computing similarity and relatedness shall be integrated in our tools and utilities.

References

Satanjeev Banerjee and Ted Pedersen. 2003. *Extended gloss overlaps as a measure of semantic relatedness*. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pages 805–810, Acapulco, August.

[4] http://wn-similarity.sourceforge.net/
[5] http://corpus.byu.edu/bnc/
[6] https://code.google.com/p/ws4j/
[7] http://www.nltk.org/howto/wordnet.html

Pushpak Bhattacharyya. 2010. *IndoWordnet*, Lexical Resources Engineering Conference (LREC 2010), Malta.

Jay Jiang and David Conrath. 1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings on International Conference on Research in Computational Linguistics, pages 19–33, Taiwan.

Claudia Leacock and Martin Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 265–283. MIT Press.

Michael Lesk. 1986. *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone*. In Proceedings of SIGDOC '86, 1986.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. *Wordnet::Similarity - Measuring the relatedness of concepts*. In Proceedings of AAAI04, Intelligent Systems Demonstration, San Jose, CA, July 2004.

Philip Resnik. 1995. *Using information content to evaluate semantic similarity in a taxonomy*. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal, August.

Zibiao Wu and Martha Palmer. 1994. *Verb semantics and lexical selection*. ACL, New Mexico.

Multilingual Sense Intersection in a Parallel Corpus with Diverse Language Families

Giulia Bonansinga
Filologia, Letteratura e Linguistica,
Università di Pisa, Italy
giuliauni@gmail.com

Francis Bond
Linguistics and Multilingual Studies,
Nanyang Technological University
bond@ieee.org

Abstract

Supervised methods for Word Sense Disambiguation (WSD) benefit from high-quality sense-annotated resources, which are lacking for many languages less common than English. There are, however, several multilingual parallel corpora that can be inexpensively annotated with senses through cross-lingual methods. We test the effectiveness of such an approach by attempting to disambiguate English texts through their translations in Italian, Romanian and Japanese. Specifically, we try to find the appropriate word senses for the English words by comparison of all the word senses associated to their translations. The main advantage of this approach is in that it can be applied to any parallel corpus, as long as large, high-quality inter-linked sense inventories exist for all the languages considered.

1 Introduction

Cross-lingual Word Sense Disambiguation (CL-WSD) is an approach to Word Sense Disambiguation (WSD) that exploits the similarities and the differences across languages to disambiguate text in an automatic fashion. Using existing multilingual parallel corpora for this purpose is a natural choice, as shown by a long series of works in the literature; see for instance Brown and Mercer (1991), Gale et al. (1992), Ide et al. (2002), Ng et al. (2003), Chan and Ng (2005), and Khapra et al. (2011) more recently.

As Diab and Resnik (2002) showed, the translation correspondences in a parallel corpus provide valuable semantic information that can be exploited to perform WSD. For instance, Tufiş et al. (2004) used parallel corpora to validate the interlingual alignments in different WordNets (WNs). Specifically, they looked at the sense intersection between the lexical items found in all the reciprocal translations of a parallel corpus.

Gliozzo et al. (2005) showed how CL-WSD can help to sense-annotate a bilingual corpus by looking at the semantic differences in a language pair. Bentivogli and Pianta (2005), on the other hand, focused on how meaning is somehow preserved despite those differences, which allows us to transfer the semantic annotation of a text in a certain language to its translation in another language. The *sense projection* procedure that they used is simple yet powerful, but it can only be applied on corpora in which at least one parallel text is annotated with senses. Nevertheless, any way to produce sense-annotated data is of great benefit to WSD given the difficulty to come across such data. This *knowledge acquisition bottleneck* is still a challenge to address for most languages.

Given the task of annotating an ambiguous word in a multilingual parallel corpus, some valuable information can be derived through the comparison of the *set of senses* of each of the word's translations. If fewer senses (or one only, in the optimal case) are retained across languages, then the cross-lingual information has helped reducing (or solving) the ambiguity.

In previous work (Bond and Bonansinga, 2015) *sense intersection (SI)*, to annotate a trilingual parallel corpus in English, Italian and Romanian built upon SemCor (SC) (Landes et al., 1998). We summarize the data used and our findings in Section 2.

In Section 3 we continue investigating in the same strand by introducing a further language, Japanese, to disambiguate English text. In Section 4 we show how an annotation task can benefit from coarser sense distinctions. In Section 5 we examine thoroughly how and how much each additional language helps the automatic sense annotation. We conclude in Section 6 and suggest some future work.

2 Multilingual Sense Intersection

In Bond and Bonansinga (2015) we explored the cross-lingual approaches pioneered by Gliozzo et al. (2005) and Bentivogli and Pianta (2005) to annotate the SC corpus (Landes et al., 1998) and two corpora built upon it from its Italian and Romanian translations. This parallel corpus, though rather small (see Subsection 2.1), is ideal for the task as it is sense-annotated in all its translations, thus making the evaluation of alternative sense annotation methods straightforward. We briefly present the data used back then and introduce the last component of the corpus, the Japanese SemCor (Bond et al., 2012), which is included in the analysis presented in this paper.

2.1 Data

Developed at Princeton University, SC is a subset of the Brown Corpus of Standard American English (Kučera and Francis, 1967) enriched with sense annotations referring to the WN sense inventory (see Section 2.2).

Bentivogli and Pianta (2005) manually translated 116 SC texts and automatically aligned them to their English counterparts. Then the sense annotations of the English words were automatically transferred following the word alignment, thus leading to the creation of a sense-annotated English-Italian corpus, MultiSemCor (MSC).

With the purpose of providing a Romanian version of SC, Lupu et al. (2005) developed the Romanian SemCor (RSC) (Lupu et al., 2005; Ion, 2007), which shares 50 texts with MSC. Unfortunately, RSC is not word-aligned to any other component of the parallel corpus, which is a requirement to perform sense mapping with any of the mentioned procedures. On the other hand, the sentence alignment is available and we are only interested in content words, so we attempted a word alignment based upon the information already available. First, we aligned all reciprocal translations in the same sentence pair with identical sense annotation. Then, we aligned the remaining content words, if any, using heuristics that exploit PoS information and path similarity in the WN ontology. Finally, we manually checked a sample of the alignment found this way and we observed a precision of 97%; of course, errors can only be introduced in the second step, when using heuristics to align the remaining unaligned content words.

Bond et al. (2012) built a Japanese SemCor (JSC) matching the texts covered in MSC, after porting the sense annotations to WN 3.0 using the mappings provided by Daude et al. (2003). The sense annotation was carried out through sense projection by exploiting the word alignment, similarly to what Bentivogli and Pianta (2005) did for Italian. 58,265 annotations were automatically transferred to Japanese content words.

JSC follows the Kyoto Annotation Format (KAF) (Bosma et al., 2009) and is released under the same license as SC.[1]

In Table 1 we remind the basic statistics of each corpus. For English and Italian we also specify the number of the target words after the migration to WordNet 3.0 (WN 3.0). In Table 2.1, we show more clearly, in terms of number of sentences, the alignments available for each language pair.

	Texts	Tokens	Target words	After mapping
EN	116	258,499	119,802	118,750
IT	116	268,905	92,420	92,022
RO	82	175,603	48,634	=
JP	116	119,802	150,555	=

Table 1: Statistics for each component of the multilingual parallel corpus built from SemCor.

2.2 Sense Inventories

When MSC was released, MultiWordNet[2] (MWN) (Pianta et al., 2002), a multilingual WordNet aligned to Princeton WN 1.6, was used. As described in Bond and Bonansinga (2015), we ported all senses annotations in MSC to WN 3.0, so to make it possible a comparison between

[1] Both the Japanese WordNet and the Japanese SemCor are available at the following address: `http://compling.hss.ntu.edu.sg/wnja/index.en.html`

[2] `http://multiwordnet.fbk.eu/`

Language	Aligned sentences
EN-IT	12,842
EN-RO	4,974
EN-JP	12,781
IT-RO	4,974
IT-JP	12,781
RO-JP	4,913

Table 2: Number of aligned sentences for each language pair.

the different components of the parallel corpus. To this aim, we used automatically inferred mappings (Daudé et al., 2000; Daudé et al., 2001). However, the changes occurred between WN versions 1.6 and 3.0 led to the loss of 4,631 sense annotations (1,204 types, half of which are adjective satellites).

The Romanian WordNet (RW), created within the BalkaNet project (Stamou et al., 2002) and then consistently grown independently (Barbu Mititelu et al., 2014), includes synsets mapped to WN 3.0 with precision of 95% (Tufiş et al., 2013).

The Japanese WN (JWN) (Isahara et al., 2008; Bond et al., 2009a; Bond et al., 2009b), originally developed by the National Institute of Information and Communications Technology (NICT) and firstly released in 2009, is a large-scale semantic dictionary of Japanese available under the WordNet license.

	Synsets	Senses
English	117,659	206,978
Italian	34,728	69,824
Romanian	59,348	85,238
Japanese	57,184	158,069

Table 3: Coverage of the WNs used.

In Table 3 we give basic coverage statistics for the WNs of our target languages. The Open Multilingual WordNet (OMW)[3] is an open-source multilingual database that connects all open WNs linked to the English WN, including Italian (Pianta et al., 2002) among the 28 languages supported (Bond and Paik, 2012; Bond and Foster, 2013). A convenient interface to OMW is provided by the Python module NLTK[4] (Bird et al., 2009).

2.3 Findings

For the sake of completeness, in previous work we performed sense projection on the Italian and Romanian corpora using English as pivot, scoring a precision of over 90% in both cases. As for SI, we report the previous precision and coverage scores obtained through trilingual SI in Table 4, along with the Most Frequent Sense (MFS) baseline, that assigns each word its most frequent sense. In this step, sense frequency statistics (SFS) are therefore necessary, but unfortunately there are very few sense-annotated corpora from which we can derive such statistics. In the case of SC the issue is even more crucial, because WN SFS are computed on SC itself. So, whenever the first sense of a lemma follows a ranking order, we are using biased statistics.

Generally speaking, the coverage scores were quite good and higher with the baseline MFS. As for precision, the gap between SI and the baseline is smaller, probably due to the bias just mentioned. On the other hand, in languages other than English, the contribution of SFS is not as decisive and SI performs better than the baseline, and particularly so in the case of Italian.

3 Multilingual Sense Intersection with languages from different families

The theoretical justification behind Multilingual Sense Intersection (SI) is in that an ambiguous word will often be translated in different words in another language. As a consequence, the knowledge of all the senses associated to its translation can help detect the sense actually intended in the original text. More commonly, such a comparison will help reduce the ambiguity, but it will not identify one single, shared sense. On the other hand, a text whose ambiguity was progressively reduced through automatic methods can be completely disambiguated by a human annotator at a lesser cost. Moreover, the more the languages available for comparison in the parallel corpus, the more likely is that SI actually manages to discern the correct sense in context.

Differently from our previous work, where we disambiguated all texts aligned with at least one other language, in the following section we show results computed over 49 texts. Those are the subset of the corpus shared across all four components and for which we have alignments. The result is an even smaller corpus, but it can show more effectively the contribution of up to three languages.

Given an ambiguous word, all its translations provide their "set of senses", as retrieved from the shared sense inventory. Then, intersection is performed over every non-empty set and successes when the final *overlap* contains only one sense, meaning that the target word has been disambiguated. Otherwise, the overlap is further intersected with the top most frequent senses available for the target lemma, and we take note whether the sense selected was the most frequent one. As be-

[3] http://compling.hss.ntu.edu.sg/omw/summx.html

[4] http://www.nltk.org

Method	English		Italian		Romanian	
	Precision	Coverage	Precision	Coverage	Precision	Coverage
MFS (baseline)	**0.761**	**0.998**	0.599	**0.999**	0.531	**1**
3-way Intersection	0.750	0.778	**0.653**	0.915	**0.590**	1
Coarse-grained MFS	**0.850**	**0.998**	0.687	**0.999**	**0.794**	**1**
Coarse-grained SI	0.849	0.778	**0.761**	0.915	0.661	1

Table 4: Comparison of the results scored with SI and MFS baseline.

fore, we resort to sense frequency statistics (SFS) whenever the target word is not yet disambiguated after SI. These frequencies were calculated over all texts in the corpus **except** the one being annotated.

4 Introducing coarse-grained senses

Sense inventories are a crucial part of this approach. Not only a sufficient coverage and the alignment to the Princeton WN are necessary; when it comes to deciding how to define close, very specific senses, a trade-off between the detail of the sense description and its actual usability in real contexts is desirable.

The fine granularity of WN senses can occasionally, depending on the application, be more of a practical disadvantage than a quality. In this analysis, for instance, error analysis suggested that the senses found through SI were often very close, but it may happen that they are discarded as wrong outputs just because one language has a WN more developed and granular than another. We should also bear in mind that the correct senses against which we evaluate were picked by trained human annotators in the first place, and human annotators tend to describe a word as precisely as possible.

Conscious of this limit, Navigli (2006) devised an automatic methodology to find a reasonable sense clustering for the senses in WN 2.1. Sense clustering can be of great help when minor sense distinctions can be ignored, allowing a coarse-grained evaluation.

They found 29,974 main clusters, some of which were manually validated by an expert lexicographer for the Semeval all-word task.

We mapped the senses in the clusters found to WN 3.0, losing 101 of them in the process (typically one-element clusters). When evaluating the results of SI, we performed a coarse-grained evaluation; in particular, whenever the sense found by SI was not correct, we checked whether it was part of a sense cluster and whether the correct sense was in it. If so, we considered the output of the algorithm correct.

Table 4 displays the difference in performance when coarse-grained evaluation is employed.

Method	English	
	Precision	Coverage
Coarse-grained MFS	0.851	**0.998**
Coarse-grained 4-SI	**0.854**	0.788

Table 5: Coarse-grained evaluation of the results scored with 4-way SI and MFS baseline, computed over the shared subset (49 texts).

5 Evaluation

In Table 4 we show the improvement in precision obtained thanks to coarse-grained evaluation with respect to the results in (Bond and Bonansinga, 2015). English and Italian show respectively a significant improvement of 0.1 and 0.11. In the case of Romanian, the improvement is not as big, but still meaningful (0.07). Of course, coarse-grained evaluation causes the MFS baseline to improve as well. In the case of English - which, again, is the component most subjected to the bias introduces by SFS - the difference between decreases a little in perspective, ma MFS still beats SI. The case of Italian is unique, in that in both cases, with fine and coarse-grained senses, SI obtains better precision scores. For Romanian, on the other hand, SI performs better until coarse-grained evaluation is employed, and the improvement achieved by MFS is striking.

In Table 5 we show our latest attempt to disambiguate English text by using the semantic information of its aligned translation in a parallel corpus. The languages that contribute to the disambiguation process are Italian, Romanian and Japanese, and all together they manage to beat MFS, if coarse-grained senses are considered.

6 Conclusions

For future work, it is important to analyze the progressive improvement that we can achieve by taking into account semantic information from one language at the time, so as to verify if it is true that very diverse languages contribute the most to the disambiguation process.

As for the sense inventories, it would be interesting to compare different lexical resources for Italian, that is MWN and ItalWordNet (ITW) (Roventini et al., 2002). ITW was born as the EuroWordNet Italian database, but even though compatible to a certain extent with EuroWordNet, it is released in XML format. ITW includes about 47.000 lemmas, 50.000 synsets and 130.000 semantic relations and is currently maintained by the Computational Linguistics Istitute (ILC) at the National Research Council (CNR). An updated version is freely available online. [5]

Finally, we could easily address, at least for English, the lack of unbiased sense frequency statistics by computing them over the WordNet Gloss Corpus, in which glosses are sense-annotated.[6] This corpus alone would provide sense frequencies for 157,300 lemma-pos pairs.

Acknowledgments

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union (2010-5094-7) and the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Verginica Barbu Mititelu, Stefan Daniel Dumitrescu, and Dan Tufiş, 2014. *Proceedings of the Seventh Global Wordnet Conference*, chapter News about the Romanian Wordnet, pages 268–275.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(03):247, September.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

[5] http://datahub.io/dataset/iwn

[6] http://wordnet.princeton.edu/glosstag.shtml

Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In *Second Italian Conference on Computational Linguistics CLiC-it 2015*. to appear.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2012. A Survey of WordNets and their Licenses. In *GWC 2012*, pages 64–71.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009a. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 1–8. Association for Computational Linguistics.

Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009b. Extending the japanese wordnet. In *15th Annual Meeting of The Association for Natural Language Processing*.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63.

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009*, Pisa, Italy.

Stephen A. Della Pietra Vincent J Della Pietra Brown, Peter F. and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the ACL*, Morristown, NJ.

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *AAAI*, volume 5, pages 1037–1042.

Jordi Daudé, Lluís Padró, and German Rigau. 2000. Mapping wordnets using structural information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*, Hong Kong.

Jordi Daudé, Lluís Padró, and German Rigau. 2001. A complete WN1.5 to WN1.6 mapping. In *Proceedings of NAACL Workshop" WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA*.

Jordi Daude, Luiss Padro, and German Rigau. 2003. Validation and tuning of wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods.

Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava. 2005. Crossing parallel corpora and multilingual lexical databases for WSD. In *Computational Linguistics and Intelligent Text Processing*, pages 242–245. Springer.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 61–66. Association for Computational Linguistics.

Radu Ion. 2007. *Metode de dezambiguizare semantica automata. Aplicat ii pentru limbile englezas i romana ("Word Sense Disambiguation methods applied to English and Romanian").* Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, Bucharest.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the japanese wordnet.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for wsd. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 561–569, Stroudsburg, PA, USA. Association for Computational Linguistics.

Henry Kučera and W. Nelson Francis. 1967. Computational analysis of present-day American English.

Shari Landes, Claudia Leacock, and Randee I Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press, Cambridge, MA.

Monica Lupu, Diana Trandabat, and Maria Husarciuc. 2005. A Romanian SemCor aligned to the English and Italian MultiSemCor. In *1st ROMANCE FrameNet Workshop at EUROLAN*, pages 20–27.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Rita Marinelli, Bernardo Magnini, Manuela Speranza, and Antonio Zampolli. 2002. Italwordnet: a large semantic database for the automatic treatment of the italian language. In *First International WordNet Conference*.

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. 2002. Balkanet: A multilingual semantic network for the balkan languages. *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Word sense disambiguation as a wordnets validation method in balkanet. In *Proceedings of the 4th LREC Conference*, pages 741–744.

Dan Tufiş, Verginica Barbu Mititelu, Dan Ştefănescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language Resources and Evaluation*, 47(4):1305–1314, December.

CILI: the Collaborative Interlingual Index

Francis Bond,[♠] **Piek Vossen,**[♢] **John P. McCrae**[♣] and **Christiane Fellbaum**[♡]
[♠]Nanyang Technological University, Singapore
[♢]VU University Amsterdam, The Netherlands
[♣]Insight Centre for Data Analytics, NUI Galway, Galway, Ireland
[♡]Princeton University, U.S.A.
<bond@ieee.org,piek.vossen@vu.nl,john@mccr.ae,fellbaum@princeton.edu>

Abstract

This paper introduces the motivation for and design of the Collaborative InterLingual Index (CILI). It is designed to make possible coordination between multiple loosely coupled wordnet projects. The structure of the CILI is based on the Interlingual index first proposed in the EuroWordNet project with several pragmatic extensions: an explicit open license, definitions in English and links to wordnets in the Global Wordnet Grid.

1 Introduction

Within 10 years of the release of Wordnet (Miller, 1990) researchers had started to extend it to other languages (Vossen, 1998). Currently, the Open Multilingual Wordnet (OMW: Bond and Paik, 2012; da Costa and Bond, 2015) has brought together wordnets for 33 languages that have released open data,[1] and automatically produced data for 150. There are even more wordnets than this: some large projects have released non-open data, notably German (Kunze and Lemnitzer, 2002) and Korean (Yoon et al., 2009) and many projects have yet to release any. This activity shows that the structure of wordnets is applicable to many languages.

All the wordnets are based on the basic structure of the Princeton wordnet (PWN: Fellbaum, 1998): synonyms grouped together into **synsets** and linked to each other by semantic relations.

The majority of wordnets have been based on the **expand** approach, that is adding lemmas in new languages to existing PWN synsets (Vossen, 1998, p83), boot-strapping from the structure of English. 28 out of 33 of the wordnets in OMW take this approach. A few wordnets are based on the **merge** approach, where independent language specific structures are built first and then some synsets linked to the PWN. In OMW, only five projects take this approach: Chinese (Taiwan), Danish, Dutch, Polish and Swedish (Huang et al., 2010; Pedersen et al., 2009; Postma et al., 2016; Piasecki et al., 2009; Borin et al., 2013).

To investigate meaning across languages, we need to link synsets cross-lingually. It is easy to link expand-style wordnets: they all link to PWN and it can be used as a **pivot** to link them together. This is one of the attractions of using the expand approach, you immediately gain multilingual links. The disadvantage is that concepts not in PWN (either because they are not lexicalized in English or just because they have not been covered yet) cannot be expressed. Because of this, many expand-style wordnets also define some new, language-specific synsets, typically a few tens or hundreds (Arabic, Chinese, Italian, Japanese, Catalan, Spanish, Galician, Finnish, Malay/Indonesian, Bulgarian, Greek, Romanian, Serbian and Turkish all do so)(Pianta et al., 2002; Tufiş et al., 2004; Elkateb and Fellbaum, 2006; Gonzalez-Agirre et al., 2012; Wang and Bond, 2013; Bond et al., 2014; Seah and Bond, 2014; Postma et al., 2016).

It is harder to link merge-style wordnets. The projects need to somehow identify links to PWN, and as a result, only a small subset of the language specific synsets are linked to PWN. Examining the unlinked synsets, this seems to be principally due to the lack of resources to link them than semantic incompatibility. For example, Danish and Polish (Pedersen et al., 2009; Piasecki et al., 2009) have many synsets which can be linked but are not currently.

Currently, when projects create their own synsets, there is no coordination between these projects. This means that similar or even identical concepts may be introduced in multiple places.

[1]We use the definition from the Open Knowledge Foundation: `http://opendefinition.org/`: ``anyone is free to use, reuse, and redistribute it --- subject only, at most, to the requirement to attribute and/or share-alike".

For example, most South East Asian languages distinguish between cooked and uncooked rice: these concepts have been added independently to the Korean and Japanese wordnets. Typically, clusters of projects have tried to coordinate, such as EuroWordNet, the Multilingual Central Repository for Basque, Catalan, Galician and Spanish (Gonzalez-Agirre et al., 2012), the MultiWordNet for Italian and Hebrew, (Pianta et al., 2002), Balkanet (Tufiş et al., 2004), the Wordnet Bahasa for Malay and Indonesian (Bond et al., 2014), the IndoWordnet project (Bhattacharyya, 2010).

Clearly, there is a need for a single shared repository of concepts. In this paper, we propose to build one: the **Collaborative InterLingual Index** (CILI). We base the index on the technical foundations laid down in EuroWordNet: a single list that is the union of all the synsets in all the wordnets (Peters et al., 1998; Vossen et al., 1999). To this we add ideas from the best-practice of the Semantic Web: a shared easily accessible resource with a well defined license; from open-source software: build a community of users who will co-develop the resource; and from experiences in many multilingual lexical projects: accept the *de facto* use of English as a common language of communication.

In the following sections we discuss the motivation further (§ 2), then describe in detail the structure of the CILI (§ 3), list some open issues (§ 4) and finally conclude.

2 Motivation

Wordnets have been built with different methods and from different starting points: expand or merge, manually or semi-automatically and based on pre-existing monolingual resources or using available bilingual resources to translate English synsets to words in the target language. Furthermore, it is up to the wordnet builders to make decisions about which words are synonyms, what are the semantic relations between the synsets and how to interpret each semantic relation. We can observe very large synsets in one wordnet being linked through PWN to small synsets in another language. Different granularities of synsets brings into questions the notion of the same concept existing across these wordnets. PWN uses 44 semantic relations (if separated by part-of-speech) but in EuroWordNet 71 relations were defined that partially overlap. Even if two wordnets use the same relation name, there is no guarantee that it is interpreted in the same way. In fact, different wordnet editors and algorithms may interpret relations differently. Even the symbols used for parts of speech differ in different projects (**adverb** is 'r' in PWN but 'b' in some projects). Finally, one can observe large differences in coverage of the vocabulary and in the degree of polysemy. Vocabularies and concepts differ in size but also in terms of genre, pragmatics, the inclusion of multiword expressions as ``phrase sets" (Bentivogli and Pianta, 2003) and specific domains and areas. Choices for distinguishing senses lead to fine-grained and coarse-grained polysemy, where the latter may lead to multiple hypernyms that can be modeled as complex types (Pustejovsky, 1995). Finally, the glosses for synsets play an underestimated role in addition to the synsets and the relations, but no formal structuring is defined for these glosses. As a result, glosses are not sufficiently descriptive to precisely identify the meaning of a concept. Such differences across wordnets make it difficult to establish the proper relations to the ILI and thus to compare and exploit wordnets across languages. Further, if a synset is not realized in a language it is not clear if that is because the concept is not lexicalized in that language, or if it is merely not realized yet (the compilers may just not have got round to it).

To solve these problems, we need to not just define an interlingual index, but also shared guidelines for relations, how to write definitions, standard data formats and so forth.

3 The Collaborative Interlingual Index

In this section we describe the core properties of CILI. To coordinate an index among all the different wordnet projects, we propose that it should, ideally, have the following properties (building on 1--5 from Fellbaum and Vossen, 2008):

1. The Interlinear Index (ILI) should be a flat list of concepts.
2. The semantic and lexical relations should mean the same things for all languages.
3. Concepts should be constructed for salient and frequent lexicalized concepts in all languages.
4. Concepts linked to Multiword units (MWUs) in wordnets should be included.
5. A formal ontology could be linked to but separate from the wordnets.

6. The license must allow redistribution of the index

7. ILI IDs should be persistent: we never delete, only **deprecate** or **supercede**; we should not change the meaning of the concept

8. Each new ILI concept should have a definition in English, as this is the only way we can coordinate across languages. The definition should be unique, which is not currently true, and preferably also parse and sense tag information should be included. Definition changes will be moderated.

9. Each new ILI concept should link to a synset in an existing project that is part of the GWG with one of a set of known relations (`hypernymy`, `meronomy`, `antonymy`, ...)

10. This synset should link to another synset in an existing project that is part of the GWG and links to an ILI concept.

    $\Rightarrow$ each concept is linked to another concept through at least one wordnet in the grid

11. Any project adding new synsets should first check that they do not already exist in the CILI

    - New concepts are added through their existing in a wordnet
    - If something fulfills the criteria is proposed
    - If no objections after three months then it is added

Property 6, an open license, is a necessary condition for groups to be able to use the ILI within their own project. To be maximally compatible, the license should place as few restrictions as possible, ideally requiring only that the source of the resource be mentioned: it should be either the wordnet license itself, Creative Commons Attribution (CC-BY) or the MIT license. We choose to use CC-BY, as the license has been well written and documented and is widely used.

Property 7, persistent identifiers, is an important criterion for stability. If the ILI changed its IDs, projects without the resources to maintain compatibility would fall behind. If a project changes its hierarchy, then it will need to add new nodes and delink the old ones. To keep backwards compatibility, even if a concept is deemed problematic, it will remain in the CILI, and marked as **deprecated**, preferably with a link to the concept that **supercedes** it.

Property 8, that all synsets should have a definition in English, recognizes that, in practice, the only language shared by all groups is English. Here we are inspired by experience with the CICC project, a multilingual machine translation project linking Thai, Chinese, Japanese, Malay and Indonesian (but not English) (CICC, 1994). No members spoke all five languages, but someone in each group spoke English, so all dictionary entries also had an English translation or definition. Having a universally understood definition is a prerequisite in avoiding redundant creation of new senses. This creates a burden on non-English speakers, which we will try to lighten by giving clear guidelines for writing definitions (see section 3.3). Note, that while the definition must be in English, the concept is not necessarily lexicalized in English, in contrast to Princeton WordNet.

Properties 9 and 10 make sure that all new concepts link to something, there should be no orphaned concepts. Exactly which links are acceptable is still a matter of research.[2]

The final point (11) is about coordination. Practically, it will not be possible to have a single moderator who can check new synsets in every language. We therefore propose that the burden of checking for duplication with existing synsets should be placed on the project wanting to add new synsets. As new concepts should be linked to existing concepts through relational links in a wordnet, and definitions in English will exist for all entries, checking for a compatible entry in the ILI should not be too burdensome. Project members with wordnets in the shared multilingual index would gain write privileges to the ILI, of course anyone should be able to read it. We will build automated tools that warn if definitions are too similar (for details see Vossen et al., 2016).

For the ILI to be successful there will be an initial cost to combine all existing non-English synsets, adding English definitions for all and merging duplicates. It would also require buy-in from all participating projects, but fortunately most non-English wordnets contain few synsets that do not correspond to an English synset, so this first step should not be too burdensome. For wordnets

[2] Many wordnets, including PWN, currently contain some orphans (e.g. ***uphill***$_{r:1}$), these would not be added to the ILI unless they are linked to something.

built with the merge approach there will be many more new synsets, these should be checked carefully and validated against corpora before being included in the ILI. We will support this with workshops at relevant conferences (such as the 16th Global Wordnet Conference).

In the long run, we hope that external resources will link to the ILI's persistent IDs (things like SUMO, TempoWordnet (Dias et al., 2014), the many Sentiment wordnets (Baccianella et al., 2010; Cruz et al., 2014).

3.1 Format

The ILI will be represented as RDF. Our reference implementation will be in Turtle (Terse RDF Triple Language: W3C, 2012) a compact format for RDF graphs.

It includes its own metadata, based on the Dublin Core, shown in Figure 1. As far as possible, triples are defined using existing schema (referenced in the preamble). The individual entries are designed to be extremely simple. Unlike synsets in individual wordnets, ILI concepts do not have explicit parts-of-speech. No further semantics is imposed within the ILI.

Each concept in the ILI has the following simple structure:

- A unique ID: `i1, i2, i3, …`
- A type: `Concept` or `Instance`
- A gloss in English: `skos:definition`
- A link to the synset that first motivated the ILI concept: `dc:source`
- Links to all current wordnets in the GWG that use this concept: `owl:sameAs`
- Optionally a `deprecate/supercedes` link

We give an example in Figure 2, which also shows the relevant prefixes.

Information about provenance (who added the entry, when it was made and so forth) are left to the version control system, for which we have chosen to use (`git`: `http://git-scm.com/`). When commits are made, the project will be added as the **author** so a record is kept of who is responsible for which change without making it visible in the ILI.

Note that the concept is defined not just by the written definition but by the links to the wordnets and the lemmas in those wordnets: the definition is a crucial tool for coordinating across languages, but is not meant to be the sole determiner of the ILI concept's meaning. The ILI concepts will always be linked to the global wordnet grid (Fellbaum and Vossen, 2007; Vossen et al., 2016).

Labels for the concepts can be produced automatically, as it is probably that different languages would want different labels. The easiest approach would be to take the most frequent lemma in the language of choice, backing off to the most frequent lemma in the language that introduced it (which can be obtained from the `dc:source`).

3.2 The WordNet Schema

In order to ensure that WordNets may be submitted in a form that is compatible with the ILI, we have developed two specific schemas, namely an XML schema based on the Lexical Markup Framework (Vossen et al., 2013, LMF) and the second in JSON-LD (Sporny et al., 2014) using the Lexicon Model for Ontologies (McCrae et al., 2012, *lemon*). These models are structured as follows:

LexicalResource The root element of the resource is the *lexical resource*

Lexicon Each WordNet has a *lexicon* for each resource, which has a name, an ID and a language. The language is given as a BCP 47 tag .

Lexical Entry Each 'word' is termed a *lexical entry*, it has exactly one lemma, at least one sense and any number of syntactic behaviors.

Lemma The lemma has a *written form* and part-of-speech, which may be one of **n**oun, **v**erb, **a**djective, adve**r**b, **p**hrase, **s**entence or **u**nknown.

Sense The *sense* has any number of *sense relations* and a synset.

Synset The *synset* has an optional definition and any number of *sense relations*.

Definition The definition is given in the language of the WordNet it came from as well as the ILI definition (in English). A definition may also have a *statement* that gives an example

Synset/Sense Relation A relation from a given list of relations such as synonym, hypernym, antonym. This list defines the relations used

```
<> a voaf:Vocabulary ;
  vann:preferredNamespacePrefix "ili" ;
  vann:preferredNamespaceUri "http://globalwordnet.org/ili" ;
  dc:title "Global Wordnet ILI"@en ;
  dc:description "The shared Inter-Lingual Index for the global wordnets.
   It consists of a list of concepts of instances with definitions,
   and their links to open wordnets."@en ;
  dc:issued "2015-07-30"^^xsd:date ;
  dc:modified "2015-07-30"^^xsd:date ;
  owl:versionInfo "0.1.1"@en ;
  dc:rights "Copyright Global Wordnet Association" ;
  cc:license <http://creativecommons.org/licenses/by/4.0> ;
  cc:attributionName "Global Wordnet Association";
  cc:attributionURL <http://globalwordnet.org>;
  dc:contributor  <http://www3.ntu.edu.sg/home/fcbond/>,  <http://john.mccr.ae> ,
                 <http://vossen.info/> ;
  dc:publisher <http://globalwordnet.org> .
```

Figure 1: ILI metadata

```
@prefix pwn30: <http://wordnet-rdf.princeton.edu/wn30/> .
@prefix jwn12: <http://compling.hss.ntu.edu.sg/omw/wns/jpn/> .
@prefix ili: <http://globalwordnet.org/ili/> .
@base <http://globalwordnet.org/ili/ili#>.

<i71370> a <Concept> ;
        dc:source       pwn30:06639428-n ;
        skos:definition "any of the machine-readable lexical databases
                         modeled after the Princeton WordNet"@en ;
        owl:sameAs      jwn12:jpn-06639428-n ;
        owl:sameAs      pwn30:06639428-n .
```

Figure 2: Example ILI entry for the concept of a *wordnet*

by the Global Wordnet Grid, and all the relations are documented on the Global Wordnet Association website.

Syntactic Behavior A syntactic behavior (verb frame) gives the subcategorization frame in plain text, such as ``Sam and Sue %s the movie".

Meta Dublin Core properties may be added to lexicons, lexical entries, senses and synsets.

Either format can be used to describe a WordNet and it is simple to convert between either. An example of the LMF form is given in figure 3 and in WN-JSON in figure 4

3.3 Guidelines for Definitions

In any given wordnet, the definition is only one of the things that helps to tell the meaning of a word, it is accompanied by the semantic relations, part of speech information, examples and so forth. The ILI is situated in the global wordnet grid, so this information should also be available. However the definition is the only thing guaranteed to be in the ILI, and the accompanying information may only be from a wordnet whose language is not comprehensible to another user. Moreover, as these definitions are given in natural language it is important to ensure that they are as unambiguous as possible, and can clearly identify the concepts, without the additional mechanisms of semantic relations. For these reasons strong guidelines for definitions are of primary importance.

There are already good general guidelines for writing dictionary definitions (Landau, 1989, Chapter Four). Almost all of these apply to wordnets in general, and the CILI in particular, with the exception that **brevity** is less important in an electronic resource.

There are some extra constraints for the CILI. First, definitions should be unique and there should be enough information to minimally distinguish one concept from all others. This was not the case in the wordnets, PWN has over 1,629 synsets with non-unique definitions, and there are similar numbers in other wordnets (1,362 in Japanese, 418 in Indonesian, 211 in Greek, 104 in Albanian and so on). For example it would not be sufficient to describe ***paella***$_{n:1}$ as ``a Spanish dish" as

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "http://globalwordnet.github.io/schemas/WN-LMF.dtd">
<LexicalResource>
    <Lexicon label="Princeton WordNet" language="en">
        <LexicalEntry id="w1">
            <Lemma writtenForm="wordnet" partOfSpeech="n"/>
            <Sense id="106652077-n-1" synset="106652077-n"/>
        </LexicalEntry>
        <Synset id="106652077-n" ili="s35545">
            <Definition
              gloss="any of the..."
              iliDef="any of the..."/>
            <SynsetRelation relType="hypernym" target="106651393-n"/>
        </Synset>
        <Meta publisher="Princeton University"
              rights="http://wordnet.princeton.edu/wordnet/license/"/>
    </Lexicon>
</LexicalResource>
```

Figure 3: Example of WordNet entry in WN-LMF

```
{
  "@context": [ "http://globalwordnet.github.io/schemas/wn-json-context.json",
                { "@language": "en" } ],
  "@id": "pwn30",
  "label": "Princeton WordNet",
  "language": "en",
  "publisher": "Princeton University",
  "rights": "wordnetlicense:",
  "entry": [{
    "@id" : "w1",
    "lemma": { "writtenForm": "wordnet" },
    "partOfSpeech": "wn:noun",
    "sense": [{
      "@id": "106652077-n-1",
      "synset": {
        "@id": "106652077-n",
        "ili": "s35545",
        "definition": {
          "gloss": "any of the..." ,
          "iliDef": "any of the..."
         },
         "hypernym": ["106651393-n"]
      }
    }]
  }]
}
```

Figure 4: Example of an entry in WN-JSON

this is not sufficiently distinctive. For the wordnets, the combination of definition and lemmas is normally enough to distinguish a word, but for the ILI, if necessary, one of the English lemmas must be included in the definition (for example, including the species name in the definition). This conflicts somewhat with the best practice for individual wordnets, where in general we want to avoid redundancy: if the synset is linked through domain-category to e.g. mathematics, we would normally not start the definition with ``(mathematics)". A case in point is the definitions for PWN30:13223710-n *ground fir, princess pine, tree clubmoss, Lycopodium obscurum* and and PWN:13223588-n *ground cedar, staghorn moss, Lycopodium complanatum* which are both defined as ``a variety of club moss". In this case, amending the definition to ``a variety of club moss (Lycopodium obscurum)" and ``a variety of club moss (Lycopodium complanatum)" makes the definitions unique (at the cost of some redundancy. We propose using some of the wide array of brackets available to show the redundant information in the ILI definition: ``⟪plant⟫ a variety of club moss ⟦Lycopodium complanatum⟧". Doing this reduces the number of non-unique definitions by over 50%. The ILI definitions are thus produced automatically from PWN 3.0, without always being identical to them.

We also place some limitations on the format.

The definition should consist of one or more short utterances, separated by semicolons. Semicolons should not be used within each utterance, use comma or colon instead. Definitions will be split on semicolons before being parsed, so it is important to be consistent here. We also do not allow the use of ASCII double quotes instead preferring Unicode left and right (double) quotes to aid parsing.

In general, we need to be very conservative in changing the definitions of concepts in the ILI. When first written, we should try not to make the definition too restricted, for example, prefer for *angel, backer* instead of ``invests in a theatrical production", prefer ``someone who invests in something, typically a theatrical production". This makes it easier to avoid having to make multiple very similar synsets.

Definitions should use standard patterns, especially for the first utterance in a definition. Ideally, the definition should consist of a **genus** (the hypernym, not necessarily the immediate hypernym) and **differentiae**, e.g.,

> ***wordnet*** *(lemma)* ``any of the machine-readable lexical databases *(genus)* modeled after the Princeton WordNet" *(differentiae)*

Adjectives and adverbs are exceptions, in that they are often defined using prepositional phrases.

Finally we make a simple requirement that definitions have a minimum length of 20 characters or 5 words.

In future work we will produce a tool to parse the definition and automatically identify the hypernym (Nichols et al., 2005), sense tag the definition (Moldovan and Novischi, 2004) and report on this to the definition writer, as well as compare the definition to definitions from similar concepts. This can help identify infelicitous definitions.

4 Open Issues

There are a few cases where it was hard to decide whether a concept should be represented in the InterLingual Index.

One example is named entities. Roughly 6.6% of the entries in PWN are linked by the `instance` relation (including the names of people, places, planets, gods and many more). Named entities are much more numerous than words and these concepts and their relations are better captured by other kinds of resources. However, some named entities can be considered part of the lexicon as well as names for objects, for example ***Glaswegian***$_{a:1}$ ``of or relating to or characteristic of Glasgow or its inhabitants", which is also used in the definition of other concepts. Thus, we retain a small number of named entities, especially geographic terms but further discussion is required to refine an exact policy.

It could also be argued that some of the derived forms (for example ***quickly***$_{r:1}$ from ***quick***$_{a:1}$) are unnecessary: as the meaning change is generative, there is no point in having two concepts. These kind of changes can be applied later by means of superceding other concepts, and for the moment we apply the distinctions made by Princeton WordNet.

5 Conclusions

We have introduced and motivated the collaborative interlingual index (CILI). Its simple design allows us to link wordnets with a minimum of extra work. Once concepts are added to the CILI, they will get a persistent ID and thereafter should not be deleted or change in meaning. We propose that the task of checking the validity of new concepts is taken up by the individual wordnet projects, with only a light layer of moderation.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7thh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta.

Luisa Bentivogli and Emanuele Pianta. 2003. Beyond lexical units: Enriching wordnets with phrasets. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 03)*, pages 67--70. Budapest.

Pushpak Bhattacharyya. 2010. Indowordnet. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. La Valletta. URL `http://www.cfilt.iitb.ac.in/indowordnet/`.

Francis Bond, Lian Tze Lim, Enya Kong Tan, and Hammam Riza. 2014. The combined wordnet Bahasa. *Nusa: Linguistic studies of languages in and around Indonesia*, 57:83--100.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64--71.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191--1211. URL `dx.doi.org/10.1007/s10579-013-9233-4`.

CICC. 1994. Research on Malaysian dictionary. Technical Report 6---CICC---MT54, Center of the International Cooperation for Computerization, Tokyo.

Fermín L Cruz, José A Troyano, Beatriz Pontes, and F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*.

Luís Morgado da Costa and Francis Bond. 2015. OMWEdit - the integrated open multilingual wordnet editing system. In *ACL-2015 System Demonstrations*.

Gaël Harry Dias, Mohammed Hasanuzzaman, Stéphane Ferrari, and Yann Mathet. 2014. Tempowordnet for sentence time tagging. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 833--838. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland. URL `http://dx.doi.org/10.1145/2567948.2579042`.

William Black Horacio Rodríguez Musa Alkhalifa Piek Vossen Adam Pease Elkateb, Sabri and Christiane Fellbaum. 2006. Building a wordnet for Arabic. In *In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Christiane Fellbaum and Piek Vossen. 2007. Connecting the universal to the specific: Towards the global grid. In *First International Workshop on Intercultural Collaboration (IWIC-2007)*, pages 2--16. Kyoto.

Christiane Fellbaum and Piek Vossen. 2008. Challenges for a global wordnet. In Webster J., Nancy Ide, and A.Chengyu Fang., editors, *Online Proceedings of the First International Workshop on Global Interoperability for Language Resources (ICGL 2008)*, pages 75--82. City University of Hongkong.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14--23. (in Chinese).

C. Kunze and L. Lemnitzer. 2002. Germanet --- representation, visualization, application. In *LREC*, pages 1485--1491.

Sidney I. Landau. 1989. *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press, Cambridge, UK.

John McCrae, Philipp Cimiano, and Elena Montiel-Ponsoda. 2012. Integrating wordnet and wiktionary with lemon. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellman, editors, *Linked Data in Linguistics*. Springer.

George A. Miller. 1990. WordNet: An online lexical database. *International Journal of Lexicography*, 3(4). (Special Issue).

Dan Moldovan and Adrian Novischi. 2004. Word sense disambiguaition of WordNet glosses. *Computer Speech and Language*, 18:301--317.

Eric Nichols, Francis Bond, and Daniel Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, pages 1111--1116. Edinburgh.

BoletteSandford Pedersen, Sanni Nimb, Jørg Asmussen, NicolaiHartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet --- the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269--299.

Wim Peters, Piek Vossen, Pedro Díez-Orzas, and Geert Adriens. 1998. Cross-linguistic alignment of wordnets with an inter-lingual-index. In Vossen (1998), pages 149--251.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293--302. Mysore, India.

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. URL `http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`, (ISBN 978-83-7493-476-3).

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch wordnet. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. (this volume).

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, and Niklas Lindström. 2014. Json-ld 1.0: A json-based serialization for linked data. W3C recommendation, The World Wide Web Consortium.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology*, 7(1--2):9--34.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. (this volume).

Piek Vossen, Wim Peters, and Julio Gonzalo. 1999. Towards a universal index of meaning. In *Proceedings of ACL-99 Workshop, Siglex-99, Standardizing Lexical Resources*, pages 81--90. Maryland.

Piek Vossen, Claudia Soria, and Monica Monachini. 2013. LMF - lexical markup framework. In Gil Francopoulo, editor, *LMF - Lexical Markup Framework*, chapter 4. ISTE Ltd + John Wiley & sons, Inc.

World Wide Web Consortium W3C. 2012. Turtle --- terse RDF triple language. `http://www.w3.org/TR/2012/WD-turtle-20120710/`.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10--18. Nagoya.

Aesun Yoon, Soonhee Hwang, Eunroung Lee, and Hyuk-Chul Kwon. 2009. Construction of Korean wordnet KorLex 1.5. *Journal of KIISE: Software and Applications*, 36(1):92--108.

YARN: Spinning-in-Progress

Pavel Braslavski
Ural Federal University
Yekaterinburg, Russia
pbras@yandex.ru

Dmitry Ustalov
Ural Federal University
Yekaterinburg, Russia
dmitry.ustalov@urfu.ru

Mikhail Mukhin
Ural Federal University
Yekaterinburg, Russia
mfly@sky.ru

Yuri Kiselev
Yandex
Yekaterinburg, Russia
yurikiselev@yandex-team.ru

Abstract

YARN (Yet Another RussNet), a project started in 2013, aims at creating a large open WordNet-like thesaurus for Russian by means of crowdsourcing. The first stage of the project was to create noun synsets. Currently, the resource comprises 100K+ word entries and 46K+ synsets. More than 200 people have taken part in assembling synsets throughout the project. The paper describes the linguistic, technical, and organizational principles of the project, as well as the evaluation results, lessons learned, and the future plans.

1 Introduction

The Global WordNet Association website lists 76 wordnets for 70 different languages[1], including multilingual resources. Although the table mentions as many as three wordnets for Russian, unfortunately no open Russian thesaurus of an acceptable quality and size is still available.

The Yet Another RussNet (YARN) project[2] started in 2013. It aims at creating a comprehensive and open thesaurus for Russian. From the linguistics point of view, the proposed thesaurus has rather a traditional structure: it consists of synsets—groups of near-synonyms corresponding to a concept, while synsets are linked to each other, primarily via hierarchical hyponymic/hypernymic relations.

YARN intends to cover Russian nouns, verbs and adjectives. Following the divide and conquer approach, we treat synset assembly and relationship establishing separately.

The main difference between YARN and the previous projects is that YARN is based on crowdsourcing. We hope that the crowdsourcing approach will make it possible to create a resource of a satisfactory quality and size in the foreseeable future and with limited financial resources. Our optimism is based both on the international practice and the recent examples of successful Russian NLP projects fueled by volunteers. Another important distinction is that the editors do not build the thesaurus from scratch; instead, they use "raw data" as the input. These "raw data" stem from pre-processed dictionaries, Wiktionary, Wikipedia, and text corpora. More than 200 people have taken part in the synset assembly in the course of the project. Currently, the resource comprises 100K+ word entries and 46K+ synsets that are available under CC BY-SA license.

The paper describes the main linguistic and organizational principles of YARN, the tools developed, and the results of the current content evaluation. We also point to some pitfalls of the chosen crowdsourcing methodology and discuss how we could address them in the future.

2 Related Work

In this section, we briefly survey projects aimed at creation of WordNet-like semantic resources for Russian, describe peculiarities of other thesauri for Slavic languages, and systematize different crowdsourcing approaches to building lexicographic resources.

[1] http://globalwordnet.org/wordnets-in-the-world/

[2] http://russianword.net/en/, not to be confused with a Hadoop subsystem.

2.1 Russian Thesauri

The **RussNet** project[3] was launched in 1999 at Saint-Petersburg university (Azarova et al., 2002). According to the RussNet developers, the resource currently contains about 40K word entries, 30K synsets, and 45K semantic relations. However, this data is not encoded in a uniform format and cannot be published or used in a NLP application in its current form.

RuThes is probably the most successful WordNet-like resource for Russian (Loukachevitch, 2011). It has been developing since 2002, and now contains 158K lexical units constituting 55K concepts. RuThes is a proprietary resource; however a subset of it was published recently[4]. The main hurdle for a wider use of the resource is a restrictive license and the fact that the data in XML format can be obtained by request only.

Another resource—**RussianWordNet**—was a result of a fully automatic translation of the Princeton WordNet (PWN) into Russian undertaken in 2003 and is freely available[5] under the PWN license. The approach based on bilingual dictionaries, parallel corpora, and dictionaries of synonyms resulted in the translation of about 45% of the PWN entries. The thesaurus contains 18K nouns, 6K adverbs, 5.5K verbs, and 1.8K adverbs; no systematic quality assessments of the obtained data were performed (Gelfenbeyn et al., 2003). Another attempt to translate the PWN into Russian, in this case—in a semi-automatic fashion—is the **Russian Wordnet** project (Balkova et al., 2004) started in 2003, but its deliverables are not available to the general public.

Russian Wiktionary[6] can be seen as an ersatz of a proper thesaurus, since along with definitions it contains—though marginally—semantic relations. Wikokit project[7] allows handling Wiktionary data as a relational database (Krizhanovsky and Smirnov, 2013). Russian Wiktionary contains about 190K word entries and 70K synonym relations as of September, 2015.

The **Universal Networking Language**[8] project is dedicated to the development of a computer language that replicates the functions of natural languages. The Russian version of its semantic network—the Universal Dictionary of Concepts—contains approximately 62K universal words (UWs) and 90K links between them and is available[9] under CC BY-SA license.

One of the recent trends is the creation of semantic resources in a fully automatic manner, where collaboratively created resources like Wikipedia and Wiktionary are used as the input. A striking example of this approach is **BabelNet**, a very large automatically generated multilingual thesaurus (Navigli and Ponzetto, 2012); the Russian part of BabelNet consists of 2.37M lemmas, 1.35M synsets, and 3.7M word senses[10]. The data is accessible through an API under CC BY-NC-SA 3.0 license. No evaluation of the Russian data has been performed yet.

As can be seen from the survey, no open human-crafted wordnet for Russian is available so far. Automatically created resources are freely available and potentially have very good coverage, but their quality is disputed.

2.2 Thesauri of Other Slavic Languages

Slavic languages are highly inflectional and have a rich derivation system. The survey of wordnets for Czech (Pala and Smrž, 2004), Polish (Maziarz et al., 2014) and Ukrainian (Anisimov et al., 2013) shows that in each case a special attention is paid to dealing with the morphological characteristics. For instance, plWordNet features a versatile system of relations with dozens of subtypes of relations between synsets and lexical units, many of which reflect derivational relations.

2.3 Crowdsourcing Language Resources

Crowdsourcing, a human-computer technique for collaborative problem solving by online communities, has gained high popularity since its inception in the mid 2000's (Kittur et al., 2013). Creation and expansion of linguistic resources using crowdsourcing became a trend in recent years as shown by Gurevych and Kim (2013).

Despite the ongoing unabated discussions about the types, merits and limitations of crowdsourcing (Wang et al., 2013), we consider the following genres of crowdsourcing: *wisdom of the crowds* (WOTC), *mechanized labor* (MLAB) and *games with a purpose* (GWAPS).

[3] http://project.phil.spbu.ru/RussNet/
[4] http://labinform.ru/pub/ruthes/
[5] http://wordnet.ru/
[6] http://ru.wiktionary.org/
[7] https://github.com/componavt/wikokit
[8] http://www.undl.org/
[9] https://github.com/dikonov/Universal-Dictionary-of-Concepts
[10] http://babelnet.org/stats

In the WOTC genre, the resource is constructed *explicitly* by a crowd of volunteers that collaborates in an online editing environment. Their participation is mostly altruistic and a participant's benefit is either self-exaltation or self-promotion of any kind. Successful examples of this genre are Wikipedia and Wiktionary. The primary issues of such resources are vandalism and "edit wars", which are usually resolved by edit patrolling and edit protection.

In the MLAB genre, the resource is created *implicitly* by the workers who submit answers to simple tasks provided by the requester. This genre is proven to be effective in many practical applications. For instance, Lin and Davis (2010) extracted ontological structure from social tagging systems and engaged workers in evaluation. Rumshisky (2011) used crowdsourcing to create an empirically-derived sense inventory and proposed an approach for automated assessment of the obtained data. Biemann (2013) described how workers can contribute to thesaurus creation by solving simple lexical substitution tasks. Most of these studies have been conducted on the commodity platforms like Amazon Mechanical Turk[11] (MTurk) and CrowdFlower[12]. Unfortunately, MTurk can hardly be used for tasks implying the knowledge of Russian because: (1) there are virtually no workers from Russia presented on the platform (Pavlick et al., 2014), and (2) a requester must have a U.S. billing address to submit tasks[13]. Having no access to the global online labor marketplaces is a serious obstacle to paying the workers due to the requirements of the local legislation of Russia. However, projects like OpenCorpora are trying to work around this problem by developing custom crowdsourcing platforms and effectively appealing to *altruism* instead of money reward (Bocharov et al., 2013). Since such altruistic mechanized labor does not imply money reward, it is not prone to spam, where an unfair worker may permanently submit random answers instead of sensible ones.

In the GWAPS genre, the crowdsourcing process is embedded into a multi-player game, in which the players have to accomplish various goals by creating new data items to win the game. Although such games are attractive and entertaining, game development is an expensive and complex kind of activity that may be feasible only for large-scale annotation projects. The examples here are Phrase Detectives[14] and JeuxDeMots[15].

3 YARN Essentials

YARN is conceptually similar to Princeton WordNet (Fellbaum, 1998) and its followers: it consists of synsets—groups of quasi-synonyms corresponding to a concept. Concepts are linked to each other, primarily via hierarchical hyponymic/hypernymic relationships.

3.1 YARN Structure

Each single-word entry in YARN is characterized by the grammatical features (the types of POS and inflection) according to Zaliznyak's dictionary (1977). Synsets may include single-word entries {суффикс (*suffix*)}, multi-word expressions {подводная лодка (*submarine*)}, and abbreviations {ПО (программное обеспечение, *software*)}. Synsets may contain a definition (*gloss* in terms of PWN). Additionally, definitions can be attached to individual words in a synset—these definitions are inherited from the dictionary data and specify a word meaning, but cannot serve as a good definition for the whole synset. "Empty synsets" (i.e. containing no words) that correspond to a non-lexicalized concept are legitimate and help to create a more harmonious hierarchy of synsets.

Each word in a synset can be accompanied by one or more usage examples. Words within synsets can attach labels from the five categories: *emotional*, *stylistic*, *chronological*, *domain/territorial*, and *semantic* (28 labels in total). This list is a result of the systematization of large and diverse Wiktionary label set. One of the synset words can be marked as the head word. Its sense is stylistically neutral, and it encompasses the meanings of the whole synset, e.g. {армия (*army*), войска (*troops*), вооружённые силы (*armed forces*)}. Each synset may belong to a domain, e.g. {кино (*movie*), кинофильм (*movie picture*), фильм (*film*)} → "Arts", {думать (*to think*), размышлять (*to ponder*)} → "Intellect".

The vertical, hypo-/hypernymic relations between synsets are decisive for the hierarchical

[11] https://www.mturk.com/mturk/welcome

[12] http://crowdflower.com/

[13] https://requester.mturk.com/help/faq#can_international_requesters_use_mturk

[14] https://anawiki.essex.ac.uk/phrasedetectives/

[15] http://www.jeuxdemots.org/

macrostructure of the thesaurus. The root of the YARN hierarchy is {предмет (*entity*), объект (*object*), вещь (*thing*)}; the second level is represented by {физическое явление (*physical phenomenon*)}, {отвлечённое понятие, абстрактное понятие, абстракция (*an abstraction*)}, {совокупность, набор (*set*), группа (*group*)}, {воображаемое, представляемое (*imaginary*)}. We elaborated 4–5 top levels for each part of speech.

The vertical links in YARN are also formed by the meronymy relations (the part-whole relations): ноздря (*nostrill*)—нос (*nose*)—лицо (*face*)—голова (*head*). The antonymy relationship connects specific words in the context of corresponding synsets. For example, the verb прибыть (*to arrive*) is the antonym of the verb отбыть (*to depart*), but not of направиться (*to head somewhere*) and the other words in the synset.

In the future, YARN will reflect the cross-POS relations between derivates: {двигаться (*to move*), движение (*movement*)}, {лес (*forest*), лесной ($forest_{adj}$)}. It will be significant for the word pairs with a minimum difference in senses.

3.2 Raw Data

As the "raw data" for the thesaurus construction we employed existing resources such as Wiktionary (which constituted the core of the input data), Wikipedia (redirects), the aforementioned result of the automatic translation of the PWN, the Universal Dictionary of Concepts, and the data from two dictionaries in the public domain. We also implicitly use the data from the Russian National Corpus (RNC) so that the corpus statistics influence the queue of words presented to the editors. Wikipedia and RNC were also used to compile the list of multi-word expressions to be included in the resource.

3.3 User Interface

Our initial approach to synset building is based on the WOTC inspired by the highly successful examples of Wikipedia and Wiktionary: our editors assemble synsets using word lists and definitions from dictionaries as the "raw data". Technically, virtually everybody can edit the YARN data—one needs only to login using a social network account. However, the task design implies minimal lexicographical skills and is more complicated than an average task offered for instance to MTurk workers. Our target editors are college or university students, preferably from the linguistics departments, who are native Russian speakers. It is desirable that students receive instructions from a university teacher and may seek their advice in complex cases. YARN differentiates the two levels of contributors—line editors and moderators. Moderators are authorized to approve thesaurus elements thus excluding them being modified by line editors.

The current synset editing interface can be accessed online[16]; its main window is presented in Figure 1. The "raw data" are placed on the left-hand side of the interface: definitions of the initial word and examples, and possible synonyms for each of the meanings, with definitions and examples for each of the synonyms. The right-hand part represents the resulting synsets including words, definitions, and examples. In principle, an editor can assemble a "minimal" synset from the dictionary "raw data" simply with several mouse clicks, without any typing.

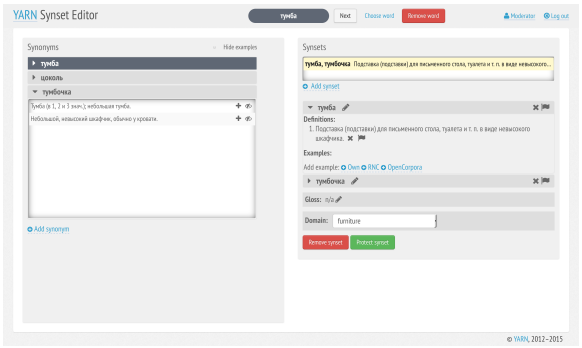


Figure 1: YARN synset assembly interface (the interface captions are translated into English for the convenience of the readers; originally all interface elements are in Russian).

Synset assembly begins with a word, or "synset starter". The editor selects an item from the list of words ranked by decreasing frequency; the already processed words are shaded. The editor can go through the words one after another or choose an arbitrary word using the search box. The top-left pane displays definitions of the initial word and usage examples if any. The possible synonyms of the initial word are listed on the bottom-left pane; they in turn contain their definitions and examples. The top-right pane displays a list of synsets containing the initial word. The editor can copy definitions and usage examples of the initial word

[16] http://russianword.net/editor

```
<synsetEntry id="s9439" author="122" version="29" timestamp="2014-11-17T07:49:46Z">
  <word ref="w9244">
    <definition source="ru.wiktionary" url="http://ru.wiktionary.org/wiki/суп">
      Жидкое кушанье, обычно представляющее собой отвар с приправами и употребляемое как первое блюдо.
    </definition>
    <example source="'Путешествие в седьмую сторону света', 2000, НКРЯ.">
      Был обед - овощной суп и курица на второе.
    </example>
  </word>
  <word ref="w40078"/>
  <word ref="w2893"/>
</synsetEntry>
```

Figure 2: XML representation of the synset {суп, бульон, похлёбка (*soup*)}.

from the top-left pane of the interface to the current synset by clicking the mouse. From the synonyms pane one can transfer words along with their definitions and examples. The editor can add a new word to the list of synonyms; it will appear with dictionary definitions and examples if presented in the parsed data. If the editor is not satisfied with the collected definitions, they can create a new one—either from scratch or based on one of the existing descriptions. Using search in the Russian National Corpus[17] and OpenCorpora[18], the editor can add usage examples. Additionally, a word or a definition within a synset can be flagged as "main", and be provided with labels. All synset edits are tracked and stored in the database along with the timestamps and the editor ID.

As a pilot study showed, editors spent about two minutes on average to compile a non-trivial synset, i.e. containing more than a single word. The top contributors demonstrated a learning effect: the average time per synset tended to decrease as the editor proceeded through the tasks, see Braslavski et al. (2014) for details.

Our next goal is to lower the threshold of participation in the data annotation and thus—to increase the number of participants. To do this, we are developing a mobile application in the MLab genre that is aimed at gathering "raw synsets": users are presented with a series of sentences with highlighted words and lists of possible contextual substitutes. This approach is similar to the experiment described in (Biemann, 2013).

3.4 Implementation Details

The YARN data are stored in a centralized database that can be accessed through a web interface. In addition, distributed teams can work directly with the database through an API. The database is periodically exported to XML format. Although the original dictionaries and thesauri were coming in different formats, we decided to develop a custom XML schema for data export[19]. We believe that XML format provides sufficient flexibility and preserves the connection to the internal data representation. The developed format is modular, as different types of objects (lexical units, synsets, and relationships) are described separately. The proposed format is somewhat similar to the Lexical Markup Framework (LMF)[20] approach, although the YARN format does not refer to the latter directly. All editing actions (in fact, aggregated "action chunks") are stored in the database. The YARN format stores the revision history analogously to the OpenStreetMap XML format[21]. A synset structure is illustrated in Figure 2.

The YARN software is implemented using Ruby on Rails framework. All data are stored in a PostgreSQL database. The user interface is implemented as a browser JavaScript application, which interacts with the back-end via JSON API. User authentication is performed through an OAuth endpoint provided by Facebook, VK and GitHub. The entire source code of the project is available in a GitHub repository[22].

3.5 Current State and Problems

The current version of the the YARN (September 2015) contains 44K synsets that consist of 48K words and 5.4K multi-word expressions; 838 words carry labels; 2.6K words are provided with at least one usage example (there are 4.2K examples in total). The resource contains 2.5K synset-level and 8.3K word-level definitions. The synset size distribution is presented in Figure 3.

[17] http://ruscorpora.ru/en/
[18] http://opencorpora.org/
[19] https://github.com/russianwordnet/yarn-formats/
[20] http://www.lexicalmarkupframework.org/
[21] http://www.openstreetmap.org/
[22] https://github.com/russianwordnet

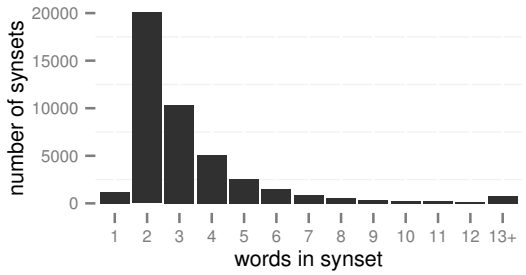


Figure 3: Synset distribution by size.

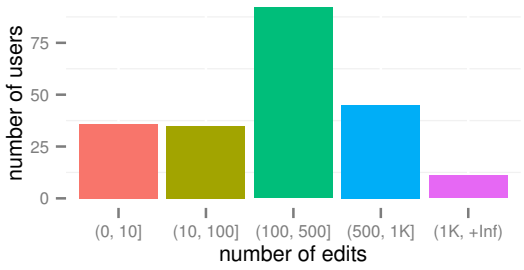


Figure 4: Distribution of users by edit count.

More than 200 people have taken part in editing YARN in the course of the project; the distribution of users by activity is shown in Figure 4. Whereas we consider the early experiment under a controlled crowd to be successful, we found the three significant problems replicating over time: organization issues, synset duplication and hyponymy/synonymy confusion.

Organization Issues. The number of synsets was growing rapidly and moderators were not able to assess all the incoming edits. In order to work around this problem, we are experimenting with MLAB workflows.

Synset Duplication. Participants do not consult the other people's work, which results in creation of duplicate synsets like {авто (*auto*), автомобиль (*automobile*), машина (*car*)} and {машина (*car*), тачка (*ride*)}.

Hyponymy Confusion. In some cases the participants mix hyponymy and synonymy, which results in strange synsets like {мультфильм (*cartoon*), мультик (*cartoon*), аниме (*anime*)}.

4 Evaluation

We compared YARN with other Russian thesauri (Kiselev et al., 2015), which have been described in Section 2.1 (Table 1). Besides YARN, the only resource available for use is RuThes-lite, the commercial use of which requires licensing. It should be noted that although the lexicon of YARN represents 100K+ words, only half of them are included in synsets. Thus, we provide the latter number.

The number of concepts indicates that crowdsourcing is a promising approach for thesauri creation for the Russian language. Interestingly, YARN contains more concepts than RussNet, a project started in 1999. However, when comparing YARN and RuThes-lite, one may notice, that they have an approximately equal number of concepts, yet the number of words in the latter is twice bigger than in YARN. This implies the hypothesis that expert-built thesauri include richer lexis that could be covered by non-expert users. Hence, the YARN synset quality requires more thorough evaluation.

4.1 Synset Quality

Since YARN is created using crowdsourcing, it seems reasonable to apply this technique for evaluation purposes, too. In our experiments we used an open source engine for MLAB workflows (Ustalov, 2015). In order to estimate the quality of the current YARN synsets, we retrieved the 200 most frequently edited synsets. We asked four experts to assess the quality of each synset by rating them on the following scale: *Excellent*—the synset completely represents a concept, *Satisfactory*—the synset is related to the concept, but some words are missing or odd words are present, and *Bad*—the synset is either ambiguous or it does not represent any sensible concept.

We aggregated the 800 obtained answers using the majority voting strategy, where the ties are resolved by choosing the worst of two answers, e.g. given the same number of votes for both *Good* and *Bad*, the latter will be selected. This resulted in 103 synsets of *Excellent*, 70 of *Satisfactory* and 27 of *Bad* quality. The results are shown in Table 2. Values in column **MV** are the numbers of synsets per each of the three grades, values in the last three columns are the numbers of synsets grouped by answer diversity—*all the answers are the same* in **1**, *two different answers present* in **2**, and *the expert opinions divided* in **3**.

We also computed the alpha annotator reliability coefficient for ordinal values to estimate the

Table 1: Russian thesauri comparison.

	# of concepts	# of relations	# of words	Availability	Commercial Usage
RussNet	5.5K	8K	15K	No	No
Russian Wordnet	157K	—	124K	No	No
RuThes	55K	210K	158K	No	No
RuThes-lite	26K	108K	115K	Yes	No
YARN	44K	0	48.6K	Yes	Yes

Table 2: YARN synset quality.

	MV	1	2	3
Excellent	103	37	62	21
Satisfactory	70	3	43	11
Bad	27	0	12	11
Total	200	40	117	43

inter-rater reliability (Krippendorff, 2013). The Krippendorff's alpha is $\alpha = 0.202$ due to the skewness of the answer distribution: more than half of the answers (434) are *Excellent*, the numbers of *Satisfactory* and *Bad* answers are 253 and 113 correspondingly. Given these results, we treat the top 200 YARN synsets as sufficiently good. These evaluation results define the upper bound for the average quality of the resource in its current state. Ustalov (2014) showed that revision count is a good proxy for quality in the Russian Wiktionary that is created in a similar fashion.

4.2 Duplicate Synsets

Sometimes users create new synsets without investigating the current synsets presented in YARN. The main problem with this is the presence of multiple entries for the same concept in the resource. Detecting such concepts requires special effort because they are not described with identical synsets but with similar ones.

Hence, we had to develop a method for automatically retrieving duplicate synsets. It was based on the heuristics suggesting that any two synonyms uniquely define a concept. This is not always true, but it lets us discover duplicate synsets with a very good recall. To estimate it, we compared the senses of random 200+ synsets having two or more common words. It turned out that more than in 85% of the cases these pairs described the same sense.

However, we found out that non-linguists do not recognize subtle nuances of meaning that are noticeable to experts, so the non-linguists cannot significantly improve the quality of duplicate extraction. Thus, this method—considering any synsets having more than two common words as duplicates—allows to detect and merge identical concepts with a quality that is comparable to what can be achieved by volunteers.

5 Conclusion

The deliverables of YARN are available under the CC BY-SA 3.0 license on the project website[23] in XML, CSV, and RDF formats. So far, we have the following plans for the future work.

- Creating verb and adjective synsets.
- Establishing hierarchical links between synsets through validation of the relationships imported from Wiktionary and other resources.
- Development of automatic methods for generating hypotheses based on Wikipedia and large text corpora.
- Development of automatic methods for preparing "raw data", as well as for post-processing of annotation results produced by the crowd.
- Widening the audience of the project's participants through mobile applications and simpler tasks.
- Development of crowd management methods, such as automatic methods for evaluation of workers, task difficulty, and annotation results, the system of incentives, etc.

Acknowledgments

This work is supported by the Russian Foundation for the Humanities project no. 13-04-12020 "New Open Electronic Thesaurus for Russian". We are grateful to Yulia Badryzlova for proofreading the text. We would also like to thank the three anonymous reviewers, who offered very helpful suggestions.

[23] http://russianword.net/data

References

Anatoly Anisimov, Oleksandr Marchenko, Andrey Nikonenko, et al. 2013. Ukrainian WordNet: Creation and Filling. In *Flexible Query Answering Systems*, volume 8132 of *Lecture Notes in Computer Science*, pages 649–660. Springer Berlin Heidelberg.

Irina Azarova, Olga Mitrofanova, Anna Sinopalnikova, Maria Yavorskaya, and Ilya Oparin. 2002. RussNet: Building a Lexical Database for the Russian Language. In *Proc. of Workshop on WordNet Structures and Standardisation, and How These Affect WordNet Applications and Evaluation*, pages 60–64, Gran Canaria, Spain.

Valentina Balkova, Andrey Sukhonogov, and Sergey Yablonsky. 2004. Russian WordNet. In *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 31–38, Brno, Czech Republic. Masaryk University Brno, Czech Republic.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Victor Bocharov, Svetlana Alexeeva, Dmitry Granovsky, et al. 2013. Crowdsourcing morphological annotation. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, volume 12(19), pages 109–124, Moscow, Russia. RGGU.

Pavel Braslavski, Dmitry Ustalov, and Mikhail Mukhin. 2014. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 101–104, Gothenburg, Sweden. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Database*. MIT Press.

Ilya Gelfenbeyn, Artem Goncharuk, Vlad Lekhelt, et al. 2003. Automatic translation of WordNet semantic network to Russian language. In *Proceedings of Dialog-2003*.

Iryna Gurevych and Jungi Kim, editors. 2013. *The People's Web Meets NLP*. Springer Berlin Heidelberg.

Yuri Kiselev, Sergey V. Porshnev, and Mikhail Mukhin. 2015. Current Status of Russian Electronic Thesauri: Quality, Completeness and Availability. *Programmnaya Ingeneria*, (6):34–40.

Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, et al. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA. ACM.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE, Thousand Oaks, CA, USA, 3rd edition.

Andrew A. Krizhanovsky and Alexander V. Smirnov. 2013. An approach to automated construction of a general-purpose lexical ontology based on wiktionary. *Journal of Computer and Systems Sciences International*, 52(2):215–225.

Huairen Lin and Joseph Davis. 2010. Computational and Crowdsourcing Methods for Extracting Ontological Structure from Folksonomy. In *The Semantic Web: Research and Applications*, volume 6089 of *Lecture Notes in Computer Science*, pages 472–477. Springer Berlin Heidelberg.

Natalia Loukachevitch. 2011. *Thesauri in information retrieval tasks*. Moscow University Press, Moscow, Russia.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proceedings of the Seventh Global Wordnet Conference*, pages 304–312, Tartu, Estonia.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2):79–88.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Anna Rumshisky. 2011. Crowdsourcing Word Sense Definition. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 74–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dmitry Ustalov. 2014. Words Worth Attention: Predicting Words of the Week on the Russian Wiktionary. In *Knowledge Engineering and the Semantic Web*, volume 468 of *Communications in Computer and Information Science*, pages 196–207. Springer International Publishing.

Dmitry Ustalov. 2015. A Crowdsourcing Engine for Mechanized Labor. *Proceedings of the Institute for System Programming*, 27(3):351–364.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31.

Andrey Zaliznyak. 1977. *Grammatical dictionary of Russian*. Russky yazyk, Moscow, USSR.

Word Substitution in Short Answer Extraction: A WordNet-based Approach

Qingqing Cai, James Gung, Maochen Guan, Gerald Kurlandski, Adam Pease
IPsoft / New York, NY, USA
[qingqing.cai|james.gung|maochen.guan|gerald.kurlandski|adam.pease]@ipsoft.com

Abstract

We describe the implementation of a short answer extraction system. It consists of a simple sentence selection front-end and a two phase approach to answer extraction from a sentence. In the first phase sentence classification is performed with a classifier trained with the passive aggressive algorithm utilizing the UIUC dataset and taxonomy and a feature set including word vectors. This phase outperforms the current best published results on that dataset. In the second phase, a sieve algorithm consisting of a series of increasingly general extraction rules is applied, using WordNet to find word types aligned with the UIUC classifications determined in the first phase. Some very preliminary performance metrics are presented.

1 Introduction

Short Answer Extraction refers to a set of information retrieval techniques that retrieve a short answer to a question from a sentence. For example, if we have the following question and answer sentence

(1) Q: Who was the first president of the United States?
A: George Washington was the first president of the United States.

we want to extract just the phrase "George Washington". But what if we have a mismatch in language between question and answer? What is an appropriate measure for word similarity or substitution in question answering? If we have the question answer pair

(2) "Bob walks to the store."

(3) "Who ambles to the store?"

we probably want to answer "Bob", because "walk" and "amble" are similar and not inconsistent. In isolation, a human would likely judge "walk" and "amble" to be similar, and by many WordNet-based similarity measures they would be judged similar, since "walk" is found as WordNet synsets 201904930, 201912893, 201959776 and 201882170, and "amble" is 201918183, which is a direct hyponym of 201904930.

We can use Resnik's method (Resnik, 1995) to compute similarity. In particular we can use Ted Pedersen's (et al) implementation (Pedersen et al., 2004), which gives the result of `walk#n#4 amble#n#1 9.97400037941652` . Word2Vec (Mikolov et al., 2013a) using their 300-dimensional vectors trained on Google News, also gives a relatively high similarity score for the two words

```
> model.similarity('walk', 'amble')
0.525
```

2 Is Similarity the Right Measure?

But what about if we have

(4) "Bob has an apple."

(5) "Who has a pear?"

We find that this pair is even more similar than "walk" and "amble"

```
> model.similarity('apple', 'pear')
0.645
```

and from Resnik's algorithm

```
Concept #1: apple
Concept #2: pear
apple  pear
apple#n#1  pear#n#1  10.15
```

and yet clearly 4 is not a valid answer to 5. One possibility is that synset subsumption as a measure of word substitution (Kremer et al., 2014; Biemann, 2013)[1] [2] may be the appropriate metric,

[1] https://dkpro-similarity-asl.googlecode.com/files/TWSI2.zip
[2] http://www.anc.org/MASC/coinco.tgz

rather than word similarity.

3 Question Answering

Our approach starts with the user's question and the sentence that is most likely to contain the answer, which is selected with the BM25 algorithm (Jones et al., 2000). Then we identify the incoming question as a particular question type according to the UIUC taxonomy[3]. To this taxonomy we have added the yes/no question type. Then we pass the sentence and the question to a class written specifically to handle a particular UIUC question type. Generally, all the base question types behave differently from one another. Within a base question type, subtypes may be handled generically or with code specially targeted for that subtype. For this paper, we first discuss the approach to question classification, and then to answer extraction with a focus on the question subtypes that are amenable to a WordNet-based approach.

4 Question Classification

This section presents a question classifier with several novel semantic and syntactic features based on extraction of question foci. We use several sources of semantic information for representing features for each question focus. Our model uses a simple margin-based online algorithm. We achieve state-of-the-art performance on both fine-grained and coarse-grained question classification. As the focus of this paper is on WordNet, we leave many details to a future paper and primarily report the features used, the learning algorithm and results, without further justification

4.1 Introduction

Question analysis is a crucial step in many successful question answering systems. Determining the expected answer type for a question can significantly constrain the search space of potential answers. For example, if the expected answer type is *country*, a system can rule out all documents or sentences not containing mentions of countries. Furthermore, accurately choosing the expected answer type is extremely important for systems that use type-specific strategies for answer selection. A system might, for example, have a specific unit for handling *definition* questions or *reason* questions.

[3] http://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html
http://cogcomp.cs.illinois.edu/Data/QA/QC/

In the last decade, many systems have been proposed for question classification (Li and Roth, 2006; Huang et al., 2008; Silva et al., 2011). Li and Roth (Li and Roth, 2002) introduced a two-layered taxonomy of questions along with a dataset of 6000 questions divided into a training set of 5000 and test set of 500. This dataset (henceforth referred to as the UIUC dataset) has since become a standard benchmark for question classification systems.

There have been a number of advances in word representation research. Turian et al. (Turian et al., 2010) demonstrated the usefulness of a number of different methods for representing words, including word embeddings and Brown clusters (Brown et al., 1992), within supervised NLP application such as named entity recognition and shallow parsing. Since then, largely due to advances in neural language models for learning word embeddings, such as WORD2VEC (Mikolov et al., 2013b), word vectors have become essential features in a number of NLP applications.

In this paper, we describe a new model for question classification that takes advantage of recent work in word embedding models, beating the previous state-of-the-art by a significant margin.

4.1.1 Question Focus Extraction

Question foci (also known as *headwords*) have been shown to be an important source of information for question analysis. Therefore, their accurate identification is a crucial component of question classifiers. Unlike past approaches using phrase-structure parses, we use rules based on a dependency parse to extract each focus.

We first extract the question word (how, what, when, where, which, who, whom, whose, or why) or imperative (name, tell, say, or give). This is done by naively choosing the first question word in the sentence, or first imperative word if no question word is found. This approach works well in practice, though a more advanced method may be beneficial in more general domains than the TREC (Voorhees, 1999) questions of the UIUC dataset.

We then define specific rules for each type of question word. For example, *what*/*which* questions are treated differently than *how* questions. In *how* questions, we identify words like *much* and *many* as question foci, while treating the heads of these words (e.g. *feet* or *people*) as a separate type known as **QUANTITY** (as opposed to **FOCUS**. Furthermore, when the focus of a *how* question

is itself the head (e.g. *how much did it cost?* or *how long did he swim?*), we again differentiate the type using a **MUCH** type and a **SPAN** type that includes words like *long* and *short*.

A head chunk such as *type of car* contains two words, *type* and *car*, which both provide potentially useful sources of information about the question type. We refer to words such as *type*, *kind*, and *brand* as **specifiers**. We extract the argument of a specifier (*car*) as well as the specifier itself (*type*) as question foci.

In addition to head words of the question word, we also extract question foci linked to the root of the question when the root verb is an **entailment** word such as *is*, *called*, *named*, or *known*. Thus, for questions like *What is the name of the tallest mountain in the world?*, we extract *name* and *mountain* as question foci. This can result in many question foci in the case of a sentence like *What **relative** of the **racoon** is sometimes known as the **cat-bear**?*

4.1.2 Learning Algorithm

We apply an in-house implementation of the multi-class Passive-Aggressive algorithm (Crammer et al., 2006) to learn our model's parameters. Specifically, we use PA-I, with

$$\tau_t = \min\left\{C, \frac{l_t}{\|x_t\|^2}\right\}$$

for $t = 1, 2, ...$ where C is the aggressiveness parameter, l_t is the loss, and $\|x_t\|^2$ is the squared norm of the feature vector for training example t. The Passive-Aggressive algorithm's name refers to its behavior: when the loss is 0, the parameters are unchanged, but when the loss is positive, the algorithm aggressively forces the loss to return to zero, regardless of step-size. τ (a Lagrange multiplier) is used to used to control the step-size. When C is increased, the algorithm has a more aggressive update.

4.2 Experiments

We replicate the evaluation framework used in (Li and Roth, 2006; Huang et al., 2008; Silva et al., 2011). We use the full, unaltered 5500-question training set from UIUC for training, and evaluate on the 500-question test.

To demonstrate the impact of our model's novel features, we performed a feature ablation test (Table 2) in which we removed groups of features from the full feature set.

Feature Set	**Fine**	**Coarse**
All	**92.0**	**96.2**
-clusters	90.2	96
-vectors	90	95.4
-clusters, vectors	89.8	95.2
-lists	88	94
-clusters, vectors, lists	86.2	92.8
-definition disambiguation	91	94.8
-quantity focus differentiation	90.2	96

Table 2: Feature ablation study: accuracies on coarse and fine-grained labels after removing specific features from the full feature set.

System	Fine	Coarse
Li and Roth 2002	84.2	91.0
Huang et al. 2008	89.2	93.4
Silva et al. 2011	90.8	95.0
Our System	**92.0**	**96.2**

Table 3: System comparison of accuracies for fine (50-class) and coarse (6-class) question labels.

4.3 Discussion

Our model significantly outperforms all previous results for question classification on the UIUC dataset (Table 3). Furthermore, we accomplished this without significant manual feature engineering or rule-writing, using a simple online-learning algorithm to determine the appropriate weights.

5 Answer Extraction

In this section we discuss techniques for short answer extraction once questions have been classified into a particular UIUC type. We employ a "sieve" approach, as in (Lee et al., 2011), that has seen some success in tasks like coreference resolution and is creating a bit of a renaissance in rule-based, as opposed to machine learning, approaches in NLP. We provide in this paper one example of how instead of taking an either/or approach, both methods can be combined into a high performance system. We focus below on the sieves that are specific to question types where we have been able to profitably employ WordNet for finding the right short answer. Preliminary results have been positive employing this approach.

We have two strategies that are used across the base question types: employing semantic role labels and recognizing appositives.

Feature Type	guitar	Cup
Lemma	guitar	cup
Shape	x+	Xx+
Authority List	instrument	sport
Word Vector*	vocals, guitars, bass, harmonica, drums	champions, championship, tournament
Brown Cluster Prefix	0010, 001010, 0010101100, ...	0111, 011101, 0111011000, ...

Table 1: Features used for head words. Each dimension of the corresponding word vector was used as a real-valued feature. *Nearest neighbors of the corresponding word vector are shown.

5.1 Corpus

Our current testing corpus consists of three parts. The first is an open source Q&A test set developed at Carnegie Mellon University (Smith et al., 2008)[4] consisting of roughly 1000 question and answer pairs on Wikipedia articles. The second is a proprietary Q&A test set developed at IPsoft consisting of a growing set of question answer pairs currently numbering roughly 2000 pairs and conducted on short sections of Wikipedia articles. The third test set is TREC-8 (Voorhees, 1999).

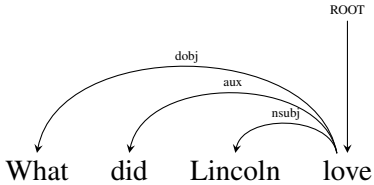
5.2 Semantic Role Labels

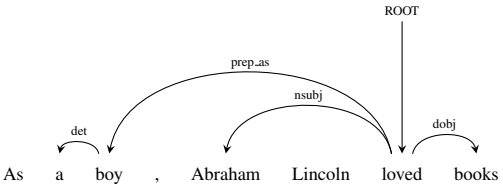
We employ the semantic role labeling of ClearNLP (Choi, 2012)[5]. While the labels are consistent with PropBank (Palmer et al., 2005), ClearNLP fixes the definition of several of the labels (A2-A5) that are left undefined in PropBank. A0 is the "Agent" relation, which is often the subject of the sentence. A1 is the "Patient" or object of the sentence. The remainder can be found in (Choi, 2012).

Let's look at an example and the list the steps followed in the code to analyse the question and answer.

(6) Q: What did Lincoln love?
A: As a boy, Abraham Lincoln loved books.

We have the following dependency graphs among the tokens in each sentence:

(7)


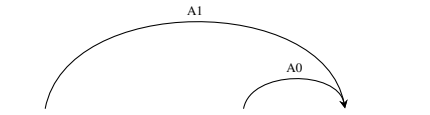
(8)


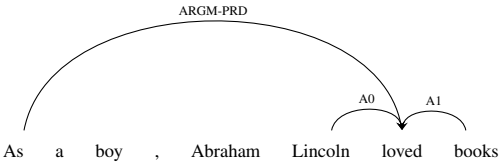
and part of speech labels

(9) What did Lincoln love?
WP VBD NNP VB

(10) As a boy, Abraham Lincoln loved books.
IN DT NN NNP NNP VBD NNS

and semantic role labels

(11)


(12)


1. We collect basic information from the question and answer sentence
   (a) find the question word, e.g. "what", "when", "where", etc. In Example 6 it is "what-1"
   (b) Locate the verb node nearest to the question word. In Example 6 it is "love-4"
   (c) Find the semantic relations in the question. We find an Agent/A0 relationship

[4] download from http://www.cs.cmu.edu/~ark/QA-data/

[5] http://www.clearnlp.com

between Lincoln-3 and the verb love-4. We find a Patient/A1 relationship between the question word What-1 and the verb love-4. (See Examples 11 and 12).

(d) Find semantic relations in the answer sentence. We find an Agent/A0 relationship between Lincoln-6 and the verb loved-7. We find an ARGM-PRD relationship between As-1 and the verb loved-7. We find a Patient/A1 relationship between books-8 and the verb loved-7. (See Examples 11 and 12).

(e) Perform a graph structure match between the question and answer graphs formed by the set of their semantic role labels. Find the parent graph node in the answer that matches as many nodes in the question as possible. In our example, loved-7 is the best match. (See Examples 11 and 12).

2. Collect and score candidate answer nodes. Score each semantic child for best parent found in the previous step, based on part of speech, named entity, dependency relations from Stanford's CoreNLP (Manning et al., 2014), and semantic role label information. We initialize each child to a value of 1.0 and then penalize it by 0.01 for the presence of any out of a set of possible undesirable features, as follows:

   - The candidate's semantic role label starts with "ARGM", meaning that its semantic role is something other than A0-A5. (See Examples 11 and 12). Note that this is only applied in cases where the question type has been identified as "Human" or "Entity"
   - The node's dependency label = "prep*" indicating that it is a prepositional relationship. Note that this is only applied in cases where the question type has been identified as "Human" or "Entity"
   - If the candidate node is the same form (word spelling) as in the question, or its WordNet hyponym
   - If the candidate node is the same root (lemma) as in the question, or its WordNet hyponym
   - If the candidate node is lower case. Note that this is only applied in cases where the question type has been identified as "Human" or "Entity"
   - If the candidate node has a child with a different semantic role label than in the question
   - If the candidate node is an adverb or a Wh- quantifier as marked by its part of speech label

3. Pick the dependency node with highest confidence score as the answer node. In our example we have As-1 = 0.97, Lincoln-6 = 0.96 and books-8 = 0.99.

Note that the step of scoring the answer nodes enumerates a small feature set with hand-set coefficients. We expect in a future phase to enumerate a much larger set of features, and then set the coefficients based on machine learning over our corpus of question-answer pairs. One simple experiment to show the value of semantic role labeling was conducted on a portion of our testing corpus. Using semantic role labels we achieved total of 638 correct answers out of 1460 questions (which was the total number in the IPsoft internal Q&A test set at the time of the test), for a correctness score of 43.7%. Without semantic role labels the result was 462 out of 1460, or 31.6%.

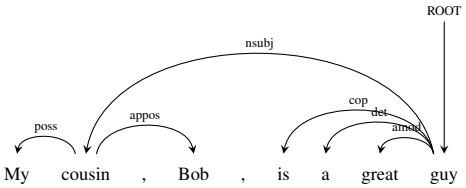
5.3 Appositives

The appositive is a grammatical construction in which one phrase elaborates or restricts another. For example,

(13) My cousin, Bob, is a great guy.

"Bob" further restricts the identity of "My cousin".

(14)



We use the appositive grammatical relation to identify the answers to "What" questions.

5.4 Entity Question Type

Short answer extraction for the Entity question type has some specialized rules for some subtypes, and some rules which are applied generally to all

the other subtypes. We are also exploring using WordNet (Fellbaum, 1998) synsets to get word lists that are members of each Entity subtype (see Table 4). This appears to have a significant effect, since 10 questions are answerable with this approach just addressing two of the 22 Entity subtypes. More work is needed to get comprehensive statistics.

5.4.1 Entity.animal Subtype

1. First try to find an appositive relationship. If there is one, use it as the answer. For example 14, if we ask "Who is a great guy?" we have a simple answer with "Bob" as the appositive. If that fails:

2. try the approach described above in subsection 5.2 and keep the candidate with the highest confidence score

5.4.2 Entity.creative Subtype

1. First try to find an appositive relationship. If there is one, use it as the answer. If that fails:

2. try the approach described above in subsection 5.2 and keep the candidate with the highest confidence score. If that fails:

3. find the first capitalized sequence of words and return it

5.4.3 All Other Entity Subtypes

1. First try to find an appositive relationship. If there is one, use it as the answer. If that fails:

2. try the approach described above in subsection 5.2 and keep the candidate with the highest confidence score

5.5 Example

Take for example the following

(15) Q: What shrubs can be planted that will be safe from deer?
A: Three old-time charmers make the list of shrubs unpalatable to deer: lilac, potentilla, and spiraea. Short Answer: Lilac, potentilla, and spiraea.

Knowing from WordNet that 112310349:{lilac}, and 112659356:{spiraea, spirea} (although not potentilla) are hyponyms of shrub makes it easy to find the right dependency parse subtree for the short answer.

Similarly for

(16) Q: What athletic game did dentist William Beers write a standard book of rules for?
A: In 1860, Beers began to codify the first written rules of the modern game of lacrosse. Short Answer: Lacrosse.

knowing that 100455599:{game} is a hypernym of 100477392:{lacrosse} makes finding the right answer in the sentence easy.

6 UIUC Question Types and Synsets

Table 4 lists all the types and subtypes in the UIUC taxonomy and the WordNet (Fellbaum, 1998) synset numbers that correspond to semantic types for the UIUC types. These are used to get all words that are in the given synsets as well as all words in the synsets that are more specific in the WordNet hyponym hierarchy than those listed. Note that below we prepend to the synset numbers a number for their part of speech. In the current scheme all are nouns, so the first number is always a "1". We only elaborate subtypes of Entity, Human, and Location as the other categories do not use WordNet for matching.

7 Conclusion

Using a WordNet-based word replacement method appears to be better for question answering than using word similarity metrics. In preliminary tests 10 questions in a portion of our corpora are answerable with this approach just addressing two of the 22 Entity subtypes with WordNet based matching. While more experimentation is needed, the results are intuitive and promising. The current approach should be validated and compared against other approaches on current data sets such as (Peñas et al., 2015).

Class	Definition	Synsets
ABBREVIATION	abbreviation	
ENTITY	entities	
animal	animals	100015388
body	organs of body	105297523
color	colors	104956594
creative	inventions, books and other creative pieces	102870092, 103217458, 103129123
currency	currency names	113385913, 113604718
dis.med.	diseases and medicine	114034177, 114778436
event	events	100029378
food	food	100021265
instrument	musical instrument	103800933
lang	languages	106282651
letter	letters like a-z	
other	other entities	
plant	plants	100017222
product	products	100021939
religion	religions	108081668, 105946687
sport	sports	100433216, 100523513, 103414162
substance	elements and substances	100020090
symbol	symbols and signs	
technique	techniques and methods	
term	equivalent terms	
vehicle	vehicles	103100490
word	words with a special property	
DESCRIPTION	description and abstract concepts	
HUMAN	human beings	
group	a group or organization of persons	107950920
ind	an individual	102472293
title	title of a person	
description	description of a person	
LOCATION	locations	
city	cities	108226335, 108524735
country	countries	108168978
mountain	mountains	109359803, 109403734
other	other locations	108630039
state	states	108654360
NUMERIC	numeric values	

Table 4: UIUC class to WordNet synset mappings

References

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Peter Brown, Peter Desouza, Robert Mercer, Vincent dellaPietra, and Jenifer Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Jinho D. Choi. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. thesis, University of Colorado at Boulder, Boulder, CO, USA. AAI3549172.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 927–936. Association for Computational Linguistics.

K. Sparck Jones, S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779 – 808.

Gerhard Kremer, Katrin Erk, Sebastian Pad, and Stefan Thater. 2014. What Substitutes Tell Us – Analysis of an "All-Words" Lexical Substitution Corpus. In *Proceedings of EACL*, Gothenburg, Sweden.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(03):229–249.

Chris Manning, John Bauer, Mihai Surdeanu, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*. Now Pub.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity: Measuring the Relatedness of Concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anselmo Peñas, Christina Unger, Georgios Paliouras, and Ioannis A. Kakadiaris. 2015. Overview of the CLEF question answering track 2015. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 539–544.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95*, pages 448–453. Morgan Kaufmann.

Joao Silva, Luísa Coheur, Ana Cristina Mendes, and Andreas Wichert. 2011. From symbolic to subsymbolic information in question classification. *Artificial Intelligence Review*, 35(2):137–154.

Noah A. Smith, Michael Heilman, , and Rebecca Hwa. 2008. Question Generation as a Competitive Undergraduate Course Project. In *NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, VA, September.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Ellen M. Voorhees. 1999. Overview of the TREC 2002 Question Answering Track. In *In Proceedings of the 11th Text Retrieval Conference (TREC)*, pages 115–123.

An overview of Portuguese WordNets

Valeria de Paiva
Nuance Communications, USA
valeria.depaiva@nuance.com

Livy Real
IBM Research Brazil
livyreal@gmail.com

Hugo Gonçalo Oliveira
CISUC, DEI, Univ. Coimbra, Portugal
hroliv@dei.uc.pt

Alexandre Rademaker
IBM Research and FGV/EMAp, Brazil
alexrad@br.ibm.com

Claudia Freitas
PUC-Rio, Brazil
claudiafreitas@puc-rio.br

Alberto Simoẽs
Universidade do Minho, Portugal
ambs@ilch.uminho.pt

Abstract

Semantic relations between words are key to building systems that aim to understand and manipulate language. For English, the "de facto" standard for representing this kind of knowledge is Princeton's WordNet. Here, we describe the wordnet-like resources currently available for Portuguese: their origins, methods of creation, sizes, and usage restrictions. We start tackling the problem of comparing them, but only in quantitative terms. Finally, we sketch ideas for potential collaboration between some of the projects.

1 Introduction

Semantic relations are a key aspect when developing computer programs capable of handling language – they establish (labeled) associations between words and can be integrated into lexical-semantic knowledge bases. Available since the beginning of the 1990s, Princeton's WordNet (Fellbaum, 1998), henceforth PWN, is a paradigmatic lexical resource. Originally created for English, its model is now a "de facto" standard, due to its wide use in applications and its adaptation to different languages.

For Portuguese, the first resource of this kind, WordNet.PT (Marrafa, 2001), was announced in 2001 but, unlike PWN, was never free to use. This meant that, in practice, there was still no open Portuguese wordnet. In parallel, a few alternatives to the wordnet model arose, some of which were compared in (Santos et al., 2010). But if those alternatives proved themselves useful for some tasks, they were not enough to enable all of the standard uses of a wordnet in Natural Language Processing (NLP), including similarity computation or word sense disambiguation. As the need for a Portuguese wordnet was keenly felt, in the early 2010s, several projects sprung up aiming to develop free Portuguese wordnets. We describe some of those wordnets, while indicating where they were created, their construction process, their availability and, when possible, their size.

We recall the wordnet model, its adaptation to other languages, and how these adaptations may be expanded through content alignment. Then, we describe the Portuguese wordnets we are aware of, alternative lexical-semantic resources, and go on to focus on the open wordnets. After that, we briefly compare the previous along a set of relevant features for processing Portuguese. Then, we suggest work leveraging what is already planned for these wordnets, as well as some ideas for collaboration. Knowing where we are in terms of our wordnets is an essential first step in establishing lexical resources, which are vital to the computational processing of the Portuguese language. [1]

2 WordNet and Alternatives

Lexical knowledge bases are organized repositories of lexical items, usually including information about the possible meanings of words, relations

[1] This paper is a shorter English version of our previous article, in Portuguese (Gonçalo Oliveira et al., 2015).

between them, definitions, and phrases that exemplify their use. The Princeton WordNet model, with English as its target language, is probably the most popular representative of this type of lexical knowledge base. Its flexibility has led not only to its growing use by the NLP community, but also to the adaptation of the model to other languages.

PWN was created manually in the early 1990s and has been updated several times since then. Initially based on psycholinguistic principles, it combines traditional lexicographic information, similar to that in a dictionary, with an appropriate organization for computational use, which facilitates its application as a basis for lexical-semantic knowledge. Like a thesaurus, PWN is organized in groups of synonymous lexical items, called *synsets*, which can be seen as the possible lexicalizations for the concepts in the language. Besides synonymy, inherent to synsets, PWN covers other types of semantic relation between synsets. For example, hypernymy – a concept is a generalization of another – or meronymy – a concept is a part of another. In addition, each synset has a part-of-speech (noun, verb, adjective or adverb); a gloss, similar to a definition in a dictionary; and it may still have phrases that illustrate its use. The inclusion of a lexical item in a synset indicates a sense of that item.

Both its free availability and the flexibility of its model were crucial to the success and widespread use of PWN. This made it possible to integrate PWN into a large number of NLP or knowledge management projects, making it virtually the standard model of a lexical resource for several languages. The popularity of the PWN knowledge base model led to the creation of the Global WordNet Association (GWA), a non-commercial organization that provides a platform for discussion, sharing and linking the wordnets of the world.

2.1 Multilingual Wordnets

Many people have studied the possibility of aligning, as far as possible, wordnets of different languages, given their similarities. Thus, the unveiling of multilingual wordnets such as EuroWordNet (Vossen, 1997) or MultiWordNet (Pianta et al., 2002), which nonetheless follow very different approaches. In EuroWordNet, wordnets are created independently for each language, and only after that they are aligned, relying on similarities or, indirectly, using Princeton WordNet as a pivot, through the so-called Inter-Language Index (ILI). In MultiWordNet, the first step was to translate, as much as possible, one wordnet, usually Princeton's, into the other languages. Among the multilingual wordnets aligned with PWN, there are, for instance, BalkaNet (Stamou et al., 2002), dedicated to the languages of the Balkans, and the Multilingual Central Repository (Gonzalez-Agirre et al., 2012) (henceforth, MCR) dedicated to the languages of Spain.

Open Multilingual WordNet (Bond and Foster, 2013), henceforth OMWN, is an initiative to facilitate access to different wordnets, for different languages. To this end, wordnets, created independently, were normalized using PWN, and then connected to each other and accessed through a common interface. Another initiative that should be mentioned is the Universal WordNet (de Melo and Weikum, 2009) (henceforth, UWN), a multilingual lexical knowledge base automatically built from PWN and the alignment of multilingual versions of Wikipedia.

There are also several projects on the alignment of PWN with other lexical resources or knowledge bases. These include, for instance, YAGO (Suchanek et al., 2007), UBY (Gurevych et al., 2012), BabelNet (Navigli and Ponzetto, 2012), SUMO (Pease and Fellbaum, 2010) and DOLCE (Gangemi et al., 2010).

2.2 Closed Portuguese WordNets

There is no doubt that the open-source character of PWN was key in its wide acceptance. Still, not all resources that followed on the footsteps of PWN have chosen to make their results freely available. We describe three projects that resulted in Portuguese wordnets that are not free to use.

WordNet.PT (Marrafa, 2001), henceforth WN.PT, was the first Portuguese wordnet, in development since 1998. Its construction is essentially manual and it follows the EuroWordNet (Vossen, 1997) model, which means WN.PT is created from scratch for Portuguese. WN.PT 1.6, released in 2006, covers a wide range of semantic relations, including: hypernym, whole/part, equivalence, opposition, categorization, instrument-for, or place-of. More recently, WN.PT was expanded to *Global WordNet.PT* (Marrafa et al., 2011), which contains 10,000 concepts, including nouns, verbs and

adjectives, their lexicalizations in different variants of Portuguese and their glosses, in a network of more than 40,000 relation instances. An approach to expand the WN.PT semi-automatically with relations extracted from a corpus (Amaro, 2014) was recently presented, which shows that the project is still active.

WordNet.BR (henceforth, WN.BR) aimed to be a wordnet for Brazilian Portuguese. In its first development phase (Dias-da-Silva et al., 2002), a team of linguists analyzed five Portuguese dictionaries and two corpora to collect information on synonymy and antonymy. This resulted in the manual creation of synsets and antonymy relations between them, and writing some glosses and example sentences. In a second phase (Dias-da-Silva, 2006), its synsets were manually aligned with PWN, in a similar process to that followed in the EuroWordNet project, using bilingual dictionaries. After this alignment, the semantic relations between synsets with equivalents in Portuguese and English were inherited. It is assumed that the full version of WN.BR covers relations of hyperonymy, part-of, cause and implication (entailment). However, this version is not available online. One can view and download the results of phase one, available under the name of Electronic Thesaurus of Portuguese (TeP) (Maziero et al., 2008). TeP includes more than 44,000 lexical items, organized into 19,888 synsets, which in turn are connected through 4,276 antonymy relations.

MultiWordNet.PT, commonly referred to as MWN.PT, is the Portuguese section of the MultiWordNet project (Pianta et al., 2002), which can be purchased through the European Language Resources Association catalog. MWN.PT includes 17,200 manually validated synsets, which correspond to approximately 21,000 senses and 16,000 lemmas, covering both European and Brazilian variants of Portuguese. As a resource established under the MultiWordNet project, its synsets are derived from the translation of their PWN equivalents. Transitively, this resource turns out to be also aligned with the MultiWordNets of Italian, Spanish, Hebrew, Romanian and Latin.

The manual creation of a wordnet is a complex task, which requires much effort and time. When it was not possible to use an open Portuguese wordnet, researchers working on the processing of Portuguese felt the need to develop free alternatives which, in most cases, were also simpler. Those include OpenThesaurus.PT, typically used to suggest synonyms in word processors; PAPEL (Gonçalo Oliveira et al., 2008), a lexical-semantic network, automatically extracted from a Portuguese dictionary, with words connected through a wide range of semantic relationships; the Port4Nooj lexical resources (Barreiro, 2010), which include a set of definitions and semantic relations between words; and the Dicionário Aberto (Simões et al., 2012), an open electronic dictionary which includes also several explicit relationships between words.

3 Open Portuguese Wordnets

Open wordnets for Portuguese finaly appeared in the early 2010s. They were created by automatic or semi-automatic means and all assume that lexical-semantic resources must be open-source to be really useful to the community. We present four wordnets that fall in this category.

3.1 Onto.PT

The Onto.PT (Gonçalo Oliveira and Gomes, 2014) project begun in 2008. To create a new wordnet in a completely automatic fashion, Onto.PT used several lexical resources available at the time, with special focus on those of the project PAPEL (Gonçalo Oliveira et al., 2008), including grammars to extract relations from dictionaries. Other exploited resources include Wiktionary.PT, Dicionário Aberto (Simões et al., 2012), TeP (Maziero et al., 2008), OpenThesaurus.PT and, more recently, OpenWN-PT (de Paiva et al., 2012).

The creation of Onto.PT follows the ECO approach (Gonçalo Oliveira and Gomes, 2014), tailored to for this project, but flexible enough to integrate words and relations extracted from different sources. ECO is different from other approaches because it tries to learn the whole structure of a wordnet, including the contents and boundaries of synsets, as well as the synsets involved in semantic relations. Hence, despite exploring, automatically, handcrafted resources, the authors refer to ECO as a "fully automatic" approach. It consists of three main phases: (i) relation extraction between words; (ii) synset discovery from the clusters of the extracted synonymy network (an initial set of synsets, such as those of TeP, may be used as a starting point); (iii) mapping word arguments of remaining relations to the discovered synsets. In

Onto.PT 0.6 (Gonçalo Oliveira et al., 2014), dictionary definitions were also assigned to synsets, automatically.

Onto.PT is different from the typical wordnet, not only for its creation process, but also because it includes a wide range of semantic relations that are not in PWN. Those relations are the same as the ones in PAPEL, extracted from dictionaries, and include causation, purpose, location or manner, among others.

On the one hand, ECO allows for the creation of a large knowledge base with little effort – Onto.PT 0.6 covers ≈169,000 distinct lexical items, organized in ≈117,000 synsets, which in turn are related through ≈174,000 relation instances. On the other hand, there are reliability consequences. For example, in Onto.PT 0.35, 74% of synsets were correct, in 18% there was no agreement between two judges, and the remaining had at least one incorrect word. The quality of relationships also varies dramatically depending on the type. Considering that relations between incorrect synsets are also wrong, the hypernymy connections were just 65% correct and between 78%-82% in a set with other relation types. These evaluation efforts are described in (Gonçalo Oliveira and Gomes, 2014). Nevertheless, Onto.PT was used, for instance, in the expansion of synonyms for information retrieval (Rodrigues et al., 2012) or for creating lists of causal verbs (Drury et al., 2014).

Due to its design, Onto.PT is a dynamic resource and, from release to release, may have significant changes in the number and size of its synsets. Thus, it is not planned to be aligned with PWN. Onto.PT is freely available in RDF/OWL[2], following an existing PWN model (van Assem et al., 2006), expanded to cover all its relation types.

3.2 OpenWordNet-PT

OpenWordNet-PT (de Paiva et al., 2012) abbreviated to OpenWN-PT, is a wordnet originally developed as a syntactic projection of the Universal WordNet (UNW). Its long-term goal is to serve as the main lexicon for a NLP system, focused on logical reasoning, based on representation of knowledge, using an ontology such as SUMO. The process of creating OpenWN-PT uses machine learning techniques to build relations between graphs representing lexical information from versions in multiple languages of Wikipedia entries and open electronic dictionaries. OpenWN-PT has constantly been improved through linguistically motivated additions, either manually or from evidence in large corpora. This is also the case for the lexicon of nominalizations, NomLex-PT, tightly integrated with the OpenWN-PT (Freitas et al., 2014).

OpenWN-PT employs three language strategies in its lexical enrichment process: (i) translation; (ii) corpus extraction; (iii) dictionaries. Regarding translations, glossaries and lists produced for other languages, such as English, French and Spanish, are used, automatically translated and manually revised. The addition of data from corpora contributes with words or phrases in common use, which may be specific to Portuguese or do not appear in other wordnets. The first corpora experiment in OpenWN-PT was the integration of NomLex-PT. The use of a corpus, while useful for specific conceptualizations in the language, brings additional challenges for the mappings alignment, since it is expected that there will be expressions for which there is no synset in the English wordnet. As for the information in dictionaries, this was used indirectly through PAPEL (Gonçalo Oliveira et al., 2008).

Like Onto.PT, OpenWN-PT is available in RDF/OWL (Real et al., 2015), following and expanding, when necessary, the mapping proposed by (van Assem et al., 2006). Both the OpenWN-PT data and schema of the RDF model are freely available for download. The philosophy of OpenWN-PT is to keep a close connection with PWN, but try to fix the biggest mistakes created by the automated methods, through language skills and tools. A consequence of this close connection is the ability to minimize the impact of lexicographical decisions on splitting/grouping the senses in a synset. While such decisions are, to a great extent, arbitrary, the practical criterion of following the multilingual alignment behaves as a pragmatic and practical guiding solution.

OpenWN-PT was chosen by the developers of Freeling (Padró and Stanilovsky, 2012), OMWW (Bond and Foster, 2013), BabelNet and Google Translate, as the representative Portuguese wordnet in those projects, respectively, due to its comprehensive coverage of the language and its accuracy. OpenWN-PT currently has 43,925 synsets, of which 32,696 correspond to nouns, 4,675 to verbs, 5,575 to adjectives and 979 to ad-

[2] http://ontopt.dei.uc.pt

verbs. Besides being available for download, the data can be retrieved via a SPARQL endpoint[3] and can be consulted and compared with other wordnets both through the OMWN interface and its own interface[4].

3.3 PULO

PULO (Simões and Guinovart, 2014), short for Portuguese Unified Lexical Ontology, intends to incorporate resources from open publicly available wordnets into a free Portuguese wordnet, perfectly aligned and included in the MCR project (Gonzalez-Agirre et al., 2012), which already includes wordnets for Spanish, Catalan, Basque and Galician, in addition to PWN.

The beginning of this project, in late 2014, involved some experiments on the translation and alignment between the English, Spanish and Galician wordnets. Beyond those, this process used probabilistic translation dictionaries (Simões and Almeida, 2003), a dynamic Portuguese-Galician translation dictionary (Guinovart and Simões, 2013), and the official Orthographical Vocabulary of the Portuguese Language. This resulted in ≈50,000 word meanings, but only ≈17,000 were actually added to PULO. This was due to the statistical nature of the approach and the cutoff line established. The scoring value obtained for each meaning was properly stored on the database and may serve as a measure of relevance or quality of each meaning.

Currently, as the other wordnets of MCR, the ontological structure of PULO is the same as PWN. Despite this similarity, the internal structure of the database allows each individual wordnet to be easily extended to new concepts. PULO is available for download and has currently 25,711 senses, corresponding to 17,854 synsets. In a second stage of the process, a machine translation of glosses was produced using the MyMemory API[5]. Through the same interface, it is possible to consult the other languages of the MCR, as well as to browse through the base ontology.

3.4 Ufes WordNet

The Ufes WordNet (Gomes et al., 2013) (UfesWN.BR) aims at building a Brazilian Portuguese database with a similar structure to PWN, based on automatic translation. For this, a tool based on the Google Translate API was developed to translate the contents of PWN. UfesWN.BR covers 34,979 words, grouped in 48,981 synsets, connected by 238,413 relations. However, only 31,6% of the English synsets were translated and these translations are not very reliable. In the scope of this project, the glosses of PWN were also translated. They could be useful for other projects, depending on the quality and easiness of alignment, which has not been investigated.

4 Comparing Open WordNets

Table 1 summarises the main properties of the Portuguese wordnets. The most common alternative to the creation of a wordnet for Portuguese is based on translation, manual (MWN.PT), automatic (UfesWN.BR), based on a syntactic projection (OpenWN-PT), or on triangulation between resources (PULO). Within these four approaches, PULO stands out for using as a "pivot", not only the English wordnet, but also the wordnets for Spanish and Galician. Unlike all others, the structure of Onto.PT is learned fully automatically, based on the extraction of relationships from other textual resources or wordnets, and discovering clusters of synonyms, used as synsets. Among the advantages of a completely manual approach is the creation of a resource with an accuracy of virtually 100%. On the other hand, with an automatic approach, a larger resource can be created in a shorter time, avoiding tedious and time-consuming work, prone to raise issues. A semi-automatic method where expediency can be reigned in by accuracy would seem the best approach.

Name	Creation		Update	Usage
	Synsets	Relations		
WN.PT	manual	manual	manual	closed
WN.BR	manual	transitivity	manual?	free synsets
MWN.PT	manual? translation	transitivity	?	paid license
Onto.PT	RE,*clustering*	RE,*clustering*	automatic	free
OpenWN-PT	UWN projection	transitivity	semi-autom	free
UfesWN.BR	machine translation	transitivity	?	free
PULO	triangulation	transitivity	semi-autom	free

Table 1: Properties of Portuguese wordnets. A '?' is shown for fields we could not fill.

We also made a superficial comparison of their lastest versions, that should not be seen as more than a purely quantitative tabling. We do not consider the consistency nor the usefulness of the con-

[3] http://logics.emap.fgv.br:10035/repositories/wn30

[4] http://wnpt.brlcloud.com/wn/

[5] http://mymemory.translated.net/

tents of the various Portuguese wordnets.

On the number of covered lexical items, Onto.PT stands out for including more than three times more lexical items than the second largest wordnet, OpenWN-PT. This confirms that a fully automatic construction approach leads to a larger resource. Equally important for the size of Onto.PT, is the amount (currently six) and the type of resources used, including: resources that cover different variants of Portuguese, which can lead to minor spelling variations; and dictionaries, which already have a wide coverage of the language. Either manually or automatically, it is common to exploit dictionaries in the construction of a wordnet. Still, their automatic exploitation results in many different words and meanings that exist and are valid, but a large slice are of little use in colloquial Portuguese.

On the number of word senses, synsets and relation instances, Onto.PT also stands out from the rest. But it should be noted that there is an intrinsic trade-off between the size of a wordnet and the accuracy and usefulness of the resource under scrutiny. One of the difficulties in developing a wordnet is precisely to decide, on the one hand, if two words are to be regarded as synonymous and thus placed within the same synset and, on the other hand, which words should be in different synsets. These are typical lexicography challenges to which there is probably no final unique answer. But there seems to be a consensus that a very large number of synsets is a sign of "noise" in the process of grouping words and/or in the discrimination process. Correction/accuracy is undoubtedly one of the bottlenecks of building wordnets. If, on the one hand, size and coverage are a quantitative comparison, which is relatively simple, the same cannot be said about the quality assessment. PWN, built manually, may even reflect questionable decisions, but does not contain "errors" as such, as we are using it as a baseline for comparison. As for the wordnets built automatically, or semi-automatically, for languages other than English, quality assessment will always be an issue, since there is no golden reference available – this is precisely what they want to become. From this perspective, resources that rely on human labor have an advantage, although we do not know exactly how this advantage can and should be measured. An alignment with PWN may be important for obtaining additional knowledge, mostly from other resources aligned with it In addition to relation inheritance, an alignment allows access to knowledge of other extensions of PWN, such as WordNet-domains, SentiWordNet or TempoWordNet. On the other hand, a blind alignment does not consider that different languages represent different socio-cultural realities, do not cover the same part of the lexicon and, even where they seem to be common, several concepts are lexicalized differently (Hirst, 2004).

Both WN.PT and Onto.PT cover a wide range of relation types, some not typically present in wordnets. We recall that, for Onto.PT, their extracted was possible due to the regularities in dictionary definitions.

5 Building on Open WordNets

We presented and compared various wordnets that currently exist for Portuguese. Among them, four are freely available; until recently, one synset base (TeP) was also freely available; one (MWN.PT) may be purchased; and another can be explored online (WN.PT). The creation of these wordnets followed different approaches, from completely manual labour, through translation-based approaches with more or less manual labour, to an approach in which the whole structure is populated automatically. We hope to have shown that, currently, it makes no sense to regret that there is no Portuguese wordnet. In fact, the use of a wordnet in a project targeting Portuguese is becoming less of a problem of finding a workaround solution, and increasingly more one of choosing the most suitable within the available alternatives. This selection should consider, among other things, the need to align with other wordnets, the error tolerance, the coverage needs – both with regards to the lexical items and to relationships between them – and even the available budget. Since each wordnet has distinct characteristics, one should not discard the use of more than one wordnet in the same project.

It is sensible to ask whether all these alternatives make sense or if it would be preferable to focus on a single effort to build a single Portuguese wordnet, trying to harness the strong points of each of the projects described. The authors of this article, responsible for Onto.PT, OpenWN-PT and PULO, believe that there are advantages both on converging into a single wordnet and on keeping separate projects. Thus, in the short term, the development of each wordnet will remain a respons-

ability of its original team, but there will be closer monitoring of each other's work. The idea is that each project may reuse what is done by the others, this way minimizing duplicate work, but without losing sight of specific goals.

In a near future, Onto.PT will become a fuzzy wordnet, based on the redundancy accross several Portuguese computational lexical resources, including the other open wordnets, whose further updates will be welcome by this new initiative. Following ECO, confidence degrees will be assigned to each decision taken, including the membership of words to synsets (first experiments in Santos and Gonçalo Oliveira (2015)), or the attachment of relations to synsets. This will enable the users to select between a larger but less reliable resource and a smaller one with fewer issues.

Similarly to Onto.PT, the other open wordnets will devise the integration of the contents of each other, or replicate their enrichment approaches.

6 Conclusions

We presented a collection of Portuguese wordnets. While none feels as mature as Princeton WordNet, some have already been used in applications. Joint efforts, as we started doing and hope to do more, seem the only way of making progress in this hard problem. Clearly, the envisaged applications will lead to slightly different strong points in our resources, but the common denominator remains to provide a wordnet that is open, large coverage and as reliable as possible.

References

Raquel Amaro. 2014. Extracting semantic relations from portuguese corpora using lexical-syntactic patterns. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14, Reykjavik, Iceland, May. ELRA.

Anabela Barreiro. 2010. Port4NooJ: an open source, ontology-driven portuguese linguistic system with applications in machine translation. In *Proceedings of the 2008 International NooJ Conference (NooJ'08)*, Budapest, Hungary. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. ACL Press.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Bento C. Dias-da-Silva, Mirna F. de Oliveira, and Helio R. de Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. In *Advances in Natural Language Processing (PorTAL 2002)*, LNAI, pages 189–196, Faro, Portugal. Springer.

Bento C. Dias-da-Silva. 2006. Wordnet.Br: An exercise of human language technology research. In *Proceedings of 3rd International WordNet Conference (GWC)*, GWC 2006, pages 301–303, South Jeju Island, Korea, January.

Brett Drury, Paula C.F. Cardoso, Janie M. Thomas, and Alneu de Andrade Lopes. 2014. Lexical resources for the identification of causative relations in portuguese texts. In *Proceedings of the 1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp)*, ToRPorEsp, pages 56–63, São Carlos, SP, Brasil, October. BDBComp.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. 2014. Extending a lexicon of Portuguese nominalizations with data from corpora. In *Proceedings of Computational Processing of the Portuguese Language - 11th International Conference (PROPOR 2014)*, São Carlos, Brazil, oct. Springer.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2010. Interfacing WordNet with DOLCE: towards OntoWordNet. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 3, pages 36–52. Cambridge University Press.

Marcelo Machado Gomes, Walber Beltrame, and Davidson Cury. 2013. Automatic construction of brazilian portuguese WordNet. In *Proceedings of X National Meeting on Artificial and Computational Intelligence*, ENIAC 2013.

Hugo Gonçalo Oliveira and Paulo Gomes. 2014. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2):373–393.

Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. 2008. PAPEL: A dictionary-based lexical ontology for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pages 31–40, Aveiro, Portugal, September. Springer.

Hugo Gonçalo Oliveira, Inês Coelho, and Paulo Gomes. 2014. Exploiting Portuguese lexical knowledge bases for answering open domain cloze questions automatically. In *Proceedings of the 9th Language Resources and Evaluation Conference*, LREC 2014, Reykjavik, Iceland, May. ELRA.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2525–2529. ELRA, LREC'12.

Hugo Gonçalo Oliveira, Valeria de Paiva, Cláudia Freitas, Alexandre Rademaker, Livy Real, and Alberto Simões. 2015. As wordnets do português. In Alberto Simões, Anabela Barreiro, Diana Santos, Rui Sousa-Silva, and Stella E. O. Tagnin, editors, *Linguística, Informática e Tradução: Mundos que se Cruzam*, volume 7(1) of *OSLa: Oslo Studies in Language*, pages 397–424. University of Oslo.

Xavier Gómez Guinovart and Alberto Simões. 2013. Retreading Dictionaries for the 21st Century. In José Paulo Leal, Ricardo Rocha, and Alberto Simões, editors, *2nd Symposium on Languages, Applications and Technologies*, volume 29 of *OpenAccess Series in Informatics (OASIcs)*, pages 115–126. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource. In *Proceedings of 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, pages 580–590, Avignon, France. ACL Press.

Graeme Hirst. 2004. Ontology and the lexicon. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–230. Springer.

Palmira Marrafa, Raquel Amaro, and Sara Mendes. 2011. WordNet.PT Global – extending WordNet.PT to Portuguese varieties. In *Proceedings of 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 70–74, Edinburgh, Scotland. ACL Press.

Palmira Marrafa. 2001. *WordNet do Português: uma base de dados de conhecimento linguístico*. Instituto Camões.

Erick G. Maziero, Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e Linguagem Humana*, TIL, pages 390–392.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Adam Pease and Christiane Fellbaum. 2010. Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 2, pages 25–35. Cambridge University Press.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of 1st International Conference on Global WordNet*, GWC 2002.

Livy Real, Fabricio Chalub, Valeria de Paiva, Claudia Freitas, and Alexandre Rademaker. 2015. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of 53rd Annual Meeting of the ACL and 7th International Joint Conference on NLP of Asian Federation of NLP - 4th Workshop on Linked Data in Linguistics: Resources and Applications*, Beijing, China, jul.

Ricardo Rodrigues, Hugo Gonçalo Oliveira, and Paulo Gomes. 2012. Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática*, 4(1):31–39, April.

Fábio Santos and Hugo Gonçalo Oliveira. 2015. Descoberta de synsets difusos com base na redundância em vários dicionários. *Linguamática*, page accepted for publication, December.

Diana Santos, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, and Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*, APL 2009, pages 681–700. APL.

Alberto Simões and Xavier Gómez Guinovart. 2014. Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain*, volume 8854 of *LNCS*, pages 239–248. Springer.

Alberto Simões, Álvaro Iriarte Sanromán, and José João Almeida. 2012. Dicionário-Aberto: A source of resources for the Portuguese language processing. In *Proceedings of Computational Processing of the Portuguese Language, 10th International Conference (PROPOR 2012), Coimbra Portugal*, volume 7243 of *LNCS*, pages 121–127. Springer, April.

Alberto M. Simões and J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September.

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. 2002. BalkaNet: A multilingual semantic network for the balkan languages. In *Proceedings of 1st Global WordNet Conference*, GWC'02.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW 2007, pages 697–706, Alberta, Canada. ACM Press.

Mark van Assem, Aldo Gangemi, and Guus Schreiber. 2006. RDF/OWL representation of WordNet. W3c working draft, World Wide Web Consortium, June.

Piek Vossen. 1997. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of DELOS workshop on Cross-Language Information Retrieval*, Zurich.

Towards a WordNet based Classification of Actors in Folktales

Thierry Declerck
DFKI GmbH
Saarbrücken, Germany &
Austrian Centre for
Digital Humanities (ACDH)
Vienna, Austria
declerck@dfki.de

Tyler Klement
Saarland University
Saarbrücken, Germany
klement.tyler@gmail.com

Antonia Kostova
Saarland University
Saarbrücken, Germany
akostova@coli.uni-saarland.de

Abstract

In the context of a student software project we are investigating the use of WordNet for improving the automatic detection and classification of actors (or characters) mentioned in folktales. Our starting point is the book "Classification of International Folktales", out of which we extract text segments that name the different actors involved in tales, taking advantage of patterns used by its author, Hans-Jörg Uther. We apply on those text segments functions that are implemented in the NLTK interface to WordNet in order to obtain lexical semantic information to enrich the original naming of characters proposed in the "Classification of International Folktales" and to support their translation in other languages.

1 Introduction

This short paper reports on the current state of a student software project aiming at supporting the automatized classification of folktales along the line of the classification proposed by Hans-Jörg Uther (2004). This classification scheme is considered as a central source for the analysis work of folklorists. It builds on former work by Antti Aarne (1961) and Stith Thompson (1977). In the following, we are using the acronym ATU for referring to (Uther, 2004): ATU standing for Aarne-Thompson-Uther.

We focus in the current work on the detection of common superclasses to the naming of the main actors (or characters) that are mentioned in the various types of folktales listed by Uther (2004). In doing this we are able to propose more generic classes of characters and an extended vocabulary, and so to link to other classification systems, like the Motif-Index of Folk-Literature proposed by Stith Thompson[1]. In general, we are aiming at a WordNet[2] based generation of lexical semantic relations for building a terminology network of actors/characters mentioned in folktales. Our work is anchored in the field of Digital Humanities (DH), where there is an increased interest in applying methods from Natural Language Processing (NLP) and Semantic Web (SW) technologies to literary work.

In the following sections we will present first the data we are dealing with and the transformations we applied on those for being able to use the NLTK interface to WordNet[3]. We describe then the functions of NLTK we are using and how we can benefit from those for building a more generic vocabulary and extending the basic terminology for classifying actors/characters in folktales.

Related work on this topic is presented in Declerck (2012), which is more focused on the use of Wiktionary for translation and also dealing rather with the formal representation of the terminology used in ATU.

2 The Data Source

We are taking the ATU classification scheme as our starting point. Just below we display the initial part of a *type* of folktale, which in ATU is marked using an integer, possibly followed by a letter. In this example we deal with type 2, which is included in the list of types "Wild Animal" (from type 1 to type 99), and more specifically within the list "The Clever Fox (Other Animal)" (from type 1 to type 69)[4].

[1] See the online version of the index: http://www.ruthenia.ru/folklore/thompson/index.htm.

[2] See (Fellbaum, 1998) and (Miller, 1995).

[3] NLTK is described in (Bird et al., 2009), with an updated online version: http://www.nltk.org/book/. At http://www.nltk.org/howto/wordnet.html the WordNet interface is described in details.

[4] See also https://en.wikipedia.org/wiki/Aarne-Thompson_classification_systems,

2 *The Tail-Fisher.* A bear (wolf) meets a fox who has caught a big load of fish. He asks him where he caught them, and the fox replies that he was fishing with his tail through a hole in the ice. He advises the bear to do likewise and the bear does. When the bear tries to pull his tail out of the ice (because men or dogs are attacking him), it is frozen in place. He runs away but leaves his tail behind [K1021]. Cf. Type 1891.
Combinations: This type is usually combined with episodes of one or more other types, esp. 1, 3, 4, 5, 8, 15, 41, 158, and 1910.

In this example, we can see the number of the type ("2"), its label ("The Tail-Fisher") and a text summarizing the typical motifs of this type of folktale. At the end of this "script", a link to a corresponding Thompson Motif-Index is provided ("[K1021]"). Finally, types are indicated, with which the current type is usually combined.

For us, a very interesting pattern in the description part of the type entry is "A bear (wolf)". This way (and also using more complex patterns), the author specifies variants of actors/characters that can play a role within a folktale type. We found this pattern interesting because our assumption is that in most of the cases only semantically related actors/characters can be mentioned in this text construct. And those pairs of variants give us a promising basis for trying to generate more generic terms from WordNet for classifying actors in folktales and so to support the linking of ATU to other classification schemes.

Our work consisted first in extracting from ATU the relevant text segments corresponding to such patterns and then to query WordNet in order to see if the characters named in such text segments are sharing relevant lexical semantic properties.

2.1 Pre-Processing the ATU Catalogue

In order to be able to apply functions of the WordNet interface of NLTK to the ATU classification scheme, we first had to transform the original document into a punctuation separated text format, using for this a Python script. For the type 6, just to present another example of an ATU type, we have now the following text format:

```
6˜Animal Captor Persuaded to
Talk.˜ A fox (jackal, wolf)
catches a chicken (crow, bird,
hyena, sheep, etc. ) and is
about to eat it. The weak animal
asks a question and the fox
answers. Thus he releases the
prey and it escapes. ˜K561.1
```

With this new format, where the sign " ˜ " is used as the separator, it is very easy to write code that is specialized for dealing with parts of the ATU entries. For our work, we concentrate only on the third field of the " ˜ " separated input file. This way we avoid the "noise" that could be generated if considering the use of parentheses in the second field (the label of the type), like:

```
Torn-off Tails (previously The
Buried Tail).
```

which is used in the label of type 2A.

2.2 Pattern Extraction

On the basis of a manual analysis of the ATU entries, regular expressions for detecting the formulation of variants of actors/characters have been formulated and implemented in Python. Below we show some examples of extracted text segments, on the basis of the Python script:

- A master (supervisor)
- an ox is so big that it takes a bird a whole day (week, year)
- A sow (hare)
- A giant has sixty daughters (sons)
- a brook (sea)
- A man puts a pot with hot milk (chocolate)
- A man who has recently been married meets a friend (neighbor, stranger)
- A wolf (bee, wasp, fly)
- A suitor (suitors)

with more details given in the French or German corresponding pages.

- a flea (fly, mouse)
- a series of animals (hen, rooster, duck, goose, fox, pig)
- a person (animal)
- An ant (sparrow, hare)

As the reader can see, each text segment starts with an indefinite Nominal Phrase (NP) and ends with a closing parenthesis. This pattern is consistently used in ATU, and corresponds to our intuition that a referent in discourse is mostly introduced by an indefinite NP. For the first step of our investigation of the use of WordNet for generating more generic terms for the mentioned actors, we decided to concentrate on the simple sequence "A/An Noun (Noun)", like for example "A fox (wolf)".

2.2.1 Accessing WordNet with the NLTK Interface

NLTK provides for a rich set of functions for accessing WordNet. The first function we applied was the one searching for the least common hypernym for the two words used in the pattern "A/An Noun (Noun)". Some few results on such a search for all the synsets of the considered noun-pairs are displayed below for the purpose of exemplification, where we indicate the least common hypernym with the abbreviation LCH:

- Synset(man.n.01) & Synset(fox.n.05) => LCH(Synset(person.n.01))
- Synset(fox.n.01) & Synset(jackal.n.01) => LCH(Synset(canine.n.02))
- Synset(fox.n.01) & Synset(cat.n.01) => LCH(Synset(carnivore.n.01))
- Synset(raven.n.01) & Synset(crow.n.01) => LCH(Synset(corvine_bird.n.01))

It is for sure interesting to see that depending on the word they are associated with, synsets of "fox", for example, can be related to a different hypernym. In the case of "fox.n.05" and "man.n.01" sharing the hypernym "person.n.01", we have to check if this case should be filtered out, since the hypernym is too generic. We tested for this the NLTK function "path_similarity", which computes a measure on the basis of the respective length of the path needed for each synset to the shared LCH. For "man.n.01" and "fox.n.05" the function "path_similarity" gives '0.2', while for "fox.n.01" and "jackal.n.01" it gives '0,33'. We might have '0,33' as a threshold for accepting the selected hypernym as a relevant generalization of the words used in the patterns of ATU we are investigating. Or allowing also lower similarity measures, but filtering out the selected hypernym on the basis of the length of the path leading from it to the root node. The LCH "canine.n.02' has a much longer path to "entity" as does the LCH "person.n.01". Our first experiments seem to indicate that the longer the path of the hypernym to the root node, the more informative is the generalization proposed by querying WordNet for the least common hypernym.

Additionally to those two functions of the NLTK interface to WordNet, we make use of the possibility to extract from WordNet all the hyponyms of the involved synsets. This can offer an extended word base for searching in folktale texts for relevant actors/characters. While this assumption seems reasonable in certain cases, like for example for the synset "overlord.n.01" for which we can retrieve hyponyms like "feudal_lord", "seigneur' and "seignior", it is not clear if it is beneficial to retrieve all the scientific names listed as hyponyms of the synset "fox.n.01", like "Urocyon_cinereoargenteus" or "Vulpes_fulva". But in any case, the terminology basis of the words used in ATU can this way be extended.

Last but not least, we take advantage of the multilingual coverage of WordNet, using for this another function implemented in NLTK. As an example, for the following pairs mentioned in ATU, we get from WordNet the French equivalents:

- Synset(fox.n.01) & Synset(wolf.n.01) => ['renard'] & ['loup', 'louve']
- Synset(dragon.n.02) & Synset(monster.n.04) => ['dragon'] & ['démon', 'monstre', 'diable', 'Diable']
- Synset(enchantress.n.02) & Synset(sorceress.n.01) => ['sorcière'] & ['enchanteur', 'ensorceleur', 'sorcière']

As part of future work, we are considering those multilingual equivalents provided by WordNet as a starting point for providing for a multilingual extension of the ATU classification.

3 An Ontology for ATU

In order to store all the results of the work described above, including the multilingual correspondences of the English terminology used in ATU, we decided to go for the creation of an ontology of ATU, a step which is also aiming at supporting the linking of this classification scheme to other approaches in the field. The ontology was generated automatically from the transformed ATU input data described in section 2.1., and encoded in the OWL and RDF(s) representation languages[5]. ATU not being a hierarchical classification, we decided to have only one class in the ontology, and to encode each type of ATU as an instance of this class. As a result, we have 2221 instances. The main class is displayed just below, using the Turtle syntax [6] for its representation:

```
:ATU
  rdf:type owl:Class ;
  rdfs:comment
  "\"Ontology Version of ATU\""@en ;
  rdfs:label "\"The Types of International
  Folktales Aarne-Thompson-Uther\""@en ;
  rdfs:subClassOf owl:Thing ;
.
```

An instance of this class, for example for the type 101, has the following syntax:

```
<http://www.semanticweb.org/tonka/
   ontologies/2015/5/tmi-atu-ontology#101>

 rdf:type :ATU ;

 linkToTMI <http://www.semanticweb.org/
   tonka/ontologies/2015/5/
 tmi-atu-ontology#K231.1.3> ;

 rdfs:comment "\"Type 101 of ATU\""@en ;

 rdfs:isDefinedBy "The Old Dog as Rescuer
   of the Child (Sheep). A farmer plans
 to kill his faithful old dog because
 it cannot work anymore. The wolf makes
 a plan to save the dog: The latter is to
 rescue the farmer's child from the wolf.
 The plan succeeds and the dog's life is
 spared. The wolf in return wants to
 steal the farmer's sheep. The dog
 refuses to help and loses the wolf's
 friendship . "@en ;

 rdfs:label "\"The Old Dog as Rescuer
   of the Child (Sheep)\""@en ;
.
```

The reader can see in this extensive example that each instance of the ATU class is named in the first line of the code by an Unique Resource Identifier (URI). The property "rdf:type" indicates that the object named by the URI is an instance of the class "ATU". The last element of the code, introduced by "rdfs:label", stores the original label in English ("en"). We will use this property "rdfs:label" to encode the multilingual correspondences. We encode the original description of the type as a value to the property "rdfs:isDefinedBy".

The property "linkToTMI" is the way we go for linking ATU types to Motifs listed in the Motif-Index of Folk-Literature (which we abbreviate with TMI). This linking is still in a preliminary stage, since we first have to finalize the corresponding TMI ontology, and also check the validity of the linking to TMI we extracted from the ATU book. This kind of linking is the one we will use for interconnecting all types of classification schemes used for folktales (and maybe also for other literary genres). We will add a property for including relevant hypernyms (and possibly hyponyms) extracted from WordNet to the current labels, contributing this way to the semantic enrichment of the original classification.

4 Conclusion and future Work

We presented work done in the context of a running student software project consisting in accessing WordNet for providing for lexical semantic information that can be used for enriching an existing classification scheme of folktales with additional terms gained from the extraction of relevant hypernyms (and to a certain extent from hyponyms) of words naming characters playing a central roles in folktales. The aim is to generate a WordNet based network of terms for the folktale domain.

As future work, an investigation will be performed in order to determine the optimal length of the path between a Lowest Common Hypernym (LCH) and the root node of WordNet as the filtering process for excluding irrelevant and noise introducing LCHs. We will also perform an evaluation of the extracted LCHs against a manually annotated set of ATU entries. And we will compare the French equivalents of the synsets proposed by WordNet with the French terms used in the French Wikipedia page for the AT. Additionally, we plan to compare our WordNet based approach as the basis for the linking between ATU and TMI to the machine learning approach to such a linking described in (Ofex et al., 2013).

[5] See `http://www.w3.org/2001/sw/wiki/OWL` and `http://www.w3.org/TR/rdf-schema/`.

[6] See http://www.w3.org/TR/turtle/ for more details.

Acknowledgments

We would like to thank Alyssa Price for providing for the manual analysis of the patterns occurring in ATU. Our gratitude goes also to the two anonymous reviewers for the very helpful comments on the previous version of this short paper.

References

Antti Aarne. 1961. *The Types of the Folktale: A Classification and Bibliography*. The Finnish Academy of Science and Letters, Helsinki.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Thierry Declerck, Karlheinz Mrth, Piroska Lendvai. 2012. Accessing and Standardizing Wiktionary Lexical Entries for the Translation of Labels in Cultural Heritage Taxonomies. *Proceedings of the Eight International Conference on Language Re-sources and Evaluation (LREC'12)*. Istanbul, Turkey.

Christiane Fellbaum (ed). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38, No. 11: 39-41.

Thierry Declerck, Karlheinz Mrth, Piroska Lendvai. 2013. Linking Motif Sequences to Tale Types by Machine Learning. *Proceedings of the 2013 Workshop on Computational Models of Narrative*, 166-182. Dagstuhl, Germany

Stith Thompson. 1977. *The Folktale*. University of California Press, Berkeley.

Hans J. Uther. 2004. *The Types of the Folktale: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. The Finnish Academy of Science and Letters, Helsinki.

Extraction and description of multi-word lexical units in plWordNet 3.0

Agnieszka Dziob
Wrocław University of Technology
Wrocław, Poland
agnieszka.dziob@pwr.edu.pl

Michał Wendelberger
Wrocław University of Technology,
Wrocław, Poland

Abstract

In this paper, we present methods of extraction of multi-word lexical units (MWLUs) from large text corpora and their description in plWordNet 3.0. MWLUs are filtered from collocations of the structural type Noun+Adjective (NA).

1 Introduction

Our focus in this paper are multi-word lexical units (henceforth, MWLUs), derived from collocations (automatically extracted from corpora). As in the case of many linguistic terms, there is no agreement among scholars on their common defining criteria. Two main approaches are distinguished. The first one treats as collocations all expressions that tend to co-occur in the immediate syntactic neighbourhood (Firth 1957). This approach is followed by the constructors of corpora (cf. Przepiórkowski 2012). The second approach puts the emphasis on the linguistic properties of collocations such as non-compositionality and impossibility of modification and substitution (Evert 2004). In this approach the term *collocation* is close to the term *multi-word expression* (henceforth, MWE), used in computational linguistics for the linkage of words of the established meaning, analysed as a whole (Sag et al. 2002) and to our understanding of the term MWLU. In the present paper we define MWLU by reference to *lexical unit* (henceforth, LU), a central element of a wordnet (Fellbaum 1998), a whole attributed with meaning and morphosyntactic properties (Derwojedowa et al. 2008). Thus, MWLU will be an LU, consisting of more than one word and constituting a semantic and morpho-syntactic whole. It is close in spirit to Maziarz et al. 2015 proposal saying that MWLU is *"built from more than one word, associated with a definite meaning somehow stored in one's mental lexicon and immediately retrieved from memory as a whole"* (Maziarz et al. 2015). Such a definition forces one to perceive MWLUs as having defined structure and semantics which makes the connection "behave like the single individual" (Calzolari et al. 2002).

2 Data preparation

In the work on extracting MWEs, IPI PAN Corpus[1] and the plWordNet corpus of the Wrocław University of Technology (Piasecki et al. 2014) corpora were used. The extraction was carried out using the set of MWeXtractor tools, developed for the purposes of the CLARIN[2] project. MWeXtractor is a package of tools, which was created for the purposes of the construction of MWLU's network in plWordNet and their syntactic description. It is the part of a bigger infrastructure for aimed for the work with text corpora. The package user has the access to the data cloud, where they record their own corpora (or uses the existing corpora available on the open licence). MWeXtractor tools package is available on the open licence. Sketch Engine is a tool for the work with corpora, which allows for the extraction of collocations on the basis of their grammatical relations (Kilgarriff et al. 2004). In many respects Sketch Engine and MWeXtractor do not differ from each other. For the purposes of the development of

[1] http://korpus.pl/
[2] http://clarin-pl.eu/

MWeXtractor package new statistical measures were implemented, described in this Section. Those measures, which are compilations or modifications of the known measures, improved extraction results, described in Sections 2 and 3.

In the first phase, the authors defined initial data (sets of corpora, tagset, WCCL's operators describing relations within a collocation (Radziszewski et al. 2011)). In addition, the order of candidates for MWLU can be changed and the continuity of the elements of a collocation does not have to be preserved. The next stage was a dispersion of collocations, through which candidates whose syntactic traits were regarded interesting, are being promoted. In the MWeXtractor package, apart from available measures that are present in the subject literature, the measures designed for the purposes of the present work and presented in Sections 2.1, 2.2, 2.3 were also implemented.

2.1 W Specific Exponential Correlation

The function W Specific Exponential Correlation is a compilation of a few other associative measures, of Specyfic Exponential Correlation among others described above. She is represented by the following pattern:

$$y = p(x,y)\log_2 \frac{p(x,y)^e}{p(x)p(y)}$$

And for her the described generalization is used the pattern:

$$y = p(x_1,x_2,...,x_n)\log_2 \frac{p(x_1,x_2,...,x_n)^e}{\prod_{2=1}^{n} p(x_i)}$$

2.2 W Order

W Order is the function based on the assumption, that for them the chic more peculiar to the given connection in which storage connections are appearing, with it more interesting, more certain collocation. The function is disregarding interpretation of the order of the chic, examining only their number and the frequency distribution in chics and from the frequency riots of the collocation for the given candidate, and studying only their attitude.

$$y = \frac{1}{\prod_{i=1}^{n}(1+\frac{f(S(t)i))}{maks(f(S(t)))+1)}}$$

2.3 W Term Frequency Order

This function W Term Frequency Order includes the frequency of appearing of the candidate which many associative measures are using assessed as good.

$$y = \mathrm{f}\ (t)\mathrm{WOrder}(t)$$

Two types of files are final data - files with lists k-best of candidates for MWE, and files with evaluations of these lists. The number of generated files in the ranking is equal ((and + V + C) ∗ R ∗ F), where and, V and are indicating C one by one number of exploited functions of associative, vector associative measures and classifiers, however R and F are one by one a number of rounds and folds of cross validation. Additionally for every file with the ranking generated is being Q of files of the evaluation of this ranking, where Q is a number of exploited functions of the evaluation of lists k-bests.
The final list of extracted collocations also contained collocaltions being already Lexical Units in plWordNet. Last filtering consisted in removing proper names and determined descriptions and these LU's.

2.4 Results

Table 1 presents the 20 bests of extracted collocations (of the k-best list). The list included forms of lemma according part of speech:

String of lemma of corpus
N:*link* A:*zewnętrzny* ('external link')
N:*raz* A:*pierwszy* ('first time')
N:*wojna* A:*światowy* ('word war')
N:*to* A:*sam* ('the same')
N:*samorząd* A:*terytorialny* ('local government')
N:*piłka* A:*nożny* ('football')
N:*porządek* A:*dzienny* ('agenda')
N:*papier* A:*wartościowy* ('security')
N:*sprawa* A:*wewnętrzny* ('affairs')
N:*igrzyska* A:*olimpijski* ('Olimpic Games')

N:*strona* A:*drugi* (‘other side’)
N:*podatek* A:*dochodowy* (‘income tax’)
N:*minister* A:*właściwy* (‘minister responsible’)
N:*finanse* A:*publiczny* (‘public finance’)
N:*rada* A:*nadzorczy* (‘supervisory board’)
N:*opieka* A:*zdrowotny* (‘healt care’)
N:*rok* A:*ubiegły* (‘last year’)
N:*ciąg* A:*daleki* (‘string far’)
N:*działalność* A:*gospodarczy* (‘bussines activity’)
N:*projekt* A:*rządowy* (‘government project’)

Table 1: Bests of extracted collocations

3 Syntactically non-compositional MWE’s

Automatic evaluation was the first phase of verification of the extracted collocations. We verified syntactic non-compositionality for NA-type collocations (noun and a postposed Adjective), for which we defined syntactic idiosyncrasies, attesting the stability of the connection (in such a form) in the corpus. Based on a statistical analysis, we argue that MWLUs syntactic non-compositionality must have the following features:

1. established word order
2. separability.

What we understand by the established word order is the ratio of neutral word order (Adjective in postposition) occurrence in the corpus to the alternative word order (Adjective in preposition). We took the established word order as the main criterion, and if its occurrence was lower than 87.09%, the algorithm suggested abandoning further procedure (Maziarz et al. 2015). In the case of reaching more than 87.09 % of occurrence, the algorithm tested separability defined as the ratio of occurrence in the word order with the Adjective in preposition and postposition divided by at least one other text word to the sum of occurrences in both word orders, but without no text word between elements of the collocation.

Finally, by using this method we extracted 607 collocations – potential MWLUs. From this list, we rejected several proper names and incomplete phrases. The rest of collocations was automatically accepted.

Table 2 shows chosen syntactically non-compositional MWLUs.

gra losowa (‘game of chance’)
energetyka odnawialna (‘renewable energy industry’)
klęska żywiołowa (‘natural disaster’)
kodeks celny (‘customs code’)
linie papilarne (‘fingerprint’)
medycyna weterynaryjna (‘veterinary medicine’)
obszar wiejski (‘rural area’)
oficer prasowy (‘Press officer’)
pole golfowe (‘golf course’)
pojemność skokowa (‘engine displacement’)

Table 2: Syntactially non-compositional MWLU’s

4 Verification of extracted collocations

At this stage, we gave linguists the list of extracted collocations for verification. At the preliminary stage of verification, linguists removed (i) combinations which were proper names (and were eliminated during the automatic verification), (ii) combinations with incomplete phrases or (iii) peculiar metaphorical uses (rare in accessible sources). Next, linguists assessed the remaining combinations in accordance to the following criteria:

1. a word cannot appear outside the given collocation (imprisoned meaning),
2. terminology,
3. paraphraseability,
4. free word order (in case of the type NA) (Maziarz et al. 2015a)

By a phrase “a word cannot appear outside the given collocation” we understood the word, for which a given collocation is specific, i.e. the word does not appear in any other collocation in Polish or it does not appear in predicative position. An example of such a collocation is *linia naboczna* (‘lateral line’).

As “terms”, we recognised these collocations, which are precisely and explicitly specified in one or more sources (Polański et al. 1999). In the case of mathematical-natural sciences, technical sciences, law, econometrics or linguistics one source, e.g. encyclopaedia (specialist), the specialist dictionary or the specialist lexicon, was enough for positive verification of the collocation. In the case of

other disciplines (especially social sciences or humanities) to do the positive verification two sources of the types listed above were needed. Universal encyclopedias and normative legal texts (acts, regulations) were treated as sufficient sources for term status confirmation of the selected units (Maziarz et al. 2015a). We also took into account other sources (e.g. scientific texts, institutional regulations) whose status is confirmed by some organization (e.g. scientific unit, association). In such cases, to do the positive verification it was essential for the candidate to occur in two sources.

"Paraphraseability" means the possibility of occurrence of a collocation in transformations, in which the collocation becomes separated, or one of its elements is replaced by another word or phrase, without the change in meaning. At this stage the following transformations were allowed:

1. a subordinate clause instead of an Adjective or a participle: *niebieska teczka = teczka, która jest niebieska* ('blue file = file, which is blue');
2. a noun or a prepositional phrase instead of an Adjective (with the force of semantic transposition): *tekst prawny = tekst prawa* ('legal test = text of law'), *drewniana podłoga = podłoga z drewna* ('wooden floor = floor made of wood');
3. a synonym or a dictionary definition in the place of any element of a collocation: gra zespołowa = zabawa towarzyska, która ma określone zasady, może wymagać rekwizytów[3] (team game = team sociable fun, which has particular rules, can need requisites).

In the case of the NA-type, an additional criterion, i.e. word order, was taken into account. On the basis of corpus data, linguists judged whether it was possible to change word order in a collocation without changes in its meaning. In addition, we decided that for the change in word order to be unacceptable, the ratio of NA word order to AN word order has to be greater than 100:1 (Maziarz et al. 2015a).

5 Applications

MWLUs are collected in the MWE dictionary, in which the following description of candidates is applied:

1. MWE's syntactic scheme,
2. MWE's part of speech,
3. MWE's base form,
4. MWE's syntactic head,
5. base form of each MWE's component,
6. part of speech for each MWE's component.

At present, the dictionary contains 45 thousand MWLUs, mainly of nouns and bigrams. MWLU's are grouped together according to syntactic schemes described according to the WCCL formalism (Radziszewski et al. 2011a). The dictionary is systematically enlarged.

Acknowledgements

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

References

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catharine MacLeod, & Antonio Zampolli. 2002. *Towards best practice for multiword expressions in computational lexicons*. W: Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC-2002). Las Palmas, Canary Islands - Spain.

Derwojedowa Magdalena, Szpakowicz Stanisław, Zawisławska Magdalena i Piasecki Maciej. 2008. *Lexical units as the centrepiece of a wordnet.* Proceedings of Intelligent Information Systems, Zakopane Poland. Institute of Computer Science PAS.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences Word Pairs and Collocations*, University of Stuttgart.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

John Firth. 1957. *The synopsis of linguistic theory 1930-1955. In studies of linguistic analysis*. The Philological Society, Oxford.

Adam Kilgarriff, Pavel Rychly, Pavrl Smrz, David Tugwell. 2004. *The Sketch Engine*. Proceedings of the 11th EURALEX International Congress. France.

[3] Source: plWordNet (`http://plwordnet.pwr.wroc.pl/wordnet/`)

Marek Maziarz, Stan Szpakowicz, Maciej Piasecki. 2015. *A Procedural Definition of Multi-word Lexical Units*. Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria.

Marek Maziarz, Stanisław Szpakowicz, Maciej Piasecki, and Agnieszka Dziob. 2015a. *Jednostki wielowyrazowe. Procedura sprawdzania leksykalności połaczeń wyrazowych* [‘*Multi-word units. A procedure for testing the lexicality of collocations*’]. Technical Report PRE-11, Faculty of Computer Science and Management, Wroclaw University of Technology.

Maciej Piasecki, Marek Maziarz, Stanisław Szpakowicz, and Ewa Rudnicka. 2014. *PlWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources.* Proc. 7th International Global Wordnet Conference, Tartu, Estonia, 25-29 January.

Krzysztof Polański (ed.). 1999. *Encyklopedia językoznawstwa ogólnego*. [‘*Encyclopedia of general linguistics*’], Ossoliński National Institute, Wroclaw.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus - preliminary version*. Institute of Computer Sciences, PAS, Warsaw.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk (ed.). 2012. *National Corpus of Polish.* Polish Scientific Publishers PWN, Warsaw.

Adam Radziszewski, Adam Wardyński and Tomasz Śniatowski. 2011. WCCL: A Morpho-syntactic Feature Toolkit. Text, Speech and Dialogue. Volume 6836 of Lecture Notes in Computer Science. Springer.

Adam Radziszewski, Michał Marcińczuk, Adam Wardyński. 2011a. *Specyfikacja języka WCCL* [‘*Specification of WCCL language*’]. Faculty of Computer Science and Management, Wroclaw University of Technology. Source: http://nlp.pwr.wroc.pl/redmine/projects/joskipi/wiki/Specyfikacja.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. 2012. *Multiword Expressions: A Pain in the Neck for NLP*. Proceedings of the 3rd Intenational Conference on Computational Linguistics and Intelligent Text Processing. Mexico City.

Establishing Morpho-semantic Relations in FarsNet (a focus on derived nouns)

Nasim Fakoornia
Shahid Beheshti University
Tehran, Iran
nasim_fakournia@yahoo.com

Negar Davari Ardakani
Shahid Beheshti University
Tehran, Iran
n_davari@sbu.ac.ir
na34@soas.ac.uk

Abstract

This paper aims at a morpho-semantic analysis of 2461 Persian derived nouns, documented in FarsNet addressing computational codification via formulating specific morpho-semantic relations between classes of derived nouns and their bases. Considering the ultimate aim of the study, FarsNet derived nouns included 12 most productive suffixes have been analysed and as a consequence 45 morpho-semantic patterns were distinguished leading to creation of 17 morpho-semantic relations. The approach includes a close examination of beginners, grammatical category and part of speech shifts of bases undergoing the derivation process. In this research the morpho-semantic relations are considered at the word level and not at the synset level which will represent a cross-lingual validity, even if the morphological aspect of the relation is not the same in the studied languages. The resulting morpho-semantic formulations notably increase linguistic and operative competence and performance of FarsNet while is considered an achievement in Persian descriptive morphology and its codification.

1 Introduction

A comprehensive and detailed description of the relevant linguistic levels is a prerequisite for achieving progress in natural language processing (NLP). Wordnets are very popular lexical ontologies, relying on morphological, semantic and morpho-semantic descriptions and formulations. FarsNet which is a Persian wordnet has been established in 2009 by NLP research lab of Shahid Beheshti University. It goes closely in lines and principles of Princeton WordNet, EuroWordNet and BalkaNet (shamsfard et al. 2010). The latest version of FarsNet (2.0) contains 22180 nouns (including 2756 derived nouns), 5691 verbs, 6560 adjectives and 2014 adverbs. Besides semantic relations (synonymy, hypernymy, hyponymy, meronymy and antonymy) and morphological relations (derivation), some additional conceptual relations such as domain and related to, have been devised in FarsNet. At present (2015), it consists of more than 36000 entries, organized in almost 2000 synsets. The present study which is aimed at formulating morpho-semantic relations of FarsNet's derived nouns provides the wordnet with the basic required information for automation of the relations.

According to Deléger et al. (2009), a morpho-semantic process decomposes derived, compound and complex words into their base and associates such process to their semantic interpretation. Through morpho-semantic analysis derived and compound words are analysed morphologically and relations between base and derivational form are interpreted semantically (Namer & Baud 2007). Raffaelli & Kerovec (2008) consider "morphosemantics" as the best expression describing studies which deal with links between form and meaning at the word level.

Derivation and compounding are the two main word formation processes. Persian derivational morphology consists of an affixal system in which the number of suffixes is more than prefixes. Persian derivational morphological processes include suffixation, prefixation, only a single case of circumfixation and no infixation (Davari and Arvin 2015). Affixation patterns in this language are generally regular however in some cases there are few exceptions (Megerdoomian 2000). According to Keshani (1992) Persian derivational processes are relying on almost 56 suffixes. The aim of the present study is to neatly explore, formulate and classify

the morpho-semantic patterns of derived nouns by analysing the relevant data in FarsNet. It is worth noting that the present article originates from a wider scope research by Fakoornia (2013), in which all FarsNet derived nouns (2756) were analysed in order to establish morpho-semantic relations between derived nouns and their bases. The derived nouns under study included 26 different suffixes. In this study the derivatives of 12 most productive noun marker suffixes (2461) have been focused. This study enriches FarsNet while improves morpho-semantic codification of Persian.

After a brief introduction to FarsNet word entries in general and noun entries in particular, the process of morpho-semantic pattern formulation will be elaborated for the selected suffixes.

2 FarsNet Word Entries

Entries include phonological transcription, part of speech, synonyms and their classifications in to a synset, word meaning and an example. A beginner will be selected for each lexeme. According to Miller et al. (1990) a beginner is a primitive semantic component of any word in its hierarchically structured semantic field. Beginners could be used in the recognition of domains synsets. Different syntactic types can be related to each other in FarsNet; mapping each entry to its corresponding concept in Princeton WordNet 3.0 is also possible (Shamsfard et al., 2010). Using this information is essential in establishing morpho-semantic relations. Table 1 shows the prevailing noun beginners in FarsNet.

Noun beginners				
1. act	6. cognition	11. location	16. plant	21. shape
2. animal	7. communication	12. motive	17. possession	22. state
3. artifact	8. event	13. object	18. process	23. substance
4. attribute	9. feeling	14. person	19. quantity	24. time
5. body	10. group	15. phenomenon	20. relation	25. food

Table 1: list of noun beginners in FarsNet

The synsets which do not fall into any of the above categories will be tagged by the label *nothing*. The semantic relations are also established among the synsets with the same POS. Synsets with different POS will be tagged by labels such as "related to". There are 3 choices for mapping a synset to the correspondent one in Princeton WordNet 3.0: equivalence mapping, near-equivalence mapping and no-mapping. Finally, the morphological relations among senses, such as derivational relations are marked.

Besides specifying a noun type (such as common, proper, countable, uncountable, pronoun, number or infinitive), a classification on the basis of some more general semantic features (such as belonging to human, animal, location or time) is provided.

3 Data Analysis

For the purpose of this study the noun corpus of FarsNet (22180) were thoroughly explored. First of all, the list of derived nouns (2756) was prepared. Then they were broken into their roots and affixes. From among 26 suffixes, in this paper, the 12 most frequents were selected (2461 derivatives), described and analysed. They are listed in table 2, the morphological descriptions are compatible with Keshani's (1992) description of Persian suffixes.

	Suffix	**POS**	**Semantic load**
1	"-i"	n-n a-n n-d d-d	Any type of noun, adjective & adverb
2	"-e"	n-n v-n a-a	Any type of noun & adjective
3	"-æk"	n-n a-n v-n	Any type of noun
4	"-ʧe"	n-n	Diminution similarity
5	"-gah"	n-n v-n d-d	Location Body part time
6	"-dan"	n-n	Location Body part dish

7	"-gær"	v-n n-n n-a	Profession object
8	"-ban"	n-n	Similarity
9	"-ænde"	v-a v-n d-a a-a	Any type of noun & adjective
10	"-ar"	v-n v-a n-n n-a	Any type of noun & adjective
11	"-eʃ"	v-n a-n	Any type of noun
12	"-ane"	n/ a- a/ d n/ d- n	Any type of adjective & adverb food

Table 2: A list of selected suffixes

The following information is required to link each noun to its base:

- Morphological information of nouns; including POS of the base and derivative as well as other noun types such as; proper, common, number, etc.
- Semantic category; including human, animal, location, time or nothing.
- Beginner; such as act, person, feeling, event, etc., (table 1).
- The derivational relation between derived noun and its base.

4 Morpho-semantic Analysis of Selected Suffixes

In this part we will scrutinize our 12 most productive selected noun marker suffixes from a morpho-semantic point of view. More information about the other Persian suffixes could be found in Fakoornia (2013).

4.1 "-i"

In FarsNet, 4125 nouns ends in letter /i/ among which 1880 nouns are considered to be derivatives of suffix "-i".

"-i" is an extremely productive Persian suffix. It has the potential for connecting to bases with different grammatical category, to compound words and even to syntactic phrases.

a. "-i" connects to nouns and adjectives and makes abstract noun, expressing an attribute or a state. The process is highly productive in Persian. Thus if "-i" connects to a noun or an adjective with different types of beginners, the resulting derivative beginner will be *attribute* or *state*. Considering the mentioned regularity the relation could be expressed as follows: "derivative attribute of base", for example "bideGat**i** attribute of bideGat", (carelessness attribute of careless). FarsNet includes 802 tokens of such nouns.

b. "-i" connects to agent nouns and present participles, describing a job or an act and makes noun infinitive referring to a field, a job or an act. In Persian the beginner of agent noun is *person* and the beginner of gerund is *act* or *cognition*. So if "-i" connects to a noun belonging to *person* or to present participle, the beginner of derivative will be *act* or *cognition.* Following this the relation "base agent of derivative" is predictable. For example; "mohændes agent of mohændes**i**" (engineer agent of engineering). FarsNet includes 890 tokens of such nouns.

c. "-i" connects to agent noun and makes nouns referring to location or territory. So if "-i" connects to a base which is *person*, and makes a derivative referring to *location*, we will have the relation "derivative location of base" for example "tælaforuʃ**i** location of tælaforuʃ" (jewelry location of jeweler). FarsNet includes 15 tokens of such nouns.

d. Other structures include the use of "-i" to refer to colors. Colors inherited from *property*. Thus if the base beginner is anything and the derivative beginner is *property*, the relation "base the same color as derivative" will be established. For example: "porteGal the same color as porteGal**i**" (orange the same color as orange). FarsNet includes 15 tokens of such nouns.

e. "-i" connects to some other nouns, verbs and adjectives (excluding the above mentioned ones) and makes derivatives, referring to *feeling*, *process*, *event*, *act*, *person*, *object*, *nothing* etc. So if the base POS is verb, noun, adjective (other than present participles) and the derivative beginner could be anything, we will have the relation "derivative related to base". For

example; "baran**i** related to baran" (raincoat related to rain). FarsNet includes 144 tokens of such nouns.

f. There are also 14 derivatives of "-i" in FarsNet which can be classified in both (a) and (b). In this case relations of "derivative attribute of base" and "base agent of derivative" can be established. For example: "bædæxlag**i** attribute of bædæxlag" (irritability attribute of irritable) and also "bædæxlag agent of bædæxlag**i**".

A summary of what has been explicated is listed in the table 3:

	input			**output**			
	Base POS	suffix	Base beginner	derivative POS	derivative beginner	morpho-semantic relation	number
a	n/adj	"-i"	anything	n	attribute/state	derivative attribute of base	802
b	n/pres. part.	"-i"	person	n	act/ cognition	base agent of derivative	890
c	n	"-i"	person	n	location	derivative location of base	15
d	n	"-i"	anything	n	Property	derivative the same color as base	15
e	v/n/adj	"-i"	anything except above	n	anything	derivative related to base	144
f	n/ pres. part.	"-i"	Person	n	attribute/state/act/cognition	"derivative attribute of base" and "base agent of derivative"	14
Total				1880			

Table 3: morpho-semantic patterns of suffix "-i" derivatives

According to the above patterns, "- i"'s word formation processes are formulated. The beginners are given in parenthesis and the frequency of each pattern is given in bracket.

a. Noun (person)/ present participle + "-i" = noun (act/ cognition) → base agent of derivative <890>.
b. Noun (anything)/ adjective + "-i" = noun (attribute/ state) → derivative attribute of base <802>.
c. Verb/ noun (other) / adjective + "-i" = noun (anything) → derivative related to base <144>.
d. Noun (person) + "-i" = noun (location) → derivative location of base <15>.
e. Noun (anything) + "-i" = noun (property) → derivative the same color as base <15>.
f. Noun (person) + "-i" = noun (attribute/ state/ act/ cognition) → derivative attribute of base/ base agent of derivative <14>.

As can be seen, "-i" is frequently involved in forming derivatives with beginners such as act, cognition and attribute. Few numbers of its derivatives are categorized under location and property. Formula (3) shows those patterns not covered in other structures.

4.2 "-e"

7 morpho-semantic patterns have been distinguished for suffix "-e":

a. Verb + "-e" = noun (anything except act) → derivative related to base verb form: "soru**e** related to sorudan" (song related to sing) <47>.

b. Noun (anything) + "-e"→ noun (other) → derivative related to base: "ruz**e** related to ruz" (fast[1] related to day) <42>.
c. Adjective + "-e" = noun (anything) → base attribute of derivative: "jævan attribute of jævan**e**" (young attribute of sprout) <23>.
d. Noun (object/ body) + "-e" = noun (anything) → derivative similar to base: "dæhan**e** similar to dæhan" (opening similar to mouth) <16>.
e. Noun (quantity) + "-e" = noun (time) → base quantity of derivative: "dæh quantity of dæh**e**" (ten quantity of decade) <4>.
f. Verb + "-e" = noun (act) → derivative act of base verb form: "xænd**e** act of xændidæn" (laughter (n.) act of laugh (v.)) <3>.
g. Diminutive noun (person) + "-e"= noun (person) → derivative pejorative sense of base: "doxtæræk**e** pejorative sense of doxtæræk" (bad girl pejorative sense of little girl <1>.

As can be seen, "-e" often links to verbs and creates derivatives with different types of beginners; it seldom results in pejorative nouns.

4.3 "-æk"

Suffix "-æk"* shows 8 morpho-semantic patterns in Persian:

a. Noun (anything) + "-æk" = noun (anything except food) → derivative similar to base: "suræt**æk** similar to suræt", (mask similar to face) <22>.
b. Noun (anything except person/ animal/ food) + "-æk" = noun (anything except person, animal and food) → derivative similar to base and derivative diminutive of base: "ʃæhr**æk** similar to ʃæhr" and "ʃæhr**ak** diminutive of ʃæhr", (town similar to city) and (town diminutive of city) <11>.
c. Adjective + "-æk" = noun (anything) → base attribute of derivative: "sorx attribute of sorx**æk**", (red attribute of measles) <6>.
d. Verb + "-æk" = noun (anything) → derivative related to base verb form: "gælt**æk** related to gæltidæn", (roller related to roll) <4>.
e. Noun (anything) + "-æk" = noun (food) → derivative similar to base "pæʃm**æk** similar to pæʃm", (cotton candy similar to wool) <3>*.
f. Noun (person/ animal) + "-æk" = noun (person/ animal) → derivative diminutive of base: "doxtær**æk** diminutive of doxtær", (little girl diminutive of girl) <2>.
g. Noun (body) + "-æk" = noun (act) → base agent of derivative: "naxon agent of naxon**æk**", (nail agent of pick) <1>.
h. Noun (body) + "-æk" = noun (body) → derivative related to base: "guʃ**æk**** related to guʃ", (eardrum related to ear) <1>.

* Formula (a) and (e), however similar cannot be merged into a single category as in (a) although the beginner of both derivative and base can be anything, the tokens of each category are exclusive. It should be mentioned that in pattern (e) the beginner of derivative can be the same as the base.

** As the POS and the beginner of the word "guʃ**æk**" (eardrum), do not change in the derivation process, during computational codification it is classified in second formula, however, according to its meaning it cannot entered in that group, thus it should be manually excluded and entered in a general relation (derivative related to base) formulated for it.

4.4 "-ʧe"

Suffix "-ʧe" shows 2 morpho-semantic patterns:

a. Noun (anything) + "-ʧe" = noun (anything) → derivative diminutive of base and derivative similar to base: "dæryaʧe diminutive of dærya" and "dærya**ʧe** similar to dærya", (lake diminutive of sea) and (lake similar to sea) <28>.

"-ʧe" in some nouns does not refer to similarity or diminution but it merely indicates a vague relatedness, an example is "ʔænbær**ʧe**", (sachet). In such situations the relation "derivative related to base" is formulated, but during computational codification derivatives belonging to this

[1] abstain from certain foods, as for religious or medical reasons (especially during the day)

structure, automatically classified in the previous structure which should be manually removed from it. In FarsNet there was only one derivative of this type. Thus the formula would be:

b. Noun (anything) + "-ʧe" = noun (anything) → derivative related to base: "ʔænbærʧe related to ʔænbær", (sachet related to ambergris) <1>.

4.5 "-gah"

Suffix "-gah" shows 3 morpho-semantic patterns:

a. Noun (anything)/ verb + "-gah" = noun (location) → derivative location of base: "dærman**gah** location of dærman", (health centre location of treatment) <83>.
b. Noun (anything) + "-gah" = noun (body) → derivative related to base: "gij**gah** related to gij", (temple related to dizzy) <6>.
c. Verb + "-gah" = noun (anything) → derivative related to base verb form: "did**gah** related to didæn", (viewpoint related to view) <1>.

The above shows that the number of derivatives, having location as their beginner is more than the other beginners. Moreover the suffix rarely connects to a verb.

4.6 "-dan"

Suffix "-dan" shows a single morpho-semantic pattern in Persian:

a. Noun (anything) + "-dan" = noun (anything) → derivative location of base: "gol**dan** location of gol", (vase location of flower) <11>.

4.7 "-gær"

Suffix "-gær" shows 4 morpho-semantic patterns:

a. Noun (act) + "-gær" = noun (person) → derivative agent of base: "arayeʃ**gær** agent of arayeʃ", (stylist agent of makeup) <28>.
b. Noun (anything except act) + "-gær" = noun (person) → derivative related to base: "ahæn**gær** related to ahæn", (blacksmith related to iron) <13>.
c. Noun (anything) + "-gær" = noun (object) → derivative instrument of the base: "næmayeʃ**gær** instrument of næmayeʃ", (monitor instrument of display) <3>.
d. Verb + "-gær" = noun (object) → derivative agent of base verb form: "rofte**gær** agent of roftæn", (dustman agent of sweep) <1>.

4.8 "-ban"

Suffix "-ban" shows 3 morpho-semantic patterns:

a. Noun (anything) + "-ban" = noun (person) → derivative protector of base: "jængæl**ban** protector of jængæl", (woodsman protector of wood) <17>.
b. Noun (anything) + "-ban" = noun (object) → derivative related to base: "saye**ban** related to saye", (sunshade related to shade) <3>.
c. Verb + "-ban" = noun (person) → derivative agent of base verb form: "dide**ban** agent of didæn", (sentinel agent of guard) <2>.

4.9 "-ænde"

Suffix "-ænde" shows a single morpho-semantic pattern:

a. Verb + "-ænde" = noun (anything) → derivative agent of base verb form: "ʔafarin**ænde** agent of ʔafæridæn", (creator agent of create) <76>.

4.10 "-ar"

Suffix "-ar" shows 4 morpho-semantic patterns:

a. Noun (anything) + "-ar" = noun (anything) → derivative related to base: "dad**ar** related to dad", (God related to justice) <5>.
b. Verb + "-ar" = noun (act) → derivative act of base verb form: "goft**ar** act of goftæn", (speech act of say) <2>.
c. Verb + "-ar" = noun (person) → derivative agent of base: "xærid**ar** agent of xæridæn", (buyer agent of buy) <2>.
d. Verb + "-ar" = noun (anything except act and person) → derivative related to base

verb form: "saxt**ar** related to saxtæn", (structure related to construct) <2>.

4.11 "-eʃ"

Suffix "-eʃ" shows 3 morpho-semantic patterns:

a. Verb + "-eʃ" = noun (act) → base act of derivative verb form: "Gorridæn act of Gorr**eʃ**", (roar (v.) act of roar (n.)) <68>.
b. Verb + "-eʃ" = noun (anything except act) → derivative related to base verb form: "deræxʃ**eʃ** act of deræxʃidæn", (shine act of shine) <15>.
c. Noun (anything) + "-eʃ" = noun (anything) → derivative related to base: "yon**eʃ** related to yon", (ionization related to ion) <8>.

4.12 "-ane"

Suffix "-ane" shows 3 morpho-semantic patterns:

a. Noun (anything)/ adverb + "-ane" = noun (food) → derivative food of the base: "sobh**ane** food of sobh", (breakfast food of morning) <7>.
b. Verb + "-ane" = noun (object) → derivative instrument of base verb form: "res**ane** instrument of resandæn", (media instrument of broadcast) <6>.
c. Noun (anything) + "-ane" = noun (anything except food) → derivative related to base: "ʔængoʃt**ane** related to ʔængoʃt", (thimble related to finger) <5>.

The 2 represented exceptions; "guʃ**æk**" (eardrum) and "ʔænbær**ʧe**" (sachet) will naturally and respectively fall in the formulated relations "derivative similar to base" or "derivative diminutive of base" and "derivative diminutive of base" or "derivative similar to base", however considering the meaning of their bases and the resulting derivatives, they do not belong to the mentioned relations, thus some other relations should be formulated to include them.

5 Conclusion

Morpho-semantic analysis of a selection of 2461 derived nouns in FarsNet showed 45 morpho-semantic patterns and 17 morpho-semantic relations (such as "derivative agent of base", "derivative location of base", etc.) for 12 most productive suffixes. Considering that only 2 words out of 2461 (0.08%) did not fall into the patterns, it could be concluded that the patterns have successfully provided the foundations for establishing automatic relations between derived or complex nouns and their bases in FarsNet. The coincident consideration of the words' morphological features such as their POS, their semantic and grammatical category (e.g. agent noun, participle noun, present participle, etc.) as well as recognizing the beginners of the bases (e.g. act, person, food, etc.) and their change after the affixation process have been the key criteria in formulating the relations which were especially crucial for the majority of studied suffixes that were polysemous. Defining and codifying these morpho-semantic patterns leads us to coherent establishment of morpho-semantic relations in FarsNet and hence has a remarkable developing impact on the applicability of the data base in machine translation, question answering systems, etc. Although In this research the morpho-semantic relations are considered at the word level and not at the synset level, mapping the results to the relations formulated in other languages wordnets will provide a cross-lingual validity, even if the morphological aspect of the relation is not the same in the mapped languages.

References

Davari Ardakani, Negar and Mahdiyeh Arvin. 2015. Persian. In N. Grandi and L. Kortvelyessy, editors. *Edinburgh Handbook of Evaluative Morphology*. Edinburgh University Press, Edinburgh, pages 287-295.

Deléger, Louise, Fiammetta Namer and Pierre Zweigenbaum. 2009. Morphosemantic Parsing of Medical Compound Words: Transferring a French Analyzer to English. *International Journal of Medical Informatics*, 78 (1): 48-55.

Fakoornia, Nasim. 2013. Morphosemantic Analysis of Nouns in Persian and English Aiming at Computational Codification. Master's thesis, Shahid Beheshti University, June.

Farshidvard, Khosrow. 2007. *Derivation and Compounding in Persian*. Zavar press, Tehran.

Keshani, khosrow. 1992. *Suffix Derivation in Contemporary Persian*. Iran University Press, Tehran.

Megerdoomian, Karine. 2000. *Persian Computational Morphology: A unification-based Approach*. NMSU,

CRL, Memoranda in Computer and Cognitive Science (MCCS-00-320).

Miller, George A. et al. 1990. Introduction to Wordnet: An Online Lexical Database. *Journal of Lexicography*, 3 (4): 235-244. doi:10.1093/ijl/3.4.235

Namer, Fiammetta and Robert Baud. 2007. Defining and Relating Biomedical Terms: Towards a Cross-language Morphosemantics-based System. *International Journal of Medical Informatics*, 76(2-3): 226-233.

Raffaelli, Ida and Barbara Kerovec. 2008. Morphosemantic fields in the Analysis of Croatian Vocabulary. *Jezikoslovlje*, 9 (1-2): 141-169.

Shamsfard, Mehrnoush et al. 2010. *Semi-Automatic Development of FarsNet; The Persian WordNet*. 5th Global WordNet Conference (GWA8020), Mumbai, India.

Using WordNet to Build Lexical Sets for Italian Verbs

Anna Feltracco
Fondazione Bruno Kessler
Università di Pavia, Italy
feltracco@fbk.eu

Lorenzo Gatti
Fondazione Bruno Kessler
Università di Trento, Italy
l.gatti@fbk.eu

Simone Magnolini
Fondazione Bruno Kessler
Università di Brescia, Italy
magnolini@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
Povo-Trento, Italy
magnini@fbk.eu

Elisabetta Jezek
Università di Pavia
Pavia, Italy
jezek@unipv.it

Abstract

We present a methodology for building lexical sets for argument slots of Italian verbs. We start from an inventory of semantically typed Italian verb frames and through a mapping to WordNet we automatically annotate the sets of fillers for the argument positions in a corpus of sentences. We evaluate both a baseline algorithm and a syntax driven algorithm and show that the latter performs significantly better in terms of precision.

1 Introduction

In this paper we present a methodology for building lexical sets for argument slots of Italian verbs. Lexical sets (Hanks, 1996) are paradigmatic sets of words which occupy the same argument positions for a verb, as found in a corpus. For example, for the verb *read*, the following set can be built by observing the lexical fillers of the object position in the BNC corpus:

(1) *read* {book, newspaper, bible, article, letter, poem, novel, text, page, passage, ...}

To collect lexical sets for Italian verbs, we use the lexical resource T-PAS (Jezek et al., 2014), an inventory of typed predicate argument structures for Italian manually acquired from corpora through inspection and annotation of actual uses of the analyzed verbs. In the current version of the T-PAS resource, only the verb is tagged in the annotated corpus, while the lexical items for each argument slots are not. Thus, the annotation of the lexical sets will enrich the actual version of the resource and will open to experiments for automatically extending its coverage.

A relevant step in our methodology is the annotation of the lexical items for argument positions in sentences. A previous work (Jezek and Frontini, 2010) has already outlined an annotation scheme for this purpose, and highlighted its benefits for NLP applications. In that work, however, the annotation of lexical sets was intended as manual, whereas the methodology we propose here is conceived for automatic annotation, and exploits an existing external resource. Under this perspective our work is related to semantic role labeling (Palmer et al., 2010).

This paper is organized as follows. Section 2 introduces the T-PAS resource; in Section 3 the lexical set population task is defined, and in Section 4 the experimental setting is presented. Section 5 discusses the results and is followed by the error analysis in Section 6. Finally, Section 7 provides some conclusions and directions for future work.

2 Overview of the T-PAS Resource

T-PAS, Typed Predicate Argument Structures, is a repository of verb patterns acquired from corpora by manual clustering of distributional information about Italian verbs (Jezek et al., 2014).

The resource has been developed following the lexicographic procedure called Corpus Pattern Analysis, CPA (Hanks, 2004). In particular, in the resource T-PASs are semantically motivated and are identified by analysing examples found in a corpus of sentences, i.e. a reduced version of ItWAC (Baroni and Kilgarriff, 2006).

After analyzing a sample of 250 concordances of the verb in the corpus, the lexicographer defines each T-PAS recognising its relevant structure and identifying the Semantic Types (STs) for each argument slots by generalizing over the lexical sets observed in the concordances; as an exam-

2 [[Human]] **divorare** [[Document]]
[[Human]] legge [[Document]] con grande interesse

Figure 1: T-PAS#2 for the verb *divorare*.

Figure 2: Lexical Set identification for T-PAS#2 for the verb *divorare*.

ple, Figure1 shows the T-PAS#2 of the verb *divorare*: [[Human]] divorare [[Document]] (Eng. to devour), where [[Document]] stands for {*libro, romanzo, saggio*} (Eng. {book, newspaper, essay}) (Figure 2). STs are chosen among a list of about 230 corpus-derived semantic classes compiled by applying the CPA procedure to the analysis of concordances for about 1500 English and Italian verbs (Jezek et al., 2014)[1]. If no generalization is possible, the lexical set is listed. Finally, the lexicographer associates the instances in the corpus to the corresponding T-PAS and adds a free-text description of its sense (Figure 1). The T-PAS resource thus lists the analyzed verbs[2], the identified T-PASs for each verb, the annotated instances for the T-PAS in the corpus.

In the next Sections, we will define the lexical set population task and describe the experiment we ran and its evaluation.

3 Task Definition

The aim of our system is to automatically derive lexical sets corresponding to the STs in the T-PAS resource. The task is defined as follows. The system receives as input *(i)* a T-PAS of a certain verb and *(ii)* a sentence associated to that T-PAS in the resource. The system should correctly mark (where present) the lexical items or the multiword expressions correspondent to the STs of each argument position specified by the T-PAS (i.e. sentence annotation step). By replicating this annotation for all the sentences of a T-PAS, the system will build the lexical set for a specific ST in a specific T-PAS (i.e. lexical set population step).

[1]Labels for STs in T-PAS are in English, as in the corresponding English resource PDEV (Hanks and Pustejovsky, 2005).

[2]The current version of T-PAS contains 1000 analyzed average polysemy verbs, selected on the basis of random extraction of 1000 lemmas out of the total set of fundamental lemmas of Sabatini Coletti (2007).

For instance, example (2) shows the T-PAS#1 of the verb *preparare* (Eng. to prepare) and a sentence associated to it.

(2) [[Human]] **preparare** [[Food | Drug]]
"La nonna, prima di infornare le patate, **prepara** una torta"
(Eng. "the grandmother, before baking the potatoes, **prepares** a cake")

In this case, the system should identify *nonna* (Eng. grandmother) as a lexical item for [[Human]]-SUBJ and *torta* (Eng. cake) for [[Food]]-OBJ. If this annotation is repeated for all the sentences of the T-PAS#1 of the verb *preparare*, the system will build the lexical set for the ST [[Human]] in Subject position in the T-PAS, such as {*nonna, chef, Gino, bambina, ..*}, and for [[Food]] in object position, such as {*torta, zuppa, pasta, panino, ..*}.

4 Experimental Setting

In order to identify possible candidate items for a ST, the system uses information from MultiWordNet (Pianta et al., 2002)(from now on MWN); e.g. to derive that "grandmother" is a human being and associate it to the ST [[Human]] and that "cake" is a type of food and associate it to the ST [[Food]]. The task, thus, required an initial mapping between the T-PAS resource and MWN. Then, we compared a naive Baseline algorithm and a more elaborated algorithm that we called LEA, Lexical Set Extraction Algorithm. Finally, to evaluate the performance of our methodology we also created a gold standard.

ST to Synset mapping. For our experiment, the list of STs used in the T-PAS resource was automatically mapped onto corresponding WordNet 1.6 synsets. For instance, the ST [[Human]] was mapped to all the synsets for the noun *human* (i.e. *human#n*). Manual inspection was limited to the case in which there is no exact match between a ST and a synset (e.g. by associating "atmospheric-phenomenon" to [[Weather Event]]).

The Baseline algorithm. The Baseline algorithm identifies possible candidate members of the lexical set corresponding to a certain ST for a certain T-PAS by *(i)* lemmatizing each sentence using TextPro (Pianta et al., 2008), *(ii)* checking if each lemma is in MWN and *(iii)* determining whether

the lemma belongs to a synset that was mapped to the ST, or if it is an hyponym of one such synsets.

For instance, in example (2), the Baseline lemmatizes the sentence and selects as possible candidates the nouns of the sentence, i.e. *nonna*, *torta* and *patate*. The Italian lemma *nonna* is thus searched in MWN and the correspondent English lemmas *grandma#n#1, grandmother#n#1, granny#n#1, grannie#n#1* are found. Since none of these synset lemmas match with [[Human]], [[Food]] or [[Drug]], the MWN hierarchy is traversed until *human#n#1* is found, which is mapped to [[Human]]. The same is done for *torta* and *patate*, until [[Food]] is found. Thus, for (2), the Baseline identifies *nonna* as [[Human]] and *torta* and *patate* as [[Food]] (with *patate* being a misclassified item, as it is not referred to the verb *preparare*).

The LEA algorithm. Compared to the Baseline algorithm, the LEA algorithm takes into account also the dependency tree of the sentence, named entities as recognized by TextPro, and multiword expressions.

It starts by *(i)* finding the position of the verb in an example and considering as valid candidate only the chunks that are a subject, direct object or complement of the verb according to the TextPro dependency tree. With respect to the Baseline, this leads to a more precise identification of the items for the argument slots of just the verb we are considering. For instance, in (2) we expect the algorithm to correctly identify *nonna* as [[Human]] and *torta* as [[Food]], but not proposing *patate* (as the Baseline does).

The LEA algorithm also *(ii)* checks if the verb allows the same ST for subject and object, as in the T-PAS#3 of *pettinare*: [[Human1]] pettinare [[Human2]] (Eng. to comb someone's hair). In the sentence "La mamma pettina il bambino" (Eng. The mum combs the baby), LEA will correctly propose *mamma* as [[Human1]] and *baby* as [[Human2]]. In this case, it also checks if the verb is in passive form and swaps the items for subject and object position as needed, improving the precision with respect to the Baseline.

Furthermore, the algorithm *(iii)* checks if the chunk contains/overlaps with proper names related to persons, organizations and locations detected by TextPro, and, if this is the case, checks the corresponding type of named entity against the ST allowed by the T-PAS frame (e.g. *Maria Rossi* → Person → [[Human]]). Since the Baseline recognizes only named entities that are in MWN, we expect this algorithm to identify more items.

Finally, LEA *(iv)* looks for multiword expressions in a chunk by checking if the combination exists in MWN. For instance, in "La nonna prepara la conserva di frutta" (Eng.: the grandmother prepares the fruit conserve), LEA should identify *conserva di frutta* as [[Food]] (while the Baseline identifies only the token *frutta*).

The LEA algorithm, thus, should recognize as valid only the items for a certain argument slot of the analyzed verb (and not for other verbs in the sentence), solve major cases of same ST in different slots and identify named entities and multiword expressions.

Gold Standard. We created a gold standard for the task by manually annotating 500 examples. We asked three annotators to mark the lexical items or the multiword expressions that correspond to the STs, without annotating pronouns or relative clauses. We selected the 500 sentences by extracting 10 sentences for 10 different STs in 5 different T-PASs (for a total of 50 different T-PASs belonging to 47 verbs). In particular, we chose, among all the STs within the [[Inanimate]] hierarchy, 10 types that are used in at least 5 different T-PAS, each of them having at least 10 (potential) sentences associated in the corpus resource. For example, we selected [[Food]] and annotated 10 sentences for T-PAS#1 of *mangiare* "[[Human]] mangiare [[Food]]" (Eng. to eat), since *(i)* there are at least 5 verbs with a T-PAS containing [[Food]], like *mangiare* itself and *(ii)* we have at least 10 sentences available for each of these five T-PASs [3]. This selection of few STs was intended to better compare performances of the algorithms for different lexical sets.

The gold standard annotation resulted in a total of 981 annotated tokens out of 15090 (the average sentence length being 30.18 tokens).

5 Results

For what concerns sentence annotation, we evaluate overall precision, recall and F-measure, con-

[3] This is mainly a selection criteria. Considering that we analyzed a limited number of examples for each verb, and that more than one ST can be specified for each argument slot, it is also possible that none of the sentences extracted for a ST for a verb instantiate that particular ST.

sidering as a positive match when the algorithms agree with the gold standard in recognizing a token as an item (or part of the item in case of multiword expressions) instantiating a ST for a precise position.

Compared to the Baseline, the LEA algorithm registers a significant higher value for precision (see Automatic Mapping in Table 1). This is not surprising, as the Baseline considers as valid all the items in the sentence that can correspond to the ST, without taking into account if they are in the argument position required by the T-PAS or not. On the contrary, the LEA algorithm also considers the syntactic structure, thus lowering the false positives rate; the downside effect is that its recall is lower than the one of the Baseline.

Automatic mapping			
	Precision	Recall	F1
Baseline	0.28	0.42	0.34
LEA	0.70	0.25	0.37
Mapping with manual revision			
Baseline	0.30	0.52	0.38
LEA	0.72	0.32	0.44

Table 1: Results for sentence annotation for the Baseline Algorithm and the LEA Algorithm.

We also measured the similarity between the 5 most populated lexical sets in the gold standard (from 6 to 15 tokens in 10 sentences) and their correspondent lexical sets built by the two algorithms (see Table 2), by calculating the Dice's coefficient[4] (van Rijsbergen, 1979). For example, we compare the lexical set of the T-PAS#1 of *crollare*: [[Building]] crollare (Eng. to fall down) {*e.g. casa, muro, torre*} with the lexical set for the same ST in the same T-PAS derived by the Baseline and LEA.

Results show that both the Baseline and LEA do not reach high overlap. In fact, even if LEA has an high precision in identifying the members of the lexical set, the low recall penalizes the amount of items it can detect given few sentences to annotate. On the contrary, the Baseline is favored by a higher recall, but its low precision causes major differences with the gold standard sets. For these reasons, we believe that on a broader scale, the higher precision for LEA is more advisable with respect to the Baseline.

[4]Dice's coefficient measures how similar two sets are by dividing the number of shared elements of the two sets by the total number of elements they are composed by. This produces a value from 1, when both sets share all elements, to 0, when they have no element in common.

	Baseline	LEA
Cuocere#2-SBJ-[[Food]]	0.54	0.57
Crollare#1-SBJ-[[Building]]	0.40	0.25
Dirottare#1-OBJ-[[Vehicle]]	0.72	0.50
Prescrivere#2-OBJ-[[Drug]]	0.42	0.46
Togliere#4-OBJ-[[Garment]]	0.45	0.22

Table 2: Dice's value for lexical set annotation for the Baseline Algorithm and the LEA Algorithm.

6 Error Analysis

The results presented in the first part of Table 1 were manually inspected to identify sources of errors. In particular, we have noticed that many inaccuracies are due to the automatic mapping of STs to WordNet synsets. For instance, both algorithms failed to recognize *casa* (Eng.: house), corresponding to the ST [[Building]] which was automatically mapped onto *building#n*; they would have succeeded, had the ST been mapped to the more general *construction#n*.

Even when the automatic mapping works, the different structure of the two resources can lead to wrong results. For instance, vehicles such as *elicottero* (Eng.: helicopter) are frequently generalized by the ST [[Vehicle]] in T-PAS and are hyponyms of *vehicle#n* in MWN. However, while in T-PAS [[Machine]] is a hypernym of [[Vehicle]], the same is not true for *machine#n* in MWN. As a consequence, in the sentences in which vehicles are considered members of the lexical set correspondent to [[Machine]], even traversing the MWN hierarchy, the algorithms can not consider these items as valid candidates for the ST [[Machine]].

To solve at least some of these problems, we manually inspected the 40 STs of the sentences of the gold standard, and modified the automatic mapping of 11 of those; for example, we chose to translate the ST [[Building]] to *construction#n*, and mapped [[Machine]] to both *transport#n* and *machine#n*. This led to a significant improvement of the recall for both algorithms, and a minor improvement of the precision, as shown in Table 1.

This improvement is also reflected on the second part of the task (i.e. the creation of the lexical

set). For example, the Dice value for *Crollare#1-SBJ-[[Building]]* improves from 0.4 to 0.71 for the Baseline and from 0.25 to 0.6 for LEA.

Another significant aspect concerns the recognition of proper names: out of the 185 tokens that are -or are part of- proper nouns (137 are related to persons, locations or organizations), the Baseline recognized correctly only 10 (mainly common nouns that are used as proper names), while the LEA algorithm only 26.

Finally, some errors are introduced in the PoS tagging and dependency parsing steps. During the former, an incorrect tag can be assigned to a word (e.g. a noun could be mis-tagged as an adjective) and hinder both algorithms, as the word would not be checked in MWN. The latter only undermines the recall of the LEA algorithm instead. Moreover, LEA does not deal with complex syntactic structure yet (e.g. when our verb is in an infinitive phrase, which is the object of a main verb, such as "[..] e il presidente chiede agli italiani di *ipotecare* la casa [..]", Eng.: [..] and the president asks Italians to *mortgage* their houses [..]).

7 Conclusion and Further Work

In this paper we have presented an experiment for the automatic building of lexical sets for argument positions of the Italian verbs in the T-PAS resource. The method is based on the use of MWN in order to match the STs with the potential fillers of each argument position.

The experiment suggests that LEA can be used to automatically populate the lexical sets with good precision. We believe that significantly better results could be obtained with an accurate manual mapping of the STs to synsets, possibly narrowed to specific senses (e.g. mapping [[Building]] to just the third sense of *construction#n*). Furthermore, recognizing proper nouns proved a difficult task, and even using named entities recognition in addition to MWN was not enough. Therefore a resource to map these nouns to a synset in the WordNet hierarchy is needed; BabelNet (Navigli and Ponzetto, 2012) could prove useful in this sense.

Further work includes the extension of the sentence annotation and lexical set population for all T-PAS and the comparison of the same ST in different T-PASs in order to study Italian verbs' selectional preferences from the perspective of verb selectional classes (for example, all verbs that select [[Food]] as object).

References

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue française de linguistique appliquée*, 10(2):63–82.

Patrick Hanks. 1996. Contextual dependencies and lexical sets. *The International Journal of Corpus Linguistics*, 1(1).

Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the 11th EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud*, volume 1, pages 87–98.

Elisabetta Jezek and Francesca Frontini. 2010. From Pattern Dictionary to Patternbank. In G.M. De Schrijver, editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis*, pages 215–239. Kampala:Menha Publishers.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS: a resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the 1st international conference on global WordNet*, volume 152, pages 55–63.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Francesco Sabatini and Vittorio Coletti. 2007. *Dizionario della lingua italiana 2008*. Milano: Rizzoli Larousse.

CJ van Rijsbergen. 1979. *Information Retrieval. 1979*. Butterworth.

A Taxonomic Classification of WordNet Polysemy Types

Abed Alhakim Freihat
Qatar Computing Research Institute
Doha, Qatar
afraihat@qf.org.qa

Fausto Giunchiglia
University of Trento
Trento, Italy
fausto@disi.unitn.it

Biswanath Dutta
Indian Statistical Institute (ISI)
Bangalore, India
bisu@drtc.isibang.ac.in

Abstract

WordNet represents polysemous terms by capturing the different meanings of these terms at the lexical level, but without giving emphasis on the polysemy types such terms belong to. The state of the art polysemy approaches identify several polysemy types in WordNet but they do not explain how to classify and organize them. In this paper, we present a novel approach for classifying the polysemy types which exploits taxonomic principles which in turn, allow us to discover a set of polysemy structural patterns.

1 Introduction

Polysemy in WordNet (Miller, 1995) corresponds to various kinds of linguistic phenomena and can be grouped into various polysemy types (Falkum, 2011). Although WordNet was inspired by psycholinguistic and semantic principles (Miller et al., 1990), its conceptual dictionary puts greater emphasis on the lexical level rather than on the semantic one (Dolan, 1994). Lexicalizing polysemous terms without any further information about their polysemy type affects the usability of WordNet as a knowledge resource for semantic applications (Mandala et al., 1999).

In general, the state of the art approaches suggests different solutions to the polysemy problem. The most prosperous among these approaches are the regular/systematic polysemy approaches such as (Buitelaar, 1998) (Barque and Chaumartin, 2009) (Veale, 2004) (Peters, 2004). These approaches propose the semantic regularity as a basis for classification of the polysemy classes and offer different solutions that commensurate the nature of the discovered polysemy types.

Despite the diversity and depth of the state of the art solutions, no or very little attention has been given, so far, to the principles or rules used to identify polysemy types. In fact, none of these approaches can explain how to identify the polysemy types of the discovered polysemy structural patterns or how to differentiate for example, between homonymy and metaphoric structural patterns. Although Apersejan's semantic similarity criterion (Apresjan, 1974) can be used to account for regularity in polysemy, it can not predict the polysemy type of the regular polysemy types in WordNet. Our hypotheses in this paper is that identifying and differentiating between the polysemy types of the regular polysemy structural patterns requires understanding the hierarchical structure of WordNet and, thus, the criteria related to the taxonomic principles that the hierarchical structure of WordNet comply with or violates. In this paper, we show how to use two taxonomic principles as criteria for identifying the polysemy types in WordNet. Based on these principles, we introduce a semi automatic method for discovering and identifying three polysemy types in WordNet.

The paper is organized as follows. In Section two, we discuss the problem. In Section three, we introduce the formal definitions we use. In Section four, we discuss the taxonomic principles that we use to discover three of the polysemy types in WordNet. In Section five, we give an overview of our approach. In Section six, we show how to use the taxonomic principles to identify metaphoric structural patterns. In Section seven, we demonstrate how to determine specialization polysemy structural patterns. In Section eight, we describe how to discover homonymy structural patterns. In Section nine, we explain how to handle false positives in the structural patterns. In Section ten, we present the results of our approach. In Section eleven, we conclude the paper and depict our future work.

2 Problem Statement

WordNet is a machine readable online lexical database for the English language. Based on psycholinguistic principles, WordNet has been developing since 1985, by linguists and psycholinguists as a conceptual dictionary rather than an alphabetic one (Miller et al., 1990). Since that time, several versions of WordNet have been developed. In this paper, we are concerned with WordNet 2.1. WordNet 2.1. contains 147,257 words, 117,597 synsets and 207,019 word-sense pairs. The number of polysemous words in WordNet is 27,006, where 15776 are nouns.

In this paper, we deal with polysemous nouns at the concept level only. We do not consider polysemy at the instance level. After removing the polysemous nouns that refer to proper names, the remaining polysemous nouns are 14530 nouns.

WordNet does not differentiate between the types of the polysemous terms and it does not contain any information in terms of polysemy relations that can be conducted to determine the polysemy type between the synsets of a polysemous term. The researchers who attached the polysemy problem in WordNet gave different descriptions for the polysemy types in WordNet. For example, polysemy reduction approaches (Edmonds and Agirre, 2008) (Mihalcea R., 2001) (Gonzalo J., 2000) differentiate between contrastive polysemy and complementary polysemy. Regular polysemy approaches such as (Barque and Chaumartin, 2009) (Veale, 2004) (Peters, 2004) (Freihat et al., 2013) (Lohk et al., 2014) give more refined classification of the polysemy types into metonymy, metaphoric, specialization polysemy, and homonymy. In one of our recent papers, *compound noun polysemy* is introduced as a new polysemy type beside the former four polysemy types in WordNet (Freihat et al., 2015).

So far, no polysemy reduction approaches have introduced a mechanism for classifying the polysemy types into contrastive and complementary. Instead, these approaches adopt semantic and probabilistic rules to discover redundant and/or very fine grained senses. On the other hand, the regular polysemy approaches embrace a clear definition for classifying polysemous terms into regular and non regular polysemy (Apresjan, 1974). Although, the definition of regular polysemy in these approaches is useful to distinguish between regular and non regular polysemy, these approaches do not reveal the principles or the criteria used to classify polysemous terms into polysemy types.

In this paper, we explain how to use the exclusiveness property and the collectively exhaustiveness property (Bailey, 1994) (Marradi, 1990) for identifying the following polysemy types.

1 **Metaphoric polysemy:** Refers to the polysemy instances in which a term has literal and figurative meanings (Evans and Zinken, 2006). In the following example, the first meaning of the term `fox` is the literal meaning and the second meaning is the figurative.
#1 `fox: alert carnivorous mammal.`
#2 `dodger, fox, slyboots: a shifty deceptive person.`

2 **Specialization polysemy:** A type of related polysemy which denotes a hierarchical relation between the meanings of a polysemous term. In the case of abstract meanings, we say that a meaning A is a more general meaning of a meaning B. We may also use the taxonomic notations type and subtype instead of more general meaning and more specific meaning respectively. For example, we say that the first meaning of `turtledove` is a subtype of the second meaning.
#1 `australian turtledove, turtledove: small Australian dove.`
#2 `turtledove: any of several Old World wild doves.`

3 **Homonymy:** Refers to the contrastive polysemy instances, where meanings are not related. Consider for example the following polysemy instance of the term `bank`.
#1 `depository financial institution, bank: a financial institution.`
#2 `bank: sloping land (especially the slope beside a body of water).`

3 Approach Notations

We begin with the basic notations. Lemma is the basic lexical unit in WordNet that refers to the base form of a word or a collocation. Based on this definition, we define a natural language term or simply a term as a lemma that belongs to a grammatical category; i.e., noun, verb, adjective or adverb.

Definition 1 *(Term).*

A term T is a quadruple ⟨Lemma, Cat⟩, where
a) Lemma is the term lemma;
b) Cat is the grammatical category of the term.

Synset is the fundamental structure in WordNet that we define as follow.

Definition 2 *(WordNet synset).*

A synset S is defined as ⟨Cat, Terms, Gloss, Relations⟩, where
a) Cat is the grammatical category of the synset;
b) Terms is an ordered list of synonymous terms that have the same grammatical category Cat;
c) Gloss is a text that describes the synset;
d) Relations is a set of semantic relations that hold between synsets.

Now, we move to the hierarchical structure of WordNet. WordNet uses the relation direct hypernym to organize the hierarchical relations between the synsets. This relation denotes the superordinate relationship between synsets. For example, the relation direct hypernym holds between `vehicle` and `wheeled vehicle` where `vehicle` is hypernym of `wheeled vehicle`. The direct hypernym relation is transitive. In the following, we generalize the direct hypernym relation to reflect the transitivity property, where we use the notion hypernym instead of a direct hypernym.

Definition 3 *(hypernym relation).*

For two synsets s and $s^{'}$, s is a hypernym of $s^{'}$, if the following holds: s is a direct hypernym of $s^{'}$, or there exists a synsets $s^{''}$ such that s is a direct hypernym of $s^{''}$ and $s^{''}$ is a hypernym of $s^{'}$.
For example, `vehicle` is a hypernym of `car`, because `vehicle` is direct hypernym of `wheeled vehicle` and `wheeled vehicle` is a direct hypernym of `car`.
We use the following symbols to denote direct hypernym/hypernym relations:
a) $s \prec s^{'}$ if s is a direct hypernym of $s^{'}$
c) $s \prec^{*} s^{'}$ if s is a hypernym of $s^{'}$

Using the direct hypernym relation, wordNet organizes noun-synsets in a hierarchy that we define as follows.

Definition 4 *(wordNet hierarchy).*

Let $S = \{s_1, s_2, ..., s_n\}$ be the set of noun-synsets in WordNet. WordNet hierarchy is defined as a connected and rooted digraph $\langle S, E \rangle$, where
a) `entity` $\in S$ is the single root of the hierarchy;
b) $E \subseteq S \times S$;
c) $(s_1, s_2) \in E$ if $s_1 \prec s_2$;
d) For any synset s $\neq$ `entity`, there exists at least one synset $s^{'}$ such that $s^{'} \prec s$.

In this definition, point (a) defines the single root of the hierarchy and point (d) defines the connectivity property in the hierarchy.
We move now to the semantics of WordNet. We define the subset of the semantics of WordNet hierarchy that is relevant for our approach. A full definition of the WordNet semantics is described in approaches such as (Alvarez, 2000) (Rudolph, 2011) (Breaux et al., 2009).
We define the semantics of WordNet using an Interpretation $I = \langle \Delta^I, f \rangle$, where Δ^I is an non empty set (the domain of interpretation) and f is an interpretation function.

Definition 5 *(Semantics of WordNet Hierarchy).*

Let $WH = \langle S, E \rangle$ be wordNet hierarchy. We define an Interpretation of WH, $I = \langle \Delta^I, f \rangle$ as follows:
a) $entity^I = \Delta^I$;
b) $\perp^I = \varnothing$;
c) $\forall s \in S$: $s^I \subset \Delta^I$;
d) $(s_1 \sqcap s_2)^I = s_1^I \cap s_2^I$;
e) $(s_1 \sqcup s_2)^I = s_1^I \cup s_2^I$;
f) $s_1 \sqsubseteq s_2$ if $s_1^I \subseteq s_2^I$.

In points a) and b), we define the empty and universal concepts. Point c) states that Δ^I is closed under the interpretation function f. In and d) and e), we define the conjunction and disjunction operations. In f), we define the subsumption relation.

We present now the polysemy notations. A term is polysemous if it is found in the terms of more than one synset. A synset is polysemous if it contains at least one polysemous term. In the following, we define polysemous terms.

Definition 6 *(polysemous term).*

A term t = ⟨Lemma, Cat, T-Rank⟩ is polysemous if there is a term $t^{'}$ and two synsets s and $s^{'}$, $s \neq s^{'}$ such that
a) t ∈ s.Terms and $t^{'} \in s^{'}$.Terms;
b) t.Lemma = $t^{'}$.Lemma;
c) t.Cat = $t^{'}$.Cat.

In the following, we define polysemous synsets.

Definition 7 *(polysemous synset).*

A synset s is polysemous if any of its terms is a polysemous term.
It is possible for two polysemous synsets to share more than one term. Two polysemous synsets and their shared terms constitute a polysemy instance. In the following, we define polysemy instances.

Definition 8 *(polysemy instance).*

A polysemy instance is a triple $[\{T\}, s_1, s_2]$, where s_1, s_2 are two polysemous synsets that have the terms {T} in common.

For example, the term `bazaar` belongs to the following polysemy instances: $[\{bazaar, bazar\}, \#1, \#2]$, $[\{bazaar\}, \#1, \#3]$, and $[\{bazaar\}, \#2, \#3]$.
#1 `bazaar, bazar: a shop where a variety of goods are sold.`
#2 `bazaar, bazar: a street of small shops.`
#3 `bazaar, fair: a sale of miscellany; often for charity.`
We move now to the last part of our definitions. We exploit the structural properties in WordNet hierarchy to identify the polysemy types of the polysemy instances in WordNet. According to the connectivity property of WordNet hierarchy in definition 4, any two synsets in wordNet have at least one common subsumer that we define as follows.

Definition 9 *(common subsumer).*

Let s_1, s_2, and s be synsets in wordNet. The synset s is a common subsumer of s_1 and s_2 if $s \prec^* s_1$ and $s \prec^* s_2$.

The WordNet hierarchy is a DAG (directed acyclic graph). This implies that it is possible for two synsets to have more than one common subsumer. We define the least common subsumer as the subsumer with the least height.
In the following, we define structural patterns.

Definition 10 *(structural pattern).*

A structural pattern of polysemy instance I = $[\{T\}, s_1, s_2]$ is a triple $P = \langle r, p_1, p_2 \rangle$, where
a) r is the least common subsumer of s_1 and s_2;
b) $r \prec p_1$ and $r \prec p_2$;
c) $p_1 \prec^* s_1$ and $p_2 \prec^* s_2$.

We call r *the pattern root* and p_1, p_2 *the pattern hyponyms*. For example, the structural pattern of the polysemy instance $[\{bazaar, bazar\}, s_1, s_2]$ is $\langle mercantile\ \ establishment, marketplace, shop \rangle$ as shown in Figure 1, where `mercantile establishment` is the pattern root and `marketplace` and `shop` are the pattern hyponyms. A special structural pattern is the

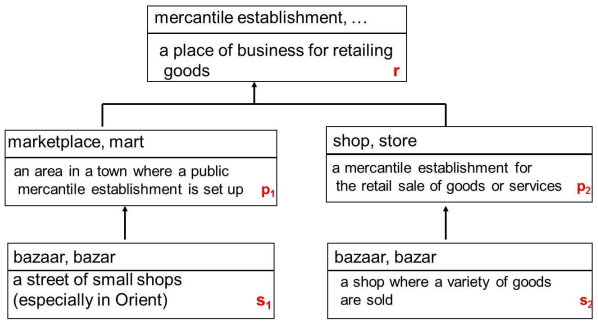


Figure 1: Example of a structural pattern

common parent structural pattern as illustrated in Figure 2. A strcutural pattern $P = \langle r, p_1, p_2 \rangle$ of a polysemy instance I = $[\{T\}, s_1, s_2]$ is a common parent structural pattern if $p_1 = s_1$ or $p_2 = s_2$.

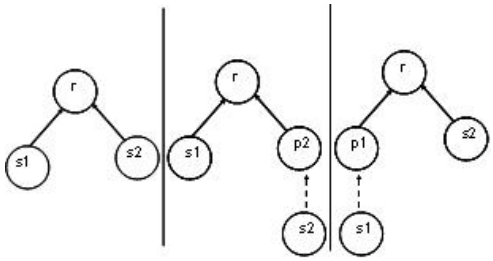


Figure 2: Common parent structural pattern

4 Taxonomic principles in WordNet

WordNet hierarchy represents a classification hierarchy where synsets are the nodes. Classification hierarchies should fulfill among other requirements the exclusiveness property and the exhaustiveness property.
We begin with the exclusiveness property.

Definition 11 *(Exclusiveness property).*

Two synsets $s_1, s_2 \in S$ fulfill the exclusiveness property if $s_1^I \sqcap s_2^I = \perp^I$. For example, `abstract entity` and `physical entity` fulfill the exclusiveness property. On the other hand `expert` and `scientist` do not fulfill this property because $expert^I \sqcap scientist^I \neq \perp^I$.
The exclusiveness property means that any two

sibling nodes n_i, nj in the hierarchy are disjoint, i.e., $n_i^I \not\sqsubseteq n_j^I$ and $n_j^I \not\sqsubseteq n_i^I$. Analyzing the structural patterns in WordNet shows that the exclusiveness property is not always guaranteed in WordNet. For example, the pattern $\langle person, expert, scientist \rangle$ shown in Figure 3 does not fulfill this property because forcing this property would result in preventing a scientist to be an expert or an expert to be a scientist. We

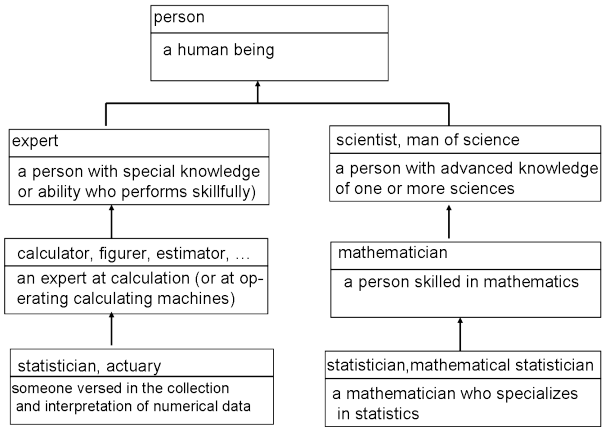


Figure 3: An example of exclusiveness property violation

are concerned with the cases, where the synsets s_1 and s_2 are not disjoint and each of them subsumes a synset of the same polysemous term such as the term *statistician* in Figure 3. The fact that the two synsets of the polysemous terms are not disjoint implies that the polysemy type of these two synsets can not be homonymy, metonymy, or metaphoric. This can be explained as follow. The polysemy type homonymy implies that the two synsets are unrelated and that the disjointness between the two synsets indicates a relation between the two synsets. Metonymy on the other hand means that one synset is a part of the other synset. Now, we explain the exhaustiveness property.

Definition 12 *(Collective Exhaustiveness).*

Two synsets $s_1, s_2 \in S$ are collectively exhaustive if it is possible to find a synset s such that $s^I = s_1^I \sqcup s_2^I$ and s_1, s_2 fulfill the exclusiveness property.

For example, `abstract entity` and `physical entity` fulfill the collectively exhaustiveness property because $entity^I = abstract\ \ entity^I \sqcup physical\ \ entity^I$. On the other hand `worker` and `female` in the pattern $\langle peron, worker, female \rangle$ do not fulfill this property because `worker` corresponds to a role and `female` to a concept. This is because `person` is a direct hypernym of the concept `organism` and the role `causal agent`.

5 Approach Overview

We exclude the structural patterns whose pattern root resides in the first and second level in WordNet hierarchy. Accordingly, any structural pattern whose root belongs to the synsets {`entity`, `abstract entity`, `abstraction`, `physical entity`, `physical object` } was automatically excluded. Our hypothesis is that the pattern hyponyms in these structural patterns in general fulfill the exclusiveness and the exhaustiveness property. These patterns are subject to our current research in discovering metonymy structural patterns. On the other hand, exclusiveness and exhaustiveness property are not guaranteed for all structural patterns whose roots reside in the third level and beyond. The input of the algorithm is the taxonomic structure of WordNet, starting from level 3, after removing lexical redundancy in compound nouns (Freihat et al., 2015). The output consists of three lists that contain specialization polysemy, metaphoric polysemy and homonymy instances. The first step of our algorithm is automatic, while the other two are manual.

S1. **Structural pattern discovery**: The input of this step is the current structure of WordNet after removing lexical redundancy. The algorithm returns structural patterns associated with their corresponding polysemy instances.

S2. **Structural pattern classification**: In this step, we manually classify the structural patterns returned in the previous step. The output consists of four lists of patterns associated with their polysemy instances. These four lists are:
Specialization polysemy patterns: This list contains the patterns whose corresponding instances are specialization polysemy candidates.
Metaphoric patterns: This list contains the patterns whose corresponding instances are metaphoric candidates.
Homographs patterns: This list contains the patterns whose corresponding instances are homonymy candidates.
Singleton patterns: The patterns in this group are those patterns that have one polysemy in-

stance only and thus cannot be considered to be regular.

S3 **Identifying false positives**: In this step, we manually process the polysemy instances in the four lists from the previous step. Our task is to decide the polysemy type for the instances in the singleton patterns list and remove false positives form the other three lists.

6 Metaphoric Structural Patterns

Identifying metaphoric patterns is based on the distinction between the literal meaning and the figurative meaning. Our idea is that it is not possible for a literal and the figurative meaning to be collectively exhaustive. Violating the exhaustiveness property in a structural pattern $\langle r, p_1, p_2 \rangle$ may be a result of the following:
a) p_1 and p_2 belong to different types and can not be subsumed by the pattern root r, or
b) $p_1 \sqsubset p_2$ or $p_2 \sqsubset p_1$.
For example $female$ and $worker$ can not be subsumed by $person$ in the pattern $\langle person, female, worker \rangle$ as shown in Figure 4. On the other hand, it is correct that

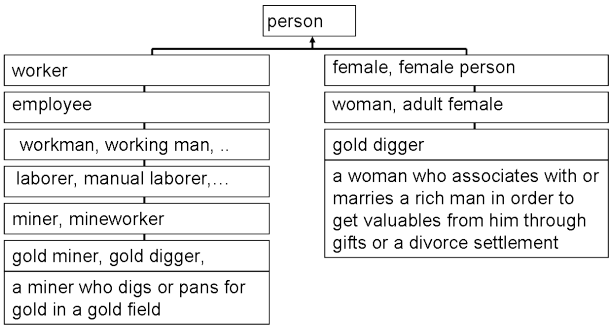


Figure 4: Example of a metaphoric polysemy instance

person and animal are organisms in the structural $\langle organism, animal, person \rangle$ but it is clear that $person^I \sqsubset animal^I$
In the following, we define metaphoric patterns structural pattern as follows.

Definition 13 *(Metaphoric structural pattern).*

A pattern $p = \langle r, p_1, p_2 \rangle$ is metaphoric if p_1 and p_2 do not fulfill the collectively exhaustiveness property.
In the following we give examples for identified metaphoric patterns. The pattern $\langle organism, animal, person \rangle$ is metaphoric. Although both synsets share the same hypernym `organism`, they are not collectively exhaustive as explained. The polysemy instances that belong to this pattern are 326 instances. Consider for example the following instance.
#1 `snake, serpent, ophidian: limbless scaly elongate reptile.`
#2 `snake, snake in the grass: a deceitful or treacherous person.`
Another example is the pattern $\langle attribute, property, trait \rangle$. Although, both synset share the same hypernym `attribute`, they are not collectively exhaustive because $trait^I$ is a special case of $property^I$ ($trait^I = property^I \sqcap person^I$). The polysemy instances that belong to this pattern are 111 instances. Consider for example the following instance.
#1 `softness:the property of giving little resistance to pressure and being easily cut or molded.`
#2 `gentleness, softness, mildness: acting in a manner that is gentle and mild and even-tempered.`

7 Specialization Polysemy Structural Patterns

We use the exclusiveness property and the pattern root in a structural pattern to discover specialization polysemy candidates indirectly. The relation between the synsets in specialization polysemy is hierarchical. The hierarchical relation between the synsets in a specialization polysemy instance indicates that the exclusiveness property does not hold between synsets and thus between the structural pattern hyponyms.
We define specialization polysemy patterns as follows.

Definition 14 *(specialization polysemy structural pattern).*

A pattern $p = \langle r, p_1, p_2 \rangle$ is a specialization polysemy pattern if a) and b) hold
a) p_1 and p_2 do not fulfill the exclusiveness property.
b) p_1 and p_2 fulfill the exhaustiveness property.
In the following we give examples for identified specialization polysemy patterns. All instances that belong to the common parent structural patterns are classified as specialization polysemy instances. The polysemy instances that belong to this pattern are 2879 instances. Consider for example the following instance.

#1 capital, working capital: assets available for use in the production of further assets.
#2 capital: wealth in the form of money or property owned by a person or business and human resources of economic value.
Another example is the pattern $\langle act, action, activity \rangle$. The polysemy instances that belong to this pattern are 406 instances. Consider for example the following.
#1 employment, work: the occupation for which you are paid.
#2 employment, engagement: the act of giving someone a job.
Another example, is the pattern $\langle animal, invertebrate, larva \rangle$. The polysemy instances that belong to this pattern are 17 instances. Consider for example the following.
#1 ailanthus silkworm, Samia cynthia: large green silkworm of the cynthia moth.
#2 cynthia moth, Samia cynthia, Samia walkeri: large Asiatic moth introduced into the United States; larvae feed on the ailanthus.

8 Homonymy Structural Patterns

We define homonymy patterns as follows.

Definition 15 *(Homonymy structural pattern).*

A pattern $p = \langle r, p_1, p_2 \rangle$ is homonymy pattern if the following condition hold.
a) p_1 and p_2 fulfill the exclusiveness property;
b) p_1 and p_2 fulfill the exhaustiveness property;
c) There is no relation between p_1 and p_2.
In the following we give examples for identified homonymy patterns. The pattern $\langle organism, person, plant \rangle$. The polysemy instances that belong to this pattern are 40 instances. Consider for example the following instance.
#1 spinster, old maid: an elderly unmarried woman.
#2 zinnia, old maid, old maid flower: any of various plants of the genus Zinnia.
Another example is the pattern $\langle organism, animal, plant \rangle$. The polysemy instances that belong to this pattern are 41 instances. Consider for example the following.
#1 red fox, Celosia argentea: weedy annual with spikes of silver-white flowers.
#2 red fox, Vulpes fulva: New World fox; often considered the same species as the Old World fox.
Another example is the pattern $\langle vertebrate, bird, mammal \rangle$. The polysemy instances that belong to this pattern are 13 instances. Consider for example the following.
#3 griffon, wire-haired pointing griffon: breed of medium-sized long-headed dogs.
#4 griffon vulture, griffon, Gyps fulvus: large vulture of southern Europe and northern Africa.

9 False Positives Identification

In this section, we describe the third step of our approach. Our task here is to process the four lists returned at the end of the pattern classification and remove false positives. These lists are the metaphoric polysemy list, the specialization polysemy list, the homonymy list, and a list of non regular (singleton patterns) list. This task can only be performed manually due to the implicit and missing information in synset glosses. Our procedure for determining the polysemy class of a polysemy instance is based on the three definitions in the previous section, where we process the polysemy instances instance by instance to determine the the relation between the synsets of the polysemy instances.
If a polysemy instance does not belong to the polysemy type it was assigned to (false positive instance), we assign it to its corresponding polysemy type.
In the following, we give examples for false positives. The common parent structural pattern which was automatically assigned to the specialization polysemy type (step 1 in Section 5) contains 180 false positive polysemy instances, 98 of them were identified as homonymy instances. One example is:
#1 cardholder: a person who holds a credit card or debit card.
#2 cardholder: a player who holds a card or cards in a card game.
Metaphoric false positives (82 instances) were also identified in the common parent class. Consider for example the following instance.
#1 game plan: (figurative) a carefully thought out strategy for achieving an objective in war.

#2 `game plan: (sports) a plan for achieving an objective in some sport.`
Another example is the pattern $\langle organism, animal, person \rangle$ which was assigned to the metaphoric polysemy type contains 326 polysemy instances, 74 of them were identified as homonyms such as the following instance.
#2 `Minnesotan, Gopher: a native or resident of Minnesota.`
#3 `ground squirrel, gopher, spermophile: any of various terrestrial burrowing rodents of Old and New Worlds.`

10 Results and Evaluation

The number of polysemy instances computed by the polysemy instances discovery algorithm is 41306. We excluded 28318 instances because the pattern roots of these instances reside in the first and the second level of the hierarchy as per the approach discussed in Section 5.The remaining number of polysemy instances is 12988. These instances are divided in two groups as follow. 12988 of these instances belong to 1028 regular type compatible patterns and 1569 instances belong to single tone patterns. The classification of the pasterns and the result of the false positive removing is shown in the following tables.

#Type	#patterns	#instances
Specialization	823	9902
Metaphoric	134	1697
Homonymy	71	1389
Total	1028	12988

Table 1: Classification of the regular structural patterns

In Table 2, we show the results removing false postive instances, where we see that the average false positives is about 17%.

#Poly Type	#Instances	#False Positives
Specialization	9902	1740
Metaphoric	1697	175
Homonymy	1389	295
Total	12988	2210

Table 2: False Positives in Pattern Classification

To evaluate our approach, 3797 polysemy instances were evaluated by two evaluators. The agreement of the evaluators with our approach was on 96.5% of the instances. In the following Table 3, a refers to our approach, e_1, e_2 refer to evaluator1 and evaluator 2 respectively.

$e_1 = e_2 = a$	3665 (96.5%)
$a = e_1$	3621 (95.3%)
$a = e_2$	3600 (94.8%)

Table 3: Evaluation of the polysemy classification

11 Conclusion and future Work

In this paper, we have presented how to use two taxonomic principles for classifying the polysemy types in WordNet. We have demonstrated the usefulness of our approach on classifying three polysemy types, namely, specialization, metaphoric and homonymy. In this approach, we were able to discover all specialization polysemy structural patterns and subsets of the metaphoric and metonymy structural patterns. We aim to continue our work to study the metonymy patterns in the upper level of WordNet hierarchy, where we generalize our structural pattern definition as follows.

Definition 16 *(generalized structural pattern).*

A structural pattern of polysemy instance I = $[\{T\}, s_1, s_2]$ is a triple $P = \langle r, p_1, p_2 \rangle$, where
a) r is the least common subsumer of s_1 and s_2;
b) $r \prec^* p_1$ and $r \prec^* p_2$;
c) $p_1 \prec^* s_1$ and $p_2 \prec^* s_2$.
Our hypothesis is that in case of metonymy structural patterns: the nodes p_1 and p_2 fulfill the exclusiveness and the exhaustiveness properties and there is a part of relation between p_1 and p_2. The conditions for metaphoric and homonymy structural patterns obtained by adapting the new structural definition remain the same as explained in this paper.

Acknowledgment

The research leading to these results has received partially funding from the European Community's Seventh Framework Program under grant agreement n. 600854, Smart Society (http://www.smart-society-project.eu/).

References

Jordi Alvarez. 2000. Integrating the wordnet ontology into a description logic system.

JU. Apresjan. 1974. Regular polysemy. *Linguistics*, pages 5–32.

Kenneth D. Bailey. 1994. *Typologies and Taxonomies: An Introduction to Classification Technique*. Sage Publications, Thousand Oaks, CA, 1994///.

Lucie Barque and Franois-Rgis Chaumartin. 2009. Regular polysemy in wordnet. *JLCL*, 24(2):5–18.

Travis D. Breaux, Annie I. Anton, and Jon Doyle. 2009. Semantic parameterization: A process for modeling domain descriptions. *ACM Transactions on Software Engineering Methodology*, 18(2).

Paul Buitelaar. 1998. Corelex: Systematic polysemy and underspecification. *PhD thesis,Brandeis University, Department of Computer Science*.

W. B. Dolan. 1994. Word sense ambiguation: clustering related senses. In *Proceedings of COLING94*, pages 712–716.

Philip Edmonds and Eneko Agirre. 2008. Word sense disambiguation. *Scholarpedia*, 3.

Vyvyan Evans and Jrg Zinken. 2006. Figurative language in a modern theory of meaning construction: A lexical concepts and cognitive models approach.

Ingrid Lossius Falkum. 2011. The semantics and pragmatics of polysemy: A relevance-theoretic account. *PhD thesis, University College London*.

Abed Alhkaim Freihat, Fausto Giunchiglia, and Bisu Dutta. 2013. Regular polysemy in wordnet and pattern based approach. *International Journal On Advances in Intelligent Systems*, 6(3&4), jan.

Abed Alhkaim Freihat, Bisu Dutta, and Fausto Giunchiglia. 2015. Compound noun polysemy and sense enumeration in wordnet. *In Proceedings of the 7th International Conference on Information, Process, and Knowledge Management (eKNOW)*, pages 166–171.

Verdejo F. Gonzalo J., Chugur I. 2000. Sense clusters for information retrieval: Evidence from semcor and the eurowordnet interlingual index. *ACL-2000 Workshop on Word Senses and Multi-linguality, Association for Computational Linguistics*, pages 10–18.

Ahti Lohk, Kaarel Allik, Heili Orav, and Leo Vhandu. 2014. Dense components in the structure of wordnet. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Complementing wordnet with roget's and corpus-based thesauri for information retrieval. In *EACL*, pages 94–101. The Association for Computer Linguistics.

Alberto Marradi. 1990. Classification, typology, taxonomy. *Quality & Quantity: International Journal of Methodology*, 24(2):129–157.

MoldovanD. I. Mihalcea R. 2001. Ez.wordnet: Principles for automatic generation of a coarse grained wordnet. *FLAIRS Conference*, pages 454–458.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Wim Peters. 2004. Detection and characterization of figurative language use in wordnet. *PhD thesis, Natural Language Processing Group, Department of Computer Science, University of Sheffield*.

Sebastian Rudolph. 2011. Foundations of description logics. In Arenas C. Polleres A., dÁmato C., editor, *Reasoning Web. Semantic Technologies for the Web of Data - 7th International Summer School 2011*, volume 6848 of *LNCS*, pages 76–136. Springer.

Tony Veale. 2004. Pathways to creativity in lexical ontologies. In *In Proceedings of the 2nd Global WordNet Conference*.

Some Strategies for the Improvement of a Spanish WordNet

Matías Herrera **Javier González** **Luis Chiruzzo** **Dina Wonsever**
Facultad de Ingeniería
Universidad de la República

Abstract

Although there are currently several versions of Princeton WordNet for different languages, the lack of development of some of these versions does not make it possible to use them in different Natural Language Processing applications. So is the case of the Spanish Wordnet contained in the Multilingual Central Repository (MCR), which we tried unsuccessfully to incorporate into an anaphora resolution application and also in search terms expansion. In this situation, different strategies to improve MCR Spanish WordNet coverage were put forward and tested, obtaining encouraging results. A specific process was conducted to increase the number of adverbs, and a few simple processes were applied which made it possible to increase, at a very low cost, the number of terms in the Spanish WordNet. Finally, a more complex method based on distributional semantics was proposed, using the relations between English Wordnet synsets, also returning positive results.

1 Introduction

The Multilingual Central Repository (Agirre, Laparra, Rigau, & Donostia, 2012) follows the model proposed by the EuroWordNet project. EuroWordNet (Vossen, 1998) is a multilingual lexical database with wordnets for several European languages, structured in the same way as Princeton's WordNet. The MCR comprises five different languages: English, Spanish, Catalan, Basque and Galician. The Inter-Lingual-Index (ILI) allows us to link the words in one language with their equivalent translation in any of the other languages, thanks to the automatically generated mappings among WordNet versions. For example: the ILI identifier "ili-30-02084071-n" corresponds both to the English synset "eng-30-02084071-n" with lemmas "dog, domestic dog", and to the Spanish synset "spa-30-02084071-n" with lemmas "can, perro". In addition, it corresponds to the Basque synset "eus-30-02084071-n" with lemmas "zakur, or, txakur", to the synset "cat-30-02084071-n" for Catalan with lemmas "ca, canis familiaris", and also "glg-30-02084071-n" for Galician with lemmas "can, Canis familiaris". The current ILI version corresponds to WordNet 3.0. All identifiers stem from the original synset in English. In the previous example there is a translation for each one of the languages, however, this is not the most common scenario. The MCR is incomplete, at least for the Spanish version. This document presents several strategies to extend the coverage of the Spanish version. An in-depth analysis of the different problems of the Spanish MCR is presented in section 2, and section 3 describes several processes to enhance it. Section 4 presents the evaluations carried out for the strategies proposed and section 5 presents final observations on the general results and the possibility to launch an enhanced version on line.

2 Problems on the MCR Spanish WordNet

2.1 Deficiencies of the current Spanish MCR: first evaluation

For the purpose of finding the deficiencies of the MCR WordNet, our initial approach was to use it and test it out. Version 3.0 was used, since this is the latest version currently available. The web interface provided by the MCR (Benítez et al., 1998) was used to fulfill this stage. The MCR was requested to provide the results both in English and Spanish for all the searches made, in order to be able to compare them. Below we provide some examples of this initial informal evaluation and the

following section presents a quantitative evaluation:

- Lack of common words

  Some common words such as "cargador" and the adverb "no" were found to be missing.

- Empty synsets

  Some Spanish synsets were available through the web interface but they were empty. For example, the synset "spa-30-00396699-r" did not contain any variants, but its English equivalent "eng-30-00396699-r" did. This shows that there were no Spanish translations in the MCR for the lemmas "meagerly", "sparingly", "slenderly" and "meagrely". When searching for the adverb "escasamente", which is a possible translation for "sparingly", it was not found.

- Very few entries for the grammatical category adverbs

  Once evaluated, it was concluded that adverb coverage of the MCR was very low. We have already mentioned the example for the adverb "no". It was also found that the adverbs "recién" (just) and "rápidamente" (quickly) were not present, although these are very commonly used in Spanish.

- Lack of glosses or phrases that show the usage of the terms in Spanish.

  No Spanish gloss was found for many of the words searched. For example, we found that the result for the noun "cuchillo", "spa-30-03623556-n" and "spa-30-03624134- n" did not include a Spanish gloss for these synsets. Additionally, a generalized lack of phrases that illustrate the use of the lemmas and synsets was found.

2.2 Deficiencies in the current MCR: evaluation on a corpus

Several MCR WordNet coverage measures were applied taking Corin corpus (Grassi, Malcuori, Couto, Prada, & Wonsever, 2001) as a baseline. Corin corpus is a synchronous corpus that comprises the years 1996-2000 and contains literary-type texts by Uruguayan authors (essays and fiction) and journalistic texts published in Montevideo (articles and interviews). Several other language processing tools were used in addition to Corin, such as Freeling (Carreras, Chao, Padró, & Padró, 2004) and the dictionaries Apertium (Armentano-Oller et al., n.d.) and Wiktionary(*Wikimedia Foundation. 2008b. Wiktionary.*, 2008).

The following aspects were studied:

1. The percentage of available lemmas in the Spanish version of WordNet.
2. The percentage of corpus lemmas for which there was a translation available.
3. The percentage of these lemmas that was not present in the Spanish MCR but did have an available translation in the English MCR.

The results obtained are presented as follows:

2.2.1 Percentage of Corin lemmas available in the Spanish version of WordNet

POS	Lemmas not found		Lemmas found		Processed lemmas
N	69,29%	2780	30,71%	1232	4012
A	51,00%	840	49,00%	807	1647
V	75,35%	1235	24,65%	404	1639
R	32,79%	121	67,21%	248	369
Total	48,70%	3734	51,30%	3933	7667

The previous chart shows the total number of lemmas processed, their Parts Of Speech and how many of them were number found on WordNet . We can see that verbs are the grammatical category with the lowest coverage at 25%. The remaining POS show a higher coverage, with adverbs showing the highest one.

2.2.2 Percentage of corpus lemmas for which there was a translation available

POS	Untranslated		Translated		Lemmas
N	12,04%	483	87,96%	3529	4012
A	21,07%	347	78,93%	1300	1647
V	18,00%	295	82,00%	1344	1639
R	16,26%	60	83,74%	309	369
Total	15,46%	1185	84,54%	6482	7667

Using the two mentioned dictionaries we were able to cover a large percentage of the lemmas present in the corpus. Even so, the results do not ensure the quality of the translations. Therefore, it is necessary to improve the resources used for this purpose.

2.2.3 Lemmas not found in the Spanish MCR but with a translation available in the English MCR

Out of the 6482 lemmas translated into English, we focused on those found in the English MCR, so it was possible to compare the lemmas which were not found in the Spanish MCR but did have a translation available in the English MCR.

POS	Lemmas not in Spanish MCR		Lemmas in Spanish MCR		Total
N	43,40%	1349	56,60%	1759	3108
A	46,37%	492	53,63%	569	1061
V	15,02%	176	84,98%	996	1172
R	69,00%	187	31,00%	84	271
Total	39,27%	2204	60,73%	3408	5612

We can conclude that verbs are the grammatical category with the widest coverage, and adverbs are the most incomplete. In addition, nouns and adjectives present a coverage of just over 50%.

3 Strategies to improve WordNet

To improve the existing Spanish WordNet we conducted tests with processes that we have called "selectors", following the terminology already used in the field (WoNeF). A selector is a mechanism that, when applied to an English synset, will choose the translation or translations for the Spanish synset based on the original in English. Previously defined selectors were tested, supported by Apertium and Wiktionary translators, and in addition, two new selectors were defined, one based on morphology and the other based on the exploitation of semantic relations between synsets, with frequentist criteria used in distributional semantics. Selectors are applied in two differentiated stages, which are separately evaluated.

3.1 Translation methods

The translation process used was key for the application of this method to create the Spanish WordNet based on the English WordNet. We used two different methods: automatic translation and dictionaries. With regard to dictionaries, Wiktionary was used as well as a dictionary created based on the XML stem files of the Apertium dictionary. The automatic translation used was the one provided by Bing Translator (*Bing Online Translator.*, 2015). These tools were chosen mainly due to their availability, since they are either free and/or open. Wiktionary and Apertium were downloaded from their respective websites, and Bing Translator was used online through its API.

Microsoft's Bing Translator does not take into account the grammatical category of the word to be translated, therefore, there were cases where if verbs were translated, it would return nouns, or even the same verb but in a different conjugated form, instead of the infinitive form used in the search. In order to solve this problem, it was decided to use the results returned by the translator, and conduct a morphological analysis applying Freeling. The procedure entails obtaining all the possible grammatical categories of the word and its lemma, to afterwards select the words with the same grammatical category as the originally translated English word.

We decided to use a dictionary created based on the XML stem files of the Apertium dictionary rather than the already processed Apertium dictionary, since, for some reason, when making a request it would only return one possible translation, even if the XML file contained more. It was possible to obtain all the available translations for each word using the XML stem files.

3.2 Phase 1: Initial selectors

Below we present the experiments conducted with simple selectors already reported in the literature: monosemy and single translation. It is surprising that these selectors are still productive over the currently available version of WordNet, as our experiments show.

Monosemy Monosemy takes those words found in a single synset. This condition seems to show that there is no ambiguity and, therefore, all translations obtained are added to the corresponding synsets in the Spanish WordNet. For example, when applying this selector to the synset "eng-30-00048268-r" whose lemma is "currently" the three possible translations obtained by the translators "hoy", "ahora" and "actualmente" are selected since "currently" is only found in one synset in the English WordNet.

Single translation This selector takes all the words that have a single translation into Spanish and places it in all corresponding Spanish WordNet synsets. For example, when applying this selector to the

synset “eng-30-00061528-r”, whose lemma is “abruptly” and the translation returned is “abruptamente”, this will be selected since it is the single translation.

Factorization The factorization selector works at synset level. It takes all synsets from the English WordNet and returns all possible translations for each lemma. Once the set of translations for each lemma is put together, the selector selects those translations found as a common translation for all the lemmas in the synset, that is, with the intersection of the translation sets for each lemma. For example, consider the synset “eng-30-01309991-a”, whose lemmas are “artless” and “ingenuous”. The translations for “artless” are: “inocente”, “ingenuo” and “cándido” and those for “ingenuous” are: “inocente” and “ingenuo”. In this case, by applying the selector we obtained “inocente” and “ingenuo”, as a common translation.

Derived Adverb This selector obtains adverbs from the English WordNet and then the adjectives from which these derive. The property “is_derived_from” provided by the MCR was used to obtain the adjectives from which these adverbs derive. Once the adjective synsets are returned, we will obtain all the variants. These are in turn translated so as to later apply the morphological derivation rules to build adverbs in Spanish. By applying this selector to the synset “eng-30-00033562-r” whose lemma is “mildly” and is linked to the POS adjective synset “eng-30-01508719-a” whose lemma is “mild”, we will obtain “suavemente” and “levemente”. The latter are generated based on both available translations for “mild”: “suave” and “leve”, and by applying the following morphological derivation rules.

If the adjective ends in an “o”, it will be replaced by the sequence “amente”, for example, “lento” resulting in “lentamente”. If the adjective ends in an “r” or “n”, then , add the sequence “amente”, for example, “encantador” and “fanfarron” and their respective results “encantadoramente” and “fanfarronamente”. The sequence “mente” will be added to the rest of the adjectives that do not fall in the categories above mentioned, for example, “educada” and “educadamente”. Since this selector builds words by applying morphological derivation rules, we observed that sometimes it would return adverbs that do not exist in Spanish. Therefore, we decided to validate them against a corpus comprised of Spanish news text. To do so, we extracted all adverbs from said corpus to put together a list of adverbs to validate the existence of the adverbs built by the selector. The weakness of such validation method lies in the fact that it may discard adverbs which are correct as they are not found in the reference corpus. However, we considered more pertinent to ensure that accurate words were added. Moreover, it is always possible to use a longer list of known adverbs to reduce the number of false negatives.

Levenshtein This selector uses Levenshtein’s edit distance, based on the assumption that, if the distance between a word in English and its translation is short, they can be considered to have the same sense. Minor modifications are made to reduce the distance between one word and its translation. One example of these transformations is the inversion of the letters “r” and “e” to be applied to the word “tiger” and corresponding translation “tigre”. After doing the transformation, Levenshtein’s distance becomes 0. When implementing the initial selectors we decided not to use it since it did not return good results during the initial experiments. A possible explanation for this is that Spanish and English do not share as many cognate terms as English and French do, as discussed in the WoNeF article.

Singular translation selectors, monosemy and single factorization Levenshtein were inspired in (Atserias, Climent, Farreres, Rigau, & Guez, 1997), while Levenshtein was used in (Pradet, de Chalendar, & Desormeaux, 2014). Derived adverbs was our own production.

3.3 Phase 2: distributional semantics

For the expansion stage we proposed a selector that would exploit the relations between synsets and frequencies of occurrence of both words within a corpus, to determine which translation is the correct one for each ambiguous synset. It

is worth noting that this selector would be used when both related lemmas in English are known, and one of them gets only translation but for the other one there are several possible translations.

A detailed explanation of the implementation of this phase is presented below:

Let's suppose that we have a synset SA associated to synset SB in WordNet through a hypernymy relation. In addition, we have two English lemmas LA and LB for SA and SB respectively. The translations for LA are TA_1 and TA_2, and the translations for LB are TB. So to decide which translation is correct for this lemma, we searched for the occurrence of each translation in a corpus. These searches are considered as a function and represented with letter Θ. This process is called disambiguation.

For example, for calculating $\Theta(TA_1, TB)$ we count all occurrences of the words TA_1 and TB that happen within the same sentence.

$$O_1 = \frac{(\Theta(TA_1, TB)}{\Theta(TA_1) + \Theta(TB)}$$

$$O_2 = \frac{(\Theta(TA_2, TB)}{\Theta(TA_2) + \Theta(TB)}$$

In case $O_1 \geq O_2 \implies TA_1$ is chosen as the translation of LA.

However, if $O_1 < O_2 \implies TA_2$ is chosen as the translation of LA.

An example of the application of this expansion phase follows:

We know that SA = "eng-30-09776346-n" and SB = "eng-30-09816771-n" are related through the hypernym relation and they have the lemmas LA = "affiliate" and LB = "associate" respectively. Furthermore, we know that TB = "asociado" and the translation candidates for "affiliate" are TA_1 = "filial" and TA_2 = "afiliado". Because $O(filial, asociado) = 0.0$ and $O(afiliado, asociado) = 8.18129755379e^{-05}$, then we know that $O(afiliado, asociado) \geq O(filial, asociado) \implies$ the word TA_2 = "afiliado" is chosen as the translation of LA.

The previous result is correct because the English gloss for SA = "eng-30-09776346-n" is: "a subordinate or subsidiary associate; a person who is affiliated with another or with an organization".

The semantic relations used for this process were hypernymy, meronimy and antonymy, and the frequency counts were performed over the Spanish news text corpus.

4 Evaluation of results

We show evaluations for the initial selectors, for the phase 2 process and a global evaluation of results within a lexical semantics effort.

4.1 Quantitative evaluation of phase 1 results

In the evaluation we randomly selected 1000 synsets for each POS (verb, adverb, noun and adjective). The translations of every lemma in all the sorted synsets were obtained and the four selectors mentioned above were applied. The results obtained were stored in a database.

POS	Translated		Untranslated	
R	82,80%	1187	17,20%	246
V	71,90%	1226	28,10%	478
A	59,50%	969	40,50%	659
N	71,20%	1036	28,80%	419
All	71,00%	4418	29,00%	1802

Table 1: Translated lemmas

As can be seen, 71 % of the lemmas processed returned a translation. When we analyze the data at grammatical category level, we see that adverbs is the category with the highest translation percentage, with over 80 %. The other categories behave in a similar way to each other, adjectives being the category with the least coverage with almost 60 % of translations returned.

The following table shows the distribution of the translation of the lemmas for each of the 4000 synsets selected. Our aim was to obtain the results returned for each selector over the total of lemmas translated, but avoiding the overlapping of results by providing an order of importance. There follows the order applied: single selector, monosemy selector, factorization selector and others. For "V", "A" and "N" POS, the others include the translations that were not selected by any selector. For "R" POS, as well as translations not selected by any selector, the translations determined by the derived adverbs selector are also included.

POS	Singulars	Monosemic and not singular	Not monosemic, not singular and factored
R	56,40%	6,10%	1,30%
V	58,00%	2,00%	0,80%
A	77,60%	5,20%	0,90%
N	72,70%	4,60%	1,40%
All	70,20%	4,70%	1,20%

Table 2: Translation by selector

As seen here, verbs and adverbs had the worst result, while adjectives had the best result: 16.3%. We must remember that these data do not consider the results of the derived adverbs selector. These were excluded from the comparison because they could not be compared with the rest of the POS.

4.2 Synsets for which the initial selectors obtained results

POS	Yes		No	
R	73,90%	739	26,10%	261
V	52,80%	528	47,20%	472
A	59,90%	599	40,10%	401
N	63,70%	637	36,30%	363
All	62,60%	2845	37,40%	1155

Table 3: Synsets for which the initial selectors obtained results

As seen here, the POS with the highest coverage by initial selectors were adverbs, with almost 74%; without distinguishing according to POS, there is a 62.60% coverage.

4.3 Comparison with current WordNet

POS	New		Existent	
R	83,80%	694	16,20%	134
V	50,40%	390	49,60%	384
A	62,50%	429	37,50%	257
N	54,80%	423	45,20%	349
All	63,30%	1936	36,70%	1124

Table 4: Comparison with current WordNet

As seen here, for each POS there was a high percentage of synsets that had translations which were not found in the current Spanish WordNet (MCR 3.0). Adverbs is the grammatical category with the highest percentage: approximately 83%. In total there were just over 63% new synsets. As only the initial selectors were applied, we concluded that we would see a significant improvement at the end of the process.

A manual qualitative evaluation was conducted to measure the accuracy of the results. We randomly selected 25 synsets for each POS (verb, adverb, noun and adjective) of the added ones, and we verified if the result was correct or not. For the selectors that work at synset level, the data in table 5 reflect the percentages of the resulting correct or incorrect synsets, and for the selectors that work at lemma level, the percentages correspond to the resulting correct or incorrect synsets.

POS	Monosemy	Single translation	Factorization	Derived adverb
V	93.48%	98.39%	100.00%	-
A	96.08%	100.00%	96.00%	-
N	93.48%	100.00%	100.00%	-
R	97.14%	94.59%	92.00%	92.00%
All	95.04%	98.25%	97.00%	-

Table 5: Accuracy for the initial selectors

Although the derived adverbs selector was the least accurate one, it returned a very good result: 92%.

As seen in the charts above, the results of the four selectors were very good: all show over 92 % of effectiveness and some reach 100 % for some POS.

5 Evaluation of phase 2 results

5.1 Lemmas processed

The 1040 synsets that were not translated in phase 1 because they were ambiguous were applied and evaluated in phase 2. As phase 2 can fail for various reasons, in this section we present detailed information about the results obtained to identify such reasons. As phase 2 exploits the relations between the existing synsets in WordNet up to the present, if the synsets are not related to any other synsets, or if they are, but such synsets are empty for Spanish, this method returns no results. Therefore three different groups can be observed on the following table.

POS	With relations	With relations and no trans.	With relations and trans.
R	83,10%	10,56%	6,34%
V	1,20%	10,40%	88,40%
A	1,79%	34,52%	63,69%
N	0,00%	22,17%	77,83%
Total	12,21%	16,92%	70,87%

As seen here, adverbs is the grammatical category that has the least connected synsets, which shows that our method does not return good results for this POS. The other grammatical categories have enough relations and they are sufficiently complete for phase 2 to return results.

5.2 Lemmas processed in phase 2 with relations and with translations for these relations

It is important to highlight that for lemmas corresponding to synsets associated to other already complete synsets, the method applied in phase 2 can fail if there were no occurrences in the corpus of the possible candidates for all lemmas. This is explained in the following results.

POS	With result		Without result	
R	33,33%	3	66,67%	6
V	63,80%	282	36,20%	160
A	60,75%	65	39,25%	42
N	70,95%	127	29,05%	52
Total	64,72%	477	35,28%	260

As can be seen here, there is margin for improvement: 35 %, which can be improved by increasing the size of the search corpus.

5.3 Comparison with current WordNet

In this section we compare the results obtained in phase 2 with the results of the current WordNet, as only the results that do not appear in the current WordNet will entail a real increase in the completeness of WordNet.

POS	Not present		Present	
R	66,67%	2	33,33%	1
V	73,05%	206	26,95%	76
A	52,31%	34	47,69%	31
N	62,20%	79	37,80%	48
Total	67,30%	321	32,70%	156

5.4 Manual evaluation of disambiguated synsets

A manual qualitative evaluation was conducted to measure the accuracy of the results. We randomly selected 25 synsets for each POS (verb, adverb, noun and adjective) and we verified if the result was correct or not. We must remember that for adverbs there were only two results. It is important to remember that most of the errors detected at this stage correspond to lemmas that had been accurately translated but whose translation was not the correct one for the synset in question. The lemma “cup” of synset “eng-30-03147901-n” with the sense of “trophy” is a good example of this. The translations obtained for the lemma were “taza” and “copa”, and when requesting disambiguation the process selected “taza”, which was not the correct meaning for this synset.

POS	Correct	Incorrect
R	100.00%	0.00%
V	68.00%	32.00%
A	84.00%	16.00%
N	68.00%	32.00%
Total	74.03%	25.97%

From these evaluations we can conclude that phase 2 was not as accurate as phase 1. These results could be improved by increasing the size of the corpus or by improving the method. A larger corpus would have more sentences, that is to say, more contexts where the meaning of candidates can be validated. The translations where the gender does not match in English could be discarded to improve the method. Doing this would discard cases like that of synset “spa-30-10129825-n”, whose gloss is “mujer joven”. For the lemma “girl”, which corresponds to said English synset, a possible translation obtained was “chico”. This is a clear example where the original lemma in English and the resulting translation do not match in gender. Another way to improve the method would be to prioritize some specific relations.

6 Evaluation of the results on Corin lexicon

To evaluate the results obtained in both phases we implemented a task to measure the semantic coverage on a small corpus, in this case Corin. For this task we obtained all the lemmas in the corpus, applied Freeling to know the grammatical category, and then searched WordNet. This process was first executed with the original WordNet, our starting point, and then with the resulting WordNet. The aim was to measure the improvement in the coverage of the existing lemmas in the corpus under study of the resulting WordNet regarding the current WordNet. We must remember that the process to improve WordNet was executed on a random set of 1000 synsets per POS. The results obtained must be weighed considering the percentage these synsets represent within the total number of synsets for each POS. These percentages are shown in the following table.

There follows a table with the percentages of

POS	Total	Processed Synsets	
V	13845	1000	7.22%
N	83090	1000	1.20%
R	3621	1000	27.62%
A	18156	1000	5.51%

coverage obtained according to each POS, for the two versions of WordNet: the original one and the one expanded by this method.

POS	Original WordNet		Expanded WordNet		In the Corpus
V	75.35%	1235	77.36%	1268	1639
N	69.29%	2780	70.09%	2812	4012
R	32.79%	121	62.87%	232	369
A	51.00%	840	54.34%	895	1647

We can conclude that adverbs was the category with the best results, reaching a coverage of almost 63 % over the original 33 %. Two reasons explain this: first, adverbs is the category least covered by the original WordNet, and it was also the POS where the strategy was implemented more times, which was executed on just over 27 % of its synsets. The coverage also improved for the other POS. Though it is true that the improvement was relatively small (between 1 % and 3 %), we must remember that in these cases the method was applied to a small percentage of the synsets in WordNet.

7 Conclusions

Different strategies were designed and implemented in order to enrich the current Spanish WordNet from the English WordNet within the context of the expansion model. The strategy was to use a series of selectors which were called "initial selectors" as a first step. We then applied a method based on the exploitation of the semantic relations of WordNet so as to add variants that the initial selectors had not been able to add. The results obtained show that the strategy used is effective as it entails a significant improvement of the current Spanish WordNet, thus complying with the initial expectations. One of the weaknesses lies in the translation methods and tools, as they provide the resources our proposals are based on. This is why they strongly condition the final results. Regarding the strategy implemented, the initial selectors are sufficient to significantly improve the current WordNet, with a 92 % accuracy, while there was a 74 % accuracy in phase 2.

References

Agirre, A. G., Laparra, E., Rigau, G., & Donostia, B. C. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. in gwc 2012 6th international global wordnet conference.

Armentano-Oller, C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Montava Belda, M. A., Ortiz-Rojas, S., ... Sánchez-Martínez, F. (n.d.).

Atserias, J., Climent, S., Farreres, X., Rigau, G., & Guez, H. R. (1997). Combining multiple methods for the automatic construction of multilingual wordnets. In *In proceedings of international conference on recent advances in natural language processing (ranlp'97), tzigov chark* (pp. 143–149).

Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., & Taulé, M. (1998). Methods and tools for building the catalan wordnet. In *in proceedings of elra workshop on language resources for european minority languages.*

Bing online translator. (2015). `https://www.bing.com/translator`.

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th international conference on language resources and evaluation (lrec'04).*

Grassi, M., Malcuori, M., Couto, J., Prada, J. J., & Wonsever, D. (2001). Corpus informatizado: textos del español del uruguay (corin). In *Slplt-2-second international workshop on spanish language processing and language technologies-jaén, españa.*

Pradet, Q., de Chalendar, G., & Desormeaux, J. B. (2014). Wonef, an improved, expanded and evaluated automatic french translation of wordnet. *Volume editors*, 32.

Vossen, P. (1998). Introduction to eurowordnet. In *Eurowordnet: A multilingual database with lexical semantic networks* (pp. 1–17). Springer.

Wikimedia foundation. 2008b. wiktionary. (2008). `http://www.wiktionary.org`.

An Analysis of WordNet's Coverage of Gender Identity Using Twitter and The National Transgender Discrimination Survey

Amanda Hicks
University of Florida
Gainesville, FL, USA
aehicks@ufl.edu

Michael Rutherford
University of Arkansas for Medical Sciences
Little Rock, AR, USA
mwrutherford@uams.edu

Christiane Fellbaum
Princeton, University
Princeton, NJ, USA
fellbaum@prinecton.edu

Jiang Bian
University of Florida
Gainesville, FL, USA
bianjiang@ufl.edu

Abstract

While gender identities in the Western world are typically regarded as binary, our previous work (Hicks et al., 2015) shows that there is more lexical variety of gender identity and the way people identify their gender. There is also a growing need to lexically represent this variety of gender identities. In our previous work, we developed a set of tools and approaches for analyzing Twitter data as a basis for generating hypotheses on language used to identify gender and discuss gender-related issues across geographic regions and population groups in the U.S.A. In this paper we analyze the coverage and relative frequency of the word forms in our Twitter analysis with respect to the National Transgender Discrimination Survey data set, one of the most comprehensive data sets on transgender, gender non-conforming, and gender variant people in the U.S.A. We then analyze the coverage of WordNet, a widely used lexical database, with respect to these identities and discuss some key considerations and next steps for adding gender identity words and their meanings to WordNet.

1 Introduction

Gender identity is richly lexicalized in American English. Nevertheless, a cursory investigation of gender identity in WordNet (Miller, 1995) suggests that coverage of non-binary gender identity is low. The goal of our research is to measure the coverage of WordNet's gender identity and to suggest steps to improve it.

There is increasing incentive to include gender identity terms and other words that are relevant to transgender, gender variant, non-binary, and gender non-conforming people in WordNet. For example, the Institute of Medicine (IOM) recently recommended (1) gathering data on sexual orientation and gender identity in Electronic Health Records (EHR) as part of the meaningful use objectives in EHRs, (2) developing standardization of sexual orientation and gender identity measures to facilitate synthesizing scientific knowledge about the health of sexual and gender minorities, and (3) supporting research to develop innovative methods of conducting research with small populations to determine the best ways to collect information on LGBT minorities. Furthermore, it is important for the medical community to use words that are common among patients and research participants since the use of language that is familiar to the participant has been shown to improve response rates in data collection (Catania et al., 1996; Institute of Medicine, 2011; Alper et al., 2013).

However, there are challenges to determining which words to include in WordNet and how to define them. Based on the limited research available, some evidence (Dargie et al., 2015; Kuper et al., 2012; Scheim and Bauer, 2015) suggests that vocabulary for self-identifying gender and sexual orientation varies by community. There is clear evidence of lexical variation associated with geography in linguistics studies (Carver, 1987; Chambers, 2001; Nerbonne, 2013). Also, through discussions with members of the trans* community and health care providers at LGBT clinics across the country, we have learned that new words are frequently coined to describe gender identity and that the connotations of existing words may vary across communities. We use 'trans*' broadly to refer to transgender, transsexual, gender non-conforming, gender variant, and non-binary individuals.

User generated content on social media, such as Twitter, is a valuable resource because it can

provide a source for gleaning information about people's daily lives to answer scientific questions. In our previous work, we produced a data set to investigate words used to discuss gender in the general population and among self-identifying trans* persons using Twitter (Hicks et al., 2015). With 'self-identifying' we refer to people who have stated that they have a trans* identity either through their tweets or in the National Transgender Discrimination Survey (NTDS) (Grant et al., 2011). We believe that we can augment our Twitter data set with the NTDS data to produce a data set that is in sync with current speakers' language, that can serve as a starting point for enriching WordNet's coverage of gender identity, and that can contribute to the medical and clinical goals outlined at the beginning of this section.

The National Transgender Discrimination Survey (NTDS) is the largest survey of the trans* population in the United States to date (Harrison et al., 2012). The survey was designed to collect information about "the broadest possible swath of experiences of transgender and gender nonconforming people" in the U.S.A., including questions about how participants identify their own gender and an option to write in one's own identity (Harrison et al., 2012). We have compiled a list of the gender-identity word forms (henceforth simply 'words') from this survey and performed a normalized frequency analysis that can be compared to our Twitter data set.

In our previous work we built a data set and visualization tools that show relative frequency and co-occurrence networks for American English trans* words on Twitter (Grant et al., 2011). Our goal in this paper is to perform a two-fold coverage analysis of WordNet with respect to American English gender identity.

Our hypothesis is that a comprehensive list of words used to self-identify gender will require examining the words trans* people use in different contexts. In order to evaluate this hypothesis, we perform a frequency analysis of words from both sets.

Our approach is as follows. First, we compare the trans* identity words that we identified in our previous work with the words from the NTDS to assess the coverage of the Twitter set. Next, we produce an updated set of words using the NTDS and compare WordNet's coverage of gender identity against this list.

2 Methods

Here we describe our language analysis of the Twitter data and the NTDS data.

2.1 Language Analysis of Twitter Data

The general idea underlying our approach is to identify tweets that are relevant to the discussion of trans* related issues and then examine the variations in language used for gender identification by different communities, that is, by population (trans* people vs. the general public) and by geographical location (U.S. states). The analysis workflow consists of five main steps, as depicted in Figure 1: 1) collect tweets that are potentially related to discussions about gender identification; 2) preprocess and geotag tweets with their corresponding U.S. state; 3) build supervised classification models based on textual features in the tweets to a) filter out irrelevant tweets and b) find people who are self-identified as trans*; 4) collect relevant (both self-identifying trans* users and users in the general public who discussed trans* related issues) users' Twitter timelines which consists of all of their tweets in chronological order; and 5) compare the usage of gender identification words by geographical locations (i.e., by U.S. states) and by population groups (self-identifying trans* people vs. the general public).

Some of the search terms are ambiguous and their meanings are context dependent. For example, the tweet 'That Hot Pocket is full of trans fats' is not related to discussions of gender identification even though it contains the keyword 'trans'. To account for this observation, we engineered a binary classifier to determine the likelihood that a tweet is relevant to the discussion of gender identification and to remove those that are unlikely to be relevant from the corpus in step 3. We also leverage a number of visualization techniques to provide straightforward and easy-to-understand visual representations, namely, word clouds, co-occurrence matrices, and network graphs to substantiate our findings. A full description of this work and analysis of terms can be found in (Hicks et al., 2015).

2.2 Language Analysis of NTDS Data

Unlike the Twitter study data processing techniques, the NTDS dataset did not require the preprocessing for language filtering, geotagging or the mining techniques for the identification of rel-

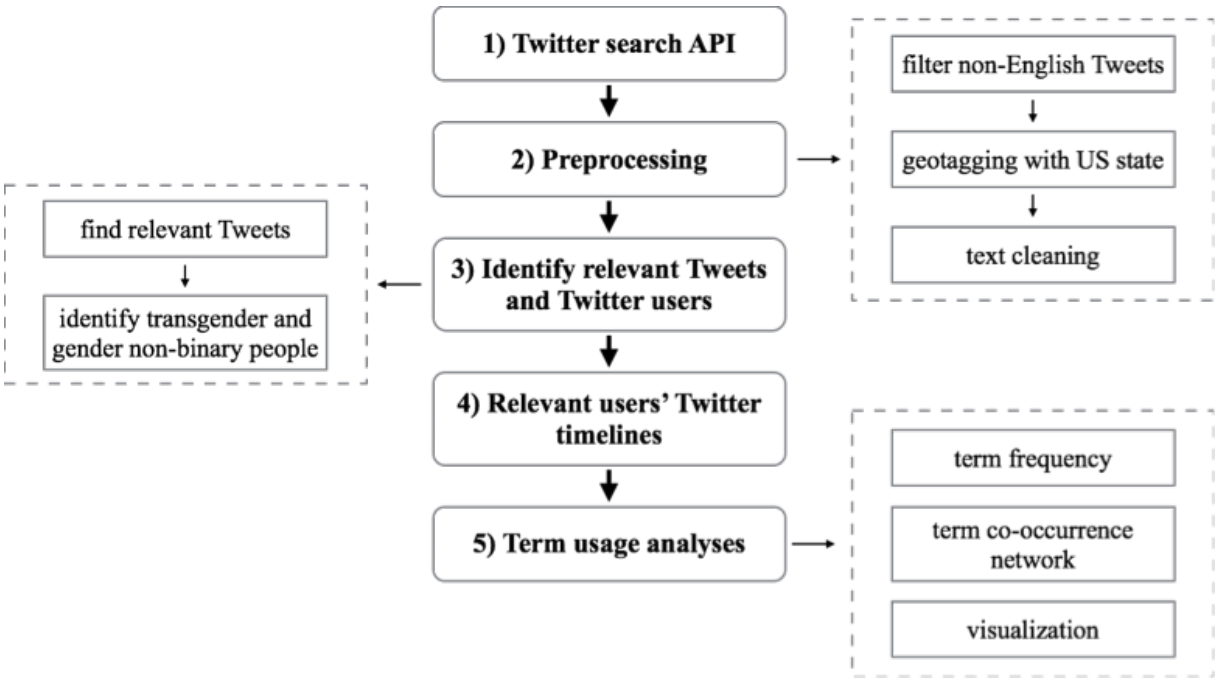


Figure 1: The analysis workflow for identifying tweets related to trans* issues

evant trans* individuals. Knowing that the records were all unique self-identified trans* individuals, we were able to skip ahead to Step 5, the term usage analysis.

The Twitter data analysis methods were duplicated and restricted to the term extraction and usage analysis, including term frequencies and word cloud generation.

We utilized questions three and four from the NTDS. These questions asked what gender identity the respondent identified with at the time of the survey and how strongly they identified with certain identities. Figure 2 shows these questions.

Term frequency analyses were generated based on all words utilized, no matter the degree with which the respondent specified (strongly, somewhat, or not at all). The frequencies were then measured both at a state and national level for coverage comparisons with the Twitter set.

2.3 Coverage Analysis of Twitter Words

We performed a coverage analysis of the words in the Twitter data set with those from the NTDS data set. We collated all of the words in the NTDS questions three and four as well as the identity words used in the write-in responses. We removed terms that were preceded by a hash tag in the Twitter set and words that were only used once in the NTDS set, and then we measured the number of common words from both the Twitter list and the NTDS list. Due to the character limit on Twitter, abbreviations are common in Tweets as are alternate spellings of words (e.g., 'gender queer' and 'gender-queer'). We also gathered words into groups consisting of alternative spellings and abbreviations. 'Genderqueer' and 'gender-queer' are in the same group. Henceforth we call these groups of word forms simply 'groups'. We measured the degree of overlap of groups in Twitter and in NTDS which is reported in the results section of this paper.

2.4 Coverage Analysis of WordNet

Our next step was to generate a list of words to use in the coverage analysis of WordNet. We removed the Twitter terms that contained a hash tag from the Twitter data set and removed word forms that only had one occurrence in the NTDS set. We then took the union of these sets to produce a set of words for evaluating the coverage of WordNet. Similarly, we produced a list of groups with alternate spellings and abbreviations by taking the union set of corresponding groups for the Twitter list and NTDS list. For example, the NTDS word groups contained the set of word forms (gender non-conforming, gender non conforming) and the Twitter word groups contained (gender non-conforming, gnc). The compiled set of groups contains (gender non-conforming, gender non conforming, gnc).

3. What is your primary gender identity today?

- ❍ Male/Man
- ❍ Female/Woman
- ❍ Part time as one gender, part time as another
- ❍ A gender not listed here, please specify _____

4. For each term listed, please select to what degree it applies to you.

	Not at all	Somewhat	Strongly
Transgender	❍	❍	❍
Transsexual	❍	❍	❍
FTM (female to male)	❍	❍	❍
MTF (male to female)	❍	❍	❍
Intersex	❍	❍	❍
Gender non-conforming or gender variant	❍	❍	❍
Genderqueer	❍	❍	❍
Androgynous	❍	❍	❍
Feminine male	❍	❍	❍
Masculine female or butch	❍	❍	❍
A.G. or Aggressive	❍	❍	❍
Third gender	❍	❍	❍
Cross dresser	❍	❍	❍
Drag performer (King/Queen)	❍	❍	❍
Two-spirit	❍	❍	❍
Other, please specify _____	❍	❍	❍

Figure 2: Questions 3 and 4 from the National Transgender Discrimination Survey that asks respondents to report their gender identity

We automatically searched for words and groups of synonymous words ('synsets') that corresponded to words and groups using the Natural Language Tool Kit's (NLTK) interface for WordNet 3.0 (Bird et al., 2009). We then manually evaluated which synsets were relevant to gender identity. We did not evaluate whether the WordNet definition accurately characterized the intended meaning of the word, in part because we do not have a reliable method for ascertaining the intended meaning of the word and also because that is outside of the scope of our coverage analysis.

Many of the groups that did not have a corresponding synset in WordNet 3.0 were compounds such as 'trans person of color'. Our next step was to produce a list of words in compounds and search for corresponding synsets in WordNet. We manually identified compounds and then generated a set of words in the compounds. We removed stop words from the set with NLTK. Once again we programmatically searched for synsets using NLTK and then manually evaluated whether the retrieved synset was relevant to gender identity. We classified the compounds into three groups: (1) those that were partially covered by WordNet, meaning they contained at least one word that corresponded to a relevant synset and at least one that did not, (2) those that were completely covered by WordNet, meaning every word in the compound (excluding stop words) was represented in WordNet, and (3) those that had no coverage in WordNet.

3 Results

First we discuss the results of analysis of our Twitter data. Then we discuss our analysis of WordNet's coverage of trans* related terms.

3.1 Language Analysis of Twitter Data

We collected over 53.8 million tweets matching the search queries during a 116-day period from January 17, 2015 to May 12, 2015 inclusive. Out of the collected tweets, about 29 million tweets (54.2%) were in English. We were able to extract location information for 368,518 tweets (1.26% of English tweets from 119,778 unique users), which we retained for further processing. We eliminated the tweets that were deemed irrelevant (15,478 tweets from 3,785 users) based on a classification model we developed (Hicks et al., 2015). From the remaining records, 115,993 Twitter users were classified as relevant, of which 1,921 users were classified as self-identifying trans*. In addition to the data we collected using the search API, we

Figure 3: Word clouds representing the relative frequency of trans* words used by self-identifying trans* people on Twitter in the U.S.A. (left) and self-identifying trans* people in questions three and four of the NTDS (right)

	Unique	Shared
NTDS Words	79.66% (141 / 177)	20.34% (36 / 177)
NTDS Groups	81.82% (117 / 143)	18.18% (26 / 143)
Twitter Words	80.65% (150 / 186)	19.35% (36 / 186)
Twitter Groups	67.50% (54 / 80)	32.50% (26 / 80)

Table 1: The percentage of overlap among NTDS and Twitter words and groups

crawled more than 337.9 million tweets from the 115,993 relevant Twitter users' timelines. Out of the 337.9 million tweets, 872,340 Twitter messages contain one or more of the keyword forms of our interest. These 872k tweets comprise the corpus we used for language usage analysis.

3.2 Coverage of Twitter Word Groups

Table 1 contains a summary of the degree of overlap between the set of Twitter trans* words and their groups and the NTDS trans* words and their groups. Only about 18% of the NTDS groups were represented in the Twitter data set. Section 4.2 contains a discussion of some of the main reasons for the most frequent word forms not being in the Twitter data set.

The word clouds in Figure 3 illustrate two interesting facts about word usage to self-describe trans* identity.

First, different words appear in different contexts. For example, 'cis' and 'shemale' are prevalent on Twitter but not in the NTDS. Second, even words that are common across contexts are used with different frequency. For example, 'genderqueer' is prominent in the NTDS word cloud but relatively small in the Twitter word cloud (top left-hand quadrant). Conversely, 'Transgender' is more prominent in the Twitter word cloud than the NTDS.

3.3 WordNet's Coverage of Gender Identities

We found that 39% of the words in our compiled list of trans* groups have a corresponding synset in WordNet 3.0. Another 28% of the words were compounds that contain at least one component word with a corresponding synset in WordNet and one without. 33% of the words did not have any corresponding entries in WordNet. These results are summarized in Figure 4. Table 2 shows a numerical analysis of WordNet's 3.0 coverage of our trans* related words.

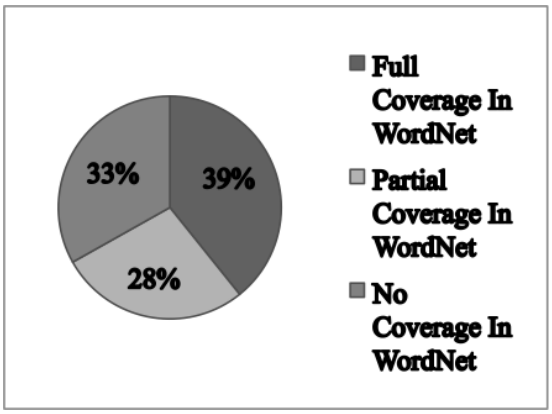


Figure 4: Summary of WordNet 3.0's coverage of trans* word groups

4 Discussion

4.1 Limitations

We note that our previous study is limited by the user demographics available on social media

	Word Groups in WordNet
Full WordNet Coverage	**71**
Non-Compounds In WordNet	39
Compound - Full Coverage	32
Compound with Partial WordNet Coverage	**50**
Non-compounds not in WordNet	61
Compound - No WordNet Coverage	13
No WordNet coverage	**60**
Total Trans* Word groups	**195**

Table 2: Analysis of trans* word groups in WordNet 3.0 reported by number

platforms. The users of social media tend to be younger; 37% of Twitter users are under 30, while only 10% are 65 or older, as of 2014 (Duggan, Ellison, Lampe, Lenhart, & Madden, 2014). There are also power users who exhibit a substantially greater level of activity than the average user (Pew Research Center, 2015). These characteristics are likely to create sample bias and impose limitations on mining meaningful information from Twitter that represents a broader population. For instance, Twitter data may not be reliable for mining information about older people who may not use Twitter.

The NTDS was published in 2011, but more current data are being collected at the time of writing this paper. The Transgender Survey 2015 was launched in August 2015 (U.S, 2015) and the PRIDE study in June 2015 (PRI, 2015). We expect these newer data sources to be completed within the next year or two. Both studies collect demographic data on trans* individuals, including identity words. This will provide insight into which words are relatively stable over time and may also reveal words that are emerging as more prevalent.

4.2 Words Excluded From Twitter Search Terms

While compiling a list of words for Twitter, we observed the distinctions among trans* identities, intersex conditions, and sexual orientation. As a result we excluded words that were specifically intersex related or that describe sexual orientation from the Twitter set. However, intersex and

NTDS Term	NTDS %	Included in Twitter Set
Aggressive	10.4%	No
Genderqueer	5.5%	Yes
Transgender	5.3%	Yes
Butch	5.2%	No
Female-to-male	5.2%	Yes
Androgynous	5.2%	No
Male-to-female	5.2%	Yes
Transsexual	5.2%	Yes
Two-Spirit	5.2%	Yes
Intersex	5.2%	No

Table 3: Ten most frequent words in NTDS

sexual orientation words were among participant responses in the NTDS so were included in our NTDS data set. The heterogeneous nature of the Twitter term lists and NTDS term lists may skew the coverage analysis of our Twitter list. However, this heterogeneity is valuable for our analysis of WordNet's coverage since it provides a more comprehensive list of words that trans* people use to describe their own identities.

An examination of tables 3 and 4 reveals three main reasons words from the NTDS term lists

were not included in the Twitter term lists: (1) Polysemy - 'Aggressive' is polysemous and would result in too many false hits in the Twitter search. Similarly 'androgynous' produced too many false hits since many people who used this word were tweeting about fashion. (2) Gender words that are not trans* specific -'male', 'female', 'woman', and 'man', are used with such prevalence that we excluded them in the Twitter set since they are unhelpful in identifying tweets about trans* issues. (3) Identity words that are not trans* specific - 'butch' and 'intersex' were deliberately excluded from the Twitter set since we were following the conceptual distinctions among sexual orientation, gender identity, and intersex. However, the NTDS data set shows that when individuals describe their gender identities, they do not limit their descriptions to these high level distinctions.

4.3 Suggestions for Integrating Gender Identity Into the WordNet Database

Approximately one third of the compounds with partial or no coverage have 'gender' as a component term. The synsets for 'gender' in WordNet are tied to biological properties and reproductive roles, and there is no synset for gender as a social role independently of reproductive features. Other words that would have a significant effect on WordNet's coverage of compounds are 'trans', 'genderqueer', and 'femme'. Some words that are relevant to the trans* issues such as 'agender', 'cisgender' (describing somebody who is not trans*), and 'binarism' are missing.

In addition to adding more words to integrate gender identity in WordNet, efforts should be made to craft informed definitions and example sentences of new words and to evaluate the accuracy of existing entries. Likewise, more work needs to be done to identify synsets. The word groups that we used for this study grouped morphologically similar words such as 'gender queer' and 'gender-queer'. However, we did not group words like 'agender' and 'genderless' into synsets. Methods for reliably detecting synonyms of gender identity words should be developed and tested.

Finally, methods also need to be developed for establishing hierarchy relations among gender identity words. Such methods may include testing established lexical patterns with English speakers who are competent with trans* vocabulary (Hearst, 1992). Another approach may include leveraging the responses in question 4 of the NTDS to detect hierarchy relations. For example, if most participants who identify strongly as transgender also identify strongly as genderqueer but not vice versa, this could indicate that 'genderqueer' is a hypernym of 'transgender'.

NTDS Term	NTDS %	Included in Twitter Set
Genderqueer	16%	Yes
Male	8.7%	No
Female	8.2%	No
Woman	4.9%	No
Queer	4.7%	No
Transgender	3.5%	Yes
Trans	2.7%	Yes
Man	2.7%	No
Butch	1.8%	No
Female-to-male	1.7%	Yes

Table 4: The ten most frequent words in the NTDS write-in fields in questions three and four

4.4 Future Work

Wordnets have been built in some seventy different languages, and each reflects the culture of the speakers. Mapping gender identity words across languages should reveal interesting similarities and differences. For example, India allows its citizens to officially identify as 'third gender', or *hijra*, a term that encompasses biological males dressing in women's clothes as well as intersex individuals. Future research within the global wordnet community could ask whether such officially sanctioned words cover distinct words used in specific communities and if so, how do they correspond to the English words identified in our work? Twitter corpora can show which terms are used in similar or identical contexts (n-grams), suggesting synonymy and shared synset membership. Additionally, questionnaires could be developed and submitted to the trans* population for input on how to accurately represent the terms. Reflecting geographic and group differences poses additional challenges, akin to dialectal variation that is cur-

rently marked in WordNet with usage flags.

5 Conclusion

Our hypothesis was that a comprehensive list of words used to describe gender identity will require sets of words taken from different contexts. To test this hypothesis we performed a coverage analysis of trans* words taken from two different contexts, Twitter and the National Transgender Discrimination Survey. We found that while there was some overlap, there was significant variation of words used between these contexts. As a result, we generated a more comprehensive list of trans* words from both sources. A second aim of this paper was to assess WordNet's coverage of trans* identity. We found that, while there is some coverage of trans* words in WordNet, there is more work to be done to ensure more comprehensive coverage.

Acknowledgements

We are grateful to the National Center for Transgender Equality (NCTE) for providing the dataset from the National Transgender Discrimination Survey. This work was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 and the University of Arkansas for Medical Sciences UL1 TR000039. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NCTE.

References

Joe Alper, Monica N Feit, Jon Q Sanders, et al. 2013. *Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records: Workshop Summary*. National Academies Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Craig M Carver. 1987. *American Regional Dialects: A Word Geography*. University of Michigan Press.

Joseph A Catania, Diane Binson, Jesse Canchola, Lance M Pollack, Walter Hauck, and Thomas J Coates. 1996. Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly*, 60(3):345–375.

Jack K Chambers. 2001. Region and language variation. *English world-wide*, 21(2):169–199.

Emma Dargie, Karen L Blair, Caroline F Pukall, and Shannon M Coyle. 2015. Somewhere under the rainbow: Exploring the identities and experiences of trans persons. *The Canadian Journal of Human Sexuality*.

Jaime M Grant, Lisa Mottet, Justin Edward Tanis, Jack Harrison, Jody Herman, and Mara Keisling. 2011. *Injustice at Every Turn: A Report of the National Transgender Discrimination Survey*. National Center for Transgender Equality.

Jack Harrison, Jaime Grant, and Jody L Herman. 2012. A gender not listed here: Genderqueers, gender rebels, and otherwise in the National Transgender Discrimination Survey. *LGBTQ Public Policy Journal at the Harvard Kennedy School*, 2(1).

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Amanda Hicks, R. Hogan, William, Michael Rutherford, Bradley Malin, Mengjun Xie, Christiane Fellbaum, Zhijun Yin, Daniel Fabbri, Josh Hanna, and Jiang Bian. 2015. Mining Twitter as a first step toward assessing the adequacy of gender identification terms on intake forms. In *Proceedings of the AMIA 2015 Annual Symposium*. American Medical Informatics Association.

Institute of Medicine. 2011. The health of lesbian, gay, bisexual, and transgender people: Building a foundation for better understanding.

Laura E Kuper, Robin Nussbaum, and Brian Mustanski. 2012. Exploring the diversity of gender and sexual orientation identities in an online sample of transgender individuals. *Journal of Sex Research*, 49(2-3):244–254.

John Nerbonne. 2013. How much does geography influence language variation? *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, pages 220–36.

2015. The Population Research in Identity and Disparities for Equality (PRIDE) Study. `http://www.pridestudy.org`. Accessed: 2015-08-24.

Ayden I Scheim and Greta R Bauer. 2015. Sex and gender diversity among transgender persons in Ontario, Canada: Results from a respondent-driven sampling survey. *The Journal of Sex Research*, 52(1):1–14.

2015. U.S. Trans Survey 2015. `http://www.transsurvey.org`.

Where Bears Have the Eyes of Currant: Towards a Mansi WordNet

Csilla Horváth[1], **Ágoston Nagy**[1], **Norbert Szilágyi**[2], **Veronika Vincze**[3]
[1]University of Szeged, Institute of English–American Studies
Egyetem u. 2., 6720 Szeged, Hungary
horvathcs@ieas-szeged.hu, nagyagoston@lit.u-szeged.hu
[2]University of Szeged, Department of Finno-Ugrian Studies
Egyetem u. 2., 6720 Szeged, Hungary
norbertszilagyi91@gmail.com
[3]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
Tisza Lajos krt. 103., 6720 Szeged, Hungary
vinczev@inf.u-szeged.hu

Abstract

Here we report the construction of a wordnet for Mansi, an endangered minority language spoken in Russia. We will pay special attention to challenges that we encountered during the building process, among which the most important ones are the low number of native speakers, the lack of thesauri and the bear language. We will discuss our solutions to these issues, which might have some theoretical implications for the methodology of wordnet building in general.

1 Introduction

Wordnets are lexical databases that are rendered according to semantic and lexical relations between groups of words. They are supposed to reflect the internal organization of the human mind (Miller et al., 1990). The first wordnet was constructed for English (Miller et al., 1990) and since that time, wordnets have been built for several languages including several European languages, mostly in the framework of EuroWordNet and BalkaNet (Alonge et al., 1998; Tufiş et al., 2004) and other languages such as Arabic, Chinese, Persian, Hindi, Tulu, Dravidian, Tamil, Telugu, Sanskrit, Assamese, Filipino, Gujarati, Nepali, Kurdish, Sinhala (Tanács et al., 2008; Bhattacharyya et al., 2010; Fellbaum and Vossen, 2012; Orav et al., 2014). Synsets within wordnets for different languages are usually linked to each other, so concepts from one language can be easily mapped to those in another language. Wordnets can be beneficial for several natural language processing tasks, be it mono- or multilingual: for instance, in machine translation, information retrieval and so on.

In this paper, we aim at constructing a wordnet for Mansi, an indigenous language spoken in Russia. Mansi is an endangered minority language, with less than 1000 native speakers. Most often, minority languages are not recognized as official languages in their respective countries, where there is an official language (in this case, Russian) and there is one or there are several minority languages (e.g. Mansi, Nenets, Saami etc.). Hence, the speakers of minority languages are bilingual, and usually use the official or majority language in their studies and work, and the language of administration is the majority language as well. However, the minority language is typically restricted to the private sphere, i.e. among family members and friends, and thus it is mostly used in oral communication, with only sporadic examples of writing in the minority language (Vincze et al., 2015). Also, the cultural and ethnographic background of Mansi people may affect language use: certain artifacts used by Mansi people that are unknown to Western cultures have their own vocabulary items in Mansi and vice versa, certain concepts used by Western people are unknown to Mansi people, therefore there are no lexicalized terms for them.

The construction of a Mansi wordnet help us explore how a wordnet can be built for a minority language and also, an endangered language. Thus, we will investigate the following issues in this paper:

- What are the specialties of constructing a wordnet for a minority language?
- What are the specialties of constructing a word-

net for an endangered language?

- What are the specialties of constructing a wordnet for Mansi?

The paper has the following structure. First, the Mansi language will be shortly presented from linguistic, sociolinguistic and language policy perspectives. Then our methods to build the Mansi wordnet will be discussed, with special emphasis on specific challenges as regards endangered and minority languages in general and Mansi in particular. Later, statistical data will be analysed and our results will be discussed in detail. Finally, a summary will conclude the paper.

2 The Mansi Language

Mansi (former term: Vogul) is an extremely endangered indigenous Uralic (more precisely Finno-Ugric, Ugric, Ob-Ugric) languages, spoken in Western Siberia, especially on the territory of the Khanty-Mansi Autonomous Okrug. Among the approximately 13,000 people who declared to be ethnic Mansi according to the data of the latest Russian federal census in 2010 only 938 stated that they could speak the Mansi language.

The Mansi have been traditionally living on hunting, fishing, to a lesser extent also on reindeer breeding, they got acquainted with agriculture and urban lifestyle basically during the Soviet period. The principles of Soviet linguistic policy according to which the Mansi literary language has been designed kept changing from time to time. After using Latin transcription for a short period, Mansi language planners had to switch to the Cyrillic transcription in 1937. While until the 1950s the more general tendency was to create new Mansi words to describe the formerly unknown phenomena, later on the usage of Russian loanwords became more dominant. As a result of these tendencies some of the terms describing contemporary environment, urban lifestyle, the Russian-dominated culture are Russian loanwords, while others are Mansi neologisms created by Mansi linguists and journalists. It is not uncommon to find two or even three different synonyms describing the same phenomena (for example, hospital): by the means of borrowing the word from Russian (больница), or using the Russian loanword in a form adapted to the Mansi phonology (пӯльница), or using a Mansi neologism to describe it (мāхум пусмалтан кол, 'a house for healing people, hospital', as opposed to няврам пусмалтан кол 'children hospital, children's clinic' or ӯйхул пусмалтан кол 'veterinary clinic').

3 Semi-automatic construction of the Mansi WordNet

In this section, we will present our methods to construct the Mansi WordNet. We will also pay special attention to the most challenging issues concerning wordnet building.

3.1 Low number of native speakers

The first and greatest problem we met while creating the Mansi wordnet was that only a handful of native speakers have been trained in linguistics. Thus, we worked with specialists of the Mansi language who have been trained in linguistics and technology, but do not have native competence in Mansi.

As it is not rentable to build a WordNet from scratch and as our annotators are native speakers of Hungarian, we used the Hungarian WordNet (Miháltz et al., 2008) as a starting point. First, we decided to include basic synsets, and the number of the synsets is planned to be expanded continuously later on. We used Basic Concepts – already introduced in EuroWordNet – as a starting point: this set of synsets contains the synsets that are considered the most basic conceptual units universally.

3.2 Already existing resources

In order to accelerate the whole task and to ease the work of Mansi language experts, the WordNet creating process was carried out semi-automatically. Since there is no native speaker available who could solve the problems requiring native competence, we were forced to utilize the available sources as creatively as possible.

First, the basic concept sets of the Hungarian WordNet XML file were extracted and at the same time, the non-lexicalized elements were filtered as in this phase, we intend to focus only on lexicalized elements.

Second, we used a Hungarian-Mansi dictionary to create possible translations for the members of

the synsets. The dictionary we use in the process is based on different Mansi-Russian dictionaries (e.g. Rombandeeva (2005), Balandin and Vahruševa (1958), Rombandeeva and Kuzakova (1982)). The translation of all Mansi entries to Hungarian and to English in the new dictionary is being done independently of WordNet developing (Vincze et al., 2015).

In order not to get all Hungarian entries of the WordNet translated to Mansi again, a program code was developed to replace the Hungarian terms with the already existing translations from the dictionary. Only literals are replaced, definitions and examples are left untouched, so that the linguists can check the actual meaning and can replace them with their Mansi equivalents. The Mansi specialists' role is to check the automatic replacement and to give new term candidates if there is no proper automatic translation.

In this workphase, as there are no synonym dictionaries or thesauri available for the Mansi language, the above-mentioned bilingual student dictionaries are used as primary resources. These dictionaries were designed to be used during school classes, they rarely contain any synonyms, antonyms or hypernyms, and hardly any phrases or standing locutions. (Most of these dictionaries were written by the same authors, thus – besides the inconsistent marking of vowel length – fortunately we do not have to pay special attention to possible contradictions or incoherence.) Hence originates the unbalanced situation in which we are either missing the Mansi translation, either the Mansi definition belonging to the same code, and we are able to present the translation, the definition and the examples of usage only in a few extraordinary instances. The sentences illustrating usage in the synset come from our Mansi corpus, built from articles from the Mansi newspaper called *Luima Seripos* published online semimonthly at `http://www.khanty-yasang.ru/luima-seripos`. In its final version, our corpus will contain above 1,000,000 tokens, roughly 400,000 coming from the online publications and the rest from the archived PDF files.

Even if based on the Hungarian WordNet, the elements of the Mansi WordNet can be matched to the English ones and those of other wordnets since the Hungarian WN itself is paired with the Princeton WordNet (Miller et al., 1990).

3.3 Bear language

Another very special problem occurred during wordnet building in Mansi, that is the question regarding the situation of the so called "bear language". The bear is a prominently sacred animal venerated by Mansi, bearing great mythical and ritual significance, and also surrounded by a detailed taboo language. Since the bear is believed to understand the human speech (and also to have sharp ears), it is respectful and cautious to use taboo words while speaking about the bear, the parts of its body, or any activity connected with the bear (especially bear hunting) so that the bear would not understand it. The taboo words of this "bear language" may be divided into two major subgroups: Mansi words which have a different, special meaning when used in connection with the bear (e.g. сосыг 'currant' but also meaning 'eye', when speaking of the bear's eyes), and those which may be used solely in connection with the bear (e.g. хащӆлы 'to be angry', as opposed to кантлы 'to be angry' speaking of a human). Even the word for bear belongs to taboo words and has only periphrastic synonyms like Вӧртӧлнӧйка 'an old man from the forest' etc.

As a first approach, taboo words were included as literals in the synsets because their usage is restricted in the sense that they can solely be used in connection with bears. Hence, first we marked the special status of these literals, for which purpose we applied the note "bear". However, it would have also been practical to well differentiate the synsets that are connected to "bears". This can be realized in many ways: for example, the "bear"-variants of the notions should be the hyponyms of their respective notions, like хащӆлы 'to be angry', which can be considered as a hyponym of кантлы 'to be angry' speaking of a human. However, this solution is not a perfect one since (i) this is not a widespread method either in WordNets of other languages and therefore it would not facilitate WordNet-based dictionaries and (ii) it is not a true hyponym, that is, a real subtype of their respective notion connected to humans. Finally, we decided to put these notions in separate synsets, which has the advantage that these notions are grouped together and it is easier to do a targeted search on these expressions.

4 Results

The manual correction of the automatically translated Basic Concept Set 1 is in progress. Currently, the online xml file contains 300 synsets. These synsets had altogether 410 literals, thus a synset had 1.37 literals in average: this proportion was 1.88 in the original Hungarian WordNet xml file. Concerning the proportion of the two part-of-speech categories, nouns prevail over verbs with 210 nouns (70%), 90 verbs.

Presumably 40% of all lexicon entries are multiword expressions, regardless of word class or derivational processes. In many case when the Russian word refers to special posts or professional person, the proper Mansi word is a roundabout phrase. For example the учитель 'schoolteacher $masc.$' could be translated as нӣврамыт ханисьтан хум built up of the element *children-teaching man* , and the feminine counterpart учительница 'schoolteacher $fem.$' as нӣврамыт ханисьтан нӭ from *children-teaching woman*. Though the multi-word expressions are highly variable in their elements, replacing the dedicated parts with synonyms, or adding new ones to enrich the layers of senses. The number of multi-word expressions in this version of the Mansi WordNet is 74, that is 18% of all literals.

Section 3.2 enumerated some challenges about transforming an already existing WordNet to Mansi. Some synsets in the Basic Concept Set also have proved to be difficult to handle. For example, the Mansi language is only occasionally (if ever) used in scientific discourse. Therefore, the terms 'unconscious process', 'physiology' or 'geographical creature' cannot have any Mansi equivalents and therefore can be included in the Mansi WordNet only as non-lexicalized items. The number of such literals is 34, that is 16% of all literals.

5 Discussion

Building a wordnet for a minority or endangered language can have several challenges. Some of these are also relevant for dead languages, however, wordnets for e.g. Latin (Minozzi, 2009), Ancient Greek (Bizzoni et al., 2014) and Sanskrit (Kulkarni et al., 2010) prove that these facts do not necessarily mean an obstacle for wordnet construction. Here we summarize the most important challenges and how we solved them while constructing the Mansi wordnet.

5.1 Wordnet construction for minority and endangered languages

First, linguistic resources, e.g. mono- and multilingual dictionaries may be at our disposal only to a limited extent and second, there might be some areas of daily life where only the majority language is used, hence the minority language has only a limited vocabulary in that respect. As for the first challenge, we could rely on the Mansi-Russian-English-Hungarian dictionary under construction, which is itself based on Mansi-Russian dictionaries (see above) and we made use of its entries in the semi-automatic building process. However, if there are no such resources available, wordnets for minority languages should be constructed fully manually. For dead languages which are well-documented and have a lot of linguistic descriptions and dictionaries (like Latin and Ancient Greek), this is a less serious problem.

As for the second challenge, we applied two strategies: we introduced non-lexicalized synsets for those concepts that do not exist in the Mansi language or we included an appropriate loanword from Russian.

Besides being a minority language, Mansi is also an endangered language. Almost none of its native speakers have been trained in linguistics, which fact rules out the possibility of having native speakers as annotators. Thus, linguist experts specialized in the Mansi language have been employed as wordnet builders and in case of need, they can contact native speakers for further assistance. This problem is also relevant for dead languages, where there are no native speakers at all, however, we believe that linguists with advanced knowledge of the given language can also fully contribute to wordnet building.

5.2 Specialties of wordnet construction for Mansi

Wordnet building for Mansi also led to some theoretical innovations. As there is a subvocabulary of the Mansi language related to bears (see above), we intended to reflect this distinction in the wordnet too. For that reason, we introduced the novel relation "bear", which connect synsets that are only used in connection with bears and synsets that in-

clude their "normal" equivalents. All this means that adding new languages to the spectrum may also have theoretical implications which contribute to the linguistic richness of wordnets.

6 Conclusions

In this paper, we reported the construction of a wordnet for Mansi, an endangered minority language spoken in Russia. As we intend to make the Mansi wordnet freely available for everyone, we hope that this newly created language resource will contribute to the revitalization of the Mansi language.

In the future, we would like to extend the Mansi wordnet with new synsets. Moreover, we intend to create applications that make use of this language resource, for instance, online dictionaries and linguistic games for learners of Mansi.

Acknowledgments

This work was supported in part by the Finnish Academy of Sciences and the Hungarian National Research Fund, within the framework of the project *Computational tools for the revitalization of endangered Finno-Ugric minority languages (FinUgRevita)*. Project number: OTKA FNN 107883; AKA 267097.

References

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti, and Wim Peters. 1998. The Linguistic Design of the EuroWordNet Database. *Computers and the Humanities. Special Issue on EuroWordNet*, 32(2-3):91–115.

A.N. Balandin and M.I. Vahruševa. 1958. *Mansijski-russkij slovar' s leksičeskimi paralelljami iz južno-mansijskogo (kondinskogo) dialekta*. Prosvešenije, Leningrad.

Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors. 2010. *Principles, Construction and Application of Multilingual Wordnets. Proceedings of GWC 2010*. Narosa Publishing House, Mumbai, India.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of ancient greek wordnet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1054.

Christiane Fellbaum and Piek Vossen, editors. 2012. *Proceedings of GWC 2012*. Matsue, Japan.

M. Kulkarni, C. Dangarikar, I. Kulkarni, A. Nanda, and P. Bhattacharya. 2010. Introducing Sanskrit WordNet. In *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, Mumbai, India. Narosa Publishing House.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 311–320, Szeged. University of Szeged.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Stefano Minozzi. 2009. The Latin WordNet project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*, pages 707–716.

Heili Orav, Christiane Fellbaum, and Piek Vossen, editors. 2014. *Proceedings of GWC 2014*. Tartu, Estonia.

E.I. Rombandeeva and E.A. Kuzakova. 1982. *Slovar' mansijsko-russkij i russko-mansijskij*. Prosvešenije, Leningrad.

E.I. Rombandeeva. 2005. *Russko-mansijskij slovar'*. Mirall, Sankt-Peterburg.

Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors. 2008. *Proceedings of GWC 2008*. University of Szeged, Department of Informatics, Szeged, Hungary.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, 7(1-2):9–43.

Veronika Vincze, Ágoston Nagy, Csilla Horváth, Norbert Szilágyi, István Kozmács, Edit Bogár, and Anna Fenyvesi. 2015. FinUgRevita: Developing Language Technology Tools for Udmurt and Mansi. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages*, Tromsø, Norway, January.

WNSpell: a WordNet-Based Spell Corrector

Bill Huang
Princeton University
yh3@princeton.edu

Abstract

This paper presents a standalone spell corrector, WNSpell, based on and written for WordNet. It is aimed at generating the best possible suggestion for a mistyped query but can also serve as an all-purpose spell corrector. The spell corrector consists of a standard initial correction system, which evaluates word entries using a multifaceted approach to achieve the best results, and a semantic recognition system, wherein given a related word input, the system will adjust the spelling suggestions accordingly. Both feature significant performance improvements over current context-free spell correctors.

1 Introduction

WordNet is a lexical database of English words and serves as the premier tool for word sense disambiguation. It stores around 160,000 word forms, or lemmas, and 120,000 word senses, or synsets, in a large graph of semantic relations. The goal of this paper is to introduce a spell corrector for the WordNet interface, directed at correcting queries and aiming to take advantage of WordNet's structure.

1.1 Previous Work

Work on spell checkers, suggesters, and correctors began in the late 1950s and has developed into a multifaceted field. First aimed at simply detecting spelling errors, the task of spelling correction has grown exponentially in complexity.

The first attempts at spelling correction utilized edit distance, such as the Levenshtein distance, where the word with minimal distance would be chosen as the correct candidate.

Soon, probabilistic techniques using noisy channel models and Bayesian properties were invented. These models were more sophisticated, as they also considered the statistical likeliness of certain errors and the frequency of the candidate word in literature.

Two other major techniques were also being developed. One was similarity keys, which used properties such as the word's phonetic sound or first few letters to vastly decrease the size of the dictionary to be considered. The other was the rule-based approach, which implements a set of human-generated common misspelling rules to efficiently generate a set of plausible corrections and then matching these candidates with a dictionary.

With the advent of the Internet and the subsequent increase in data availability, spell correction has been further improved. N-grams can be used to integrate grammatical and contextual validity into the spell correction process, which standalone spell correction is not able to achieve. Machine learning techniques, such as neural nets, using massive online crowdsourcing or gigantic corpora, are being harnessed to refine spell correction more than could be done manually.

Nevertheless, spell correction still faces significant challenges, though most lie in understanding context. Spell correction in other languages is also incomplete, as despite significant work in English lexicography, relatively little has been done in other languages.

1.2 This Project

Spell correctors are used everywhere from simple spell checking in a word document to query completion/correction in Google to context-based in-passage corrections. This spell corrector, as it is for the WordNet interface, will focus on spell correction on a single word query with the additional possibility of a user-inputted semantically-related word from which to base corrections off of.

2 Correction System

The first part of the spell corrector is a standard context-free spell corrector. It takes in a query such as $speling$ and will return an ordered list of three possible candidates; in this case, it returns the set $\{spelling, spoiling, sapling\}$.

The spell corrector operates similarly to the Aspell and Hunspell spell correctors (the latter which serves as the spell checker for many applications varying from Chrome and Firefox to OpenOffice and LibreOffice). The spell corrector we introduce here, though not as versatile in terms of support for different platforms, achieves far better performance.

To tune the spell corrector to WordNet queries, stress is placed on bad misspellings over small errors. We will mainly use the Aspell data set (547 errors), kindly made public by the GNU Aspell project, to test the performance of the spell corrector. Though the mechanisms of the spell corrector are inspired by logic and research, they are included and adjusted mainly based on empirical tests on the above data set.

2.1 Generating the Search Space

To improve performance, the spell corrector needs to implement a fine-tuned scoring system for each candidate word. Clearly, scoring each word in WordNet's dictionary of 150,000 words is not practical in terms of runtime, so the first step to an accurate spell corrector is always to reduce the search space of correction candidates.

The search space should contain all possible reasonable sources of the the spelling error. These errors in spelling arise from three separate stages (Deorowicz and Ciura, 2005):

1. Idea → thought word
   i.e. $distrucally \rightarrow destructfully$
2. Thought word → spelled word
   i.e. $egsistance \rightarrow existence$
3. Spelled word → typed word
   i.e. $autocorrecy \rightarrow autocorrect$

The main challenges regarding search space generation are:

1. Containment of all, or nearly all, possible reasonable corrections
2. Reasonable size
3. Reasonable runtime

There have been several approaches to this search space problem, but all have significant drawbacks in one of the criteria of search space generation:

- The simplest approach is the lexicographic approach, which simply generates a search space of words within a certain edit distance away from the query. Though simple, this minimum edit distance technique, introduced by Damerau in 1964 and Levenshtein in 1966, only accounts for type 3 (and possibly type 2) misspellings. The approach is reasonable for misspellings of up to edit distance 2, as Norvig's implementation of this runs in $\sim$0.1 seconds, but time complexity increases exponentially and for misspellings such as $funetik \rightarrow phonetic$ that are a significant edit distance away, this approach will not be able to contain the correction without sacrificing both the size of the search space and the runtime.

- Another approach is using phonetics, as misspelled words will most likely still have similar phonetic sounds. This accounts for type 2 misspellings, though not necessarily type 1 or type 3 misspellings. Implementations of this approach, such as using the SOUNDEX code (Odell and Russell, 1918), are able to efficiently capture misspellings such as $funetik \rightarrow phonetic$, but not misspellings like $rypo \rightarrow typo$. Again, this approach is not sufficient in containing all plausible corrections.

- A similarity key can also be used. The similarity key approach stores each word under a key, along with other similar words. One implementation of this is the SPEED-COP spell corrector (Pollock and Zamora, 1984), which takes advantage of the usual alphabetic proximity of misspellings to the correct word. This approach accounts for many errors, but there are always a large number of exceptions, as the misspellings do not always have similar keys (such as the misspelling $zlphabet \rightarrow alphabet$).

- Finally, the rule-based approach uses a set of common misspelling patterns, such as $im \rightarrow in$ or $y \rightarrow t$, to generate possible sources of the typing error. The most complicated

approach, these spell correctors are able to contain the plausible corrections for most spelling errors quite well, but will miss many of the bad misspellings. The implementation by Deoroicz and Ciura using this approach is quite effective, though it can be improved.

Our approach with this spell corrector is to use a combination of these approaches to achieve the best results. Each approach has its strengths and weaknesses, but cannot achieve a good coverage of the plausible corrections without sacrificing size and runtime. Instead, we take the best of each approach to much better contain the plausible corrections of the query.

To do this, we partition the set of plausible corrections into groups (not necessarily disjoint, but with a very complete union) and consider each separately:

- Close mistypings/misspellings:

  This group includes typos of edit distance 1 ($typo \rightarrow rypo$) and misspellings of edit distance 1 ($consonent \rightarrow consonant$), as well as repetition of letters ($mispel \rightarrow misspell$). These are easy to generate, running in $O(n \log n \alpha)$ time, where n is the length of the entry and α is the size of the alphabet, to generate and check each word (though increasing the maximum distance to 2 would result an significantly slower time of $O(n^2 \log n \alpha^2)$.

- Words with similar phonetic key:

  We implement a precalculated phonetic key for each word in WordNet, which uses a numerical representation of the first five consonant sounds of the word:

  0: (ignored) a, e, i, o, u, h, w, [gh](t)
  1: b, p
  2: k, c, g, j, q, x
  3: s, z, c(i/e/y), [ps], t(i o), (x)
  4: d, t
  5: m, n, [pn], [kn]
  6: l
  7: r
  8: f, v, (r/n/t o u)[gh], [ph]

  Each word in WordNet is then stored in an array with indices ranging from $[00000]$ (no consonants) to $[88888]$ and can be looked up quickly.

  This group includes words with a phonetic key that differs by an edit distance at most 1 from the phonetic key of the entry ($funetik \rightarrow phonetic$), and also does a very good job of including typos/misspellings of edit distance greater than 1 (it actually includes the first group completely, but for pruning purposes, the first group is considered separately) in very little time $O(Cn)$ where $C \sim 5^2 \times 9$.

- Exceptions:

  This group includes words that are not covered by either of the first two groups but are still plausible corrections, such as $lignuitic \rightarrow linguistic$. We observe that most of these exceptions either still have similar beginning and endings to the original word and are close edit distance-wise or are simply too far-removed from the entry to be plausible. Searching through words with similar beginnings that also have similar endings (through an alphabetically-sorted list) proves to be very effective in including the exception, while taking very little time.

As many generated words, especially from the later groups, are clearly not plausible corrections, candidate words of each type are then pruned with different constraints depending on which group they are from. Words in later groups are subject to tougher pruning, and the finding of a close match results in overall tougher pruning.

For instance, many words in the second group are quite far removed from the entry and completely implausible as corrections (e.g. $zjpn \rightarrow [00325] \rightarrow [03235] \rightarrow suggestion$), while those that are simply caused by repetition of letters (e.g. $lllooolllll \rightarrow loll$) are almost always plausible, so the former group should be more strictly pruned.

Finally, since the generated search space after group pruning can be quite large (up to 200), depending on the size of the search space, the search space may be pruned, repetitively, until the size of the search space is of an acceptable size.

Some factors considered during pruning include:

- Length of word
- Letters contained in word
- Phonetic key of word

- First and last letter agreement
- Number of syllables
- Frequency of word in text (COCA corpus)
- Edit distance

This process successfully generates a search space that rarely misses the desired correction, while keeping both a small size in number of words and a fast runtime.

2.2 Evaluating Possibilities

The next step is to assign a similarity score to all of the candidates in the search space. It must be accurate enough to discern that $disurn \longrightarrow discern$ but $disurn \not\longrightarrow disown$ and versatile enough to figure out that $funetik \longrightarrow phonetic$.

Our approach is a modified version of Church and Gale's probabilistic scoring of spelling errors. In this approach, each candidate correction c is scored following the Bayesian combination rule:

$$P(c) = p(c) \max\Big(\prod_i p(t_i \mid c_i)\Big)$$

$$C(c) = c(c) + \min\Big(\sum_i c(t_i \mid c_i)\Big)$$

Where $P(c)$ is the frequency of the candidate correction, $P(t_i \mid c_i)$ the cost of each edit distance operation in a sequence of edit operations that generate the correction. The cost is then scored logarithmically based on the probability, where $c(t_i \mid c_i) \propto -\log\big(p(t_i \mid c_i)\big)$. The correction candidates are then sorted, with lower cost meaning higher likelihood.

We use bigram error counts generated from a corpora (Jones and Mewhort, 2004) to determine the values of $c(t \mid p)$. Two sets of counts were used:

- Error counts:
  - Deletion of letter β after letter α
  - Addition of letter β after letter α
  - Substitution of letter β for letter α
  - Adjacent transposition of the bigram $\alpha\beta$
- Bigram/monogram counts (log scale):
  - Monograms α
  - Bigrams $\alpha\beta$

First, we smooth all the counts using add-k smoothing (where we set $k = \frac{1}{2}$), as there are numerous counts of 0. Since the bigram/monogram counts were retrieved in log format, for sake of simplicity of data manipulation, we only smooth the counts of 0, changing their values to -0.69 (originally undefined). We then calculate $c(t_i \mid c_i)$ as:

$$c(t_i \mid c_i) = k_1 \log\left(\frac{1}{p(\alpha \to \beta)}\right) + k_2$$

Where $p(\alpha \to \beta)$ is the probability of the edit operation and k_1, k_2 factors that adjust the cost depending on the uncertainty of small counts and the increased likelihood of errors if errors are already present.

For the different edit operations, $p(x \to y)$ is:

$$p(x \to y) = \begin{cases} \text{deletion :} & \frac{\text{del}'(xy)}{N'(xy)} \\ \text{addition :} & \frac{\text{add}'(xy) \cdot N}{N'(x) \cdot N'(y)} \\ \text{substitution :} & \frac{\text{sub}'(xy) \cdot N}{N'(x) \cdot N'(y)} \\ \text{reversal :} & \frac{\text{rev}'(xy)}{N'(xy)} \end{cases}$$

And for deletion and addition of letters at the beginning of a word:

$$p(x \to y) = \begin{cases} \text{deletion :} & \frac{\text{del}'(.y)}{N'(.y)} \\ \text{addition :} & \frac{(\text{add}'(.y)) \cdot N \cdot w}{N'(y)} \end{cases}$$

To evaluate the minimum cost $\min\big(\sum_i c(t_i \mid c_i)\big)$ of a correction, we use a modified Wagner-Fischer algorithm, finds the minimum in $O(mn)$ time, where m, n are the lengths of the entry and correction candidate, respectively. This is done over for candidate corrections in the search space generated in (3.1).

Now, the probabilistic scoring by itself is not always accurate, especially in cases such as $funetik \longrightarrow phonetic$. Thus, we modify the scoring of each candidate correction to significantly improve the accuracy of the suggestions:

- Instead of setting $c(c) = -\log(p(c))$, we find that using $c(c)$ as multiplicative constant as a function $f(c)^\gamma$, where $f(c)$ is the frequency of the word in the corpus and γ an empirically-determined constant, yields significantly more accurate predictions.
- We add empirically-determined multiplicative factors λ_i pertaining to the following factors regarding the entry and the candidate correction:

- Same phonetic key (not restricted to first 5 consonant sounds)
- Same aside from repetition of letters
- Same consonants (ordered)
- Same vowels (ordered)
- Same set of letters
- Similar set of letters
- Same number of syllables
- Same after removal of es

(Note that other factors were considered but the factors pertaining to them were insignificant)

The candidate corrections are then ordered by their modified costs $C'(c) = C(c) \prod_i \lambda_i$ and the top three results, in order, are returned to the user.

3 Semantic Input:

The second part of the spell corrector adds a semantic aspect into the correction of the search query. When users have trouble entering the query and cannot immediately choose a suggested correction, they are given the option to enter a semantically related word. WNSpell then takes this word into account when generating suggestions, harnessing WordNet's vast semantic network to further optimize results.

This added dimension in spell correction is very helpful for the more severe errors, which usually arise from the "idea → thought word" process in spelling. These are much harder to deal with than conventional mistypings or misspellings, and are exactly the type of error WNSpell needs to be able to handle (as mistyped or even misspelled queries can be fixed without too much trouble by the user). The semantic anchor the related word provides helps WNSpell establish the idea" behind the desired word and thus refine the suggestions for the desired word.

To incorporate the related word into the suggestion generation, we add some modifications to the original context-free spell corrector.

3.1 Adjusting the Search Space:

One of the issues in search space generation in the original is that a small fraction of plausible corrections are still missed, especially in more severe errors. To improve the coverage of the search space, we modify the search space to also include a nucleus of plausible corrections generated semantically, not just lexicographically. Since the missed corrections are lexicographically difficult to generate, using a semantic approach would be more effective in increasing coverage.

The additional group of word forms is generated as follows:

1. For each synset of the related word, we consider all synsets related to it by some semantic pointer in WordNet.
2. All lemmas (word forms) of these synsets are evaluated.
3. Lemmas that share the same first letter or the same last letter and are not too far away in length are added to the group.

The inclusion of the additional group is indeed very effective in capturing the missed corrections and remains relatively small in size.

Some examples of missed words captured in this group from the training set are (entry, correct, related):

- *autoamlly, automatically, mechanically*
- *conibation, contribution, donation*

3.2 Adjusting the Evaluation:

We also modify the scoring process of each candidate correction to take into account semantic distance. First, each candidate correction is assigned a semantic distance d (higher means more similar) based on Lesk distance:

$$d = \max_i \max_j s(r_i, c_j)$$

Which takes the maximum similarity over all pairs of definitions of the related word r and candidate c where similarity s is measured by:

$$s(r_i, c_j) = \sum_{w \in R_i \cap C_j, w \notin S} k - \ln(n_w + 1)$$

Which considers words w in the intersection of the definitions that are not stopwords and weights them by the smoothed frequency n_w of w in the COCA corpus (as rarity is related to information content) and some appropriate constant k.

Additionally, if r or c is found in the other definition, we also add to the similarity s of two definitions $a\big(k - \ln(n_{r/c} + 1)\big)$ for some appropriate constant $a > 1$. This resolves many issues that come up with hypernyms/hyponyms (among others) where two similar words are assigned a low

score since the only words in common in their definitions may be the words themselves.

We also consider the number n of shared subsequences of length 3 between r and c, which is very helpful in ruling out semantically similar but lexicographically unlikely words.

We then adjust the cost function C' by:

$$C'' = \frac{C'}{(d+1)^{\alpha}(n+1)^{\beta}}$$

For some empirically-determined constants α and β. The new costs are then sorted and the top three results returned to the user.

4 Results

We used the Aspell data set to train the system. The test set consists of 547 hard-to-correct words. This is ideal for our purposes, as we are focusing on correcting bad misspellings as well as the easy ones. Most of the empirically-derived constants from (3.2) were determined based off of results from this data set.

4.1 Without Semantic Input

We compare the results of WNSpell to a few popular spellcheckers: Aspell, Hunspell, Ispell, and Word; as well as with the proposition of Deorowicz and Ciura, which seems to have the best results on the Aspell test set so far (other approaches are based off of unavailable/uncompatible data sets).

Ideally, for comparison, it would be ideal to run each spell checker on the same lexicon and on the same computer for consistent results. However, due to technical constraints, it is rather infeasible to do so. Instead, we will use the results posted by the authors of the spell checkers, which, despite some uncertainty, will still yield consistent and comparable results.

First, we compare our generated search space with the lists returned by Aspell, Hunspell, Ispell, and Word (Atkinson). We use a subset of the Aspell test set containing all entries whose corrections are in all five dictionaries. The results are shown in Table 1.

Search Space Results

Method	% found	Size (0/50/100%)		
WNSpell	97.4	1	10	66
Aspell (0.60.6n)	90.1	2	12	100
Hunspell (1.1.12)	83.2	1	4	15
Ispell (3.1.20)	54.8	0	1	29
Word 97	75.4	0	2	20

Table 1

Compared to these three spell correctors, WNSpell clearly does a significantly better job containing the desired correction than Aspell, Hunspell, Ispell, or Word within a set of words of acceptable size.

We now compare the results of the top three words returned on the list with those returned by Aspell, Hunspell, Ispell, Word. We also include data from Deorowicz and Ciura, which also uses the Aspell test set. Since the dictionaries used were different, we also include Aspell results using their subset of the Aspell test set. The results are shown in Table 2, and a graphical comparison is shown in Figure 1.

Once again, WNSpell significantly outperforms the other five spell correctors.

Aspell Test Set Results (% Identified)

Method	Top 1	Top 2	Top 3	Top 10
WNSpell	77.5	88.5	91.2	96.1
Aspell (0.60.6n)	54.3	63.0	72.9	87.1
Hunspell (1.1.12)	58.2	71.5	76.6	82.3
Ispell (3.1.20)	40.1	47.9	50.4	54.1
Word 97	62.6	69.4	72.7	75.4
Aspell (n)	56.9	66.9	74.7	87.9
DC	66.3	75.5	79.6	85.5

Table 2

Aspell Test Set Results (% Identified)

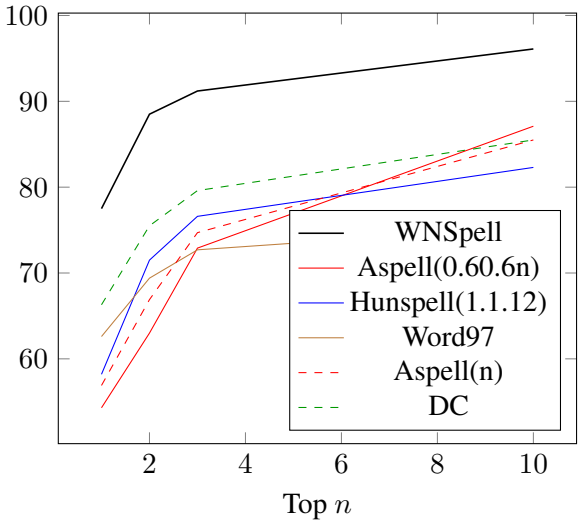


Figure 1

We also test WNSpell on the Aspell common misspellings test set, a list of 4206 common misspellings and their corrections. Since the word corrector was not trained on this set, it is a blind comparison. Once again, we use a subset of the Aspell test set containing all entries whose corrections are in all five dictionaries. The results are

shown in tables 3 and 4, and a graphical comparison is shown in Figure 2.

Blind Search Space Results

Method	% found	Size (0/50/100%)		
WNSpell	98.4	1	4	50
Aspell (0.60.6n)	97.7	1	9	100
Hunspell (1.1.12)	97.3	1	5	15
Ispell (3.1.20)	85.2	0	1	26

Table 3

Blind Test Set Results

Method	Top 1	Top 2	Top 3	Top 10
WNSpell	91.4	96.3	97.6	98.3
Aspell (0.60.6n)	73.6	81.2	92.0	97.0
Hunspell (1.1.12)	80.8	92.0	95.0	97.3
Ispell (3.1.20)	77.4	82.7	84.3	85.2

Table 4

Blind Test Set Results (% Identified)

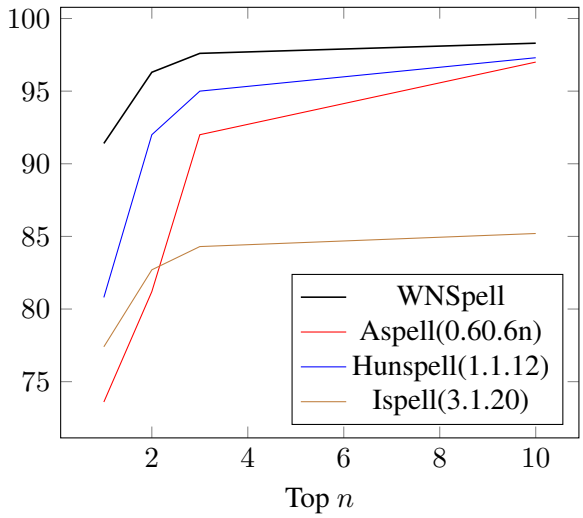


Figure 2

Additionally, WNSpell runs in decently fast time. WNSpell takes ∼13ms per word, while Aspell takes ∼3ms, Hunspell ∼50ms, and Ispell ∼0.3ms. Thus, WNSpell is a very efficient standalone spell corrector, achieving superior performance within acceptable runtime.

4.2 With Semantic Input

We test WNSpell with the semantic component on the original training set, this time with added synonyms. For each word in the training set, a human-generated related word is inputted.

With the addition of the semantic adjustments, WNSpell performs considerably better than without them. The results are shown in Table 5 and a graphical comparison in Figure 3:

Semantic Results (% Identified)

Method	Top 1	Top 2	Top 3	Top 10
with	87.4	93.0	96.5	99.1
without	77.5	88.5	91.2	96.1

Table 5

Semantic Results (% Identified)

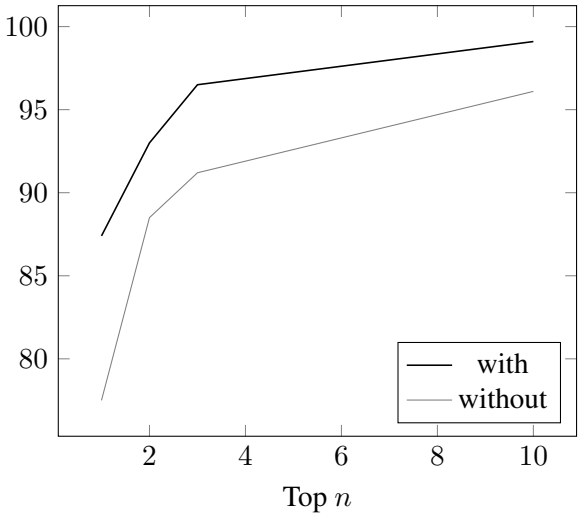


Figure 3

The runtime for WNSpell with semantic input, however, is rather slow at an average of ∼200ms.

5 Conclusions:

The WNSpell algorithm introduced in this paper presents a significant improvement in accuracy in correcting standalong spelling corrections over other systems, including the most recent version of Aspell and other commercially used spell correctors such as Huspell and Word, by approximately 20%. WNSpell is able to take into a variety of factors regarding different types of spelling errors and using a carefully tuned algorithm to account for much of the diversity in spelling errors presented in the test data sets. There is a efficient sample space pruning system that restricts the number of words to be considered, strongly improved by a phonetic key, and an accurate scoring system that then compares the words. The accuracy of WNSpell in correcting hard-to-correct words is quite close that of most peoples' abilities and significantly stronger than other methods.

WNSpell also provides an alternative using a related word to help the system find the desired correction even if the user is far off the mark in terms of spelling or phonetics. This added feature once again significantly increases the accuracy of

WNSpell by approximately 10% by directly connecting the idea word the user has in mind to the word itself. This link allows for the possibility of users who only know what rough meaning their desired word has or context it is in to actually find the word.

5.1 Limitations:

The standalone algorithm currently does not take into consideration vowel phonetics, which are rather complex in the English language. For instance, the query $spoak$ would be corrected into $speak$ rather than $spoke$. While a person easily corrects $spoak$, WNSpell would not be able to use the fact that $spoke$ sounds the same while $speak$ does not. Rather, all three have consonant sounds s, p, k and have one different letter from $spoak$. But an evaluation of edit distance finds that $speak$ is clearly closer, so the algorithm chooses $speak$ instead.

WNSpell, a spell corrector targeting at single-word queries, also does not have the benefit of contextual clues most modern spell correctors use.

5.2 Future Improvements:

As mentioned earlier, introducing a vowel phonetic system into WNSpell would increase its accuracy. The semantic feature of WNSpell can be improved by either pruning down the algorithm to improve performance or possibly using/incorporating other closeness measures of words into the algorithm. One possible addition is the use of some distributional semantics, such as using pre-trained word vectors to search for similar words (such as Word2Vec).

Additionally, WNSpell-like spell correctors can be implemented in many languages rather easily, as WNSpell does not rely very heavily on the morphology of the language (though it requires some statistics of letter frequencies as well as simplified phonetics). The portability is quite useful as WordNet is implemented in over a hundred languages, so WNSpell can be ported to other non-English WordNets.

References

D. Jurafsky and J.H. Martin. 1999. *Speech and Language Processing*, Prentice Hall.

R. Mishra and N. Kaur. 2013. "A survey of Spelling Error Detection and Correction Techniques," *International Journal on Computer Trends and Technology*, Vol. 4, No. 3, 372-374.

K. Atkinson. "Spell Checker Test Kernel Results," http://aspell.net/test/.

S. Deorowicz and M.G. Ciura. 2005. "Correcting Spelling Errors by Modeling their Causes," *Int. J. Appl. Math. Comp. Sci.*, Vol. 15, No. 2, 275-285.

P. Norvig. "How to Write a Spell Corrector," http://norvig.com/spell-correct.html.

K.W. Church and W.A. Gale. 1991. "Probability Scoring for Spelling Correction," AT&T Bell Laboratories

M.N. Jones and J.K. Mewhort. 2004. "Case-Sensitive Letter and Bigram Frequency Counts from Large-Scale English Corpora," *Behavior Research Methods, Instruments, & Computers*, 36(3), 388-396.

Corpus of Contemporary American English. (n.d.). http://corpus.byu.edu/coca/.

Sophisticated Lexical Databases - Simplified Usage: Mobile Applications and Browser Plugins For Wordnets

Diptesh Kanojia
CFILT, CSE Department,
IIT Bombay,
Mumbai, India
diptesh@cse.iitb.ac.in

Raj Dabre
School of Informatics,
Kyoto University,
Kyoto, Japan
prajdabre@gmail.com

Pushpak Bhattacharyya
CFILT, CSE Department,
IIT Bombay,
Mumbai, India
pb@cse.iitb.ac.in

Abstract

India is a country with 22 officially recognized languages and 17 of these have WordNets, a crucial resource. Web browser based interfaces are available for these WordNets, but are not suited for mobile devices which deters people from effectively using this resource. We present our initial work on developing mobile applications and browser extensions to access WordNets for Indian Languages.

Our contribution is two fold: (1) We develop mobile applications for the Android, iOS and Windows Phone OS platforms for Hindi, Marathi and Sanskrit WordNets which allow users to search for words and obtain more information along with their translations in English and other Indian languages. (2) We also develop browser extensions for English, Hindi, Marathi, and Sanskrit WordNets, for both Mozilla Firefox, and Google Chrome. We believe that such applications can be quite helpful in a classroom scenario, where students would be able to access the WordNets as dictionaries as well as lexical knowledge bases. This can help in overcoming the language barrier along with furthering language understanding.

1 Introduction

India is among the topmost countries in the world with massive language diversity. According to a recent census in 2001, there are 1,365 rationalized mother tongues, 234 identifiable mother-tongues and 122 major languages[1] . Of these, 29 languages have more than a million native speakers, 60 have more than 100,000 and 122 have more than 10,000 native speakers. With this in mind, the construction of the Indian WordNets, the IndoWordNet (Bhattacharyya, 2010) project was initiated which was an effort undertaken by over 12 educational and research institutes headed by IIT Bombay. Indian WordNets were inspired by the pioneering work of Princeton WordNet(Fellbaum, 1998) and currently, there exist WordNets for 17 Indian languages with the smallest one having around 14,900 synsets and the largest one being Hindi with 39,034 synsets and 100,705 unique words. Each WordNet is accessible by web interfaces amongst which Hindi WordNet(Dipak et al., 2002), Marathi WordNet and Sanskrit WordNet(Kulkarni et al., 2010) were developed at IIT Bombay[2]. The WordNets are updated daily which are reflected on the websites the next day. We have developed mobile applications for the Hindi, Marathi and Sanskrit WordNets, which are the first of their kind to the best of our knowledge.

This paper is organized as follows: Section 2 gives the motivations for the work. Section 3 contains the descriptions of the application with screen-shots and the nitty gritties. We describe the browser extensions in Section 4, and we conclude the paper with conclusions, and future work in Section 5. At the very end, some screen-shots of the applications and browser extensions are provided.

[1] http://en.wikipedia.org/wiki/Languages_of_India
[2] http://www.cfilt.iitb.ac.in/

2 Motivation

According to recent statistics, about 117 million Indians[3], are connected to the Internet through mobile devices. It is common knowledge that websites like Facebook, Twitter, Linkedin, Gmail and so on can be accessed using their web browser based interfaces but the mobile applications developed for them are much more popular. This is a clear indicator that browser based interfaces are inconvenient which was the main motivation behind our work. We studied the existing interfaces and the WordNet databases and developed applications for Android, iOS and Windows Phone platforms, which we have extensively tested and plan to release them to the public as soon as possible.

Our applications and plugins are applicable in the following use cases:

1. Consider an educational classroom scenario, where students, often belonging to different cultural and linguistic background wish to learn languages. They would be able to access the WordNets as dictionaries for multiple Indian languages. This would help overcome the language barrier which often hinders communication, and thus, understandability. The cost effective and readily available "Aakash" tablet device[4] will be one of the means by which our application will be accessed by educational institutes over India.

2. Tourists traveling to India can use the WordNet mobile apps for basic survival communication, because Indian language WordNets contain a lot of culture and language specific concepts, meanings for which may not even be available on internet search.

3. People who read articles on the internet may come across words they do not understand and can benefit from our plugins which can help translate words and give detailed information about them at the click of a button.

[3]"Internet trends 2014 report" by Mary Meeker, Kleiner Perkins Caufield & Byers (KPCB)

[4]http://www.akashtablet.com/

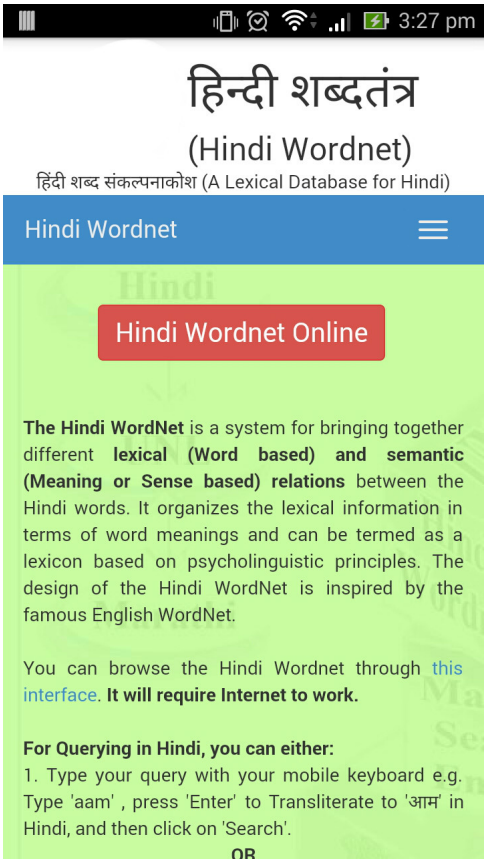


Figure 1: Home Screen

4. Linguists who happen to be experts at lexical knowledge can use the WordNet apps as well as plugins to acquire said knowledge irrespective of whether they have mobile phones or PCs.

Apart from the cases mentioned above, there are many other cases where our apps and plugins can be used effectively.

3 Mobile WordNet Applications

In the subsections below we describe the features of the applications accompanied by screen-shots.

3.1 Home Screen

When the user starts the application, the home screen (Figure: 1) is shown with a brief description of how to use it, the link which takes the user to search interface.

3.2 Search Interface

We have provided the user with two types of input mechanisms, Phonetic Translitera-

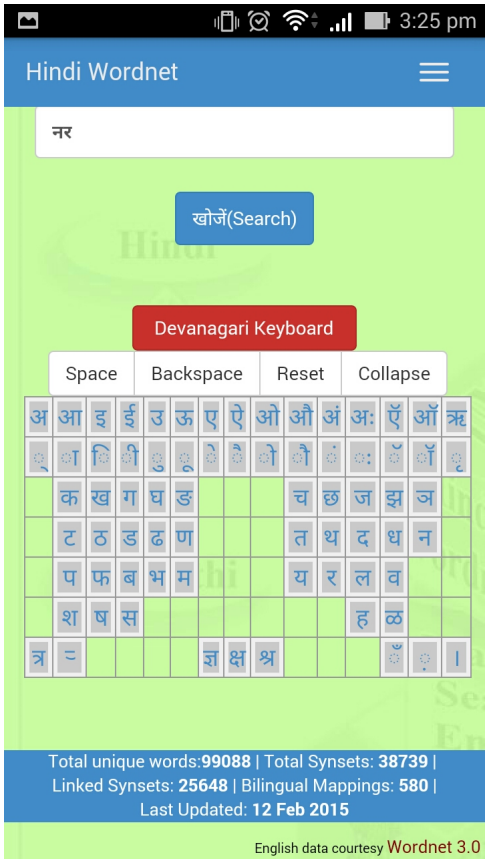


Figure 2: Devanagari Keyboard

tion using Google Transliteration API[5], and a JavaScript based online keyboard (Figure: 2) for input of Hindi Unicode characters. Transliteration for a native user is very convenient. In case, the user does not know the right combination of keys then the keyboard for Devanagari is provided. These two methods ensure that all words can be easily entered for searching. Thereafter, by touching / clicking on "Search", the synsets with all relevant information are retrieved.

3.3 Search Process

Indian languages are fairly new to the web, and despite standard UTF encoding of characters, there remain a few steps to be taken to sanitize the input for WordNet search. The steps taken by us are given below:

3.3.1 *Nukta* Normalization for Hindi

Hindi Characters such as क (ka), ख (kha), ग (ga), ज (ja), ड (ḍa), ढ (ḍha), फ (pha), झ (jha), take up *nukta* symbol to form क़ (qa), ख़ (kha),ग़ (ġa),ज़ (za),ड़ (da),ढ़ (ṛha),फ़ (fa),झ़ (zha), respectively. These characters occur twice in the Unicode chart, both with *nukta* as a separate unicode character, and adjoining the parent character. We normalize the input for standard unicode encoding with nukta as a separate character before search.

3.3.2 Morphological Analysis

Before searching in the databases the word is first passed to a morphological analyzer to obtain its root form. We use Hindi Morph Analyzer (Bahuguna et al., 2014) to return the root form of the input word for Hindi language, since by principle, WordNet only contains root forms of the words.

Due to non availability of other language Morphological Analyzers, we may not be able to include them in the search process. Though, in the future, we can use a fully automated version of the "Human mediated TRIE base generic stemmer"(Bhattacharyya et al., 2014) for obtaining root forms for other languages later.

3.3.3 Handling Multiple Root forms

Our interface also requests the user to select the preferred root, if more than one root forms are returned post morphological analysis. The user can then just select one and then the synset retrieval process is initiated on the server. It gives the user more control, and choice over results. We assume that while searching the WordNet, a user may not be familiar with all the senses of the words, or all the morphology of the word. It may be possible that the user came across the word over the internet, and is using our plugin to search the WordNet. This feature enables the user to select the appropriate root, or check all the possibilities for the correct answer.

3.4 Application Design

We have used the WebView class, and URL loading from the Android SDK[6], and Windows Phone SDK[7] to display a responsive layout of the WordNets. WebView renders the application pages seamlessly onto the mobile / handheld devices, thus making the application usable for mobile, tablet, and other handheld

[5]https://developers.google.com/transliterate/

[6]https://developer.android.com/

[7]https://dev.windows.com/en-us/develop/download-phone-sdk

device of any size.

Similarly, for iOS, we have used the UIWebView class with some scaling measures to render the pages with a responsive layout onto the device screen. Our application is compatible with all iOS devices. It will be deployed to Apple App Store soon.

A preliminary check on internet connection is done before connecting to the web interface, and retry button is provided on the front, in case an internet connection is not detected.

3.5 Search Results

The results returned by the server are interpreted by the application pages and displayed in a very simplistic manner. We display all synsets for each part of speech and all senses of that word and initially showing the synset words, gloss and example.These senses are categorized by their part of speech categories. We have conformed to the principles of good User Interface design and provided for an incremental information display.

3.5.1 Additional Information

If the user wishes to see the synset relations and the translations of that word in other synsets the link "Relations and Languages" should be clicked to give a list of all additional information that can be displayed. Relations like Hypernymy and Hyponomy and the relevant synset in the other 16 languages can be displayed. Please refer to figure 3 for an example.

3.5.2 Current Drawbacks

Current version of Android OS (Lollipop 5.0) deployed on most of the smartphones, does not support rendering of Gujarati, Punjabi, and Nepali languages, on all devices. The language support also depends on the device manufacturer. Hence, they are currently disabled from the interface.

Also, Our applications are currently online, and can only be used if the user is connected to the internet. We plan to implement an offline version of our applications.

4 Browser Extensions

Major WordNets of the world are available via web interfaces, enabling a user to search for the senses using a web browser on a computer or mobile. The process commonly involves a user navigating to a web page, and searching the required 'word' for its senses. In a world where getting things done in one click is important, we feel that the process of searching needs to be simplified. We develop browser extensions to ease this process. Google Chrome and Mozilla Firefox are the most popular web browsers among the web users[8]. Our approach makes the search quite simple and is summarized in the following 3 steps:

- User highlights the word of interest and right-clicks the page or clicks on the plugin shortcut.
- They click the context menu option for 'Search <relevant> WordNet for . . .'
- A new tab opens up showing the information from the relevant WordNet.

We present the sample context menu screenshots, post installation in Figures 6 and 7, respectively.

5 Conclusions and Future Work

In this era of handheld mobile devices, there is a great need to make available traditional web services as mobile applications which are extremely popular. Our success in developing mobile applications for Hindi, Marathi and Sanskrit WordNets along with browser plugins for English, Hindi, Marathi and Sanskrit to simplify word look-up is the first step in providing people with easy access to such important knowledge bases. We have described a variety of use cases for our apps and plugins which are quite realistic, especially in India where language and cultural diversity is quite varied. These can have a huge impact on language education, especially in the rural areas, along with enabling people to understand a multitude of languages.

We plan to make available offline search in our apps. Also, we plan to make efforts towards improving this application to enable searching for words belonging to all languages which have a common interface via language detection. We also plan to inculcate Word Suggestions as they are being typed so that the

[8] http://gs.statcounter.com/#all-browser-ww-monthly-201506-201506-bar

user is presented with better lexical choices. Plugins like PeraPera[9] for Japanese and Chinese are quite popular since they simply provide lexical information when the user hovers the mouse over words. Implementing such a feature is something we plan to do in the immediate future. Also, We would publish our application, and browser plugin source codes publicly for research purposes.

6 Acknowledgment

We gratefully acknowledge the support of the Department of Electronics and Information Technology, Ministry of Communications and IT, Government of India. We also thank Ravi Nambudripad, for replicating the application in other languages and the entire computational linguistics group at Centre For Indian Language Technology, IIT Bombay, which has provided its valuable input and critique, helping us refine our work.

References

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and Pushpak Bhattacharyya. 2002. *An Experience in Building the Indo WordNet - a WordNet for Hindi.* In *First International Conference on Global WordNet*, (GWC 2002), Mysore, India.

Christiane D. Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* Published by *Bradford Books*

Pushpak Bhattacharyya. 2010. *IndoWordNet.* In *Proceedings of Lexical Resources Engineering Conference*, May, 2010, Malta.

Ankit Bahuguna, Lavita Talukdar, Pushpak Bhattacharyya and Smriti Singh. 2014. *HinMA: Distributed Morphology based Hindi Morphological Analyzer.* In *Proceedings of the 11th International Conference on Natural Language Processing* (ICON 2014), December, 2014.

Pushpak Bhattacharyya, Ankit Bahuguna, Lavita Talukdar, and Bornali Phukon. 2014. *Facilitated Multi-Lingual Sense Annotation: Human Mediated Lemmatizer.* In *Proceedings of the Global Wordnet Conference 2014* (GWC 2014), January, 2014.

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. *Introducing Sanskrit Wordnet.* In *Proceedings of the Global Wordnet Conference 2010* (GWC 2010), January, 2010.

[9]http://www.perapera.org/

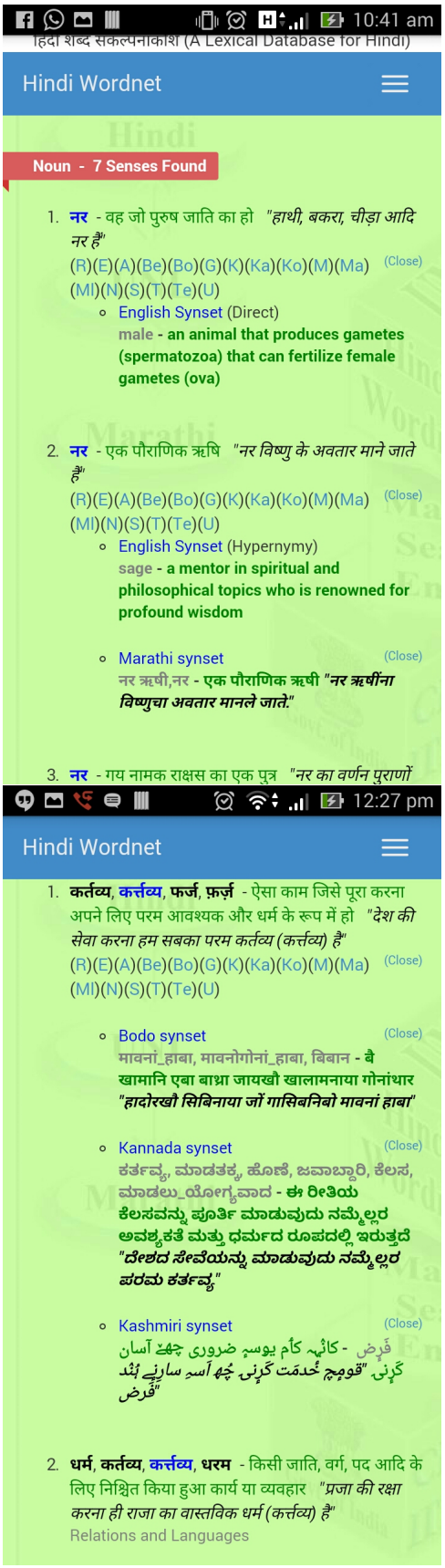


Figure 3: Screen-shots of Search Results

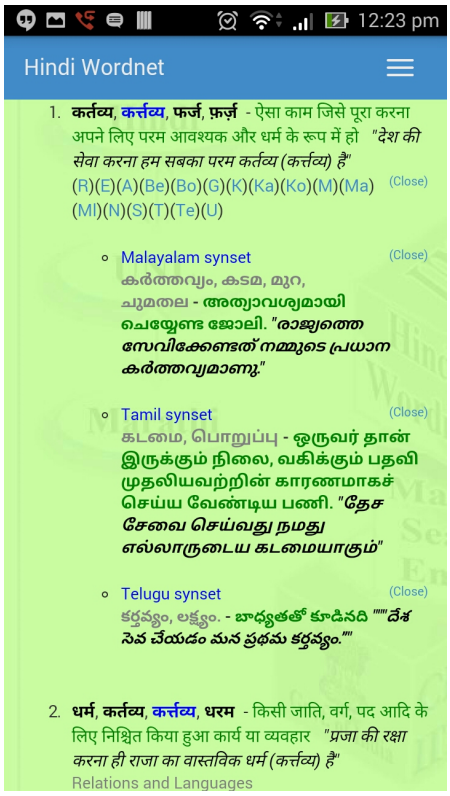


Figure 4: Search Results with Malayalam, Tamil, and Telugu Synsets

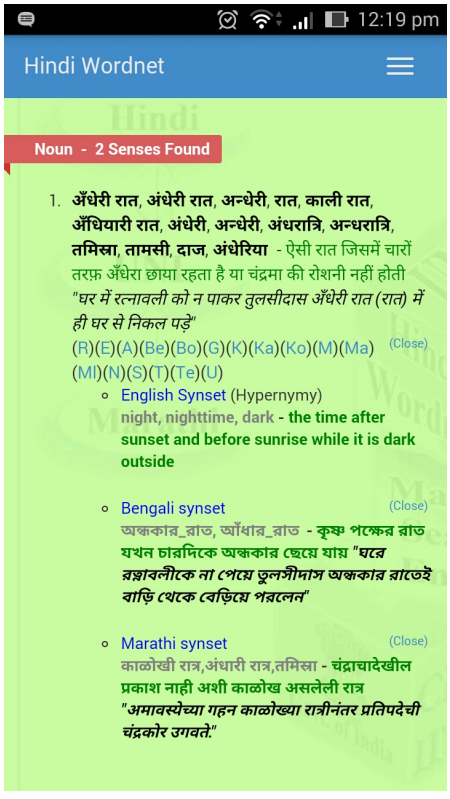


Figure 5: Search Results with English, Bengali, and Marathi Synsets

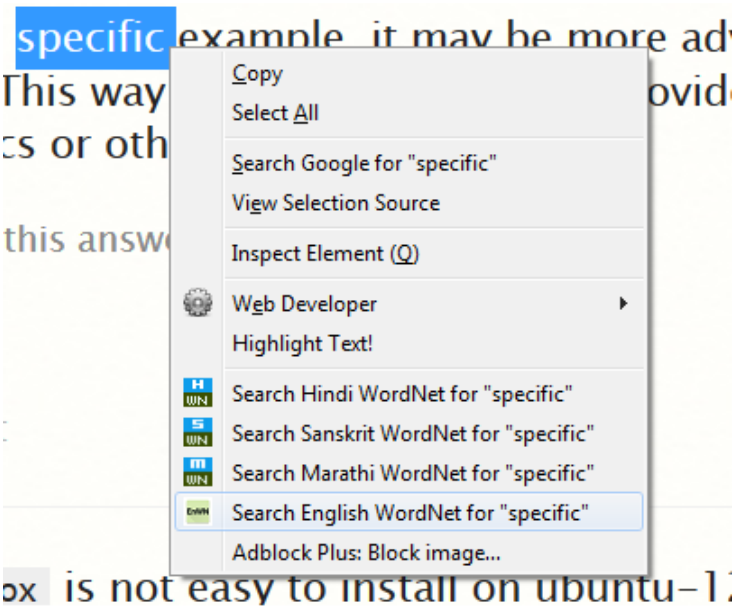


Figure 6: Browser Extensions Context Menu for word 'specific'

Figure 7: Browser Extensions Context Menu for word 'हिंसा' (hiMsaa) translated to 'violence'

A picture is worth a thousand words: Using OpenClipArt library to enrich IndoWordNet

Diptesh Kanojia, Shehzaad Dhuliawala, and Pushpak Bhattacharyya
Centre for Indian Language Technology,
Computer Science and Engineering Department,
IIT Bombay,
Mumbai, India
{diptesh,shehzaadzd,pb}@cse.iitb.ac.in

Abstract

WordNet has proved to be immensely useful for Word Sense Disambiguation, and thence Machine translation, Information Retrieval and Question Answering. It can also be used as a dictionary for educational purposes. The semantic nature of concepts in a WordNet motivates one to try to express this meaning in a more visual way. In this paper, we describe our work of enriching IndoWordNet with image acquisitions from the OpenClipArt library. We describe an approach used to enrich WordNets for eighteen Indian languages.

Our contribution is three fold: **(1)** We develop a system, which, given a synset in English, finds an appropriate image for the synset. The system uses the OpenclipArt library (OCAL) to retrieve images and ranks them. **(2)** After retrieving the images, we map the results along with the linkages between Princeton WordNet and Hindi WordNet, to link several synsets to corresponding images. We choose and sort top three images based on our ranking heuristic per synset. **(3)** We develop a tool that allows a lexicographer to manually evaluate these images. The top images are shown to a lexicographer by the evaluation tool for the task of choosing the best image representation. The lexicographer also selects the number of relevant images. Using our system, we obtain an Average Precision (P @ 3) score of 0.30.

1 Introduction

Our goal is to enrich the semantic lexicon of various Indian languages by mapping it with images from the OpenClipArt library(Phillips, 2005). India is currently experiencing a major enhancement in the digital education sector with its vision of the 'Digital India' program[1]. In this paper, we introduce an approach to enrich the IndoWordNet[2](Bhattacharyya, 2010), with images, which can help students and language enthusiasts alike. We envision the use of WordNets in the education sector to promote language research among young students, and provide them with a multilingual resource which eases their study of languages. WordNets have proven to be a rich lexical resource for many NLP sub tasks such as Machine Translation (MT) and Cross Lingual Information retrieval.

India has 22 official languages, written in more than 8 scripts. When a user reads a concept in a language that is not known to them, and moreover in an unknown script, an image can provide helpful insight into the concept. Language learners in a multilingual country like this often face difficulty mainly due to: **(a)** Not being able to find a mapping of the concept in the language being studied and their native language and **(b)** Not being able to decipher the script in the language being learnt. In such cases a pictorial representation of a concept will be very useful.

Finally, systems for Automatic image captioning and Real time video summarization can leverge the power of image enriched WordNets.

[1]http://www.digitalindia.gov.in/
[2]http://www.cfilt.iitb.ac.in/indowordnet

1.1 WordNets and IndoWordNet

WordNets are lexical structures composed of synsets and semantic relations(Fellbaum, 1998). Such a lexical knowledge base is at the heart of an intelligent information processing system for Natural Language Processing and Understanding. IndoWordNet is one such rich online lexical database containing more than twenty thousand parallel synsets for eighteen languages, including English. It uses Hindi WordNet as a pivot to link all these languages. The first WordNet was built in English at Priceton University[3]. Then, followed the WordNets for European Languages[4](Vossen, 1998), and then IndoWordNet. IndoWordNet has approximately 25000 synsets linked to Princeton WordNet. We use these linkages to mine English words from the Princeton WordNet which form the basis of our query for the OpenClipArt API. We download the images via their URLs, and store them locally, to map them to Hindi WordNet[5](Dipak Narayan and Bhattacharyya, 2002) synset IDs later.

The paper is organized as follows. In section 2, we describe our related work. In section 3 and 4, we describe our architecture, and the retrieval procedure along with the scoring algorithm. We describe the results obtained in Section 5. We describe the evaluation tool and qualitative analysis in sections 6 and 7, respectively. We conclude in section 8.

2 Related Work

Bond et al. (2009) used OCAL to enhance the Japanese WordNet, and were able to mine 874 links for 541 synsets. On the basis of manual scoring they found 62 illustrations which were best suited for the sense, 642 illustrations to be a good representation, and 170 suitable, but imperfect illustrations. We extend their work for IndoWordNet, and use OCAL to mine the illustrations. Imagenet(Deng et al., 2009) is a similar project for Princeton WordNet which provides images/URLs for a concept. It contains 21841 synsets indexed with 14,197,122 images. We present a much simpler methodology of collecting images from the web, and then using the synset words to find overlaps with image tags, and then map them.

[3]http://www.wordnet.princeton.edu
[4]http://www.illc.uva.nl/EuroWordNet/
[5]http://www.cfilt.iitb.ac.in/wordnet/webhwn/

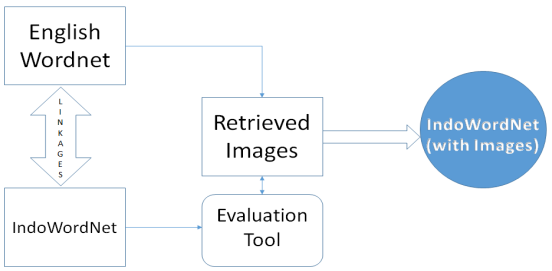


Figure 1: Our Architecture

3 Our Architecture

The following section gives the detailed architecture of our system. A diagrammatic representation is shown in figure 1. Also, we discuss the structure of the IndoWordNet and talk about how we link it to the retrieved set of images.

3.1 Dataset

A linked Hindi - English synset mapping is required to mine the image-synset mapping for Hindi. OpenClipArt contains URL tags in English, and thus a linked Hindi - English synset data structure was required. For our work, we use the following data sets:

3.1.1 Hindi Database

The latest version of Hindi WordNet is available for download at: `http://www.cfilt.iitb.ac.in/wordnet/webhwn/downloaderInfo.php`, which provides an offline interface along with the database, in text format.

3.1.2 English Database

The latest version of Princeton WordNet is available for download at: `https://wordnet.princeton.edu/wordnet/download/`. It provides both the latest database, and standalone installers for WordNet

3.1.3 Hindi-English Linkage database

WordNets have been built for around 100 different languages. Efforts towards mapping synsets across WordNets have been going on for a while in various parts of the world. IndoWordNet contains 28,446 synsets linked to the Princeton WordNet, out of which 21,876 are Nouns. Those concepts in Hindi for which there are no direct linkages in the English WordNet, it was decided to link them to a

hypernymy synset in English. The idea was that instead of having no linkage at all there would be at least a super-ordinate concept and lexical item/items with which the Hindi concept could be linked to provide *weak* translation candidates which could be exploited for various NLP tasks. IndoWordNet has 11,582 direct linkages, and 8184 hypernymy linkages. We use only 11,582 directly linked noun concepts to mine OCAL.

4 Retrieval procedure and scoring

We use the OpenClipArt API[6] to retrieve a set of results using the head word from a synset as the query, since OpenClipArt is a free to use resource, unlike Google Search results which might retrieve copyright data. The API provides a JSON output which can be easily parsed using any programming language. We use JAVA for this purpose. The result for each image provides the following data:

- The title of the 'image'
- The tags for the 'image'
- The URL of the 'image'

To rank the results, we calculate a score based on overlaps between the synsets and image meta-data. The score is derived as a weighted overlap between the words in the *Title* and *Tags* of the result image with the words of the synset. Words from each part are given a different weight owing to how useful the feature is in describing the image. For example, words from the Title are given a higher weight as compared to words from the image Tags. The algorithm increases the score if an overlap occurs and decrements the score otherwise. The magnitude of this increase and decrease depends on the weights of the words being compared. Our system allows for all these weights to be tweaked.

After the result images are scored, they are sorted based on this score. Only the top three scoring images are downloaded. These downloaded images are then evaluated by lexicographers.

5 Results

Using the methodology described above, we map several synsets of the Indian language WordNets to the available images. A total of **8,183** Hindi synsets for directly linked nouns were mapped to their corresponding images. We perform manual evaluation of the data using the tool mentioned above and have evaluated approx. 3,000 synsets for each of the languages. We continue with the manual evaluation for mapping as of now.

Table 1 describes the number of synsets of the WordNets of the following languages for which images have been found, the number of evaluated images out of these, the correctly mapped images, and the precision score for each language.

The top three images are shown to a trained linguist who decides the winner image and also calculates the precision (P@3) of results for that synset. Over a set of 8183 images, we obtain a precision (P@3) of 0.30.

Algorithm 1 Image scoring algorithm

```
1: procedure IMAGE–SCORING
2:     score:= 0
3:     weight(ImageTags) := w
4:     cost(ImageTags) := c
5:     for each token i ∈ ImageTags do
6:         for each token j ∈ Synset do
7:             if i = j then
8:                 score:= score + w
9:             else
10:                score:= score - c
11:            end if
12:        end for
13:    end for
14: end procedure
```

6 Evaluation Tool

We create a PHP[7] based interface, and provide it to lexicographers and linguists for evaluation of the images obtained. The tool uses MySQL[8] database at the back-end to store both Hindi and English WordNet databases, and uses synset ID as a pivot to display the images obtained. The tool provides with a Hindi synset words, its concept, and the English words to help the lexicographer identify its proper sense. The lexicographer chooses a winner image out of the top three, or none of

[6] https://openclipart.org/search/json/

[7] http://php.net/

[8] https://www.mysql.com/

Languages	Images Obtained	Evaluated Images	Accurate Images	Precision
Hindi	8183	3851	1154	**0.3**
Assamese	5198	2860	771	0.27
Bengali	7823	3851	1154	**0.3**
Bodo	5138	2835	765	0.27
Gujarati	7736	3787	1134	**0.3**
Kannada	5695	2883	870	**0.3**
Kashmiri	6705	3470	1043	**0.3**
Konkani	7548	3686	1110	**0.3**
Malayalam	6504	3427	954	0.28
Manipuri	5299	2907	780	0.27
Marathi	6863	3452	1031	**0.3**
Nepali	3959	2163	584	0.27
Sanskrit	7812	3851	1154	**0.3**
Tamil	7272	3611	1083	**0.3**
Telugu	5728	2980	834	0.28
Punjabi	5889	3186	896	0.28
Urdu	5096	2683	684	0.25
Oriya	7412	3660	1034	0.28

Table 1: No. of synsets linked to images

Figure 2: Screen-shot of the Evaluation Tool

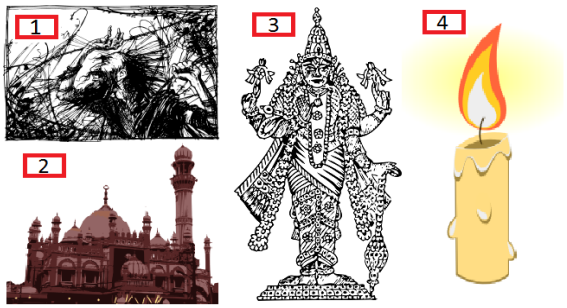


Figure 3: Accuractely acquired images

these, in case of no relevant image. They were also requested to tick the relevant images. A screen-shot for our interface is shown in Figure 2.

7 Qualitative Analysis

In this section, we explain the work done to evaluate the resultant images and the analysis of the results.

7.1 No images found

From the 11,573 sysets that were chosen to be tagged with images, We were unable to retrieve images for 3390 synsets from OpenClipArt, due to unavailability in the source. Our analysis shows that most of the synsets for which a suitable image could not be retrieved fell into two major categories:

Abstract nouns: Several of the synsets for which no images could be retrieved fell into the category of abstract nouns. For example the synset "गुलछर्रा" (*"gUlchharra"*) - 4939 which translates to "profligacy, extravagance" returned no images.

Complex synsets: Apart from abstract nouns, several complex synsets returned no results. For example, the synset "शारीरिक तरल पदार्थ" (*"shAririk TaRal PadArth"*) - 1644 which translates to "Liquid body substance" was unable to fetch any results.

We believe that synsets falling into the first category, *i.e* Abstract nouns, were too vague for an image to do justice to the concept. However, synsets falling into the second category display the limitedness of the OpenClipArt database and further the need of looking in more than one image source.

7.2 Images found

Amongst the synsets for which some images were retrieved, a link was noticed between the class of the noun and how well the image was able to explain the synset.

7.2.1 Common Nouns

Our methodology performs well in this case, and most of the images obtained were able to correctly and almost completely explain the concept. For example, the synset "मोमबत्ती" (*"momBatti"*) - 9866 meaning "candle" and synset "मस्जिद" (*"maszid"*) - 2900 meaning "mosque" retrieved very accurate results as shown in figures 3.4 and 3.2, respectively.

7.3 Proper Nouns

Our retrieval performs well for proper nouns. We were able to obtain pictures for most of the synsets which represent a country. The country flag and map was retrieved for each country name. Several Indian monuments obtained good images along with several Hindu deities. The illustration for synset "विष्णु" ("viShnU") - 2185 translating to a named entity "Vishnu" is shown in figure 3.3.

7.4 Abstract Nouns

Several images were unable to illustrate their corresponding abstract nouns. A few cases of good images were obtained such as synset "हड़कंप" (*"HaDKamp"*) - 3366 meaning "panic" was illustrated by the image 3.1.

8 Conclusion and Future Work

We successfully identified images for synsets of Indian languages and described our work on enriching IndoWordNet. Many synsets could not be linked due to the lack of appropriate image availability on OCAL. We also created a tool for manual evaluation of the data, or any other such work in the future. We evaluated the images obtained, and reported the highest precision score as 0.30. As a future work, we aim to try to retrieve these images using other open source image databases, and utilizing gloss and examples for finding overlaps. Also, The concept of Content Based Image Retrieval (CBIR) appears to be a viable option of several Indian language synsets which cannot be directly linked to a single corresponding English synset. Using CBIR, we can harness resources of several untagged image databases, and thus further enich IndoWordNet as a resource.

9 Acknowledgment

We gratefully acknowledge the support of the Department of Electronics and Information Technology, Ministry of Communications and IT, Government of India. We also acknowledge the annotation work done in this task by Rajita Shukla, Jaya Saraswati, Meghna Singh, Laxmi Kashyap, Ankit, and Amisha. Also, not to be missed, is the entire computational linguistics group at CFILT, IIT Bombay, which has provided its valuable input and critique, helping us refine our task.

References

Pushpak Bhattacharyya. 2010. Indowordnet. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, ALR7, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Prabhakar Pande Dipak Narayan, Debasri Chakrabarti and Pushpak Bhattacharyya. 2002. An experience in building the indowordnet - a wordnet for hindi. In *Proceedings of the First International Conference on Global WordNet (GWC'02)*, Mysore, India, January.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Jonathan Phillips. 2005. Introduction to the openclip art library.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Using Wordnet to Improve Reordering in Hierarchical Phrase-Based Statistical Machine Translation

Arefeh Kazemi[†], **Antonio Toral**[⋆], **Andy Way**[⋆]
[†] Department of Computer Engineering, University of Isfahan, Isfahan, Iran
{kazemi}@eng.ui.ac.ir
[⋆] ADAPT Centre, School of Computing, Dublin City University, Ireland
{atoral,away}@computing.dcu.ie

Abstract

We propose the use of WordNet synsets in a syntax-based reordering model for hierarchical statistical machine translation (HPB-SMT) to enable the model to generalize to phrases not seen in the training data but that have equivalent meaning. We detail our methodology to incorporate synsets' knowledge in the reordering model and evaluate the resulting WordNet-enhanced SMT systems on the English-to-Farsi language direction. The inclusion of synsets leads to the best BLEU score, outperforming the baseline (standard HPB-SMT) by 0.6 points absolute.

1 Introduction

Statistical Machine Translation (SMT) is a data driven approach for translating from one natural language into another. Natural languages vary in their vocabularies and also in the manner that they arrange words in the sentence. Accordingly, SMT systems should address two interrelated problems: finding the appropriate words in the translation ("lexical choice") and predicting their order in the sentence ("reordering"). Reordering is one of the hardest problems in SMT and has a significant impact on the quality of the translation, especially between languages with major differences in word order. Although SMT systems deliver state-of-the-art performance in machine translation nowadays, they perform relatively weakly at addressing the reordering problem.

Phrased-based SMT (PB-SMT) is arguably the most widely used approach to SMT to date. In this model, the translation operates on *phrases*, i.e. sequences of words whose length is between 1 and a maximum upper limit. In PB-SMT, reordering is generally captured by distance-based models (Koehn et al., 2003) and lexical phrase-based models (Tillmann, 2004; Koehn et al., 2005), which are able to perform local reordering but they cannot capture non-local (long-distance) reordering. The weakness of PB-SMT systems on handling long-distance reordering led to proposing the Hierarchical Phrase-based SMT (HPB-SMT) model(Chiang, 2005), in which the translation operates on tree structures (either derived from a syntactic parser or unsupervised). Despite the relatively good performance offered by HPB-SMT in medium-range reordering, they are still weak on long-distance reordering (Birch et al., 2009).

A great deal of work has been carried out to address the reordering problem by incorporating reordering models (RM) into SMT systems. A RM tries to capture the differences in word order in a probabilistic framework and assigns a probability to each possible order of words in the target sentence. Most of the reordering models can perform reordering of common words or phrases relatively well, but they can not be generalized to unseen words or phrases with the same meaning ("*semantic generalization*") or the same syntactic structure ("*syntactic generalization*"). For example, if in the source language the object follows the verb and in the target language it precedes the verb, these models still need to see particular instances of verbs and objects in the training data to be able to perform required reordering between them. Likewise, if two words in the source language follow a specific reordering pattern in the target language, these models can not generalize to unseen words with equivalent meaning in the same context.

In order to improve syntactic and semantic generalization of the RM, it is necessary to incorporate syntactic and semantic features into the model. While there has been some encouraging work on integrating syntactic features into the RM, to the best of our knowledge, there has been no previous work on integrating semantic

Reordering Model	Features Types	Features
Zens and Ney (2006)	lexical	surface forms of the source and target words
		unsupervised class of the source and target words
Cherry (2013)	lexical	surface forms of frequent source and target words
		unsupervised class of rare source and target words
Green *et al.* (2010)	lexical	surface forms of the source words, POS tags of the
	syntactic	source words, relative position of the source words
		sentence length
Bisazza and Federico (2013)	lexical	surface forms and POS tags of the source words
and Goto *et al.* (2013)	syntactic	surface forms and POS tags of the source context words
Gao *et al.* (2011) and	lexical	surface forms of the source words
Kazemi *et al.* (2015)	syntactic	dependency relation
The proposed method	lexical	surface forms of the source words
	syntactic	dependency relation
	semantic	synset of the source words

Table 1: An overview of the used features in the SOTA reordering models

features. In this paper we enrich a recently proposed syntax-based reordering model for HPB-SMT system (Kazemi et al., 2015) with semantic features. To be more precise, we use WordNet[1] (Fellbaum, 1998) to incorporate semantics into our RM. We report experimental results on a large-scale English-to-Farsi translation task.

The rest of the paper is organized as follows. Section 2 reviews the related work and puts our work in its proper context. Section 3 introduces our RM, which is then evaluated in Section 4.2. Finally, Section 5 summarizes the paper and discusses avenues of future work.

2 Related Work

Many different approaches have been proposed to capture long-distance reordering by incorporating a RM into PB-SMT or HPB-SMT systems. A RM should be able to perform the required reorderings not only for common words or phrases, but also for phrases unseen in the training data that hold the same syntactic and semantic structure. In other words, a RM should be able to make syntactic and semantic generalizations. To this end, rather than conditioning on actual phrases, state-of-the-art RMs generally make use of features extracted from the phrases of the training data. One useful way to categorize previous RMs is by the features that they use to generalize. These features can be divided into three groups: (i) lexical features (ii) syntactic features and (iii) semantic features. Table 1 shows a representative selection of state-of-the-art RMs along with the features that they use for generalization.

Zens and Ney (2006) proposed a maximum-entropy RM for PB-SMT that tries to predict the orientation between adjacent phrases based on various combinations of some features: surface forms of the source words, surface form of the target words, unsupervised class of the source words and unsupervised class of the target words. They show that unsupervised word-class based features perform almost as well as word-based features, and combining them results in small gains. This motivates us to consider incorporating supervised semantic-based word-classes into our model.

Cherry (2013) integrates sparse phrase orientation features directly into a PB-SMT decoder. As features, he used the surface forms of the frequent words, and the unsupervised cluster of uncommon words. Green *et al.* (2010) introduced a discriminative RM that scores different jumps in the translation depending on the source words, their Part-Of-Speech (POS) tags, their relative position in the source sentence, and also the sentence length. This RM fails to capture the rare long-distance reorderings, since it typically over-penalizes long jumps that occur much more rarely than short jumps (Bisazza and Federico, 2015). Bisazza and Federico (2013) and Goto *et al.* (2013) estimate for each pair of input positions x and y, the probability of translating y right after x based on the surface forms and the POS tags of the source words, and the surface forms and the POS tags of the source context words.

[1] http://wordnet.princeton.edu/

Gao *et al.* (2011) and Kazemi *et al.* (2015) proposed a dependency-based RM for HPB-SMT which uses a maximum-entropy classifier to predict the orientation between pairs of constituents. They examined two types of features, the surface forms of the constituents and the dependency relation between them. Our approach is closely related to the latter two works, as we are interested to predict the orientation between pairs of constituents. Similarly to (Gao et al., 2011; Kazemi et al., 2015), we train a classifier based on some extracted features from the constituent pairs, but on top of lexical and syntactic features, we use semantic features (WordNet synsets) in our RM. In this way, our model can be generalized to unseen phrases that follow the same semantic structure.

3 Method

Following Kazemi *et al.* (2015) we implement a syntax-based RM for HPB-SMT based on the dependency tree of the source sentence. The dependency tree of a sentence shows the grammatical relation between pairs of head and dependent words in the sentence. As an example, Figure 1 shows the dependency tree of an English sentence. In this figure, the arrow with label "nsubj" from "fox" to "jumped" indicates that the dependent word "fox" is the subject of the head word "jumped". Given the assumption that constituents move as a whole during translation (Quirk et al., 2005), we take the dependency tree of the source sentence and try to find the ordering of each dependent word with respect to its head ($head$-dep) and also with respect to the other dependants of that head (dep-dep). For example, for the English sentence in Figure 1, we try to predict the orientation between ($head$-dep) and (dep-dep) pairs as shown in Table 2.

We consider two orientation types between the constituents: $monotone$ and $swap$. If the order of two constituents in the source sentence is the same as the order of their translation in the target sentence, the orientation is monotone and otherwise it is swap. To be more formal, for two source words (S_1,S_2) and their aligned target words (T_1,T_2), with the alignment points (P_{S1},P_{S2}) and (P_{T1},P_{T2}), we find the orientation type between S_1 and S_2 as shown in Equation 1 (Kazemi et al., 2015).

$$ori = \begin{cases} \text{if } (p_{S1} - p_{S2}) \times (p_{T1} - p_{T2}) > 0 \\ \quad monotone \\ \text{else} \\ \quad swap \end{cases} \tag{1}$$

For example, for the sentence in Figure 1, the orientation between the source words "brown" and "quick" is monotone, while the orientation between "brown" and "fox" is swap.

We use a classifier to predict the probability of the orientation between each pair of constituents to be monotone or swap. This probability is used as one feature in the log-linear framework of the HPB-SMT model. Using a classifier enables us to incorporate fine-grained information in the form of features into our RM. Table 3 and Table 4 show the features that we use to characterize ($head$-dep) and (dep-dep) pairs respectively.

As Table 3 and Table 4 show, we use three types of features: lexical, syntactic and semantic. While semantic structures have been previously used for MT reordering, e.g. (Liu and Gilda, 2010), to the best of our knowledge, this is the first work that includes semantic features jointly with lexical and syntactic features in the framework of a syntax-based RM. Using syntactic features, such as dependency relations, enables the RM to make syntactic generalizations. For instance, the RM can learn that in translating between subject-verb-object (SVO) and subject-object-verb (SOV) languages, the object and the verb should be swapped.

On top of this syntactic generalization, the RM should be able to make semantic generalizations. To this end, we use WordNet synsets as an additional feature in our RM. WordNet is a lexical database of English which groups words into sets of cognitive synonyms. In other words, in WordNet a set of synonym words belong to the same synset. For example, the words "baby", "babe" and "infant" are in the same synset in WordNet. The use of synsets enables our RM to be generalized from words seen in the training data to any of their synonyms present in WordNet.

4 Experiments

4.1 Data and Setup

We used the Mizan English–Farsi parallel corpus [2] (Supreme Council of Information and Communication Technology, 2013), which contains

[2] http://dadegan.ir/catalog/mizan

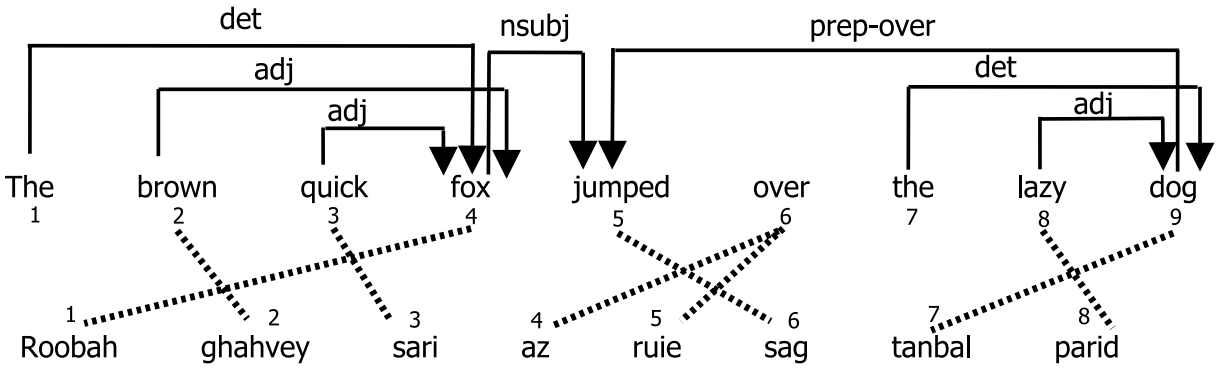


Figure 1: An example dependency tree for an English source sentence, its translation in Farsi and the word alignments

head	jumped	jumped	fox	fox	fox	dog	dog
dependant	fox	dog	the	brown	quick	the	lazy
dependant 1	fox	brown	the	the	the		
dependant 2	dog	quick	brown	quick	lazy		

Table 2: head-dependant and dependant-dependant pairs for the sentence in Figure 1.

around one million sentences extracted from English novel books and their translation in Farsi. We randomly held out 3,000 and 1,000 sentence pairs for tuning and testing, respectively, and used the remaining sentence pairs for training. Table 5 shows statistics (number of words and sentences) of the data sets used for training, tuning and testing.

	Unit	**English**	**Farsi**
Train	sentences	1,016,758	1,016,758
	words	13,919,071	14,043,499
Tune	sentences	3,000	3,000
	words	40,831	41,670
Test	sentences	1,000	1,000
	words	13,165	13,444

Table 5: Mizan parallel corpus statistics

We used GIZA++ (Och and Ney, 2003) to align the words in the English and Farsi sentences. We parsed the English sentences of our parallel corpus with the Stanford dependency parser (Chen and Manning, 2014) and used the "collapsed representation" of its output which shows the direct dependencies between the words in the English sentence. Having obtained both dependency trees and the word alignments, we extracted 6,391,956 (*head-dep*) and 5,247,526 (*dep-dep*) pairs from our training data set and determined the orientation for each pair based on Equation 1. We then trained a Maximum Entropy classifier (Manning and Klein, 2003) (henceforth MaxEnt) on the extracted constituent pairs from the training data set and use it to predict the orientation probability of each pair of constituents in the tune and test data sets. As mentioned earlier, we used WordNet in order to determine the synset of the English words in the data set.

Our baseline SMT system is the Moses implementation of the HPB-SMT model with default settings (Hoang et al., 2009). We used a 5-gram language model and trained it on the Farsi side of the training data set. All experiments used MIRA for tuning the weights of the features used in the HPB model (Cherry and Foster, 2012).

The semantic features (synsets) are extracted from WordNet 3.0. For each word, we take the synset that corresponds to its first sense, i.e. the most common one. An alternative would be to apply a word sense disambiguation algorithm. However, these have been shown to perform worse than the first-sense heuristic when WordNet is the inventory of word senses, e.g. (Pedersen and Kolhatkar, 2009; Snyder and Palmer, 2004).

4.2 Evaluation: MT Results

We selected different feature sets for (*head-dep*) and (*dep-dep*) pairs from Table 3 and Table 4 respectively, then we used them in our MaxEnt classifier to determine the impact of our novel se-

Features	Type	Description
$lex(head)$,$lex(dep)$	lexical	surface forms of the head and dependent word
$depRel(dep)$	syntactic	dependency relation of the dependent word
$syn(head)$,$syn(dep)$	semantic	synsets of the head and dependent word

Table 3: Features for ($head$-dep) constituent pairs

Features	Type	Description
$lex(head)$,$lex(dep1)$,$lex(dep2)$	lexical	surface forms of the mutual head and dependent words
$depRel(dep1)$,$depRel(dep2)$	syntactic	dependency relation of the dependent words
$syn(head)$,$syn(dep1)$,$syn(dep2)$	semantic	synsets of the head and dependent words

Table 4: Features for (dep-dep) constituent pairs

mantic features (WordNet synsets) on the quality of the MT system. Three different feature sets were examined in this paper, including information from (i) surface forms (*surface*), (ii) synsets (*synset*) and (iii) both surface forms and synsets (*both*). We build six MT systems, as shown in Table 6, according to the constituent pairs and feature sets examined.

We compared our MT systems to the standard HPB-SMT system. Each MT system is tuned three times and we report the average scores obtained with multeval[3] (Clark et al., 2011) on the MT outputs. The results obtained by each of the MT systems according to two widely used automatic evaluation metrics (BLEU (Papineni et al., 2002), and TER (Snover et al., 2006)) are shown in Table 7. The relative improvement of each evaluation metric over the baseline HPB is shown in columns $diff$.

Compared to the use of surface features, our novel semantic features based on WordNet synsets lead to better scores for both (head- dep) and (dep-dep) constituent pairs according to both evaluation metrics, BLEU and TER (except for the dd system in terms of TER, where there is a slight but insignificant increase (79.8 vs. 79.7)).

5 Conclusions and Future Work

In this paper we have extended a syntax-based RM for HPB-SMT with semantic features (WordNet synsets), in order to enable the model to generalize to phrases not seen in the training data but that have equivalent meaning. The inclusion of synsets has led to the best BLEU score in our experiments, outperforming the baseline (standard HPB-SMT) by 0.6 points absolute.

[3] https://github.com/jhclark/multeval

As for future work, we propose to work mainly along the following two directions. First, an investigation of the extent to which using a WordNet-informed approach to classify the words into semantic classes (as proposed in this work) outperforms an unsupervised approach via word clustering. Second, an in-depth human evaluation to gain further insights of the exact contribution of WordNet to the translation output.

Acknowledgments

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University, the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and by the University of Isfahan.

References

Alexandra Birch, Phil Blunsom, and Miles Osborne. 2009. A Quantitative Analysis of Reordering Phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Athens, Greece.

Arianna Bisazza and Marcello Federico. 2013. Dynamically shaping the reordering search space of phrase-based statistical machine translation. *Transactions of the ACL*, (1):327–340.

Arianna Bisazza and Marcello Federico. 2015. A survey of word reordering in statistical machine translation: Computational models and language phenomena. In *arXiv preprint arXiv:1502.04938*.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

MT System	Features
hd-surface	$Lex(head)$, $Lex(dep)$, $depRel(dep)$
hd-synset	$depRel(dep)$,$Syn(head)$, $Syn(dep)$
hd-both	$Lex(head)$,$Lex(dep)$,$depRel(dep)$,$Syn(dep)$,$Syn(head)$
dd-surface	$Lex(head)$, $Lex(dep1)$, $Lex(dep2)$, $depRel(dep1)$, $depRel(dep2)$
dd-synset	$Syn(head)$, $Syn(dep1)$, $Syn(dep2)$, $depRel(dep1)$, $depRel(dep2)$
dd-both	$Lex(head)$,$Lex(dep1)$,$Lex(dep2)$,$Syn(head)$,$Syn(dep1)$,$Syn(dep2)$, $depRel(dep1)$,$depRel(dep2)$

Table 6: Examined features for MT systems

System	BLEU ↑					TER ↓				
	Avg	diff	$\overline{s}_{\text{sel}}$	s_{Test}	p-value	Avg	diff	$\overline{s}_{\text{sel}}$	s_{Test}	p-value
baseline	10.9	-	0.6	0.0	-	80.3	-	0.8	0.0	-
dd-surface	11.4	4.58%	0.7	0.1	0.00	**79.7**	-0.74%	0.8	0.2	0.01
dd-syn	11.3	3.66%	0.6	0.2	0.01	79.8	-0.62%	0.8	0.2	0.05
dd-both	**11.5**	5.50%	0.7	0.2	0.00	79.8	-0.62%	0.8	0.5	0.02
hd-surface	11.1	2.18%	0.6	0.1	0.08	80.9	0.74%	0.8	0.3	0.01
hd-syn	11.3	3.66%	0.6	0.1	0.00	80.5	0.24%	0.8	0.2	0.40
hd-both	11.1	2.18%	0.6	0.1	0.06	81.1	0.99%	0.8	0.3	0.00

Table 7: MT scores for all systems. p-values are relative to the baseline and indicate whether a difference of this magnitude (between the baseline and the system on that line) is likely to be generated again by some random process (a randomized optimizer). Metric scores are averages over three runs. s_{sel} indicates the variance due to test set selection and has nothing to do with optimizer instability. The best result according to each metric (highest for BLEU and lowest for TER) is shown in bold.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Yang Gao, Philipp Koehn, and Alexandra Birch. 2011. Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 857–868.

Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. 2013. Distortion model considering rich context for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 155–165.

Spence Green, Michel Galley, and Christopher D. Manning. 2010. Improved models of distortion cost for statistical machine translation. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, page 867875.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT*, pages 152–159.

Arefeh Kazemi, Antonio Toral, Andy Way, Amirhassan Monadjemi, and Mohammadali Nematbakhsh. 2015. Dependency-based reordering model for constituent pairs in hierarchical smt. In *Proceedings of*

the 18th Annual Conference of the European Association for Machine Translation, pages 43–50, Antalya, Turkey, May.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–133.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 68–75.

Ding Liu and Daniel Gilda. 2010. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 716–724.

Christopher Manning and Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials*, pages 8–8.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Ted Pedersen and Varada Kolhatkar. 2009. Wordnet::senserelate::allwords: A broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, NAACL-Demonstrations '09, pages 17–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas*.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Supreme Council of Information and Communication Technology. 2013. Mizan English-Persian Parallel Corpus. Tehran, I.R. Iran.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *StatMT '06 Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63.

Eliminating Fuzzy Duplicates in Crowdsourced Lexical Resources

Yuri Kiselev
Yandex
Yekaterinburg, Russia
yurikiselev@yandex-team.ru

Dmitry Ustalov
Ural Federal University
Yekaterinburg, Russia
dmitry.ustalov@urfu.ru

Sergey Porshnev
Ural Federal University
Yekaterinburg, Russia
s.v.porshnev@urfu.ru

Abstract

Collaboratively created lexical resources is a trending approach to creating high quality thesauri in a short time span at a remarkably low price. The key idea is to invite non-expert participants to express and share their knowledge with the aim of constructing a resource. However, this approach tends to be noisy and error-prone, thus making data cleansing a highly topical task to perform. In this paper, we study different techniques for synset deduplication including machine- and crowd-based ones. Eventually, we put forward an approach that can solve the deduplication problem fully automatically, with the quality comparable to the expert-based approach.

1 Introduction

A WordNet-like thesaurus is a dictionary of a special type that represents different semantic relations between *synsets*—sets of quasi-synonyms (Miller et al., 1990). It is a crucial resource for addressing such problems as word sense disambiguation, search query extension and many other problems in the fields of natural language processing (NLP) and artificial intelligence (AI). Typical semantic relations represented by thesauri are synonymy, antonomy (primarily for nouns and adjectives), troponymy (for verbs), hypo-/hypernymic relations, and meronymy.

A good linguistic resource should not contain duplicated lexical senses, because duplicates violate the data integrity and complicate addition of semantic relations to the resource. Therefore, removing duplicated synsets from thesauri is an important problem to be addressed, especially in collaboratively created lexical resources like Wiktionary, which is known to suffer this problem (Kiselev et al., 2015). However, deduplication is rather problematic because thesauri may contain fuzzy duplicated synsets composed of different words.

The work, as described in this paper, makes the following contributions: (1) it proposes an automatic approach to synset deduplication, (2) presents a synonymic dictionary-based technique for assessing synset quality, and (3) compares the proposed approach with the crowdsourcing-based one.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 defines the problem of synset duplicates existing in thesauri. Section 4 presents a novel approach to synset deduplication. Section 5 describes the experimental setup. Section 6 shows the obtained results. Section 7 discusses the interesting findings. Section 8 concludes the paper and defines directions for future work.

2 Related Work

One of the most straightforward ways to clear a thesaurus of sense duplicates is to align its entries with another resource of proven quality, e.g. using the OntoClean methodology proposed by Guarino and Welty (2009). Consequently, synsets that will be linked with one synset from another resource represent the same concepts, and should be merged. However, such alignment can be performed only manually. It is also a time-consuming process that requires careful examination of every synset by an expert. Therefore, it is crucial to focus on methods that are either automatic or involve lesser amount of human intervention.

Many studies nowadays aim to evaluate the feasibility of crowdsourcing for various NLP problems. For instance, Snow et al. (2008) showed that non-expert annotators can produce the data whose quality may compete with the expert annotation in such tasks as word sense disambiguation and word similarity estimation (they conducted their study using Amazon Mechanical Turk[1] (AMT), a popular online labor marketplace).

Sagot and Fišer (2012) assumed that semantically related words tend to co-occur in texts. Given such an assumption, they managed to find and eliminate the words that had been added to synsets by mistake. This approach can be used to find sense duplicates, but it requires a large amount of semantic relations to be present in a resource. It should be noted that some resources that contain synsets may not contain any links between them. For instance, Wiktionary represents certain words and relations between them, but it does not explicitly link its synsets.

Sajous et al. (2013) presented a method for semi-automatic enrichment of the Wiktionary-derived synsets. First, they analyzed the contents of Wiktionary and produced new synonymy relations that had not been previously included in the resource. After that, they invited collaborators to manually process the data using a custom Firefox plugin to add missing synonyms to the data.

A similar approach was used by Braslavski et al. (2014) to bootstrap YARN (Yet Another RussNet) project, which aims at creating a large open WordNet-like machine-readable thesaurus for the Russian language by means of crowdsourcing. In this project, a dedicated collaborative synset editing tool was used by the annotators to construct synsets by adding and removing words.

The most recognized crowdsourcing workflow is the Find-Fix-Verify pattern proposed by Bernstein et al. and used in Soylent, a Microsoft Word plugin that submits human intelligence tasks to AMT for rephrasing and improving the original text (Bernstein et al., 2010). As the name implies, the workflow includes the three stages: 1) in the *Find* stage crowd workers find the text area that can be shortened without changing the meaning, 2) in the *Fix* stage the workers propose improvements for these text areas, and 3) in the *Verify* stage the workers select the *worst* proposed fixes.

Inspired by this pattern, Ustalov and Kiselev (2015) presented the Add-Remove-Confirm workflow for improving synset quality. Similarly, it contains three stages: 1) in the *Add* stage workers choose the words to be added to a synset from a given list of candidates, 2) in the *Remove* stage the workers choose the words that should be removed from a synset, 3) in the *Confirm* stage the workers choose which synset is better—the initial one or the fixed one.

3 Problem

In our study, we focus on the synsets represented in a WordNet-like thesaurus. Hence, we regard a thesaurus as a set of synsets S, where every synset $s \in S$ consists of different words and represents some sense or concept.

In lexical resources created by expert lexicographers, synsets usually correspond to different meanings, so synset duplicates never arise. Unfortunately, it is not true for the resources created by non-expert users, e.g. through the use of crowdsourcing. One approach to synset creation would be to combine manually constructed synsets with synsets that are imported from open resources. Obviously, it is going to lead to the situation where there is a plenty of synsets representing identical concepts. The crowdsourcing approach to synset creation is also prone to this drawback, as the crowd is likely to create duplicate synsets.

The following example from the Russian Wiktionary[2] shows that it contains synsets with identical meanings. For example, the synset {стоматолог (*stomatologist*), дантист (*dentist*), зубной врач (*"tooth doctor"*)} and the synset {дантист (*dentist*), стоматолог (*stomatologist*)} definitely describe the same concept "a person qualified to treat the diseases and conditions that affect the teeth". Hence, such synsets should be combined, yet they both are present in the Russian Wiktionary. Note that in this example the second synset is a full subset of the first one; however, it is possible that two synsets may intersect only partly while sharing the same meaning.

For a native speaker, it is relatively easy to detect whether two synsets share the same meanings. So, the detection may be done by non-experts via crowdsourcing. However, the key problem here is how to retrieve the pairs of synsets that presumably represent identical concepts. In the next sec-

[1] https://www.mturk.com/mturk/welcome

[2] https://ru.wiktionary.org/

tion, we propose a simple, yet effective approach.

4 Approach

Suppose the word w has several meanings. According to Miller et al. (1990), it is usually enough to provide one synonym for every meaning of w to a native speaker of a language to be able to distinguish the meanings from each other (provided that the speaker is familiar with the corresponding concepts). This phenomenon is widely exploited by explanatory dictionaries. It is also utilized in some thesauri which assume that a synset itself is enough to deduce its meaning, therefore definitions of synsets may be omitted.

Hence, we formulate the meaning deduplication problem as follows. Given a pair of different synsets $s_1 \in S$ and $s_2 \in S$, we treat them as *duplicates* if they share exactly two words:

$$\exists s_1 \in S, s_2 \in S : s_1 \neq s_2 \wedge |s1 \cap s2| = 2.$$

Obviously, this is a strong criterion that may be violated, so we propose the following two-stage workflow for synset deduplication.

Filtering. In this stage, the possible duplicates are retrieved using the above described criterion resulting in the set of synset pairs (s_1, s_2) for further validation.

Voting. In this stage, the obtained synset pairs are subject to manual verification. The pairs voted as equivalent are combined.

The assessment required in the Voting stage may be provided by expert lexicographers; in crowdsourced resources, the contributors may be invited not only to add the new data, but also to increase the quality of the created data and to deduplicate it.

5 Experiments

Since task submission to Amazon Mechanical Turk requires a U.S. billing address, this solution is not accessible to users from other countries. Although there are many other crowdsourcing platforms, e.g. CrowdFlower, Microworkers, Prolific Academic, etc., yet the proportion of Russian speakers on such platforms is still low (Pavlick et al., 2014).

Given the fact that our workers are native Russian speakers, we decided to use the open source crowdsourcing engine Mechanical Tsar[3], which is designed for rapid deployment of mechanized labor workflows (Ustalov, 2015). Inspired by the similar annotation study conducted by Snow et al. (2008), we used the default configuration, i.e. the majority voting strategy for answer aggregation, the fixed answer number per task strategy for task allocation, and the no worker ranking. The workers were invited from VK, Facebook and Twitter via a short-term open call for participation posted by us.

5.1 Stage "Filtering"

We used two different electronic thesauri for the experiments. The first one was chosen from among crowdsourced lexical resouces. Selecting between the Russian Wiktionary and YARN, we settled on the latter because it comprises one and half time more synsets, and it is easier to parse because YARN[4] synsets are available in the CSV format.

We were also interested in applying the described approach to a resource created by expert lexicographers. The current situation with electronic thesauri for the Russian language is that there is only one resource that is large enough and is available for study. This resource is RuThes-lite[5], a publicly available version of the RuThes linguistic ontology, which has been developing for many years (Loukachevitch, 2011).

We retrieved 210 presumably duplicated synsets from each resource—70 synsets with exactly two common words, 70 synsets with three, and 70 synset with four or more common words. Such a stratification is motivated by the interest in analyzing how the number of shared words correlates with their meanings.

By randomly sampling pairs of possibly duplicated synsets from YARN, we concluded that the proposed criterion for synset equivalence is very robust. It appears that for YARN this approach may be used even without the Voting stage. Thus, we decided to study whether the manual annotation does increase the quality of synset deduplication. In order to do this, we selected synsets from YARN as follows.

Since synsets in YARN are not always accompanied by sense definitions, we asked an expert to

[3] http://mtsar.nlpub.org/

[4] http://russianword.net/yarn-synsets.csv

[5] http://www.labinform.ru/pub/ruthes/

manually align the selected synsets with an expert-built lexical resource. We chose the Babenko dictionary (2011) (hereinafter referred to as BAB) as an expert-built lexical resource because it is a relatively recent dictionary with a wide language coverage. As a result of the alignment, each YARN synset s was provided with a corresponding synset s_{BAB} defined by a sense definition d.

5.2 Stage "Voting"

The goal of the Voting stage is to choose true equivalents among the prepared presumably equivalent synset. The input of this stage is a pair of synsets (s_1, s_2) from a resource, and a worker is to determine if the synsets share the same meaning (Figure 1).

Do the following synsets have the same meanings: "s_1" and "s_2"?
[ ] Yes
[ ] No

Figure 1: Task format for Voting stage (the original text was in Russian).

6 Results

6.1 Quality metrics

We use precision and recall to measure the quality of synsets in a thesaurus S. Precision $P(s)$ of a synset $s \in S$ is the fraction of the synset words with the meaning represented by s, compared to all the words in the language representing the meaning of the synset $\mathcal{L}(s)$.

$$P(s) = \frac{|s \cap \mathcal{L}(s)|}{|s|} \quad (1)$$

Recall $R(s)$ of a synset s is the fraction of all words S in the language that have the meaning that s represents.

$$R(s) = \frac{|s \cap \mathcal{L}(s)|}{|\mathcal{L}(s)|} \quad (2)$$

As may be easily noticed, it is impossible to precisely calculate the measure of synset recall $R(s)$, since the whole set of words that can correspond to a particular meaning is unknown. In order to estimate $\mathcal{L}(\cdot)$, we used the data retrieved at the Filtering stage. We combined the YARN synsets in each pair (s_1, s_2) into a new synset s. Then, we provided the resulting synset s with a corresponding definition d from the BAB and asked the same expert as in the Filtering stage to remove words from s, which do not correspond to the definition d. The fixed synsets s' were then combined with the corresponding synsets s_{BAB}. These combined synsets were used as the gold standard synsets s_{GS} for concepts, as we considered that such synsets contained all the words representing the concepts.

6.2 Example of Quality Calculation

Consider the following example in order to better understand the described process of data preparation and the further evaluations. Let say that YARN contains synset s_1={*think*, *opine*, *suppose*, *sleep*} and synset s_2={*think*, *suppose*, *reckon*}, and BAB contains synset s_{BAB}={*think*, *opine*, *suppose*, *imagine*} with definition d "expect, believe, or suppose" ($|s_1 \cap s_2| = |\{think, suppose\}| = 2$ and $|s_1 \cap s_{BAB}| = |\{think, opine, suppose\}| = 3$). Assume that the expert aligned s_1 and s_{BAB} in the Filtering stage. In that case the expert would be provided with synset $s = s_1 \cup s_2$={*think*, *opine*, *suppose*, *sleep*, *reckon*} and definition d from BAB. After fixing this synset s (by removing the wrong word *sleep*), it will be combined with the corresponding synset s_{BAB}. So the synset that will be further treated as the gold standard for this concept is s_{GS}={*think*, *opine*, *suppose*, *imagine*, *reckon*}. This set will be used as $\mathcal{L}$ for calculating (1) and (2) (for the corresponding s_1 and s_{BAB}, $\mathcal{L}(s_1) = \mathcal{L}(s_{BAB})$). According to this,

$$P(s_1) = \frac{|s_1 \cap \mathcal{L}(s_1)|}{|s_1|} = \frac{3}{4} = 0.75,$$

$$R(s_{BAB}) = \frac{|s_{BAB} \cap \mathcal{L}(s_{BAB})|}{|\mathcal{L}(s_{BAB})|} = \frac{4}{5} = 0.8.$$

Note that in the proposed evaluation method, precision P of any synset from BAB s_{BAB} is 1.0.

6.3 Quality Assessment

The procedure described in Section 6.1 allowed us to calculate the suggested quality measures for the resources (Table 1). The *BAB* row is calculated for 210 synsets from the Babenko dictionary, the YARN, *aligned* row—for 210 synsets s_1 from YARN that were aligned with the BAB by the expert, and the YARN, *machine*—for the automatically merged all 210 presumably equivalent synsets (s_1, s_2) of YARN.

The F_1-measure for YARN is expectedly lower than for the BAB, yet, after a simple merging of

Table 1: Synset quality.

	Avg P	Avg R	Avg F_1
BAB	1.000	0.661	0.796
YARN, *aligned*	0.901	0.634	0.744
YARN, *machine*	0.840	0.774	0.805

the presumably equivalent synsets, its average F_1-measure became higher than for the BAB. However, this result was due to the significant increase in the recall, while the precision dropped.

To investigate how people's participation can improve the quality of automatic merging, we conducted a crowdsourcing experiment. Every task (Figure 1) was annotated by at least three different workers. The decision about merging was made by majority voting. Table 2 shows the share of synsets that the workers decided to merge.

Table 2: Crowdsourcing synset deduplication.

# of common words	2	3	4+
YARN	$^{61}/_{70}$	$^{64}/_{70}$	$^{68}/_{70}$
RuThes-lite	$^{25}/_{70}$	$^{40}/_{70}$	$^{51}/_{70}$

Quite expectedly, the two analyzed lexical resources proved very different. Our equivalence criterion worked only in one third of the cases for RuThes-lite. And even the stronger version of the criterion (the one considering synsets that share 4+ words as sense duplicates) was true only in $\frac{2}{3}$ cases according to the annotators. However, for YARN the criterion proved to be rather robust, so that it can be applied without crowd checking, provided that the results of the merging will be verified by a moderator of the resource.

This conclusion agreed with the quality estimates of the merging performed according to human annotations (Table 3). The first row (YARN, *machine*) corresponds to the automatic merge of all 210 synsets repeats the row of Table 1 with the same name, and the second row (YARN, *crowd*) corresponds to the selective merge performed according to the human judgements. So, $61+64+68$ synset pairs (s_1, s_2) were merged (Table 2), and the 17 remained synsets we left as they were (s_1).

Table 3: YARN synset deduplication.

	Avg P	Avg R	Avg F_1
YARN, *machine*	0.840	0.774	0.805
YARN, *crowd*	0.852	0.764	0.805

7 Discussion

The F_1-measure shows no change after applying the Voting stage, yet the precision increases by 0.012 while the recall drops by 0.01. Despite the fact that the overall quality is constant regardless of the human annotations, it still presents an interesting finding, since people increase the precision of the merging. This is important because it allows to compensate, at least partially, for the reduction in the precision against the original synsets caused by the automatic merge. (Table 3).

It is also of interest that YARN contains 24.8 thousand synsets that presumably have a duplicate (58% of the synsets with two or more words), while the Russian Wiktionary has 13.2 thousand (40%), and RuThes-lite has only 6.3 thousand (28%). We may therefore conclude that the proposed approach should mainly be applied to resources that a priori are known to contain duplicate synsets rather than to improve the quality of expert-built resources.

7.1 Synset Ambiguity

The analysis of the results of the experiments and the annotations provided by our expert showed that in some cases it is almost impossible to derive a meaning from a synset. For instance, just a couple of synonyms is not enough to distinguish the meaning "a woman thought to have *evil* magic powers" from "a woman who uses magic or sorcery" (the latter definition does not imply an "*evil*" woman, which can be not obvious from a synonymy row).

Another example of such ambiguity are the concepts corresponding to "a bed *with* a back" and "a bed *without* a back". Given only a synset, it is barely possible to discern this shade of meaning and distinguish any of these two concepts from the more common one (simply "a bed"). With this observation in mind, we suggest that the authors of the wordnets for which the meanings of synsets are optional should take it into account and include definitions for vague concepts.

7.2 Pairwise Annotation

Special attention should be given to the performance of the crowd workers. In our experiment, 25 workers provided 1262 answers to 420 pairwise comparison tasks (Figure 1). The workers repeatedly reported that the tasks were time consuming due to data inconsistency. Suppose that

synset sizes are n_1 and n_2 correspondingly, and an annotator spends $O(n_1 + n_2)$ time to make a decision. Hence, even in the simplest case (Table 4) an annotator will perform $4 + 4 = 8$ operations per pair, which is inconvenient.

Table 4: Average synset sizes.

# of common words	**2**	**3**	**4+**
YARN	4.2	4.6	5.5
RuThes-lite	4.3	5.0	5.8

Further studies should avoid pairwise comparison in problems involving contextual or domain knowledge for making a decision by annotators. However, it still may be useful in various visual recognition tasks, especially when the workers are provided with an observable hint (Deng et al., 2013). We should also note that this outcome agrees well with the study conducted by Wang et al. (2012), when cluster-based task generation led to lower time spent rather than in pair-based tasks.

7.3 Agreement & Issues

We have analyzed all the cases when all the three workers gave the same answer to the task (Table 5). For YARN , the number of cases when all the workers agreed rises with the number of common words in synsets. This is quite expected considering that sharing more common words makes it more obvious that the synsets have common senses. However, we do not observe the same in RuThes-lite.

Table 5: # of merge decisions made unanimously.

# of common words	**2**	**3**	**4+**
YARN	$^{32}/_{70}$	$^{47}/_{70}$	$^{57}/_{70}$
RuThes-lite	$^{36}/_{70}$	$^{35}/_{70}$	$^{32}/_{70}$

Manual analyses of the data from RuThes-lite showed that its authors tend to discriminate meanings of synsets with common words by means of only one word, e.g. using a hyponym for a concept in one set and a corresponding hypernym in another. It is enough to emphasize the difference in meanings, but workers may find it problematic to detect the only pair of words that defines the difference in the pair of synsets. This task may become even more complicated in large synsets, as they grow in size along with the increase in the number of common words in them (Table 4).

8 Conclusion

In this study, we presented an automated approach to synset deduplication. The results were obtained from expert labels and annotations provided by crowd work. At least three different annotations per every synset pair from two different resources (YARN and RuThes-lite) were used. The approach allows to significantly increase the synset quality in crowdsourcing lexical resources. Participation of people does not notably affect the average synset quality, though the precision slightly increases when people are involved.

The results showed that two synonyms are not sufficient for defining a meaning, but three words usually give a satisfactory result. So, it is three words that should be used as a threshold value for merging duplicate synsets when using the proposed deduplication approach in a fully automatic mode. Our results, including the crowd answers and the produced gold standard, are available[6] under the terms of Creative Commons Attribution-ShareAlike 3.0 license.

As a possible future direction, we may suggest using more sophisticated similarity measures to select a threshold for fully automatic merging of synsets. Another possible way to improve the approach is to detect not just pairs, but clusters of synsets. This is hardly possible in resources that are manually crafted by a team of experts, but it is definitely worth exploring for crowdsourcing resources.

Acknowledgments

This work is supported by the Russian Foundation for the Humanities, Project No. 13-04-12020 “New Open Electronic Thesaurus for Russian”. The authors are grateful to Yulia Badryzlova for proofreading the text, and to Alisa Porshneva for labeling synsets. The authors would also like to thank all those who participated in the crowdsourced experiment.

[6] http://ustalov.imm.uran.ru/pub/duplicates-gwc.tar.gz

References

Ljudmila G. Babenko, editor. 2011. *Dictionary of synonyms of the Russian Language*. AST: Astrel, Moscow, Russia.

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 313–322, New York, NY, USA. ACM.

Pavel Braslavski, Dmitry Ustalov, and Mikhail Yu. Mukhin. 2014. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 101–104, Gothenburg, Sweden. Association for Computational Linguistics.

Jia Deng, Jonathan Krause, and Li Fei-Fei. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 580–587.

Nicola Guarino and Christopher A. Welty. 2009. An Overview of OntoClean. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 201–220. Springer Berlin Heidelberg.

Yuri Kiselev, Andrew Krizhanovsky, Pavel Braslavski, et al. 2015. Russian Lexicographic Landscape: a Tale of 12 Dictionaries. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, volume 1, pages 254–271. RGGU, Moscow.

Natalia V. Loukachevitch. 2011. *Thesauri in information retrieval tasks*. Moscow University Press, Moscow, Russia.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *Lexicography*, 3:235–244.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Benoît Sagot and Darja Fišer. 2012. Cleaning noisy wordnets. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Franck Sajous, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, and Yannick Chudy. 2013. Semi-Automatic Enrichment of Crowdsourced Synonymy Networks: The WISIGOTH System Applied to Wiktionary. *Language Resources and Evaluation*, 47(1):63–96.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dmitry Ustalov and Yuri Kiselev. 2015. Add-Remove-Confirm: Crowdsourcing Synset Cleansing. In *Application of Information and Communication Technologies (AICT), 2015 IEEE 9th International Conference on*, pages 143–147. IEEE.

Dmitry Ustalov. 2015. A Crowdsourcing Engine for Mechanized Labor. *Proceedings of the Institute for System Programming*, 27(3):351–364.

Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. 2012. CrowdER: Crowdsourcing Entity Resolution. *Proc. VLDB Endow.*, 5(11):1483–1494.

Automatic Prediction of Morphosemantic Relations

Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova,
Tsvetana Dimitrova, Maria Todorova
Department of Computational Linguistics
Bulgarian Academy of Sciences
{svetla,zarka,iva,cvetana,maria}@dcl.bas.bg

Abstract

This paper presents a machine learning method for automatic identification and classification of morphosemantic relations (MSRs) between verb and noun synset pairs in the Bulgarian WordNet (BulNet). The core training data comprise 6,641 morphosemantically related verb–noun literal pairs from BulNet. The core data were preprocessed quality-wise by applying validation and reorganisation procedures. Further, the data were supplemented with negative examples of literal pairs not linked by an MSR. The designed supervised machine learning method uses the RandomTree algorithm and is implemented in Java with the Weka package. A set of experiments were performed to test various approaches to the task. Future work on improving the classifier includes adding more training data, employing more features, and fine-tuning. Apart from the language specific information about derivational processes, the proposed method is language independent.

1 Introduction

This paper investigates a machine learning method for identification and classification of morphosemantic relations (MSRs) between verb and noun synset pairs in the Bulgarian WordNet (BulNet). It is based on the MSR dataset from the Princeton WordNet (PWN) (Fellbaum et al., 2009), automatically imported to the Bulgarian WordNet (the core dataset), the PWN semantic primitives (henceforth, semantic primes) and the derivational relations (DRs) in the Bulgarian WordNet. The derivational relations had been previously assigned automatically to the Bulgarian WordNet using a string similarity algorithm combined with heuristics (Dimitrova et al., 2014), and had been manually post-edited.

The MSRs link verb–noun pairs of synsets that contain derivationally related literals. As semantic and morphosemantic relations refer to concepts, they are universal, and such a relation must hold between the relevant concepts in any language, regardless of whether it is morphologically expressed or not. This has enabled the automatic transfer of the relations to other languages, such as Polish (Piasecki et al., 2009), Bulgarian (Koeva, 2008; Stoyanova et al., 2013; Dimitrova et al., 2014), Serbian (Koeva et al., 2008), Romanian (Barbu Mititelu, 2012; Barbu Mititelu, 2013). Other sets of MSRs have been proposed for Turkish (Bilgin et al., 2004), Czech (Pala and Hlaváčková, 2007), Estonian (Kahusk et al., 2010), Polish (Piasecki et al., 2012a; Piasecki et al., 2012b), Croatian (Šojat and Srebačić, 2014).

The study is motivated by the fact that a considerable number – 67% (7,905 out of 11,751) of the noun synsets derivationally related to verb synsets and 89% (7,962 out of 8,934) of the verb synsets derivationally related to noun synsets in the PWN 3.0. – is not labelled with an MSR. In addition, the linguistic generalisations behind the existing MSRs have been made on the basis of English derivational morphology, hence the proposed set of MSR instances may be extended based on evidence from the derivational morphology of other languages, including Bulgarian.

The present research builds on Leseva et al. (2014), where all plausible MSRs were assigned by intersecting the following pairs registered in BulNet <noun literal suffix – semantic prime of the noun synset> and <noun literal suffix – MSR between the noun and a verb synset>. Then the probability for each MSR was estimated given the frequency of occurrence of the triples <MSR – noun synset semantic prime – verb synset semantic prime> in the PWN, and was used to filter out

less probable MSRs.

In a follow-up paper (Leseva et al., 2015), a decision-tree based supervised machine learning method was designed, implemented and tested for classification of MSRs. In the present paper, we upgrade the previous research along the following lines – we propose a method designed to identify new synset pairs that have a high probability of being MSR related and to classify the respective MSRs; we test new sets of features combined in different ways (as described in the experiments), which gives us insights into possible extensions and improvements of the method.

Our task is three-fold: (i) to find out potential derivational verb–noun pairs in BulNet; (ii) for a given potential derivational pair, the classifier must determine whether a derivational relation exists (or there is just a formal coincidence); (iii) if a DR exists, decide what type of MSR connects the relevant synsets.

The first part of the task was implemented by identifying common substrings shared by noun–verb literal pairs and by mapping the resulting endings to the canonical suffixes. The implementation of (ii) and (iii) was performed using a machine learning classifier. The suffixes of the noun–verb derivational pairs and the semantic primes of the verb and noun synsets were used as features in the learning, while the types of MSR between these pairs of synsets were the classes in the classification task. Our research is focused on Bulgarian but the results are transferable across languages and the methodology can be used to enhance wordnets for other languages with semantic content.

2 Linguistic Motivation

2.1 Morphosemantic Relations

MSRs hold between synsets containing literals that are derivationally related and express knowledge additional to that conveyed by semantic relations, such as synonymy, hypernymy, etc. We use the inventory of MSRs from the PWN 3.0. morphosemantic database[1] which includes 17,740 links connecting 14,877 unique synset pairs. The MSRs were mapped to the equivalent Bulgarian synsets using the cross-language relation of equivalence between synsets.

The PWN specifies 14 types of MSRs between verbs and nouns: Agent, By-means-of (inanimate Agents or Causes but also Means and possibly other relations), Instrument, Material, Body-part, Uses ((intended) purpose or function), Vehicle (means of transportation), Location, Result, State, Undergoer, Destination, Property, and Event (linking a verb to its eventive nominalisation). These relations are assigned between verb–noun synset pairs containing at least one derivationally related verb–noun literal pair, e.g., *teacher*:2 ('a person whose occupation is teaching') is the Agent of *teach*:2 ('impart skills or knowledge to'). Most of the relations correspond to or are subsumed by eponymous semantic roles (Agent, Instrument, Location, Destination, Undergoer, Vehicle, Body-part, etc.).

2.2 Semantic Primes

All the verb and noun synsets in the PWN are classified into a number of language-independent semantic primes. The nouns are categorised into 25 groups, such as noun.act (acts or actions), noun.artifact (man-made objects), etc. The verbs fall into 15 groups, such as verb.body (verbs of grooming, dressing and bodily care), verb.change (verbs of size, temperature change, intensifying, etc.), as defined in the PWN lexicographer files.[2]

2.3 Derivational Relations

Derivational relations are language specific lexical relations (between pairs of literals in related synsets). A DR may signal the existence of a morphosemantic relation between the relevant synsets, which may or may not be defined explicitly in wordnet. A DR is formally expressed by means of a (combination of) morphological device(s), such as suffixation, prefixation, suffixation plus root vowel mutation, etc.

Most suffixes in Bulgarian can be associated with more than one MSR. Consider the suffix *-ach/-yach*. Its prototypical meaning is Agent, e.g., *polivach:1* (*waterer:2* – 'someone who waters plants or crops') but also denotes an instrumental meaning, e.g., *rezach:1* (*cutter:1; cutlery:2; cutting tool:1* – 'cutting implement; a tool for cutting') and other relations, such as: Vehicle – *prehvashtach:1* (*interceptor:1* – 'a fast maneuverable fighter plane designed to intercept enemy aircraft'); Body-part – *privezhdach:1* (*adductor:1* – 'a muscle that draws a body part toward the median line'); and others.

[1] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

[2] https://wordnet.princeton.edu/man/lexnames.5WN.html

The distinction between (part of) the meanings of a suffix corresponds to a distinction in the semantic primes of the relevant noun synsets. *Polivach:1* (Agent) has the semantic prime noun.person; *interceptor:1* (Vehicle), and *rezach:1* (Instrument) bear the semantic prime noun.artifact; *privezhdach:1* (Body-part) bears the prime noun.body. We can thus derive general rules for disambiguation or partial reduction of the number of MSRs associated with the suffix. Given a derivationally related verb–noun literal pair which has not been assigned an MSR, and a relevant suffix, we are then able to rule out the MSRs possible for that suffix but not compatible with the semantic primes of the related verb and noun synsets.

3 Linguistic Preprocessing

We performed the following consistency procedures on the wordnet structure: (i) manual inspection and disambiguation of MSRs in case of multiple relations assigned to a synset pair; (ii) validation of the consistency of the semantic primes in the hypernym–hyponyms paths; (iii) consistency check of the type of the assigned MSR against the semantic primes. The quality analysis and validation is performed only on the core dataset and is language independent, i.e., it concerns the wordnet structure, rather than any language data, and is transferrable across wordnets. This is a one-off task, ensuring the quality of the data used for machine learning, as well as for any future tasks based on these data.

3.1 Disambiguation of Multiple MSRs

We identified 450 cases of multiple MSRs assigned between pairs of synsets, which represent 50 different combinations of two (rarely three) relations. As we assume that two unique concepts are linked by a unique semantic relation, we kept only one MSR per pair of synsets to ensure the consistency of the data. The following observations served as a main point of departure.

(I) **The relations are mutually exclusive** (24 combinations of MSRs). Consider the following assignments: <Agent, Destination>, <Agent, Undergoer>. Except in a reflexive interpretation, an entity cannot be an Agent, on the one hand, and a Destination (Recipient) or an Undergoer (Patient or Theme), on the other. The actual relation is signalled by the synset gloss and usually by the suffix, e.g., the choice of Agent over Destination for the pair *pensioner:2* (*retiree:1* – 'someone who has retired from active working') – *pensioniram se:2* (*retire:7* – 'go into retirement') was based both on the gloss and on the noun suffix *-er*. In other cases, e.g. <Agent, Event>, <Agent, Instrument>, the choice of relation depends on the semantic prime, as a noun.artifact or a noun.act cannot be an Agent, and vice versa – a noun.person cannot be an Instrument or an Event.

(II) **One of the relations implies or overlaps with the other** (16 combinations of MSRs). Examples of such combinations are <Instrument, Uses>, <By-means-of, Instrument>, <Body-part, Uses>. The choice is based mainly on which relation is more informative rather than abstract. For example, Instrument is preferred instead of Uses as instruments are used for a certain purpose. The semantics of the suffix, e.g. *-tel* in *usilvatel:1* (*amplifier:1*) – *usilvam:7* (*amplify:1*), also plays a role in the choice of the relation (Instrument).

(III) **No strict distinction between the semantics of the relations** (10 combinations of MSRs), e.g., <Result, Event>, <Result, State>, <Result, Material>, <State, Event>, <Property, State>. The choice is motivated on the basis of semantic information from the synsets, such as the literals, the gloss, or the semantic primes. For instance, the eventive and the resultative meaning of deverbal nouns are not always distinguished as different senses. In such case, a noun.state synset would suggest the relation Result, while a noun.act or a noun.event synset points to Event. Definitions often give additional information about the type of MSR, e.g. 'the act of...', 'a state of...', etc. especially where the semantic prime is more specific. By inspecting the triples <verb.prime–noun.prime–MSR>, we established prime combinations that strongly indicate the type of relation, e.g., <noun.state–verb.state> points to State; <noun.event/noun.process/noun.act–verb.change> – to Event. On their own, noun.act and noun.event point to Event, noun.person – to Agent, etc.

3.2 Validation of Semantic Primes

There are many hypernym–hyponym trees in which the semantic primes shift along the tree path. For instance, the majority of the 11,574 hypernyms with the prime noun.artifact have a hyponym classified as noun.artifact, but other prime labels are also found, such as noun.substance –

for nouns denoting raw materials or synthetic substances, e.g., *pina cloth:1* ('a fine cloth made from pineapple fibers'), noun.substance, is a hyponym of *fabric:1* ('artifact made by weaving or felting or knitting or crocheting natural or synthetic fibers'), noun.artifact; etc. Moreover, some synsets are linked to two hypernyms but inherit the semantic prime of one of the two, as in: *prednisolone:1* ('a glucocorticoid (trade names Pediapred or Prelone) used to treat inflammatory conditions'), noun.substance, which is hyponym to both *glucocorticoid:1*, noun.substance, and *anti-inflammatory drug:1*, noun.artifact.

The most variation in the semantic primes of the noun synsets down a hypernym–hyponym tree is observed with: noun.state (16 other primes); noun.attribute (15); noun.group (14); etc. For example, the paths down the trees with the prime noun.group on the hypernym(s) involve noun synsets with the primes noun.person (a group of persons – for example, synsets for ethnic groups, nationalities, etc.), noun.animal (a group of animals – animal taxons, etc.), noun.plant (a group of plants – plant taxons), etc.

We analysed manually the cases where hyponyms have different semantic primes from their immediate hypernym. The primes of 33 nouns labeled as noun.Tops were changed to the prime they give name to and found predominantly in their hyponyms, e.g. *state:2* was relabelled as noun.state, *process:6; physical process:1* – as noun.process, etc. 66 hyponyms' prime labels were aligned with those of their immediate hypernym in order to reflect more precisely the semantics of the words with which they are linked. For example, *dance:2* ('move in a pattern; usually to musical accompaniment; do or perform a dance') is classified as verb.creation, its hypernym *move:14* ('move so as to change position, perform a non-translational motion') has the prime verb.motion, and *dance:2*'s hyponyms are a mix of verbs with the primes verb.creation and verb.motion. As *dance:2*'s semantics is consistent with verb.motion, the semantic prime of the verb and its hyponyms (where needed) was changed accordingly.

The majority of the shifts in the semantic primes, however, reflect specific features of the hypernym–hyponym paths – for example, the shifts between noun.substance and noun.artifact, noun.body and noun.animal or noun.plant; and so forth, especially in the cases of two hypernyms.

3.3 Cross-check of Primes and MSRs

Semantic restrictions on the combinations of semantic primes and MSRs were formulated after cross-checking their compatibility (with subsequent changes either of the semantic primes of nouns and/or verbs, or of the MSR) in order to reduce the number of possible combinations of <verb.prime–noun.prime–MSR> against those from the PWN 3.0. The purpose of the procedure is to ensure consistency of the training data.

The role Agent is associated with persons (noun.person), social entities, e.g., organisations (noun.group), animals (noun.animal) and plants (noun.plant) that are capable of acting so as to bring about a result. Instruments are concrete man-made objects (noun.artifact), but nouns with the prime noun.communication – *debugger:1* and noun.cognition – *stemmer:3* which may function as instruments are also possible.

Inanimate causes (Fellbaum et al., 2009) – non-living (and non-volitional) entities that bring about a certain effect or result – are expressed by the MSRs Body-part, Material, Vehicle, and By-means-of. The relation Body-part may be an inanimate cause that is an inalienable part of an actor and is expressed by nouns with noun.body primes (rarely noun.animal or noun.plant). The relation Material denotes a subclass of inanimate causes – substances that may bring about a certain effect (e.g. *inhibitor:1* ('a substance that retards or stops an activity'). Beside noun.substance, noun.artifacts (synthetic substances or products) also qualify for the relation, e.g. *depilatory:2* (hair removal cosmetics). The relation Vehicle represents a subclass of artifacts (means of transportation); consequently the respective synsets have the prime noun.artifact and are generally hyponyms of the synset *conveyance:3; transport:8*. Inanimate causes whose semantics differ from that of the other three relations, are assigned the generic relation By-means-of, e.g. *geyser:2* ('a spring that discharges hot water and steam') (noun.object), etc.

The relation Event denotes processual nominalisation and involves nouns such as noun.act, noun.event, noun.phenomenon, and rules out concrete entities such as animate beings, natural (noun.object) and man-made (noun.artifact) objects, etc. The relation State denotes abstract entities such as feelings, cognition, etc. The relation Undergoer denotes entities which are affected by the event or state. The relation Result involves en-

tities that are produced or have come to existence as a result of the event or state. The relation Property denotes various attributes and qualities. These relations involve nouns with various primes.

The relation Location denotes a concrete (natural or man-made) or an abstract location where an event takes place and therefore relates verbs with nouns with various primes – noun.location, but also noun.object, noun.plant, noun.artifact, noun.cognition, etc. The relation Destination is associated with the primes noun.person, noun.location and noun.artifact, which corresponds to two distinct interpretations of the relation – Recipient (noun.person) and Goal (noun.artifact, noun.location). The relation Uses denotes a function or purpose, e.g. *lipstick:1* – *lipstick:3*. The relation allows nouns with various primes, both concrete and abstract.

We examined the combinations of noun primes and MSRs in the PWN 3.0. with a view to the semantic restrictions and in some cases MSRs were modified accordingly. For instance, some noun.body nouns were originally assigned the relation Instrument, some noun.person – Event, etc. As a result, the noun primes associated with a given MSR were reduced: Agent from 17 to 4 (person, animal, plant, group); Instrument – from 9 to 3 (artifact, communication, cognition); Material – from 6 to 2 (artifact, substance); State – from 10 to 5 (state, feeling, attribute, cognition, communication); Body-part – from 4 to 3 (body, animal, plant); Event – from 24 to 13 (act, communication, attribute, event, feeling, cognition, process, state, time, phenomenon, group, possession, relation). Result, Property, By-means-of, Uses, Location, and Undergoer are more heterogeneous and few of the semantic primes were ruled out. The relations Vehicle and Destination and the corresponding semantic primes need not be subject to any changes.

The reduction of the noun.prime–verb.prime combinations for a given MSR rules out the corresponding branches in the decision trees.

The changes made in the relations and semantic primes in these validation procedures are available at: http://dcl.bas.bg/en/wordnetMSRs.

4 Training Data for the ML Task

4.1 Core data

The core training data include examples for which we are sure an MSR exists, and we know the type of the relation. The dataset comprises a total of 6,641 literal pairs in 4,016 unique synset pairs, and was compiled in two stages.

Initially, the core dataset included 6,220 instances of derivationally related verb–noun literal pairs in the BulNet verb–noun synset pairs (automatically detected and manually validated as described in Dimitrova et al. (2014)) which were assigned an MSR by automatic transfer from the PWN. We took into consideration the pairs obtained by suffixation and zero derivation.

We supplemented the core data with additional instances from BulNet extracted in the following way: (1) we identified literal pairs from BulNet which exhibited a possible DR but an MSR had not been assigned between the respective synsets; (2) after measuring the similarity of the disambiguated PWN glosses[3] for the pairs of synsets identified in step (1) using a wordnet-based measure for text similarity (Mihalcea et al., 2006), we filtered out the low similarity pairs (below threshold of 2.0); and (3) the glosses of high similarity were examined for certain structural patterns in order to determine the MSR where possible (e.g., a gloss of the type 'someone who <verb,active voice>' points to Agent, or 'instrument used for <verb>ing' – points to Instrument). As a result, 421 additional instances of morphosemantically related literal pairs were added to the core dataset.

4.2 Negative Examples Dataset

The task of determining whether an MSR holds between a given verb–noun pair is a binary classification task where the classes are *true* and *false*. To be able to train a classifier for this task, we needed a set of examples of class *false*, i.e. instances of (potentially) derivationally related verb–noun literal pairs which did not have an MSR. This can be due to various reasons: (a) one of the words has acquired an additional, usually metaphorical, meaning; (b) the similarity in the form of the noun and the verb literals is coincidental (due to historical changes in the forms, etc.) and there is no transparent DR; or (c) the relation does not fit into the pre-designed system of relations in PWN.

The negative examples were extracted automatically from BulNet and include: (i) (potentially) derivationally related verb–noun literal pairs from synsets which have mutually exclusive seman-

[3] http://wordnet.princeton.edu/glosstag.shtml

tic primes (i.e., not occurring among MSR pairs in PWN) and thus cannot be semantically related, e.g., verb.weather – noun.animal; and (ii) verb–noun literal pairs linked by a DR but not by an MSR in BulNet which formally coincide with pairs of literals that have an MSR in BulNet. For example, the literal *gotvya* is a member of the synsets *gotvya:2* (*cook:1* – 'transform and make suitable for consumption by heating', verb.change) and *gotvya:4* (*prepare:6* – 'to prepare verbally, either for written or spoken delivery', verb.creation). The noun synset *gotvach:1* (*cook:6* – 'someone who cooks food', noun.person) derived from the verb *gotvya* bears an MSR (Agent) only to *gotvya:2*, thus the pair *gotvach:1 - gotvya:4* is extracted as a negative example.

A total of over 170,000 negative instances (verb–noun literal pairs) were extracted from BulNet. As the number and quality of the negative examples (and the number of training instances in general) affect the performance of the classifier, they usually need to be balanced against the number of positive examples and only a selection of roughly the same number as positive data were applied in each task.

4.3 Preprocessing of the Data

The Bulgarian synsets connected with MSRs from the PWN were processed using previously proposed methods and datasets. The derivationally related literal pairs found in the MS-related synsets were assigned an appropriate DR, following Dimitrova et al. (2014). The particular derivational devices were automatically established and manually validated, and the variants of the affixes (suffixes in particular) were associated with a canonical suffix form, as proposed in Leseva et al. (2014).

As a first step, the word endings of each pair of verb–noun literals were identified by removing the common substring (base) shared by the two literals. In order to discard pairs that coincide in form by chance, the base was set to be at least 75% of each literal's length. Secondly, as the endings usually do not coincide with a literal's suffix (may also include part of the literal's root or stem), they were mapped to the canonical forms of the suffixes using lists of suffixes with their contextual variants. The training data contain 294 different noun endings, which were mapped to 121 canonical noun suffixes, and 172 verb endings mapped to 44 canonical verb suffixes.

In this way the number of suffix values for each MSR is reduced, while the number of examples per relation and pair of semantic primes increases, thus reducing the noise in the data that arises from the contextual suffix variants.

5 ML Method for Identification of MSRs

5.1 Features

The following features were used in the analysis of the data: (i) the canonical verb suffix; (ii) the canonical noun suffix; (iii) the semantic prime of the verb; and (iv) the semantic prime of the noun. Our data are in string format but the sets of values for both the canonical suffixes (these 121 noun and 44 verb suffixes) and the synset primes (25 semantic primes for nouns and 15 primes for verbs) are finite.

Additional features were also considered and tested such as the similarity between the glosses of the verb–noun synset pair, which was in the end disregarded due to the fact that only a limited number of instances exhibit similarity above the threshold. Instead, these examples were used to extend the training data (see section 4.1).

5.2 Implementation

The implementation of the Machine Learning is made in Java using the Weka library (Witten et al., 2011), which offers various capabilities and advanced techniques for data mining.[4]

We analysed and tested various classifiers within the Weka package in order to select the best performing one suitable for the task – decision tree algorithms, Naive Bayes classifier, K* classifier, SMO (Sequential Minimal Optimisation), linear logistic regression, etc., as well as some complex classifiers applying several algorithms in a sequence. The Naive Bayes classifier was not suitable due to the data scarcity and the fact that not all combinations of feature values were covered in the data. The K* classifier relies on an entropy-based distance measure between instances and is not particularly suitable for string and nominal data. The decision tree was considered most relevant to the task. After comparing empirically several decision tree classifiers in Weka, based on the performance evaluation using 10-fold cross-validation, we selected the algorithm of RandomTree which consistently outperformed the rest. The decision tree

[4] http://www.cs.waikato.ac.nz/ml/weka/

built by the RandomTree algorithm on each node tests a given number of random features and no pruning is performed. As a baseline, we applied on the same dataset the OneR classifier which chooses one parameter best correlating with the class value to provide best prediction accuracy, and which is particularly suited for discrete data.

Three approaches were considered with a view to the method of classification. The first one uses two separate classifiers applied in a sequence – first, a binary classifier that identifies pairs of derivationally related verb–noun literals in synsets linked via an MSR, and then, a multiclass classifier that selects the type of relation. The second approach merges the above two classifiers and applies a single multiclass classifier to assign MSRs, where the set of classes includes an additional value *null* for the instances which do not have an MSR. The third method combines a set of separate binary classifiers for each of the 14 MSRs. A verb–noun pair can be assigned more than one relation, or none (in the latter case the pair is considered unrelated). The results are presented in the following section.

5.3 Experiments

Test 1. The first experiment tests the performance of the approach which first discovers whether a verb-noun pair has an MSR, and subsequently applies a multiclass classifier to assign a particular relation to the pair. The core dataset extended with negative examples is used as training data for the binary classifier, and the classes are 'true' (there is an MSR) and 'false' (no MSR). The RandomTree classifier shows an F_1 score of 0.815 (compared to the baseline of 0.687) using 10-fold cross-validation.

The multiclass classifier is trained on the core dataset and the classes are represented by the 14 MSRs. Its F_1 score on 10-fold cross-validation is 0.842 (baseline 0.808) but varies considerably across different classes: from as high as 0.975 for Agent down to 0.333 for By-means-of (relations with less than 10 examples in the data are not considered reliable).

The F_1 score of the overall method is 0.682 since the error propagates from one phase to another. Results also show that for certain MSRs the OneR algorithm performs slightly better than the RandomTree (usually RandomTree outperforms OneR by more than 25%), which suggests that a more complex approach combining case-specific classifiers may prove more reliable.

Test 2. The second experiment tests a classifier with a list of 15 classes – the 14 MSRs and the class *null* used to label instances with no MSR. The training data include the core dataset supplemented with a limited number (6,700) of randomly selected negative examples. The results from the 10-fold cross-validation show F_1 score of 0.769 (baseline 0.654), which is significantly better than the results in Test 1. The performance also varies across relations: the highest rate is for true negatives (0.811), State (0.809), Agent (0.788), etc. In this case the RandomTree classifier significantly outperforms the baseline for all relations.

The experiment raises the question whether the negative data should be selected at random, or the training data should conform to certain selection criteria aiming at representativeness of the patterns and varieties in terms of feature values and combinations between them. Tests in this direction might be considered in the future.

Test 3. The third test examines the performance of a complex classifier combining a set of separate binary classifiers for each type of relation between a noun and a verb: there is a binary classifier (true/false) for Agent, another for Undergoer, etc. This method allows assignment of more than one relation to a given pair. In this way we can observe when uncertainty or ambiguity occurs and look for ways to tackle it. When no relation is assigned, the pair is considered unrelated. The core dataset was applied for the training of the model. In this case, for each MSR, the subset of this relation's instances constitutes the positive dataset, and the subset of instances of other relations serves as a set of negative examples.

If we look for exact matches, the results are lower: F_1 score varies from 0.81 (Agent, Event) down to $0.30 - 0.35$ (Result, By-means-of, etc.). But since in this method more than one MSR can be assigned, we can evaluate whether the correct relation is in the set of assigned relations.

The method was also tested on a dataset of 300 new examples having a DR or formally coinciding with a DR, independently extracted from BulNet (not used in the training data), preprocessed and having their class (or lack of an MSR) manually verified. Using the complex classifier, we obtained the following results: (i) exact matches are 64.00%, (ii) in another 3.33% the real class

Test	Baseline (OneR)	Random Tree
Test 1		
MSR true-false	0.687	0.815
Type of MSR	0.808	0.842
Overall	0.498	0.682
Test 2	0.654	0.769
Test 3		
Exact MSR	0.653	0.713
MSR in set	0.699	0.746
Reclassify *null*	0.710	0.781

Table 1: Evaluation results: F_1 score on the 10-fold cross-validation in Tests 1-3.

is contained in the set of guessed relations, (iii) 28.33% of the test instances are labelled as *null* while in fact they have an MSR, and (iv) the remaining 4.33% comprise incorrectly assigned relations.

The large amount of instances incorrectly labelled as *null* (28.33%) points to the need to either introduce more features to fine-tune the classifier, or to apply an additional classifier on these data using a different method, and merge results. We ran a second classifier on all data labelled by the first classifier as *null*, using only the noun semantic prime as a feature in order to assign the most probable relation according to the semantic prime of the noun. In this case the precision increased to 78.13% by taking the most frequent relation associated with each noun prime. However, in this case we assign an MSR to all test instances, thus mislabel true negatives correctly recognised by the first classifier. A more fine-tuned method and feature design, as well as training on different sets/features in each phase, may be more effective.

5.4 Follow-up

In further tests we experimented with variations in the data, i.e., addition of new training data instances exhibiting specific features. To this end, we assigned a second semantic prime to the synsets which either have two hypernyms (with two different semantic primes) and inherit the prime of only one of the two, or have a hypernym with another, different semantic prime which does not clash with the semantic prime of the hyponym – see the observations in 3.2. The purpose was to test whether the inherited semantic prime impacts the result. For instance, the assignment of a second prime noun.substance to synsets denoting synthetic substances or raw materials (noun.artifact) is expected to make the data more consistent as these noun.artifact synsets are more alike substances as regards the choice between certain relations, e.g., Material and Instrument. At present this shows only an insignificant increase in precision due to the small amount of data affected. However, with the increase of training data in the future, the number of added instances may increase as well, which can potentially yield significant improvement.

The observations on the constructed decision trees also show that the features are insufficient to fully distinguish between different MSRs as the tree structures are too shallow to achieve better results. By introducing more features, we can also test the RandomForest classification method which requires more features in order to construct a properly sized forest of RandomTree classifiers and usually outperforms the singular RandomTree method. If several learning schemes are available, it may be advantageous not to choose the best-performing one for a dataset but to use all of them and merge the results.

6 Conclusion and Future Work

Our future work will be focused on the enhancement of the method by exploring at least two mutually related directions: (i) automatic harvesting of more labelled data from other wordnets; (ii) incorporation of new features for classification and assignment of relations including heuristics derived from the WordNet structure.

Alongside the introduction of new features, it is necessary to develop techniques for reducing redundant features, as well as for correlation-based feature selection, feature ranking or principal component analysis.

We have devised experiments to extend the datasets with more data for English and Romanian. The multilingual data can contribute to the training with respect to the possible pairs of verb–noun primes and the relevant semantic restrictions.

While part of the information employed in this paper, such as the suffix lists and mappings from word endings to canonical suffixes, is language specific, the method proposed is language independent, including the linguistic processing of the data. Testing it for other languages is a task we envisage to implement in the future.

References

Verginica Barbu Mititelu. 2012. Adding morpho-semantic relations to the Romanian Wordnet. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2596–2601.

Verginica Barbu Mititelu. 2013. Increasing the effectiveness of the Romanian Wordnet in NLP applications. *Computer Science Journal of Moldova*, 21(3):320–331.

Orhan Bilgin, Ozlem Cetinoglu, and Kemal Oflazer. 2004. Morphosemantic relations in and across Wordnets – a study based on Turkish. In *Proceedings of the Second Global Wordnet Conference (GWC 2004)*, pages 60–66.

Tsvetana Dimitrova, Ekaterina Tarpomanova, and Borislav Rizov. 2014. Coping with derivation in the Bulgarian WordNet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, pages 109–117.

Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting semantics into WordNet's "morphosemantic" links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics]*, volume 5603, pages 350–358.

Neeme Kahusk, Kadri Kerner, and Kadri Vider. 2010. Enriching Estonian WordNet with derivations and semantic relations. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 195–200, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Svetla Koeva, Cvetana Krstev, and Dusko Vitas. 2008. Morpho-semantic relations in Wordnet – a case study for two Slavic languages. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 239–254.

Svetla Koeva. 2008. Derivational and morphosemantic relations in Bulgarian Wordnet. *Intelligent Information Systems*, pages 359–368.

Svetlozara Leseva, Ivelina Stoyanova, Borislav Rizov, Maria Todorova, and Ekaterina Tarpomanova. 2014. Automatic semantic filtering of morphosemantic relations in WordNet. In *Proceedings of CLIB 2014, Sofia, Bulgaria*, pages 14–22.

Svetlozara Leseva, Maria Todorova, Tsvetana Dimitrova, Borislav Rizov, Ivelina Stoyanova, and Svetla Koeva. 2015. Automatic classification of wordnet morphosemantic relations. In *Proceedings of BSNLP 2015, Hissar, Bulgaria*, pages 59–64.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press.

Karel Pala and Dana Hlaváčková. 2007. Derivational relations in Czech WordNet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81.

Maciej Piasecki, Stanislaw Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground up*. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej.

Maciej Piasecki, Radoslaw Ramocki, and Marek Maziarz. 2012a. Automated generation of derivative relations in the Wordnet expansion perspective. In *Proceedings of the 6th Global Wordnet Conference (GWC 2012)*, pages 273–280.

Maciej Piasecki, Radoslaw Ramocki, and Pawel Minda. 2012b. Corpus-based semantic filtering in discovering derivational relations. In A. Ramsay and G. Agre, editors, *Applications – 15th International Conference, AIMSA 2012, Varna, Bulgaria, September 12-15, 2012. Proceedings. LNCS 7557*, pages 14–22. Springer.

Ivelina Stoyanova, Svetla Koeva, and Svetlozara Leseva. 2013. Wordnet-based cross-language identification of semantic relations. In *Proceedings of the 4th Biennal International Workshop on Balto-Slavic Natural Language Processing*, pages 119–128.

Krešimir Šojat and Matea Srebačić. 2014. Morphosemantic relations between verbs in Croatian WordNet. In *Proceedings of the Seventh Global WordNet Conference*, pages 262–267.

Ian Witten, Eibe Frank, and Mark Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Tuning Hierarchies in Princeton WordNet

Ahti Lohk
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
ahti.lohk@ttu.ee

Christiane D. Fellbaum
Department of Computer Science
Princeton University
New Jersey, USA
fellbaum@princeton.edu

Leo Võhandu
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
leo.vohandu@ttu.ee

Abstract

Many new wordnets in the world are created and most take the original Princeton WordNet (PWN) as their starting point. This arguably central position imposes a responsibility on PWN to ensure that its structure is clean and consistent. To validate PWN hierarchical structures we propose the application of a system of test patterns. In this paper, we report on how to validate the PWN hierarchies using the system of test patterns. In sum, test patterns provide lexicographers with a very powerful tool, which we hope will be adopted by the global wordnet community.

1 Introduction and background

Many new wordnets in the world are created and most take the original Princeton WordNet (PWN) as their starting point. This arguably central position imposes a responsibility on PWN to ensure that its structure is clean and consistent. This is particularly true for hierarchical relations, which are the most frequently encoded relations and which form the backbone of the network. To validate PWN hierarchical structures we propose the application of a system of test patterns developed in (Lohk, 2015). Importantly, all instances returned by the test pattern system were manually validated by two members of the Estonian Wordnet (EstWN) team (Kadri Vare and Heili Orav). The results were encouraging, and we applied the algorithms to PWN. We propose that after couple of iterations on PWN other wordnets apply the algorithm on their resources and, after a couple of iterations, compare their structures with that of PWN, which can serve as some kind of Gold Standard for wordnets. Alternatively, the analysis is commercially available from the first author.

In this paper we report on how to validate the PWN hierarchies using the system of test patterns.

A test pattern is a description of a specific substructure in the wordnet hierarchy. The system of test patterns and the descriptions of all patterns are found in (Lohk, 2015). This system consists of ten test patterns that all involve multiple inheritance, an important property that can point to different semantic inaccuracies going back to lexicographic errors. Because it is semantic, every test pattern applies cross-lingually and sheds new light on wordnets by examining their hierarchies and helping to detect and correct possible errors.

These patterns were used to validate the semantic hierarchies of Estonian Wordnet over four years (2011–2014) and on ten versions. During this time the structure of Estonian Wordnet structure changed significantly, as described in Section 3.

The aim of this paper is to show that the same specific substructures that have been found in Estonian Wordnet also exist in Princeton WordNet. Moreover, some experiments on Princeton WordNet confirm the promising benefits of test pattern application (Section 4). Therefore, we propose test patterns as a method for validation and tuning hierarchies in PWN and all other wordnets.

This paper is structured as follows: Section 2 provides an overview of the validation methods applied to the wordnet hierarchies. Section 3 presents the results of using test patterns iteratively on EstWN. Section 4 demonstrates that the same pattern instances can be found in PWN as well as in other wordnets. Some experiments are described. We close with a conclusion and proposals for future work.

2 State of the art in validating the semantic hierarchies of wordnet

To give a better understanding of the test patterns approach we provide a short overview of the validation methods applied on the semantic hierarchies of wordnet. (Lohk 2015) argues that the methods can be divided into three groups based on two features, as shown in Table 1. These features can be formulated as questions as follows: *do they rely on corpus data and lexical resources? Do they make use the contents of a synset?*

Group of methods	use of corpus data, lexical resources	use the contents of a synset
Group I	+	+
Group II	–	+
Group III	–	–

Table 1: Features that classify a group of validating methods

Group I comprises all methods based *on lexical resources and corpora*; group II includes rules or rule-based methods, while group III consists of graph-based methods.

2.1 Corpus-based methods

The most frequently used validation methods for wordnet hierarchies rely on corpora and lexical resources. Different techniques for extracting the relevant information have been applied. Some of the well-known approaches include:

- Lexico-syntactic patterns (Hearst, 1992), (Nadig et al., 2008)
- Similarity measurements (Sagot and Fišer, 2012)
- Mapping and comparing to wordnet (Pedersen et al., others, 2013)
- Applying wordnet in NLP tasks (Saito et al., 2002)

Resources used in this group of methods are:

- Monolingual text corpora (Sagot and Fišer, 2012)
- Bilingual aligned corpora (Krstev et al., 2003)
- Monolingual explanatory dictionaries (Nadig et al., 2008)
- Wordnets (Peters et al., 1998; Pedersen et al., 2012)
- Ontologies (Gangemi et al., 2002)

2.2 Rule-based methods

These methods for validating hierarchies rely on lexical relations (word-word), semantic relations (concept-concept) and the rules among them. This includes the rules applied to the construction of WordNet (Fellbaum, 1998), and additional rules, such as the following:

- Metaproperties (*rigidity, identity, unity and dependence*) described in ontology construction (Guarino and Welty, 2002)
- Top Ontology concepts or "unique beginners" (*Object, Substance, Plant, Comestible, ...*) (Atserias et al., 2005; Miller, 1998)
- Specific rules for particular error detections (Gupta, 2002; Nadig et al., 2008). For instance, a rule proposed by (Nadig et al., 2008): *"If one term of a synset X is a proper suffix of a term in a synset Y, X is a hypernym of Y"*

2.3 Graph based methods

These methods are purely formal and do not take into account the semantics among word forms. Specific substructures of wordnet's hierarchies are checked and validated. Target substructures include:

- Cycles (Šmrz, 2004), (Kubis, 2012)
- Shortcuts (Fischer, 1997)
- Rings (Liu et al., 2004; Richens, 2008)
- Dangling uplinks (Koeva et al., 2004; Šmrz, 2004)
- Orphan nodes (null graphs) (Čapek, 2012).
- Small hierarchy (Lohk et al., 2014c)
- Unique beginners (Lohk et al., 2014c)

In addition, (Lohk, 2015) proposed different yet undiscovered substructures and shows that the application of these substructures to validate the semantic hierarchies of wordnet may improve wordnet structure significantly. While these substructures with the specific nature are used in wordnet assessment, they are called **test patterns**. Next, we explain the idea of test pattern and demonstrate their efficient use with Estonian Wordnet.

3 A case study: applying test patterns to Estonian Wordnet

Since 2011, the different type of test patterns have been developed and applied progressively to EstWN. Currently, ten test patterns exist. For

every test pattern we implemented a program to find the relevant instances. Four programs are implemented for semi-automatic application (*closed subsets, closed subset with a root, the largest closed subset and connected roots*) and six for automatic use (the test patterns shown in italics in Table 2). Instances found with test patterns using programs for semi-automatic application have been discussed in elsewhere (Lohk, 2015). Test patterns' instances found with programs for automatic use are employed in process of constant validation.

Version	Noun roots	Verb roots	Multiple inheritance cases	*Shortcut*	*Ring*	*Synset with many roots*	*Heart-shaped sub-structure*	*Dense component*	*„Compound" pattern*
60	142	24	1,296	235	3,445	1,123	1,825	104	301
61	183	22	1,592	259	3,560	1,309	1,861	121	380
62	102	16	1,700	299	3,777	1,084	1,941	128	415
63	114	16	1,815	**321**	3,831	1,137	2,103	141	447
64	149	15	1,893	**337**	3,882	1,173	2,232	149	471
65	248	14	1,717	**194**	2,171	791	**451**	132	459
66	144	4	1,677	**119**	1,796	613	259	**121**	671
67	129	4	1,164	**79**	928	477	**167**	**24**	**407**
68	131	4	691	**60**	**537**	232	**38**	**18**	**54**
69	121	4	102	**18**	**291**	35	**1**	**8**	23
70	118	4	51	**7**	**21**	**30**	**0**	**3**	**7**

Table 2: A numerical overview of EstWN spanning eleven version

Table 2 shows the number of instances that each test pattern returned after its automatic application. The first two patterns (*shortcut* and *ring*) are inspired by (Fischer, 1997; Liu et al., 2004; Richens, 2008). There are also some cases of *synset with many roots*, called *dangling uplinks* in (Koeva et al., 2004) and (Šmrz, 2004). Bold font in the table shows when the test pattern was given to a lexicographer for verification. For example, the "shortcut" cases where lexicographers who verified each instance manually in the 63rd version submitted to the EstWN. The effect, as reflected in the next version, can be clearly seen in the table. It is clear that the application of heart-shaped substructure and dense component patterns had a considerable effect on the lexicography.

As all instances of test patterns include multiple inheritance cases, the fourth column (Multiple inheritance cases) demonstrates the influence of using test patterns most clearly. For example, a comparison between versions 66 and 70 shows that the number of cases has gone down about 32 times (97%). Note that the number 118 of hierarchies has about 75% of shallow hierarchies where roots are connected to only one level of subordinates.

According to (Lohk, 2015) over ten versions of EstWN the most popular correction operation has been removing the hypernymy and hyponymy relations – 21,911 times. Secondly, 5,344 times the lexical units in synsets were changed (included deleted and added lexical units). Thirdly, 4,122 times hypernymy and hyponymy relation were replaced by another semantic relation, mainly by near *synonymy* and *fuzzynymy*.

4 Validating Princeton WordNet

Substructures connected with multiple inheritances have been used to validate PWN. (Fischer, 1997; Liu et al., 2004 and Richens, 2008) examined ***shortcuts; ring***s were suggested by (Koeva et al., 2004), and (Šmrz, 2004) examined ***dangling uplinks***. There are also some examples about ***closed subset***s in (Lohk et al., 2012) and one example of ***heart-shaped substructure*** in (Lohk and Võhandu, 2014). Lohk gives an example of a ***connected roots*** case in his poster presentation at Estonian Applied Lingvistics Conference in Tallinn on April 2013.

Next we demonstrate some examples of test patterns' instances to see their structure and how they may help to discover specific incosistencies in PWN semantic hierarhies. The complete overview of the test patterns has been given in the dissertation of first author (Lohk, 2015).

4.1 Shortcut

Shortcut is a pattern where a synset (based on **Figure 1**, {event}) is simultaneously connected to another synset ({group action}) directly and indirectly. In that case, {group action} is not an ambiguous concept. Instead of this, it just contains a redundant link (dotted line).

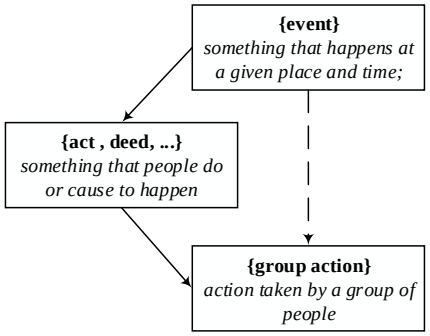


***Figure 1**. An instance of shortcut, PWN (version 3.1)*

4.2 Heart-shaped substructure

In a *heart-shaped substructure*, two nodes (based on **Figure 2**, {hard drug} and {cannabis, …}) have direct connection through an identical parent ({controlled substance}) and an indirect connection through a semantic relation {soft drug} – {narcotic}) that links their second parent.

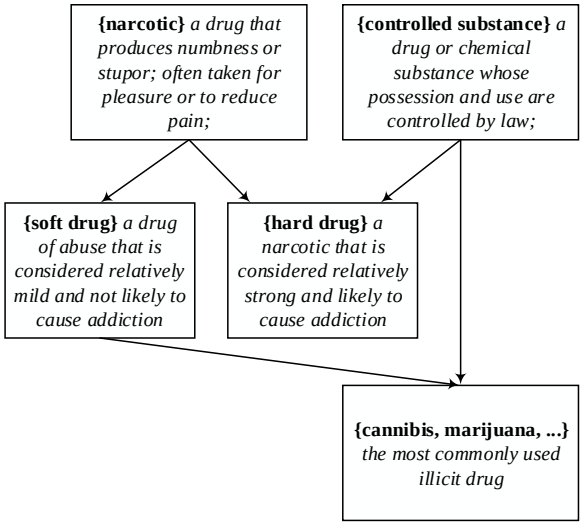


***Figure 2.** An instance of heart-shaped substructure, PWN (version 3.1)*

In case of PWN we have seen that the instances of *heart-shaped substructure* tend to show the cases where instead of *role* or *type* relation the *hypernymy* is used. That kind of example is presented in **Figure 2**, where {hard drug} is actually a certain *type* of {narcotic} and as well as in the *role* of {controlled substance}.

It is remarkable that first time when *heart-shaped substructure* was used in EstWN the number of its instances was 451 (see **Table 2**) and 5 versions later 0. Moreover, during the correction operations no *hypernymy/hyponymy* relation was changed to *role* or *type* relation (Lohk, 2015).

4.3 "Compound" pattern

"*Compound*" pattern is an exception among other test patterns while it considers the content of synsets. More precisely, that kind of substructure satisfies the following two conditions:

At first, this substructure contains a case where a lexical unit of a superordinate (based on **Figure 3** {ball}) is connected to minimum two subordinates (1-{baseball}, 2-{basketball}… 24-{volleyball}) which contain that lexical unit (*ball*).

Secondly, at least one subordinate has an extra superordinate ({baseball equipment}, {basketball equipment}, …, {golf equipment}).

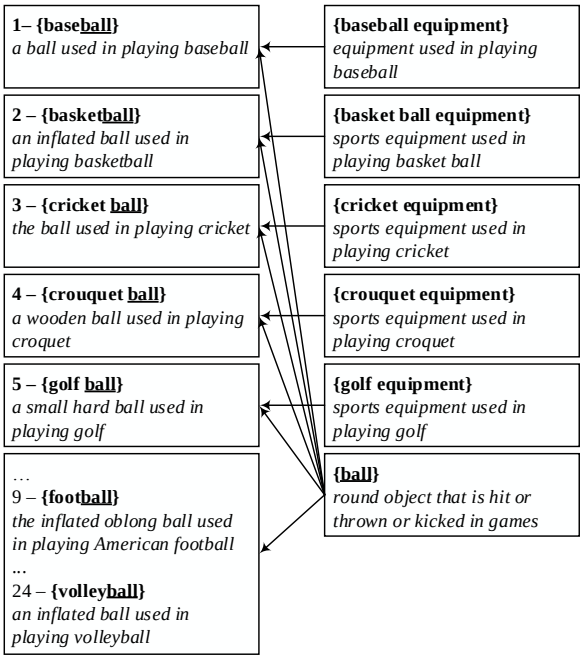


***Figure 3.** An instance of "compound" pattern, PWN (version 3.1)*

To validate that kind of instance as it is in **Figure 3**, the lexicographer has to ask if subordinates 1 to 5 have an extra superordinate, and why it is not true about subordinates from 6 to 24. Studying this figure more carefully, we see that {basketball} is a {basketball equipment}. However, {football} and {volleyball} being quite similar in their definitions do not follow the same logic.

That is to say, {football} and {volleyball} are not equipment.

4.4 Dense component

The *dense component* pattern provides the opportunity to uncover substructures where, due to the *multiple inheritance*, the density of the interrelated concepts in the semantic hierarchy is higher (Lohk et al., 2014a), (Lohk et al., 2014b). This substructure (subgraph) consists of two *synsets* (nodes) (based on Figure {manicure} and {pedicure}) with at least two identical parents (it corresponds to *complete bipartite graph*) ({beauty treatment} and {aid, attention, care, …}). The overall size of an instance of a *dense component* depends on how many *synsets* (nodes) with at least two parents are interconnected through the *multiple inheritance* and/or same parents (Lohk, 2015).

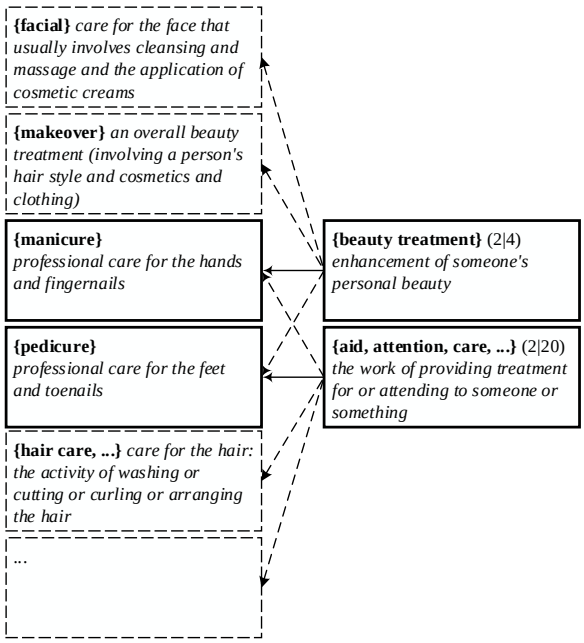


Figure 4. *An instance of dense component, PWN (version 3.1)*

In **Figure 4**, the pattern of dense component is emphasized with bold lines. While this substructure contains at least two multiple inheritance cases we see it as a case of the regularity of multiple inheritance. Herewith, the aim of the dense component is to help to detect if this regularity is justified or vice versa, if this regularity has to be expanded.

In the case of **Figure 4**, the regularity of multiple inheritance has to be expanded. Two reasons for that are concepts {facial} and {hair care, …}. In addition to {beauty treatment}, {facial} fits in with {aid, attention, care, …}. Moreover, {hair care} is a {beauty treatment} beside the {aid, attention, care, …}.

4.5 Connected roots

The *connected roots* test pattern involves different hierarchies through multiple inheritance cases.

This pattern helps to see how big and deep are the connections between POS hierarchies. Every node acts as a unique beginner is equipped with the number of hierarchy levels and the number of subordinates in the same hierarchy (Figure 1). The first number of the edge label indicates the number of common subordinates for two hierarchies. The next two numbers separated by "|" denote the hierarchy levels where the first common concept is located in both hierarchies.

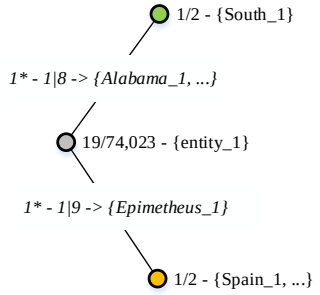


Figure 5. *An instance of connected roots, PWN (version 3.1)*

In **Figure 5**, there is only one large hierarchy with the unique beginner {entity}. It heads a 19-level hierarchy and 74,023 subordinates. By contrast, the two hierarchies ({South_1} and {Spain_1 …}) are very small. They both dominate only one additional level. The edge labels reveal that the common concepts of both hierarchies are on the first lower levels in both of the smaller hierarchy cases. Both unique beginners ({South_1} and {Spain_1}) seem to be too specific to be the highest concepts.

Table 3 presents a comparison between PWN's structure with that of other wordnets.

4.6 Short numerical overview of the test patterns' instances

In **Table 3**, it is easy to see that the wordnets are very different. Finnish Wordnet was manually translated from PWN (Lindén and Niemi, 2014) so it is not surprising that first two rows are essentially identical.

The table shows a clear need for a deep structural analysis of all wordnets. Of course, it must be remembered that the hierarchies of different

languages will never show a one-to-one correspondence, as the lexicons necessarily differ.

Version	Noun roots	Verb roots	Multiple inheritance cases	*Short cut*	*Ring*	*Synset with many roots*	*Heart-shaped sub-structure*	*Dense component*	*„Compound" Pattern*
Princeton WordNet, v3.0	12	334	1,453	40	2,991	18	155	115	358
Finnish Wordnet, v2.0	12	334	1,453	40	2,991	18	155	115	394
Cornetto, v2.0	2	2	2,438	351	5,309	62	1,226	217	549
Polish Wordnet, v2.0	637	42	10,942	553	57,887	205,254	5,037	778	541
Estonian Wordnet, v70	118	4	51	7	21	30	0	3	7

Table 3: Five wordnets in comparison

5 Conclusions

Test patterns are a unique form of validating hierarchies. They are not language-specific and can be applied cross-lingually. Their value lies in aiding lexicographers to detect and correct errors and thus provide more accurate resources.

Every test pattern has the property of multiple inheritance. In the most cases, behind the multiple inheritance there is a lexical polysemy, except the pattern of *shortcut* (Sec. 4.1).

Multiple inheritance is not always wrong. However, PWN contains so far many cases where instead of *role* or *type* relation the *hypernymy* relation has been used. This is one reason, why sometimes multiple inheritance cases are presented in PWN (see Figure 2).

In sum, the analysis of wordnet structures using test patterns provides lexicographers with a very powerful tool, which we hope will be adopted by the global wordnet community.

Reference

Atserias Batalla, J., Climent Roca, S., Moré López, J., Rigau Claramunt, G., 2005. A Proposal for a Shallow Ontologization of WordNet. Proces. Leng. Nat. Nº 35 Sept 2005 Pp 161-167.

Čapek, T., 2012. SENEQA-System for Quality Testing of Wordnet Data, in: Proceedings of the 6th International Global Wordnet Conference. Toyohashi University of Technology, Matsue, Japan, pp. 400–404.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database, MIT Press. ed. Wiley Online Library, Cambridge, USA.

Fischer, D.H., 1997. Formal Redundancy and Consistency Checking Rules for the Lexical Database WordNet 1.5, in: Workshop Proceedings of on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Association for Computational Linguistics (ACL), Madrid, Spain, pp. 22–31.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L., 2002. Sweetening Ontologies with DOLCE, in: Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Springer, pp. 166–181.

Guarino, N., Welty, C., 2002. Evaluating Ontological Decisions with OntoClean. Communications of the ACM - Ontology: Different Ways of Representing the Same Concept 45, 61–65.

Gupta, P., 2002. Approaches to Checking Subsumption in GermaNet, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Las Palmas, Canary Islands, Spain, pp. 8–13.

Hearst, M.A., 1992. Automatic Acquisition of Hyponyms from Large Text Corpora, in: Proceedings of the 14th Conference on Computational Linguistics - Volume 2, COLING '92. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, pp. 539–545.

Koeva, S., Mihov, S., Tinchev, T., 2004. Bulgarian Wordnet–Structure and Validation. Romanian J. Inf. Sci. Technol. 7, 61–78.

Krstev, C., Pavlović-Lažetić, G., Obradović, I., Vitas, D., 2003. Corpora Issues in Validation of Serbian Wordnet, in: Matoušek, V., Mautner, P. (Eds.),

Text, Speech and Dialogue, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 132–137.

Kubis, M., 2012. A Query Language for WordNet-Like Lexical Databases, in: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (Eds.), Intelligent Information and Database Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 436–445.

Liu, Y., Yu, J., Wen, Z., Yu, S., 2004. Two Kinds of Hypernymy Faults in WordNet: the Cases of Ring and Isolator, in: Proceedings of the 2nd Global Wordnet Conference. Brno, Czech Republic, pp. 347–351.

Lohk, A., 2015. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Tallinn University of Technology, Tallinn, Estonia.

Lohk, A., Allik, K., Orav, H., Võhandu, L., 2014a. Dense Component in the Structure of Wordnet, in: Proceedings of the 9th International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 1134–1139.

Lohk, A., Norta, A., Orav, H., Võhandu, L., 2014b. New Test Patterns to Check the Hierarchical Structure of Wordnets, in: Information and Software Technologies. Springer, pp. 110–120.

Lohk, A., Orav, H., Võhandu, L., 2014c. Some Structural Tests for WordNet with Results. Proceedings of the 7th Global Wordnet Conference 313–317.

Lohk, A., Vare, K., Võhandu, L., 2012. First Steps in Checking and Comparing Princeton Wordnet and Estonian Wordnet, in: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. Association for Computational Linguistics (ACL), pp. 25–29.

Lohk, A., Võhandu, L., 2014. Independent Interactive Testing of Interactive Relational Systems, in: Gruca, D.A., Czachórski, T., Kozielski, S. (Eds.), Man-Machine Interactions 3, Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 63–70.

Miller, G.A., 1998. Nouns in WordNet, in: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts, USA, pp. 24–45.

Nadig, R., Ramanand, J., Bhattacharyya, P., 2008. Automatic Evaluation of WordNet Synonyms and Hypernyms, in: Proceedings of ICON-2008: 6th International Conference on Natural Language Processing. CDAC Pune, India.

Pedersen, B.S., Borin, L., Forsberg, M., Kahusk, N., Lindén, K., Niemi, J., Nisbeth, N., Nygaard, L., Orav, H., Rögnvaldsson, E., others, 2013. Nordic and Baltic Wordnets Aligned and Compared Through "WordTies," in: The 19th Nordic Conference of Computational Linguistics. Linköping University Electronic Press, Oslo University, Norway, pp. 147–162.

Pedersen, B.S., Forsberg, M., Borin, L., Lindén, K., Orav, H., Rögnvaldsson, E., 2012. Linking and Validating Nordic and Baltic wordnets, in: Proceedings of the 6th International Global Wordnet Conference. Matsue, Japan, pp. 254–260.

Peters, W., Peters, I., Vossen, P., 1998. Automatic Sense Clustering in EuroWordNet, in: Proceedings of the 1st International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Granada, Spain, pp. 409–416.

Richens, T., 2008. Anomalies in the Wordnet Verb Hierarchy, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics (ACL), pp. 729–736.

Sagot, B., Fišer, D., 2012. Cleaning Noisy Wordnets, in: Proceedings of the 8th International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Istanbul, Turkey, pp. 23–25.

Saito, J.-T., Wagner, J., Katz, G., Reuter, P., Burke, M., Reinhard, S., 2002. Evaluation of GermaNet: Problem Using GermaNet for Automatic Word Sense Disambiguation, in: Proceedings of the LREC Workshop on WordNet Structure and Standardization and How These Affect WordNet Applications and Evaluation. European Language Resources Association (ELRA), Las Palmas, Canary Islands, Spain, pp. 14–29.

Šmrz, P., 2004. Quality Control and Checking for Wordnet Development: A Case Study of BalkaNet. Science and Technology 7, 173–181.

Experiences of Lexicographers and Computer Scientists in Validating Estonian Wordnet with Test Patterns

Ahti Lohk
Tallinn University of Technology
Akadeemia tee 15a
Tallinn, Estonia
ahti.lohk@ttu.ee

Heili Orav
University of Tartu
Liivi 2
Tartu, Estonia
heili.orav@ut.ee

Kadri Vare
University of Tartu
Liivi 2
Tartu, Estonia
kadri.vare@ut.ee

Leo Võhandu
Tallinn University of Technology
Akadeemia tee 15a
Tallinn, Estonia
leo.vohandu@ttu.ee

Abstract

New concepts and semantic relations are constantly added to Estonian Wordnet (EstWN) to increase its size. In addition to this, with the use of test patterns, the validation of EstWN hierarchies is also performed. This parallel work was carried out over the past four years (2011-2014) with 10 different EstWN versions (60-70). This has been a collaboration between the creators of test patterns and the lexicographers currently working on EstWN. This paper describes the usage of test patterns from the points of views of information scientists (the creators of test patterns) as well as the users (lexicographers). Using EstWN as an example, we illustrate how the continuous use of test patterns has led to significant improvement of the semantic hierarchies in EstWN.

1 Introduction and background

1.1 About Estonian Wordnet

The Estonian Wordnet began as a part of the EuroWordNet project (Vossen, 1998) and was built by translating basic concepts from English to allow for the monolingual extension. Words (literals) to be included were selected on a frequency basis from corpora. Extensions have been compiled manually from Estonian monolingual dictionaries and other monolingual resources. In this process, several methods have been used. For example, domain-specific methods, i.e. semantic fields like architecture, transportation, etc. have been covered. Moreover, there have been endeavors to automatically add derivatives and the results have been used in the sense disambiguation process. Version 70 of EstWN consists of 67,674 *synsets*, including 110,869 *lexical units*.

1.2 Previous experience of validation

Before the introduction of test patterns, the EstWN was validated and revised by adding new synsets and semantic relations into its semantic network. Information about new lexical concepts (synsets) originated from the Estonian language explanatory dictionary (EKSS[1]), text corpora and even from feedback on applying EstWN to the word sense disambiguation (WSD) task (Kahusk and Vider, 2002). In addition, EstWN participated in the META-NORD project, which aims to link and validate Nordic and Baltic wordnets (*Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish*) and make these resources widely available for different categories of user communities in academia and in the industry. Under this project, the preliminary task is to "*upgrade several wordnet resources to agreed standards" "and let them undergo cross-lingual comparison and validation in order to ensure that they become of the highest possible quality and usefulness*" (Pedersen et al., 2012).

The first attempt to check the structure of EstWN took place with version 55 (by the first

[1] http://www.eki.ee/dict/ekss/

author of this paper). One of the aspects studied was the number of branches a *synset* goes through before arriving at one or several *root synsets*. These results were presented at the Estonian Applied Linguistics Conference in spring 2011, where Kadri Vider[2] provided our first feedback. Her comments elucidated that EstWN requires this kind of structure checking. In the same year, the first attempt was made to validate EstWN with the test pattern[3] of *closed subset*. Test pattern instances were evaluated by Kadri Vare and some of the results were reflected in two papers (Lohk et al., 2012a), (Lohk et al., 2012b). Later Lohk et al. (2014b) discovered more test patterns, all related to multiple inheritance cases. Presently, there is a system of ten test patterns (Lohk, 2015).

This paper aims to introduce these test patterns and prove that the usage of the test patterns to validate semantic hierarchies of wordnet may significantly improve the wordnet structure. In addition, lexicographers Heili Orav and Kadri Vare share their experiences of working with these test pattern instances (Section 5).

The paper is structured as follows: Section 2 elaborates on the motivation for this work. Section 3 provides a general description of the test patterns, followed by examples of test pattern instances. Section 4 proves the efficiency of test pattern instances in validating the semantic hierarchies of wordnet. Section 5 describes the experiences of lexicographers in using test pattern instances.

2 Motivation

There are many reasons for why test patterns should be chosen as a way to validate *multiple inheritance* in the wordnet hierarchical structure (formed by its semantics). To begin with, due to the nature of *multiple inheritance*, it requires checking. More precisely, multiple inheritance is prone to semantic errors:

1) Inappropriate use of multiple inheritance (Kaplan and Schubert, 2001). There are many cases where multiple inheritance is not used as a conjunction of two properties (Gangemi et al., 2001).
2) Sometimes an IS-A relation is used instead of other semantic relations (Martin, 2003). Multiple inheritance makes it possible to compare relations that connect the parents of a synset.
3) In many cases, multiple inheritance causes topological rings (Liu et al., 2004), (Richens, 2008). According to (Liu et al., 2004), one synset cannot inherit properties from both parents.
4) Multiple inheritance may refer to a short cut problem (Fischer, 1997), (Liu et al., 2004), (Richens, 2008). One synset has a two-fold connection to another one, both directly and indirectly. The direct link is illegal.
5) Multiple inheritance may refer to dangling uplinks in the hierarchical structure (Šmrz, 2004).

Secondly, the use of test patterns has many advantages:

1) Using a test is always quicker than "*[doing] a full revision in top-down or alphabetical order*" (Čapek, 2012).
2) Use of "*manual verification and correction*" is the most reliable. (Lindén and Niemi, 2014).
3) Test pattern instances highlight substructures that refer to possible errors and they simplify the work of the expert lexicographer (Lohk et al., 2012a), (Lohk et al., 2012b), (Lohk et al., 2014b).
4) Test patterns are applicable to wordnets in any language (Lohk et al., 2014c).

3 Test patterns

3.1 General knowledge about test patterns

As mentioned above, test patterns, by their nature, are descriptions of substructures with a specific nature in the wordnet semantic hierarchy that are intended to validate its structure. All patterns have the property of *multiple inheritance*. In most cases, there is a lexical polysemy behind *multiple inheritance*. In the remaining cases, there are *synsets* that simultaneously inherit specific and general concepts (test pattern of *short cut*).

Test pattern instances help to detect possible errors in the semantic hierarchies of wordnet. Each test pattern provides a different perspective to the semantic hierarchy. Thus, they vary in their capability to discover various types of possible semantic errors. Test pattern instances are identified by programs and have to be validated by an expert lexicographer.

[2] A computational linguist from the University of Tartu.

[3] Test pattern is a description of a substructure with a specific nature in the wordnet semantic network (intended to validate the semantic hierarchies of wordnet).

Test pattern structures partially or entirely overlap with each other. However, they have different perspectives to the substructures of hierarchies and may typically point to different semantic errors therein.

There are only two ways to cover all *multiple inheritance* cases in the certain semantic hierarchy of a wordnet – by using test pattern instances of *closed subset* or test pattern instances of *ring* and *synset with many roots* together.

We developed algorithms and created programs (in the framework of the doctoral thesis of (Lohk, 2015)) to automatically find instances of the different types of test patterns. However, some algorithms and programs are implemented to semi-automatically find instances of different types of test patterns. **Table 1** gives an overview of the developed test patterns and information about the automation level of finding their instances. This table illustrates that six of the test patterns are implemented to find their instances in an automatic way and the remaining four in a semi-automatic way. In addition, it should be mentioned that the first two patterns (*short cut* and *ring*) are inspired by other authors (Fischer, 1997), (Liu et al., 2004), (Richens, 2008). Test patterns with a gray background are all the *closed subset* patterns, however, the second and third ones have a specific property. Moreover, the test pattern instances of *synset with many roots* may in some cases correspond to the substructure called *dangling uplink* noted by (Koeva et al., 2004) and (Šmrz, 2004).

Test pattern	Automation level
Short cut	*automatic*
Ring	*automatic*
Closed subset	semi-automatic
Closed subset with a root	semi-automatic
The largest closed subset	semi-automatic
Dense component	automatic
Heart-shaped substructure	automatic
Synset with many roots	automatic
"Compound" pattern	automatic
Connected roots	semi-automatic

***Table 1**: Automation level of finding test pattern instances*

Even though there exist ten test patterns (Table 1), only the instances findable in an automatic way were delivered to the lexicographer.

Below, four of them are described, while *short cut* and *ring* are considered by their authors and the main author of this paper. However, it may be useful to mention that *short cut* indicates redundancy in the semantic hierarchy and *ring* may refer to problematic synsets, which are simultaneously co-hyponyms and co-hypernyms and additions from the same domain category (Liu et al., 2004).

All of the following examples are described by the first author of this paper. Moreover, all ten test patterns are described as mathematical models (more precisely, as graphs) in the thesis of (Lohk, 2015).

In the examples, every synset is equipped with the equivalent synonyms from Princeton WordNet Version 1.5 and begins with an abbreviation "(Eq_s)". If the equivalent synonyms are unknown, free translation has been used.

3.2 Dense component

The *dense component* pattern provides an opportunity to uncover substructures where, due to *multiple inheritance*, the density of the interrelated concepts in the semantic hierarchy is higher (Lohk et al., 2014a), (Lohk et al., 2014b). This pattern contains at least two ambiguous concepts (as in Figure 1 {hotel_1} and "hostel"), which have a minimum of two identical parents ("a housing enterprise" and "accommodation building"). The benefit of this pattern is its ability to uncover all *regular polysemy* cases that reveal themselves as the *regularity of multiple inheritance*.

The lexicographer has to establish:

- whether that kind of regularity is justified, and
- whether *multiple inheritance* can be extended to another *synset*(s)

In order to better understand the semantic field of the dense component in Figure 1, the *synsets* with dotted lines are additional information to the *dense component* (*synsets* with bold lines) to grasp its content more clearly. The first number after in the brackets the *synset* indicates the number of subordinates inside the *dense component*. The second number in the brackets displays the count of all the subordinates for that *synset*.

It is a well-known fact that there are several concepts related to polysemic patterns (Langemets, 2010). Based on Figure 1, {hotel_1} and "hostel" describe that kind of pattern through *institution-building*. Checking the concept(s) additional to {hotel_1} and "hostel", {motel_1, …} is found which in its nature is quite similar to {hotel_1} and "hostel". Hence, it appears reasonable to also connect it to "accommodation building".

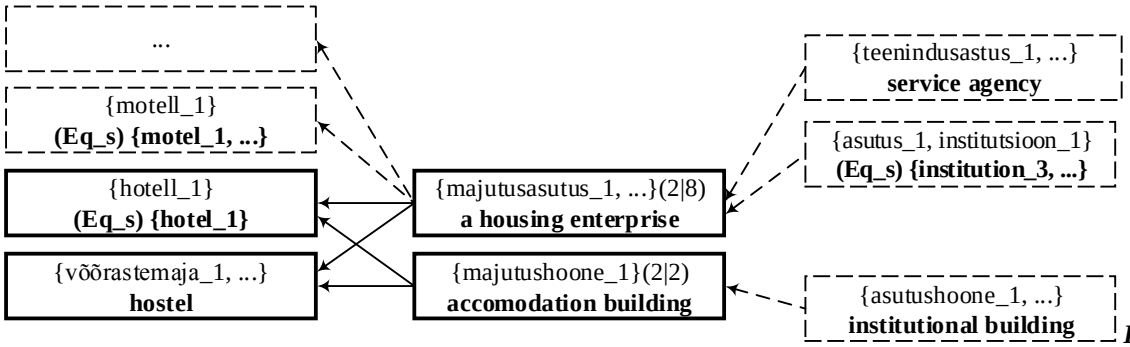


***Figure 1**. An instance of the dense component (rotated 90 degrees)*

In the latest version of EstWN, it emerged that {hotel_1} and "hostel" are no longer connected to building through a *hypernymy* relation. (Instead, the connection is through *near_synonymy*.) Meanwhile, in the current version of Princeton WordNet[4], {hotel_1} is only a building and {hostel_1} is its subordinate. For a solution, let us look at another concept similar to motel, hotel, and hostel – the hospital. EstWN organizes this concept into two *synsets*. Firstly, it denotes a *medical institution*, and secondly, a *medical building*. A similar idea is followed in Princeton WordNet. Thus, in both wordnets, *hospital* is related to an *institution* as well as a *building*. According to this example, it is advised to organize the concepts *hotel*, *motel* and *hostel* in a similar manner.

3.3 Heart-shaped substructure

The *heart-shaped substructure* pattern describes the substructure in the wordnet hierarchy where two synsets (in Figure 2, {homoepathy_1} and "mud cure, mud treatment") along with their two parents are interconnected due to a common parent ({curative_1, cure_1}) as well as through a hypernymy relationship between another one of their parents ({naturopathy_1} and {alternative medicine_1, …}).

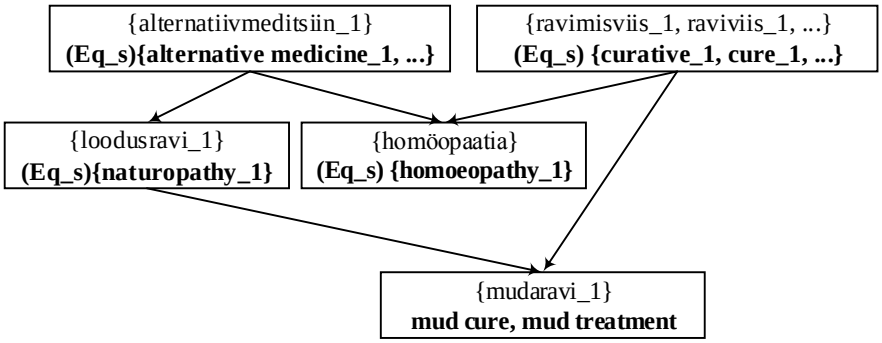


***Figure 2**. An instance of the heart-shaped substructure*

In the report file on the instances of a *heart-shaped substructure* delivered to lexicographers, additional subordinates of the two topmost nodes are shown. This helps to assess why these two *synsets* with two parents are so specific that they join superordinates but their co-members under both parents are not linked.

Secondly, this pattern indicates an instance, where a super-concept ({curative_1, cure_1, …}) seems to be connected to a sub-concept from a different taxonomy level ("mud cure, mud treatment"). On the one hand, this situation might be a particular feature of the language, but on the other hand, it might refer to an error.

An example of a *heart-shaped substructure* in Figure 2 originates from (Lohk et al., 2014b). The question arises why {homoeopathy_1} is not a subcase of {naturopathy_1}. Secondly, are "mud cure, mud treatment" and {homoeopathy_1} subcases of {alternative medicine_1} or of {curative_1, cure_1, …}? On the basis of the definitions of these concepts, the lexicographers decided that both are subcases of {curative_1, cure_1, …} and that {alternative medicine_1} is connected to them via a holonymy relation.

There is still no thorough analysis of the *heart-shaped substructure*. Therefore, there is no such

[4] http://wordnetweb.princeton.edu/perl/webwn

instance in the latest version of EstWN. In addition, as discovered in (Lohk and Võhandu, 2014), most of the cases of *heart-shaped substructure*s in Princeton WordNet pointed to situations where instead of a *hypernymy* relation there should have been a *role* or *type* relation.

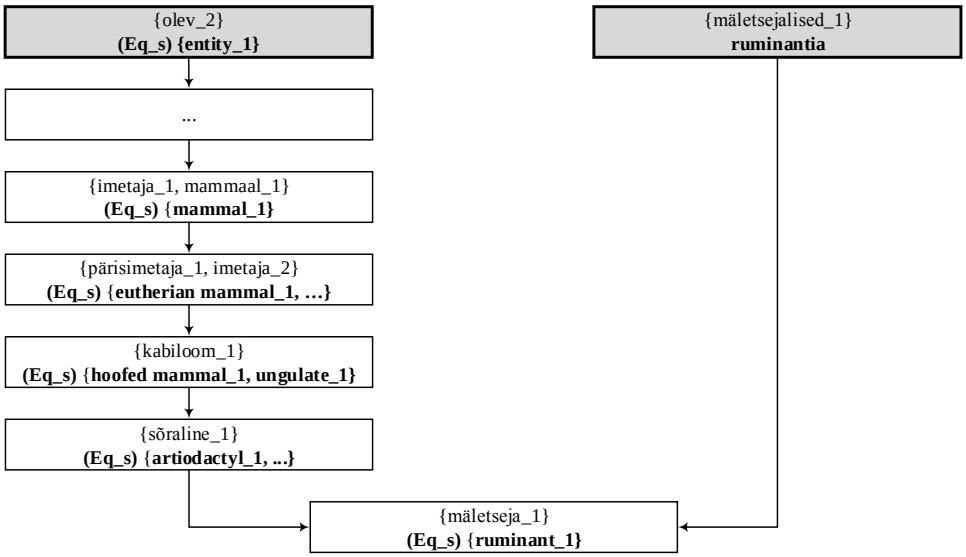


Figure 3*. An instance of the connected roots*

3.4 Synsets with many roots

Quite a similar pattern to rings is the *synset* with many roots. This pattern differs from the former one by its unconnected branches. On the one hand, this signifies that some of the detectable errors are similar to rings. On the other hand, this pattern is capable of discovering errors related to *root synsets*. Figure 3 demonstrates how one *root synset* is *a dangling uplink*[5] – "ruminant animals". It means that the *synset* ({ruminant_1}) is connected to the second parent ("ruminantia") which represents a *root synset*, but in fact, is carrying the ower-level concept. The *root synset* "ruminantia" is a taxon, i.e. it represents a group of animals with particular properties. Therefore, it was correct to change the *hypernymy* relationship between {ruminant_1} and „ruminantia" to holonymy. Thus, {ruminant_1} belongs to the group "ruminantia".

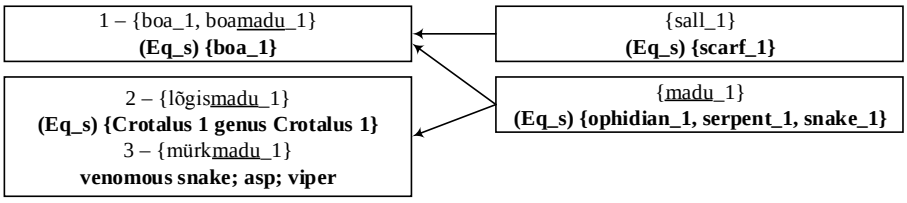


Figure 4. *An instance of the "compound" pattern*

3.5 Substructure that considers the content of synsets ("compound" pattern)

(Nadig et al., 2008) consider a relationship between *synsets* where a member of a *synset* is a suffix to the member of another *synset*. They utilize examples such as {work}, {paperwork}, and {racing}, {auto racing, car racing}. In that manner, it is possible to check whether that *synset* has a *hypernymy* relation. In this pattern, the idea of (Nadig et al., 2008) is employed to uncover all the cases where this condition is true. Additionally, we have to consider that at least one of the subordinates has an additional superordinate as in Figure 4, where {boa_1} has a superordinate {scarf_1}. In that case, the lexicographer must consider why {boa_1} with an extra superordinate did not have any connections to the other subordinates. Upon checking this additional concept ({scarf_1}), it emerges that it is totally unsuitable because while {boa_1} is a *serpent*, *scarf* is a *garment*. However, *scarf* is still related to *boa*, but in a different meaning {boa_2, feather boa_1}.

[5] *Dangling uplink* is a special case of the *synset with many roots*.

4 Experiences of lexicographers in using test pattern instances

The activities of a lexicographer are rather diverse. Compiling a thesaurus requires access to vast amounts of linguistic data (e.g. corpora, different dictionaries, databases) as well as knowledge of how to analyze these data.

Test patterns provide lexicographers with a broader overview than daily work with a lexicographic tool could ever give. All the patterns were checked individually. In many cases, additional descriptions of usage context or definitions help to ascertain the correct relations between the concepts and may also provide additional relations found to be missing.

On occasion, synsets with many hypernyms were left unaltered. For example, *morphine* is simultaneously both a narcotic and pain medicine. This illustrates a well-known problem: *"Rigidity property plays an important role when we distinguish semantic relations of* **type** *and* **role***"* because "*every type is a rigid concept and every role is a non-rigid concept*" (Hicks and Herold, 2011). It is suspected that the *hyponymy* relation may sometimes be a *role* or *type* relationship.

There were also instances where a hypernym had several hyponyms which in turn indicated a problem, namely that some hyponyms had hypernyms that were too general. Revising the hypernymy trees often reduced the amount of direct hyponyms, resulting in a more precise and systematic hierarchy.

Thus, lexicographer should also know how to use their own intuition in the decision-making process. As these test patterns only indicate possible problems, it is not sensible to apply test patterns automatically. However, it could be very useful, if the test pattern results ran simultaneously in a wordnet editing tool, so the lexicographer is provided with complementary information.

5 Iterative evolution of EstWN

Applying the test patterns to EstWN has taken place gradually. As mentioned earlier, we began validating EstWN with the *closed subset* test pattern. At that time, we studied approximately 20 instances of EstWN and Princeton WordNet. Some of the results are reflected in two joint papers with Kadri Vare (Lohk et al., 2012a) and (Lohk et al., 2012b). Later, we started to use *short cut* as well as other patterns.

In the iterative evolution of EstWN, test pattern instances were separated with help of our programs and subsequently delivered to lexicographers who validated all instances and corrected wordnet semantic hierarchies where necessary.

Table 2 reflects the number of test pattern instances over 11 EstWN versions. As background information, the noun roots, verb roots and multiple inheritance cases are also presented. Every number in this table indicates the condition of a specific version in the light of the number of test pattern instances. These numbers are found immediately after the addition of new concepts and semantic relations, and the release of the new version. Thus, the correction of semantic hierarchies is revealed in the next version of wordnet.

The bold font in Table 2 indicates the versions in which a specific pattern was applied. We may notice that in the range from 60 to 62 no test patterns are used. As a matter of fact, at that time we conducted some experiments with the *closed subset* pattern for our first two papers. Beside the numbers of test pattern instances, it is important to observe the number of multiple inheritance cases, as every test pattern instance contains at least one. The last row in this table confirms that one multiple inheritance case may be contained in many different types of test pattern instances, while the total of the last row of instances (7+21+30+0+3+7) is bigger than the multiple inheritance cases (51).

The largest changes in the number of multiple inheritance cases appear when *dense components* are taken into use in version 66. This is due to the fact that dense component contains at least two or more multiple inheritance cases in one instance. In the paper of (Lohk et al., 2014a), it was discovered that only 12% (14) of 121 dense component instances do not need any correction. Nevertheless, the next version (67) revealed 8 new instances.

The decrease in the number of multiple inheritance cases continues even after version 67 when two more patterns are applied (*heart-shaped substructure* and *"compound" pattern*). In the last version, there are only 3 *dense component* instances and 0 *heart-shaped substructure* instances. Comparing the numbers of multiple inheritance cases in versions 66 and 70, it is noted that the last number (51) is approximately 32 times smaller, i.e. multiple inheritance cases have been shrunk by approximately 97%.

Version	Noun roots	Verb roots	Multiple inheritance cases	Short cuts	Rings	Synset with many roots	Heart-shaped sub-structure	Dense component	"Compound" pattern
60	142	24	1,296	235	3,445	1,123	1,825	104	301
61	183	22	1,592	259	3,560	1,309	1,861	121	380
62	102	16	1,700	299	3,777	1,084	1,941	128	415
63	114	16	1,815	**321**	3,831	1,137	2,103	141	447
64	149	15	1,893	**337**	3,882	1,173	2,232	149	471
65	248	14	1,717	**194**	2,171	791	**451**	132	459
66	144	4	1,677	**119**	1,796	613	259	**121**	671
67	129	4	1,164	**79**	928	477	**167**	**24**	**407**
68	131	4	691	**60**	**537**	232	**38**	**18**	**54**
69	121	4	102	**18**	**291**	35	**1**	**8**	23
70	118	4	51	**7**	**21**	**30**	**0**	**3**	**7**

***Table 2**: A numerical overview of EstWN spanning 11 versions*

6 Conclusion and future works

The main collaboration between computer scientists and lexicographers in order to validate EstWN (version 60) began with the *closed subset* test pattern. The *closed subset* was successful in finding possible errors in semantic relations. Later, nine other test patterns dealing with multiple inheritance were developed (see more: Lohk, 2015). Two patterns, namely *short cut* and *ring patterns* are inspired from different authors and one pattern can in certain cases include a *dangling uplink*. In this paper, six test patterns were described but the examples covered four test patterns.

Typically, the work for using test patterns was organized as follows: the first author of this paper generated the instances of test patterns, then based on that document, the lexicographer made corrections using the EstWN editing tool.

The experience of validating Estonian Wordnet assured that the continuous usage of test patterns can significantly improve the semantic hierarchy. Multiple inheritance decreased 32 times or 97% in the last five versions of EstWN.

In the future, we plan to apply these test patterns to other types of semantic relations, for instance to *near synonymy*, *fuzzynymy* and *holonymy*. Moreover, as there are about 70 wordnets in the world, we believe that applying these test patterns to them may "*automatically characterize their modelling decisions (i.e. potential modelling errors)"*[6].

Reference

Čapek, T., 2012. SENEQA-System for Quality Testing of Wordnet Data, in: Proceedings of the 6th International Global Wordnet Conference. Toyohashi University of Technology, Matsue, Japan, pp. 400–404.

Gangemi, A., Guarino, N., Oltramari, A., 2001. Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level, in: Proceedings of the International Conference on Formal Ontology in Information Systems-Volume 2001. ACM, pp. 285–296.

Kaplan, A.N., Schubert, L.K., 2001. Measuring and Improving the Quality of World Knowledge Extracted from WordNet (No. 751). The University of Rochester Computer Science Department, Rochester, New York.

Koeva, S., Mihov, S., Tinchev, T., 2004. Bulgarian Wordnet–Structure and Validation. Romanian J. Inf. Sci. Technol. 7, 61–78.

Langemets, M., 2010. Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras. Eesti Keele Sihtasutus, Tallinn, Eesti.

[6] Comment by a reviewer.

Lindén, K., Niemi, J., 2014. Is It Possible to Create a Very Large Wordnet in 100 Days? An Evaluation. Language Resources and Evaluation 48, 191–201.

Liu, Y., Yu, J., Wen, Z., Yu, S., 2004. Two Kinds of Hypernymy Faults in WordNet: the Cases of Ring and Isolator, in: Proceedings of the 2nd Global Wordnet Conference. Brno, Czech Republic, pp. 347–351.

Lohk, A., 2015. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Tallinn University of Technology, Tallinn, Estonia.

Lohk, A., Allik, K., Orav, H., Võhandu, L., 2014a. Dense Component in the Structure of Wordnet, in: Proceedings of the 9th International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 1134–1139.

Lohk, A., Norta, A., Orav, H., Võhandu, L., 2014b. New Test Patterns to Check the Hierarchical Structure of Wordnets, in: Information and Software Technologies. Springer, pp. 110–120.

Lohk, A., Orav, H., Võhandu, L., 2014c. Some Structural Tests for WordNet with Results. Proceedings of the 7th Global Wordnet Conference 313–317.

Lohk, A., Vare, K., Võhandu, L., 2012a. Visual Study of Estonian Wordnet Using Bipartite Graphs and Minimal Crossing Algorithm, in: Proceedings of the 6th International Global Wordnet Conference. Matsue, Japan, pp. 167–173.

Lohk, A., Vare, K., Võhandu, L., 2012b. First Steps in Checking and Comparing Princeton Wordnet and Estonian Wordnet, in: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. Association for Computational Linguistics (ACL), pp. 25–29.

Lohk, A., Võhandu, L., 2014. Independent Interactive Testing of Interactive Relational Systems, in: Gruca, D.A., Czachórski, T., Kozielski, S. (Eds.), Man-Machine Interactions 3, Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 63–70.

Martin, P., 2003. Correction and Extension of WordNet 1.7, in: Conceptual Structures for Knowledge Creation and Communication. Springer, pp. 160–173.

Nadig, R., Ramanand, J., Bhattacharyya, P., 2008. Automatic Evaluation of WordNet Synonyms and Hypernyms, in: Proceedings of ICON-2008: 6th International Conference on Natural Language Processing. CDAC Pune, India.

Pedersen, B.S., Forsberg, M., Borin, L., Lindén, K., Orav, H., Rögnvaldsson, E., 2012. Linking and Validating Nordic and Baltic wordnets, in: Proceedings of the 6th International Global Wordnet Conference. Matsue, Japan, pp. 254–260.

Richens, T., 2008. Anomalies in the Wordnet Verb Hierarchy, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics (ACL), pp. 729–736.

Šmrz, P., 2004. Quality Control and Checking for Wordnet Development: A Case Study of BalkaNet. Science and Technology 7, 173–181.

Vossen, P., 1998. Introduction to EuroWordNet. Computers and the Humanities 32, 73–89.

African WordNet: A Viable Tool for Sense Discrimination in the Indigenous African Languages of South Africa

Stanley Madonsela
Department of African Languages
University of South Africa
madonfs@unisa.ac.za

Munzhedzi James Mafela
Department of African Languages
University of South Africa
mafelmj@unisa.ac.za

Mampaka Lydia Mojapelo
Department of African Languages
University of South Africa
mojapml@unisa.ac.za

Rose Masubelele
Department of African Languages
University of South Africa
masubmr@unisa.ac.za

Abstract

In promoting a multilingual South Africa, the government is encouraging people to speak more than one language. In order to comply with this initiative, people choose to learn the languages which they do not speak as home language. The African languages are mostly chosen because they are spoken by the majority of the country's population. Most words in these languages have many possible senses. This phenomenon tends to pose problems to people who want to learn these languages. This article argues that the African WordNet may the best tool to address the problem of sense discrimination. The focus of the argument will be on the primary sense of the word 'hand', which is part of the body, as lexicalized in three indigenous languages spoken in South Africa, namely, Tshivenḓa, Sesotho sa Leboa and isiZulu. A brief historical background of the African WordNet will be provided, followed by the definition of the word 'hand' in the three languages and the analysis of the word in context. Lastly, the primary sense of the word 'hand' across the three languages will be discussed.

1 Introduction

Thoughtful lexicography work for indigenous African languages of South Africa commenced just after the introduction of democratic elections in 1994. With the establishment of the Pan South African Language Board, national lexicography units were constituted in all the official languages of South Africa. The lexicography units were tasked with the duty of establishing dictionaries in the different official languages of South Africa. Although many of the dictionaries are bilingual, they give very little information regarding sense discrimination, especially for non-mother tongue speakers who are interested in learning indigenous African languages. The South African government encourages people to learn one indigenous African language in addition to their first language. Lexicography work in African languages produced so far does not address the needs of indigenous African language learners because the equivalents provided do not address the problem of sense discrimination. Similarly, indigenous African language learners take it for granted that a lexical item has the same sense across these languages, whereas sometimes the sense of a word is different in these languages even if languages are related.

This paper argues that African WordNet could be a viable tool to address problems such as those mentioned above. The equivalents of 'hand' in Tshivenḓa (Venda), Sesotho sa Leboa (Northern Sotho) and isiZulu (Zulu) are *tshanḓa*, *seatla (letsogo)* and *isandla*, respectively. Indigenous official languages of South Africa belong to the same family of languages; they are Bantu languages belonging to the Niger-Congo family. They are further divided into groups that are, to a certain extent, mutually intelligible. The Nguni language group and the Sotho language group, for example, are not mutually intelligible whereas languages within any of the two groups are. A

majority of the people in the country is multilingual but they may nevertheless not be competent in all the languages. Being a rainbow nation with a myriad of people and languages, everyday life dictates that one has some understanding or awareness, however limited, of other languages. The fact that official African languages in the country belong to the same family often tempts people, knowingly or unknowingly, to clamp them together with the saying 'if you know one you know them all' – and this is far from the truth. The lexicons and the senses reflect some similarities, overlaps and unrelatedness to an extent that they may result in miscommunication unless sense discrimination is taken care of.

We have used the English word 'hand' to demonstrate lexicalisation and sense discrimination in the languages, Sesotho sa Leboa (Northern Sotho), Tshivenda (Venda) and isiZulu (Zulu). Whilst there are other examples that could be used in the African WordNet to indicate sense discrimination across the indigenous African languages of South Africa, the choice of the word 'hand' stems from its cultural significance in the African value system. The word 'hand' has as its underpinning in the 'Ubuntu' (a value system that promotes humanity to others) element which regards humanity as a fundamental part of the eco-systems that lead to a communal responsibility to sustain life.

2 African WordNet defined

African WordNet is based on the Princeton WordNet. It is a multilingual WordNet of official indigenous languages of South Africa. WordNets for African languages were introduced with a training workshop for linguists, lexicographers and computer scientists facilitated by international experts in 2007. The development of WordNet prototypes for four official African languages started in 2008 as the African WordNet Project. This project was based on collaboration between the Department of African Languages at the University of South Africa (UNISA) and the Centre for Text Technology (CTexT) at the North-West University (NWU), as well as support from the developers of the DEBVisDic tools at the Masaryk University. The initiative resulted in first versions of WordNets for isiZulu [zul], isiXhosa [xho], Setswana [tsn] and Sesotho sa Leboa [nso], all members of the Bantu language family (Griesel and Bosch, 2014). Currently Tshivenḓa is the fifth of the nine official African languages of the country that are part of the project.

3 Word sense

Sense is defined as one of a set of meanings a word or phrase may bear especially as segregated in a dictionary entry (Miriam Webster Online). Frege (1892) argues that sense is the mode of presentation of the referent. There are multiple ways of describing and conveying information about one and the same referent; and to each of these ways correspond a distinct sense. Every word is associated with a sense, and the sense specifies the condition for being the word's referent.

According to Fellbaum (1998) in WordNet, each occurrence of a word form indicates a different sense of the word, which provides for its unique identification. A word in a synset is represented by its orthographic word form, syntactic category, semantic field and identification number. Together these items make a "sense key" that uniquely identifies each word/sense pair in the database. The sense of a word can be derived from the semantic relations that it has with other words. The manner in which word sense is viewed has a great appeal for the discussion of the word 'hand' in this article.

The underlying hypothesis of this paper relies on previous studies that used multiplicative models of composition by exploring methods to extend the models to exploit richer contexts. Studies by Gale *et al.*, (1993) and Dagan *et al.*, (1991) have used parallel texts for sense discrimination to identify semantic properties of and relations among lexemes (Dyvik, 1998). Whilst there are different approaches to sense discrimination, this paper adopts an approach by Akkaya, Wiebe and Mihalcea (2012) which is to cluster target word instances, so that the induced clusters contain instances used with the same sense.

4 The primary sense of 'hand' in the three African languages

The primary meaning of a word is its literal meaning. This section looks into the dictionary equivalents of the primary meaning of the English word 'hand' in the three languages Tshivenḓa, Sesotho sa Leboa and isiZulu. The concept under discussion in this paper is defined in WordNet as "the (prehensile) extremity of the superior limb". It is sense 1 of the domain Anatomy and SUMO Bodypart [POS: n ID: ENG 20-05246212-n BCS: 3].

4.1 Tshivenḓa

The equivalent of *hand* in Tshivenḓa is *tshanḓa.* Whereas *hand* in English refers to the part at the end of a person's arm, including the fingers and thumb (Longman Dictionary of Contemporary English, 1995), *tshanḓa* in Tshivenḓa refers to both arm and hand taken as one. Tshivenḓa does not separate between arm and hand as languages such as English do, both are taken as one.

There is a slight difference among the Tshivenḓa lexicographers in defining the lexical entry *tshanḓa.* Wentzel and Muloiwa (1982:65 and 173) define *tshanḓa* and 'hand' differently. They define *tshanḓa (pl. zwanḓa)* as *arm, hand;* whereas *hand is defined* as *tshanḓa (pl. zwanḓa).* According to these lexicographers, *tshanḓa* has got two senses, that of the whole arm, and the part at the end of a person's arm.

The same applies to *Tshivenḓa – English Ṱhalusamaipfi Dictionary (2006)*; the equivalent of *hand* is *tshanḓa* and the equivalents of *tshanḓa* are *hand* and *arm*. It would seem *Tshivenḓa – English Ṱhalusamaipfi Dictionary (2006)* adopted the definitions of the two lexical entries direct from Wentzel and Muloiwa (1982). To them both *hand* and *arm* are called *tshanḓa*. Van Warmelo (1989:388) on the other hand provides the equivalent of *tshanḓa* as *hand*. He does not differentiate between *arm* and *hand;* according to him the whole limb is *tshanḓa*. However, he also refers to the upper arm as *tshishasha.* Tshikota (2012a) and Tshikota (2012b) in his two monolingual dictionaries, *Ṱhalusamaidioma ya luamboluthihi ya Tshivenḓa* (Tshivenḓa monolingual dictionary of idioms) and *Ṱhalusamaipfi ya luamboluthihi ya Tshivenḓa* (Tshivenḓa monolingual dictionary) define tshanḓa as follows:

> **tshanḓa** *dzin* tshipiḓa tsha muvhili tshi re na minwe miṱanu tshine tsha shumiswa u fara ngatsho (Tshikota, 2012a:57)
>
> 'part of the body with five fingers, which is used to hold'
>
> **tshanḓa** (zwanḓa) *dzin 1* tshipiḓa tsha muvhili tshi re na minwe miṱanu tshine tsha shumiswa u fara ngatsho (Tshikota, 2012b:258)
>
> 'part of the body with five fingers, which is used to hold'

The definitions of the lexical entry *tshanḓa* in the two dictionaries are similar, and they refer to the English word *hand*. Lexicographers in these dictionaries were influenced by the English definition of *hand.* They do not reflect what the word *tshanḓa* refers in the spoken language. The word *tshanḓa* in spoken Tshivenḓa refers to English *arm* plus *hand.* This is attested by Wentzel and Muloiwa (1982), Van Warmelo (1989) and *Tshivenḓa – English Ṱhalusamaipfi Dictionary (2006).* The word *tshanḓa* also refers to the *palm.*

4.2 Sesotho sa Leboa

The word for 'hand' in Northern Sotho is *seatla* (plural: diatla). Ziervogel and Mokgokong's (1975) trilingual dictionary gives entries in Northern Sotho and equivalents in Afrikaans and English. The English equivalents of the word *seatla* in the dictionary are 'hand', 'palm of hand', 'handwriting'. The dictionary then continues to use the word in various linguistic contexts in order to lay bare different senses. Of the three English equivalents mentioned above, only 'handwriting' seems to be non-literal, not representing the sense under the domain - Anatomy. The first two equivalents refer to the physical part of the body. Only the first equivalent has a conceptual one-to-one with the concept defined in WordNet as "the (prehensile) extremity of the superior limb". The other equivalent 'palm of hand' is part of the whole concept defined above. Another trilingual

dictionary (Northern Sotho Language Board, 1988) gives entries in English and equivalents in Afrikaans and Northern Sotho. The latter is not only a dictionary, but a terminology and orthography standardizing document as well. The entry 'hand' has a Northern Sotho equivalent *seatla*. Following this entry is a number of English compound nouns and two-word entries which include 'hand'. Of these entries seven are clearly built on the primary meaning of 'hand'. The seven entries reflect that 'hand' is also referred to as *letsogo* in Northern Sotho. For example, the Northern Sotho equivalent of '**hand**work' is *modiro wa **diatla***, '**hand** muscle' is *mošifa wa **seatla***', '**hand** movement' is *tshepedišo ya **letsogo***, '**hand** drill' is *boro**tsogo*** , and '**hand**bag' is *sekhwama**tsogo***.

4.3 IsiZulu

Mbatha (2006: 9) in his isiZulu monolingual dictionary defines 'hand' as *isitho somuntu okuyisona abamba ngaso* 'a body part which a human uses to hold'. Mbatha's definition shows dearth of the lexicographic feature in providing the quality of definition required to give clarity. However, Doke and Vilakazi (1972: 9) in their Zulu-English dictionary define 'hand' as *forearm* (*including the hand*). From the definitions of these lexicographers, it is apparent that they define the concept not exactly the same. Mbatha seems to be focusing mostly on the functional aspect of the word 'hand' than striving to describe its meaning. Mbatha's definition shows dearth of the lexicographic feature in providing the quality of definition required to give clarity. The definition by Doke and Vilakazi on the other hand, is not detailed enough. When considering Doke and Vilakazi's definition, it lacks the defining criteria and the characteristics that are necessary to understand what the word means. What makes Doke and Vilakazi's definition incomplete is that it does not give enough information about the word. In Collins English Dictionary (1991:704) the word hand is defined as 'the prehensile part of the body at the end of the arm, consisting of a thumb, four fingers and a palm'. Considering the definitions given by Mbatha, and Doke & Vilakazi, it becomes clear the information that they have provided has a tentative validity.

5 Discussion

Across the three languages, the primary sense of 'hand' is a physical part of the human body. Lexicographers have to constantly strive to enhance the quality of definitions in monolingual dictionaries to best suit the needs and level of their target users (Gouws 2001:143). Landau (2001:162) also maintains that the definition must define and not just talk about the word or its usage. It is clear from the argument given above that they do not provide the answer to the question 'what it is' that is being defined as Gouws (Ibid) suggests. Lombard (1991:166) pinpoints defining criteria that would result in good definitions namely *completeness, clarity, accuracy, consistency, independency, objectivity* and *neutrality*. Although words for 'hand' in the three languages may refer to the different parts of the limb, starting at the end of the shoulder and ending at the fingers, the parts constitute the same limb. Whereas in Tshivenḓa and isiZulu, 'hand' is referred to as *tshanḓa* and *isandla* respectively, in Sesotho sa Leboa it is referred to as *seatla* or *letsogo*. In Tshivenḓa, *tshanḓa* is that part of the human body starting from the shoulder to the fingers. This means that the whole limb is referred to as *tshanḓa*. The sense in isiZulu is slightly different from that in Tshivenḓa because *isandla* refers to the forearm including the wrist, fingers. Whereas Tshivenḓa *tshanḓa* refers to the whole limb, isiZulu *isandla* refers part of the limb, i.e. forearm. Sesotho sa Leboa refers to the whole limb as *letsogo* 'arm', to the 'hand' as *seatla*; additionally 'hand' is referred to as *letsogo*. *Seatla* is part of the whole limb, a meronym of *letsogo* 'arm', but also used synonymously with *letsogo*. Unlike Tshivenḓa and Sesotho sa Leboa, isiZulu recognises the forearm as part of the hand, which is referred to as *isandla*. In Tshivenḓa and Sesotho sa Leboa, the palm of the hand is referred to as *tshanḓa* and *seatla,* respectively. The diagram in appendix 1 illustrates the situation sketches above.

It emerges from the Northern Sotho dictionary definitions and equivalents that the concept is lexicalized as *seatla* and/or *letsogo*. The English dictionary equivalent of Northern Sotho *letsogo* is 'arm'. *Letsogo* refers to the whole superior limb, which includes *seatla* 'hand'. The two are understood to be in a

holonym-meronym relationship, while being used as synomyms as well.

6 Conclusion

The empirical conclusion in this paper provides a new understanding of words with different senses which pose a challenge to the different speakers of the indigenous South African languages, particularly the three languages mentioned. Considering the hypothesis posed at the beginning of this paper, it can be concluded that the primary sense of *hand* in the three languages, although related, is different. People learning these languages should not conclude that because they are grouped as African languages the senses of their lexicons are similar throughout. It is also noted that the sense of *hand* in English is different from that in the African languages. WordNet is a good tool to investigate the sense of African languages' lexicons, in that the word 'arm' has a comparable sense and an ID, namely, arm: 1 [POS: n ID: ENG 20-05245410-n BCS: 3] and belongs to a specific domain: Anatomy.

The discussion in this paper has gone some way towards enhancing our understanding of the degree to which African WordNet can be a tool that can be used to differentiate word sense. This research has thrown up many questions in need of further investigation regarding the other sense such as the metaphoric use and the idiomatic expression of the word in discussion. It became evident from the discussion that the same word can have different senses in the different.

References

Akkaya, C., Wiebe, J. and Mihalcea, R. 2012. Utilizing Semantic Composition in Distributional Semantic Models for Word Sense Discrimination and Word Sense Disambiguation. Sixth IEEE International Conference on Semantic Computing (IEEE ICSC2012). Amsterdam and Philadelphia: John Benjamins *and Cognitive Processes, 6* (1), 1–28.

Dagan, I., Itai, A. and Schwall, U. 1991. Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the ACL,* 18-21 Berkeley, California, 130-137.

Doke, C.M. and Vilakazi, B.W. 1972. Zulu-English Dictionary. Johannesburg: Witwatersrand University Press.

Dyvik, H. 1998. Translations as Semantic Mirrors. *Proceedings of Workshop Multilinguality in the Lexicon II, ECAI 98*, Brighton, UK, 24-44.

Fellbaum, C. 1998. Ed. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Frege, G. 1892 "On Sense and Reference" Translated by M. Black in Geach, P. and Black, M. (eds.) (1970) *Translation*s *fr*om *th*e *Philosophica*l *Writing*s of *Gottlo*b *Frege*. Oxford: Basil Blackwell.

Gale, W. A., Church, K. W. and Yarowsky, D. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities,* 26, 415-439.

Gouws, R.H. and Prinsloo, D.J. 2005. Principles and Practices of South African Laxicography. Stellenbosch: Sun Press

Griesel, M. and Bosch, S. 2014. Taking stock of the African Wordnets project: 5 years of development. In *Proceedings of the Seventh Global WordNet Conference,* January 2014. University of Tartu, Estonia.

Landau, S.I. 2001. Dictionaries: The Art and Craft of Lexicography. Second Edition. Cambridge University Press.

Lombard, F. 1991. Die aard en aanbieding van die leksikografiese definisie. In: Lexikos 1: 158-180. Longman Dictionary of Contemporary English, 1995. 3rd Edition. Harlow: Longman Group.

Mbatha, M.O. 2006. Isichazamazwi SesiZulu. Pietermaritzburg: New Dawn Publishers.

Northern Sotho Language Board. 1988. *Sesotho sa Leboa Mareo le Mongwalo No. 4/ Northern Sotho Terminology and Orthography No. 4/ Noord-Sotho Terminologie en Spelreëls No. 4*. Pretoria: Government Printer.

Tshikota, S.L. 2012. *Ṭhalusamaidioma ya Luamboluthihi ya Tshivenḓa.* Ṭhohoyanḓou: Tshivenḓa National Lexicography Unit.

Tshikota, S.L. 2012. *Ṱhalusamaipfi ya Luamboluthihi ya Tshiveṋda.* Ṱhohoyaṋdou: Tshiveṋda National Lexicography Unit.

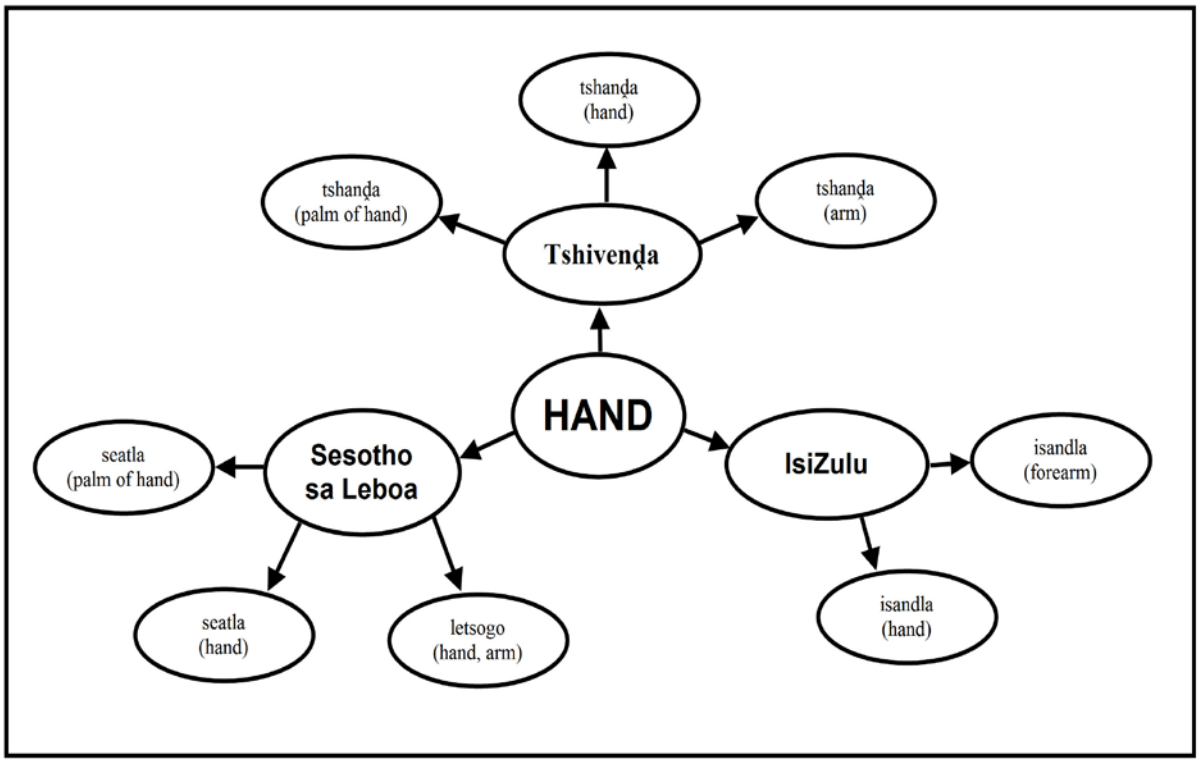
Tshiveṋda NLU. 2006. Tshiveṋda/English Ṱhalusamaipfi Dictionary. Cape Town: Phumelela Publishers.

Van Warmelo, N.J. 1989. *Venda Dictionary: Tshiveṋda – English*. Pretoria: Van Schaik.

Wentzel, P.J. 1982. *Improved Trilingual Dictionary: Venda – Afrikaans – English*. Pretoria: Univesrity of South Africa.

Ziervogel, D., Lombard, D. P and Mokgokong, P.C. 1969. *A Handbook of the Northern Sotho Language*. Pretoria: Van Schaik.

Appendix 1: Lexicalisation of ‘hand’ in the three languages



An empirically grounded expansion of the supersense inventory

Héctor Martínez Alonso† Anders Johannsen Sanni Nimb‡
Sussi Olsen Bolette Sandford Pedersen†
University of Copenhagen (Denmark)
‡Danish Society of Language and Literature (Denmark)
†alonso@hum.ku.dk, bspedersen@hum.ku.dk

Abstract

In this article we present an expansion of the supersense inventory. All new supersenses are extensions of members of the current inventory, which we postulate by identifying semantically coherent groups of synsets. We cover the expansion of the already-established supernsense inventory for nouns and verbs, the addition of coarse supersenses for adjectives in absence of a canonical supersense inventory, and supersenses for verbal satellites. We evaluate the viability of the new senses examining the annotation agreement, frequency and co-ocurrence patterns.

1 Introduction

Coarse word-sense disambiguation is a well established discipline (Segond et al., 1997; Peters et al., 1998; Lapata and Brew, 2004; Alvez et al., 2008; Izquierdo et al., 2009) that has acquired more momentum in the latter years under the name of *supersense tagging* (SST). SST uses a coarse sense inventory to label spans of variable word length (Ciaramita and Johnson, 2003; Ciaramita and Altun, 2006; Johannsen et al., 2014). This coarse sense inventory is obtained from the list of WordNet *first beginners*, i.e. the names of the *lexicographer files* that hold the synsets.

However, lexicographer files were devised for practical reasons, namely as an organization method for the development of WordNet (Miller, 1990; Gross and Miller, 1990; Fellbaum, 1990), and not as final target categories to annotate with or disambiguate from.

Nevertheless, the organization of lexicographer files is semantically motivated, and supersenses have proven useful for natural language processing such as metaphor detection or relation extraction (Ciaramita and Johnson, 2003; Tsvetkov et al., 2014a; Søgaard et al., 2015). According to Ciaramita and Altun (2006), supersenses extend the named entity recognition (NER) inventory so that the predictions of an SST model subsume the output of NER. Schneider et al. (2015) provide a full SSI for prepositions.

The current supersense inventory (henceforth SSI) enjoys *de facto* standardness, but in spite of its potential usefulness, it is used acritically. The current SSI provides 26 noun supersenes and 15 verb supersenses. Adjective and adverb lexicographer files are disregarded. We provide a revision of the SSI by an extension of its supersenses using the Danish wordnet as starting point.

This revision is empirically backed by four evaluation criteria, namely inter-annotator agreement, sense frequency after adjucation, sense cooccurrence, and NER compliance (whenever possible). Note that we do not suggest merging existing supersenses, but only extending the current SSI in a backwards-compatible manner.

We conduct our extension in three steps steps. First, we propose new supersenses when a projection between an EuroWordNet (EWN) ontological type and a supersense is not univocal (Section 2). Second, we evaluate the distribution of supersenses in terms of agreement after an annotation task, frequency and sense-sense relations (Section 4) and analyze the results across the different parts of speech (Section 5). Lastly, we suggest new supersenses (underlined in in Table 2) when large sections of the data have been assigned to backoff categories.

The main contributions of this paper are i)a set of guidelines for the inclusion of new supernses in the SSI, ii) an empirically motivated expansion of the SSI with new senses for nouns, verbs and adjectives respectively,[1] and iii) a projection from ontological types to supersenses that can be used to enrich any wordnet that is not organized in lexi-

[1] https://github.com/coastalcph/semdax

cographer files or where synsets are not fully connected to Princeton synsets.

2 Extending the supersense inventory

This section describes the extension of the SSI that results from an analysis of projections into supersenses from ontological types, ensuing both retrocompatibility with the existing inventory (i.e. all new supersenses are extensions of an existing supersense), and compability with NER tags.

We use The Danish wordnet (Pedersen et al., 2009), DanNet, as a starting point. DanNet is not organized in lexicographer files. However, its synsets are associated to ontological types (Vossen et al., 1998). We map from the ontological type of the synsets to a supersense. Table 2 provides one example for each lexical part of speech.

Ontological type	Supersense
Property+Physical+Colour	ADJ.PHYSICAL
Liquid+Natural	NOUN.SUBSTANCE
Dynamic+Agentive+Mental	VERB.COGNITION

Table 1: Supersense mapping examples.

We establish a projection into supersenses with the following steps; if an ontological type t_i:

1. does not have a straightforward 1-to-1 mapping to a supersense,
2. is the subtype of an ontological type t_j (e.g. *Liquid+Natural* is a subtype of *Liquid*),
3. and has enough support (in terms of how many synsets make up t_i),

then we propose new supersense for t_i as an extension of the supersense of t_j. We consider the support to be substantial enough when a subtype has at least 500 synsets out of the 65k synsets in DanNet and, and it makes up at least 12% of its parent supersense.

We exemplify this method by explaining how we extend DISEASE from STATE. The subtype *Property+Physical+Condition* is associated to 527 synsets and makes up 70% of the synsets of the type *Condition*. All the synsets of this subtype are diseases, and we propose the supersense DISEASE as an extension of STATE, which is otherwise the supersense translation of *Condition*.

In addition to providing new supersenses for the main three lexical parts of speech, we devise three aditional tags for verbal satellites (collocations, particles and reflexive pronouns) as aid for verbal multiwords the annotation (cf. Section 5.4). Table 2 lists the new supersenses. Underlined dupersenses marked are determined in post-annotation analysis (cf. Section 5), while the rest have been determined during the projection step described in this section. Supersenses in grey do not meet the inclusion criteria, and are thus not incorporated in our proposal for SSI extension.

New supersense	Subsumed by
Noun	
VEHICLE	} ARTIFACT
BUILDING	
CONTAINER	
DOMAIN	} COGNITION
ABSTRACT	
INSTITUTION	} GROUP
DISEASE	} STATE
LANGUAGE	} COMMUNICATION
DOCUMENT	
Verb	
ASPECTUAL	} STATIVE
PHENOMENON	} CHANGE
Adjective	
MENTAL	} ALL
PHYSICAL	
SOCIAL	
TIME	
FUNCTION	
Satellite	
COLLOCATION	} *none*
PARTICLE	
REFLPRON	

Table 2: Extensions to the sense inventory. Items in grey do not fulfill the inclusion criteria, underlined items have been suggested during post-annotation analysis.

3 Annotation task

We perform an annotation task on 5,500 sentences from a Danish contemporary corpus (Asmussen and Halskov, 2012) made up of newswire, parliamentary speech, blog posts, internet forum discussions, chatroom logs and magazine articles, plus the test section of the Danish Dependency Treebank (Buch-Kromann et al., 2003).

Any corpus choice imposes a bias, and we base the corpus choice on a twofold need: to tune the sense inventory to the needs of contemporary genres that are used for information extraction, without sacrificing its adequacy for more usual domains. Generally speaking, another corpus choice would yield a different supersense expansion.

The corpus was pre-annotated using the supersense projection list described in Section 2. Even though the size of the specific wordnet is a determining factor for the quality of the preannotation, it does not determine the coverage of the final supersense annotation, which provides full coverage because a SSI covers all content words.

Two in-house native annotators with a background in linguistics annotated the data, choosing the best pre-annotated sense or selecting a new one. A third annotator performed adjudication in case of disagreement. The overall kappa score before adjudication is 0.62. Olsen et al. (2015) provide more details on the annotation task. The resulting data has been use for automatic supersense tagging by Martínez Alonso et al. (2015).

4 Metrics

This section describes the metrics applied to the supersense-annotated corpus in order to assess the distribution of the new supersenses.

4.1 Sense-wise agreement variation

Inter-annotator agreement is a source of information on the reliability of semantic categories (Lopez de Lacalle and Agirre, 2015). In this section, we examine the variation in agreement for noun and verb supersenses. Cf. Olsen et al. (2015) for a more detailed account.

Figures 1 and 2 portray the variation of agreement across noun and verb supersenses. Each cell in the matrix indicates the probability of a token being annotated with a row-column tuple of supersenses (r_i, c_j) by the two annotators. The matrix is normalized row-wise, and each row describes the probability distribution of a certain supersense r_i to be annotated with any other supersense c_j. When r_i and c_j have the same value, annotators agree. Rows are sorted in descending order of agreement, i.e. the size of the $r_i = c_j$ box on the diagonal. The larger the box in the diagonal, the higher the agreement for a given r_i supersense.

From the standard supersenes, for instance, N.GROUP is very seldom assigned by both annotators, and there is usual disagreement with N.QUANTITY. Other senses like N.BODY have very few off-diagonal values and have near-perfect agreement.

Out of the new supersenses, N.INSTITUTION has very high agreement. However, the new supersense N.DOMAIN has very low agreement. A domain (i.e. a field of knowledge or professional discipline) is difficult to distinguish from its semantically related senses N.COGNITION and N.COMMUNICATION. Low agreement also compromises the reliability of some of the established supersenses such as NOUN.SHAPE. However, the goal of these measurements is to evaluate the new supersenses, because we do not advocate for a reduction of the canonical SSI, but an extension of the existing list of supersenses.

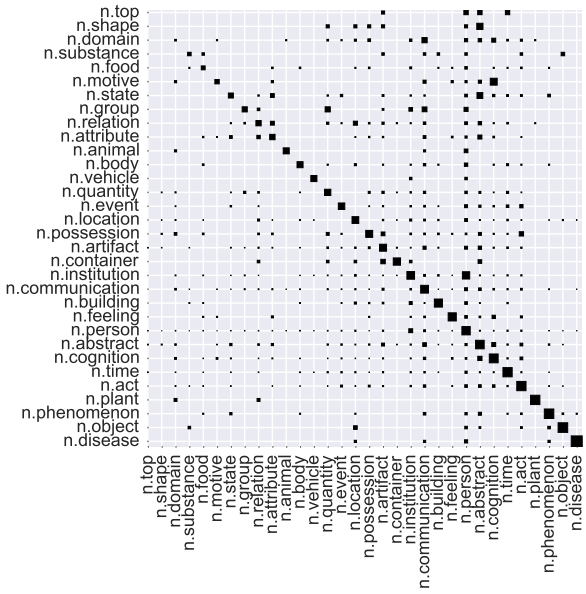


Figure 1: Agreement variation for nouns.

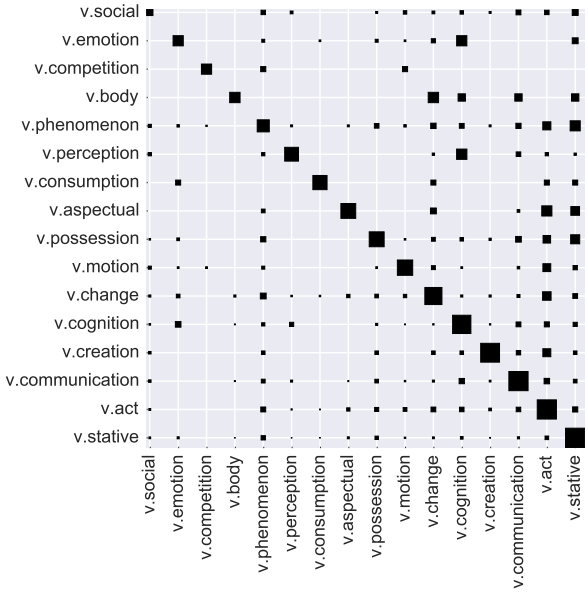


Figure 2: Agreement variation for verbs.

Agreement also varies across parts of speech. Diagonal boxes take up 69% of the probability mass of the verbs, while 58% is taken by the agreed nouns. In other words, 31% of the annotations for verbs are mismatched, whereas 42% of the nouns have mismatching annotations. We consider this difference a consequence of the size of the inventory for nouns and verbs respectively, and not an indication of verbs being *per se* easier to annotate than nouns.

4.2 Supersense frequency

Frequency is the most straightforward way of assessing whether a a certain sense has been given to enough examples to be considered relevant. If a new sense is very frequent, there is sufficient reason to consider it as a valid addition to the SSI.

Table 3 provides the absolute frequency for the 28 most frequent supersenses, namely half of the total SSI, after disagreements had been resolved by the adjudicator.

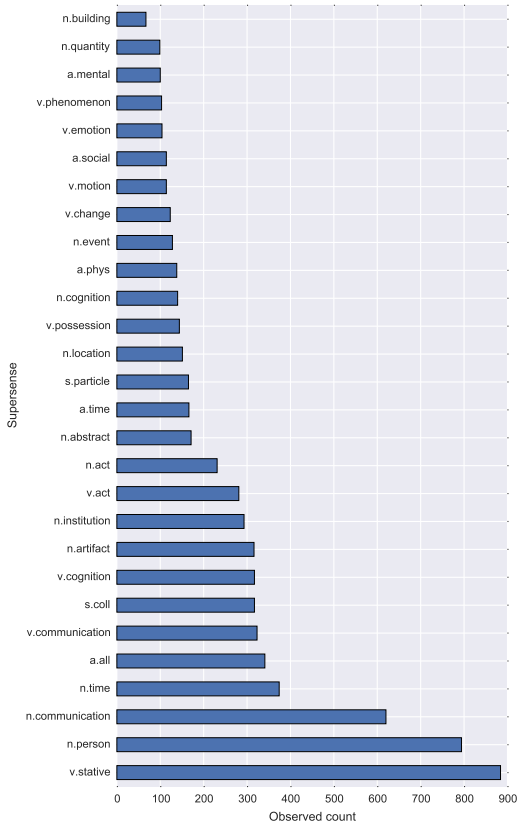


Figure 3: Distribution of frequent senses.

Presence in the top half of the sense ranking is one of the criteria for inclusion in the SSI.

4.3 Association between supersenses

A third source of information on the appropiateness of a supersense is its relation with the other established senses. This section offers an overview on how supersenses co-occur. To give account for relevant associations between senses, we use PMI (pointwise mutual information). Higher PMI values indicate stronger association, i.e. a higher conditional probability of one sense appearing in a sentence given the other, controlled for the frequency of both senses in order not to overestimate the co-occurrence of frequent senses.

Table 3 shows the twelve pairs of supersense with the highest PMI calculated across sentences. We compare the supersense-wise PMI for three corpora:

1. Danish extended (DA-EX): The Danish corpus annotated with the extended SSI described in Section 3,
2. Danish regular (DA-RG): The Danish corpus from Section 3 with regular supersenses, where the extended senses have been replaced by their subsuming original sense, e.g. all the occurrences of N.VEHICLE in DA-EX are N.ARTIFACT in DA-RG,
3. English regular (EN-RG): The English SemCor (Miller, 1990) with the regular supersense annotation.

Some of the associations are prototypical selectional restrictions like V.COMSUMPTION + N.FOOD. Other associations are topical across parts of speech, like VERB.COMPETITION and NOUN.EVENT ('They **won** the **final**'). Finally, there are associations within a part of speech, like N.DISEASE and N.BODY, or N.FOOD and N.CONTAINER. In these associations, one sense is a strong indicator for the other at the topic level (diseases are bodily, food is kept somewhere, etc).

In DA-EX we observe that three of the new nominal senses appear strongly associated with standard supersenses. These relations are topical and easy to interpret. The vehicle-substance relation is the least straightforward one and describes vehicles and the fuel they use, or the materials they are built from.

Projecting back to the regular SSI is not equivalent to annotating from scratch with it. Nevertheless, if we examine the top supersense pairs for DA-RG, we observe that the V.STATIVE sense appears three times. By ignoring the aspectual differ-

Danish (extended)		Danish (regular)		English (regular)	
v.consumption	n.food	v.consumption	n.food	v.consumption	n.food
v.contact	n.body	**v.stative**	n.plant	v.weather	n.object
n.food	**n.container**†	n.person	n.animal	v.weather	n.phenomenon
v.body	n.body	v.competition	n.relation	n.plant	n.food
n.disease†	n.body	v.competition	v.event	n.plant	n.animal
v.competition	n.event	v.change	n.substance	n.substance	n.process
v.motion	v.contact	n.state	n.feeling	v.body	n.body
v.contact	n.artifact	v.consumption	v.change	v.weather	n.substance
n.substance	n.object	v.motion	n.object	v.emotion	n.motive
n.shape	n.body	**v.stative**	v.consumption	n.plant	**n.tops**
n.vehicle†	n.substance	**v.stative**	n.substance	v.contact	n.body
v.competition	n.relation	n.substance	n.person	n.food	n.animal

Table 3: Sense pairs ranked by PMI, bold and underlined described in Section 4.3, † marks new sense.

ence, the tag receives associations with N.PLANT, V.CONSUMPTION and N.SUBSTANCE. Upon manual examination we deem these relations to be spurious, i.e. caused by the presence of the verb *være* ('be') somewhere in the sentence, except the relation between V.STATIVE and V.CONSUMPTION, which is aspectual in nature. The effect on the distribution of supersenses when projecting back to the original SSI becomes apparent for the pair V.COMPETITION + N.RELATION, which becomes the fourth highest PMI in DA-RG.

The English supersense associations of EN-RG provide an example on the effect of corpus choice when annotating. The fairly uncommon N.PLANT appears in several of the top associations, which is a sign of plant senses being used in very restricted contexts in this corpus (biology and recipes). Moreover, we also find a strong association with one of the backoff senses, namely N.TOPS, which is not desirable.

5 Supersenses across parts of speech

5.1 Nouns

This section describes the extended SSI for nouns. To the extent that nouns denote entities, they are very often of focus of interest of ontologies. To the extent that entities often have physical denotation—and thus concrete meaning—, they are the easiest concepts to categorize semantically. Indeed, many ontologies are largely nominal, cf. Suchanek et al. (2008) or Wu and Weld (2008).

WordNet lexicographer files were developed before the consolidation of NER, and named-entity coverage in wordnets is irregular. If, as stated in Section 1, NER compatibility is a favorable side effect of SST, we consider improved NER compatiblity of the new SSI as a plus.

Even though NER inventories are application dependent (cf. Nadeau and Sekine (2007) for a survey), our reference is the *de facto* standard CONLL inventory (Tjong Kim Sang and De Meulder, 2003), with the labels PERSON, LOCATION and ORGANIZATION, as well as a MISCELLANEOUS label, needed for full coverage but not present in e.g. the 7-label inventory of MUC-7 (Chinchor and Robinson, 1997).

Concrete meaning is easier to annotate (Passonneau et al., 2009) and can be the easiest to extend with new senses. As a matter of fact, the concrete N.ARTIFACT supersense is the one that yields more new supersenses in our analysis, namely N.BUILDING, N.CONTAINER and N.VEHICLE. In particular, N.BUILDING extends N.ARTIFACT because artifactual locations, already noted as a semantic type the SIMPLE ontology (Lenci et al., 2000), like houses and highways are very often predicated as locations (following locative prepositions, etc.) instead of having the typical distribution of artifacts, i.e. with the verb *use* or the preposition *with*. Moreover, N.BUILDING maps better into the *Location* type of NER. We leave the potential supersenses for instruments and machines as parts of N.ARTIFACT and do not specifiy them even further, because they hold the prototypical meaning of the supersense.

In spite of the expected higher difficulty of dealing with abstract meaning, we examine two extensions for the abstract supersense N.COGNITION

yielded by the the ontological type projection from Section 2, namely N.DOMAIN and N.ABSTRACT. The supersense N.DOMAIN covers fields of knowledge such as *philosophy*, but also other disciplines to cover sense alternations like 'I enjoyed this **dance**' (N.ACT) vs. 'I studied **dance** at the Performing Arts Academy' (N.DOMAIN). The supersense N.ABSTRACT aims at covering concepts like *idea*, and as a label for metaphorical usages of other concrete words like *pattern* in 'behavioral pattern'.

The fairly abstract supersense N.STATE yields a concrete sense DISEASE, which is much easier to annotate than its original parent supersense (cf. Figure 1). Lastly, we extend N.GROUP with N.INSTITUTION. The original sense does not map neatly into NER, as the overlap is only partial; while *ministry* would fall under the ORGANIZATION type of NER, *pack* (of rats) and *school* (of fish) would not.

5.1.1 Sense-wise evaluation

In this section we evaluate the extended noun supersenses according to four properties summarized in Table 4; whether the agreement for a supersense is high enough (Agr.), whether its frequency is high enough, whether we identify relevant associations using PMI (Assc.), and whether it potentially improves NER compliance (NER). Moreover, we suggest two new supersenses, N.LANGUAGE and N.DOCUMENT, indicated in the lower section of Table 4.

The first three properties are obtained from the metrics in Section 4. We consider agreement to be high enough when there is at least 51% agreement for a supersense. We consider frequency to be enough when the sense belongs to the first 28 senses out of 56 (i.e. the first half of the frequency-ranked SSI). None of the thresholds are particularly high, but we consider a noun supersense as a candidate for inclusion in the final SSI if two of the four properties are satisfied. In other words, none of the criteria are necessary, but fulfilling two of them is sufficient.

We observe most of the new senses fulfill at least two of the criteria, with the exception of N.DOMAIN, which fulfills none. Thus, we do not endorse using the N.DOMAIN supersense and still use N.COGNITION for fields of knowledge. Nevertheless, the N.ABSTRACT sense seems a valuable extension because it satisfies the agreement and frequency criterion.

New supersense	Agr.	Freq.	Assc.	NER
ABSTRACT	x	x		
BUILDING	x			x
CONTAINER	x		x	
DISEASE	x			x
DOMAIN				
INSTITUTION	x	x		x
VEHICLE	x		x	
LANGUAGE	–			
DOCUMENT	–	x		x

Table 4: Inclusion criteria for new noun senses.

The strongest nominal candidate for inclusion is N.INSTITUTION, which satisfies the first two initial criteria, plus improves NER compatibility.

During the annotation task, we observed that a large amount of examples of the standard N.COMMUNICATION supersense were document names, movie titles, and so on. One of the authors of this article reviewed all the N.COMMUNICATION spans and classified them in three categories, two of them mapped from the EWN top ontology, N.DOCUMENT and N.LANGUAGE, and a third back-off category for N.COMMUNICATION. Notice how, in spite of having spawned three senses (N.CONTAINER, N.VEHICLE and N.BUILDING), N.ARTIFACT is still a very frequent supersense.

The document-language distinction is a high-level type in the SIMPLE ontology (Lenci et al., 2000). Note that these two new communication subsenses do not solve the artifact-information ambiguity commonly found in lexical semantics (Pustejovsky, 1991). While N.LANGUAGE has more often an eventual reading (e.g. *conversation*, *remark*), N.DOCUMENT refers more often to works and other entities with a non-temporal denotation. We also use N.LANGUAGE for the metalinguistic usage of words (e.g. 'The word **drizzle** sounds funny'). This re-annotation produces examples like the following:

> H. C. Andersen er jo verdensberømt , fordi hans **forfatterskab**/N.DOCUMENT er blevet oversat til alle sprog/N.LANGUAGE .
>
> *H. C. Andersen is world famous, because his **writing** has been translated to all languages.*

Out of the 1513 N.COMMUNICATION cases, 360 fall under N.LANGUAGE and 928 under

N.DOCUMENT, and the remaining were left with the original label. Out of the 929 N.DOCUMENT spans, 382 are named entities, where 248 are +2 tokens in length. This metric aims at justifying having document as an NER label, where span identification is as relevant as proper labeling.

We believe the frequency of document-name named entities makes a good case for considering the N.DOCUMENT class as an addition to the SSI and to NER. However, we do not find enough support to recommend a N.LANGUAGE supersense and prefer using the original N.COMMUNICATION instead.

5.2 Verbs

Verbs are central to the theory of lexical semantics, yet their semantic characterization has been closer to the syntax-semantics interface (Levin, 1993; Kipper et al., 2000; Kipper et al., 2006). In this aspect, the wordnet SSI for verbs is very different, e.g. verbs like *jump* or *displace* are of the V.MOTION, even though their argument structures are very different. Nevertheless, verbal sense alternations are often associated with different argument structures (Grimshaw, 1990).

The V.CHANGE supersense is populated with semantically disparate categories and is very difficult to annotate, even though it is a very frequent sense, both in terms of annotated words and of synsets adscribed to it. According to Fellbaum (1990), 'the concept of change is flexible enough to accomodate verbs whose semantic description mathen them unfit for any other semantically coherent group'. In other words, the rummage box category for verbs is actually the majority class. Indeed, an expansion of change into its subsenses of CHANGE-VARY, CHANGE-STATE, CHANGE-REVERSAL, CHANGE-INTEGRITY, CHANGE-SHAPE and CHANGE-ADAPT could potentially make the supersense more useful, if one is willing to incur the cost of annotating with five more labels.

The V.PHENOMENON supersense extends V.CHANGE by delimiting events that have no agency and are not weather-related, such as *happen*, or *occur*. WordNet shows a systematic ambiguity between V.STATIVE and V.CHANGE for aspectual readings of verbs, and we also propose V.ASPECTUAL for constructions like '**start** the engine' or '**begin** to hope'.

We evaluate verb sense using the criteria we used for nouns in Section 5.1, but discarding NER compliance, which does not apply to verbs. Table 5 shows the criteria for verbs.

New supersense	Agr.	Freq.	Assc.
ASPECTUAL	x		x
PHENOMENON	x	x	

Table 5: Inclusion criteria for new verb senses.

Both new verbal supersenses satisfy two out of three of the criteria, and we can consider them candidates for the SSI extension. We leave it for further discussion whether aspectual verb reading deserves a full-fledged supersense or should be used as a satellite tag (cf. Section 5.4).

5.3 Adjectives

SST as defined by Ciaramita and Johnson (2003) only labels nouns and verbs. Adjectives have received much less attention than nouns and verbs, arguably because of the inherent difficulty of their analysis, cf. Boleda et al. (2012) for a survey on adjective classifications. In addition to the theoretical complications, adjectives are not regarded as core elements of meaning when building applications. For instance, in WordNet 3.0 there are 82k synsets for nouns, 14k for verbs, 18k for adjectives and 4k for adverbs. However, the base concepts from EWN (Vossen et al., 1998), with 4,869 synsets in total, hold 37 adjectives in contrast to 3,210 nouns and 1,442 verbs.

Moreover, the supersense-synset relation is hyponimic, but adjectives in WordNet are not taxonomically organized (Gross and Miller, 1990). For instance, there is no way to retrieve that *ashamed* and *exasperated* are emotional in nature (Tsvetkov et al., 2014b).

The meaning plasticity of adjectives makes it also hard to determine whether adjectives hold any meaning onto themselves, or their meaning is an emergent property of the relation they establish with the noun they complement. Murphy and Andrew (1993) consider adjectives monosemous elements that define their sense when predicated alongside nouns. Under this light, supersense adjectives would be superflous if adjective meaning is an epiphenomenon of noun meaning.

However, insofar adjectives can help disambiguate nominal polysemy (Tsvetkov et al.,

2014a), and have different listed synsets, we advocate for providing a set of supersenses for adjectives. This addition makes therefore SST truly all-words for the three main lexical parts of speech. Adjective classifications into supersenses or coarse classes do exist, notably in GermaNet (Hamp and Feldweg, 1997), which Tsvetkov et al. (2014b) apply to English.

When applying the projection method from Section 2, we extend A.ALL with A.MENTAL, A.PHYS, A.SOCIAL and A.TIME. These supersenses do not distinguish descriptive (i.e. extensional) from reference-modifying (intensional) adjectives, e.g. *former* is A.TIME while *imaginary* is A.MENTAL. These senses do not distinguish relational adjectives either, to the extent that *ecologic* and one of the senses of *green* should fall under the same supersense.

The new adjective SSI cannot be evaluated in the samme manner as nouns. The adjective SSI is much smaller, and the agreement and frequency metrics can be misleadingly positive. Indeed, all adjective supersenses satisfy the agreement and frequency criteria specified in Section 5.1.1.

However, A.ALL is the most frequent supersense for adjectives, and it covers 40% of the annotated adjectives. This proportion is too large, and indicates the sense inventory needs to be further specified in order to minimize how many tokens get assigned the backoff sense.

Many of the adjectives under A.ALL are function-appraisal related, such as *god* ('good'), *bedre* (better'), *stor* ('large' as in 'grand'), *vigtig* ('important'). While polarity is an important property of adjectives (Chesley et al., 2006), we do not consider it a desirable trait for supersenses, which are more oriented towards conveying sense denotation that connotation. Hence, we suggest a new supersense A.FUNCTION to give account for function-related senses, what in the terminology of Pustejovsky (1991) would be the *telic role*. We observe that the ALLGEMEIN ('general') category of GermaNet and Tsvetkov et al's MISCELLANEOUS hold similar senses.

5.4 Satellites

When annotating nouns in Section 3, we annotate continuous NER-like spans. But verb-headed multiwords pose a challenge because they are not necessarily continuous, and pose attested challenges for their annotation and automatic recognition (Hoppermann and Hinrichs, 2014; Baldwin, 2005b; Baldwin, 2005a).

We use three satellite tags; S.COLLOCATION, S.PARTICLE and S.REFLPRON (for reflexive pronouns). While the particle distinction is more relevant for satellite-framed languages (Talmy, 1985) like Germanic languages, light-verb constructions are pervasive in many languages, also characteristically verb-framed languages like Spanish or French, where we find verb-headed multiwords like *llevar a cabo* (lit. 'take to ending', 'carry out') or *avoir l'air* (lit. 'to have the air', 'seem'), respectively. A similar approach has been used by Schneider and Smith (2015).

The intention of these tags is to help isolate the head of a verb-headed multiword. We assign the sense label to the syntactic head, even though a light verb construction would be arguably best headed by its introduced noun. In this manner, *gøre grin af* ('make fun of') would be labeled as *gøre*/V.COMMUNICATION *grin*/S.COLLOCATION *af*/S.COLLOCATION', and we thus avoid giving *gøre* ('make') the V.CREATION sense.

6 Conclusions and further work

We suggest an extension of the SSI for the three main lexical parts of speech. We obtain new supersenses using a mapping from ontological types, and evaluating their distribution after an evaluation task. Most of the new suggested senses satisfy the inclusion criteria we determine. In particular, we advocate for an inclusion of the senses N.DOCUMENT and N.INSTITUTION, which improve NER compatibility.

The extension method can be applied to any wordnet where the synsets are associated to EWN ontological types. Nevertheless, the inclusion criteria might change when dealing with different languages or corpus types. Moreover, the SSI proposed in this article can be applied retroactively to any EWN-aligned synset-annotated corpus.

With regards to adjectives, the backoff A.ALL category still constitutes 40% of the annotated adjectives. In future work, we consider including senses from the GermaNet inventory, and experimenting with data-driven approaches to infer lexical categories for adjectives by means of their relations to other words in wordnets, following the work of Alonge et al. (2000), Mendes (2006), Nimb and Pedersen (2012) and corpus-based approaches like Lapata (2001).

Acknowledgements

We wish to thank Nathan Schneider and Yulia Tsvetkov for their useful comments.

References

Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Adriana Roventini, and Antonio Zampolli. 2000. Encoding information on adjectives in a lexical-semantic net for computational applications. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 42–49. Association for Computational Linguistics.

Javier Alvez, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, and German Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. In *LREC*.

Jørg Asmussen and Jakob Halskov. 2012. The CLARIN DK Reference Corpus. In *Sprogteknologisk Workshop*.

Timothy Baldwin. 2005a. Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398–414.

Timothy Baldwin. 2005b. Looking for prepositional verbs in corpus data. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 180–9.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling regular polysemy: A study on the semantic classification of Catalan adjectives. *Computational Linguistics*, 38(3):575–616.

Matthias Buch-Kromann, Line Mikkelsen, and Stine Kern Lynge. 2003. Danish dependency treebank. In *TLT*.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, page 29.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of Proceedings of EMNLP*, pages 594–602, Sydney, Australia, July.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 168–175. Association for Computational Linguistics.

Christiane Fellbaum. 1990. English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278–301.

Jane Grimshaw. 1990. *Argument structure.* the MIT Press.

Derek Gross and Katherine J Miller. 1990. Adjectives in wordnet. *International Journal of Lexicography*, 3(4):265–277.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer.

Christina Hoppermann and Erhard Hinrichs. 2014. Modeling prefix and particle verbs in GermaNet. *Global Wordnet Conference*, page 49.

Rubén Izquierdo, Armando Suárez, and German Rigau. 2009. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 389–397. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, Héctor Martínez, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. In *Lexical and Computational Semantics (*SEM 2014)*.

Karin Kipper, Hoa Trang Dang, Martha Palmer, et al. 2000. Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, et al. 2000. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Oier Lopez de Lacalle and Eneko Agirre. 2015. Crowdsourced word sense annotations and difficult words and examples. *IWCS*.

Héctor Martínez Alonso, Anders Johannsen, Anders Søgaard, Sussi Olsen, Anna Braasch, Sanni Nimb, Nicolai Hartvig Sørensen, and Bolette Sandford Pedersen. 2015. Supersense tagging for danish. In *Nodalida*.

Sara Mendes. 2006. Adjectives in wordnet.pt. In *Proceedings of the GWA 2006–Global WordNet Association Conference*.

George A Miller. 1990. Nouns in wordnet: a lexical inheritance system. *International journal of Lexicography*, 3(4):245–264.

Gregory Leo Murphy and Jane M Andrew. 1993. The conceptual basis of antonymy and synonymy in adjectives. *Journal of memory and language*, 32(3):301–319.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Sanni Nimb and Bolette Sandford Pedersen. 2012. Towards a richer wordnet representation of properties–exploiting semantic and thematic information from thesauri. *LREC 2012*, pages 3452–3456.

Sussi Olsen, Bolette Sandford Pedersen, Héctor Martínez Alonso, and Anders Johannsen. 2015. Coarse-grained sense annotation of Danish across textual domains. In *NODALIDA*.

Rebecca J Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. Dannet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299.

Wim Peters, Ivonne Peters, and Piek Vossen. 1998. Automatic sense clustering in EuroWordNet. In *LREC*. Paris: ELRA.

James Pustejovsky. 1991. The generative lexicon. *Computational linguistics*, 17(4):409–441.

Nathan Schneider and Noah A Smith. 2015. A corpus and model integrating multiword expressions and supersenses. *Proc. of NAACL-HLT. Denver, Colorado, USA. To appear*.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proc. of The 9th Linguistic Annotation Workshop*, pages 112–123, Denver, Colorado, USA, June.

Frédérique Segond, Anne Schiller, Gregory Grefenstette, and Jean-Pierre Chanod. 1997. An experiment in semantic tagging using hidden Markov model tagging. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors, *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications: ACL/EACL-97 Workshop Proceedings*, pages 78–81, Madrid, Spain, July.

Anders Søgaard, Barbara Plank, and Héctor Martínez Alonso. 2015. Using frame semantics for knowledge extraction from Twitter. In *AAAI*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Leonard Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3:57–149.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *In CoNLL*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014a. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014b. Augmenting English adjective senses with supersenses. In *Proc. of LREC*.

Piek Vossen, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. The EuroWordNet base concepts and top ontology. *Deliverable D017 D*, 34:D036.

Fei Wu and Daniel S Weld. 2008. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, pages 635–644. ACM.

Adverbs in plWordNet: Theory and Implementation

Marek Maziarz[A], **Stan Szpakowicz**[B], **Michał Kaliński**[A]
[A] Department of Computational Intelligence
Wrocław University of Technology, Wrocław, Poland
[B] School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada
[A] mawroc@gmail.com, m.kalinski.9@gmail.com
[B] szpak@eecs.uottawa.ca

Abstract

Adverbs are seldom well represented in wordnets. Princeton WordNet, for example, derives from adjectives practically all its adverbs and whatever involvement they have. GermaNet stays away from this part of speech. Adverbs in plWordNet will be emphatically present in all their semantic and syntactic distinctness. We briefly discuss the linguistic background of the lexical system of Polish adverbs. We describe an automated generator of accurate candidate adverbs, and introduce the lexicographic procedures which will ensure high consistency of wordnet editors' decisions about adverbs.

1 Adverbs in wordnets and monographs

Adverbs have yet to receive their due in wordnets.

> There are only few adverbs in WordNet (hardly, mostly, really, etc.) as the majority of English adverbs are straightforwardly derived from adjectives via morphological affixation (surprisingly, strangely, etc.)[1]

> GermaNet shares the basic division of the database into the four lexical categories noun, adjective, verb, and adverb with WordNet®, although it is not planned to implement adverbs in the current work phase.[2]

Curiously, English monographs on lexical semantics (Cruse, 1997; Geeraerts, 2010) give the adverb a short shrift. The term does not even appear in the index of either book!

Yes, most adverbs do derive from adjectives.[3] And yet, the adverb is a *bona fide* open-class part of speech. Its distinctness and its peculiarities cannot be "swept under the carpet" by making it merely an inflected adjective.

Polish morphology acknowledges the adverb grudgingly, but at least it is present in several monographs, notably in (Grzegorczykowa, 1975).

The paper presents a definition of adverbs in plWordNet (section 2), a procedure to generate candidate adverbs (section 3), a manual verification (section 4) and a few conclusions (section 5).

2 Adverbs in plWordNet

The designers of plWordNet established a spectrum of relations for nouns, verbs and adjectives (Maziarz et al., 2011a; Maziarz et al., 2011b; Maziarz et al., 2012). Table 1 lists the relations for adverbs, with examples.[4] The list is based on the adjective model (Maziarz et al., 2012); we have assumed that those relations will fit adverbs, given that most adverbs are transposition derivatives from adjectives.

Every relation type has its own test expressions. (The substitution of lexical units for variables yields correct expressions in Polish.) Language forces the tests to be polymorphic. That is because an adverb can modify a verb, an adjective or an adverb, and it can appear in a predicative position (*jest* 'to be$_{3^{rd}\ person}$' + adverb).

[1] https://wordnet.princeton.edu/

[2] http://www.sfs.uni-tuebingen.de/lsd/germanet_structure.shtml – dated 2009

[3] Calculations on dictionary material show that only 1% of all adverbs is not derived from adjectives (Grzegorczykowa, 1998, p. 524).

[4] See http://tinyurl.com/okdc5w7 for all relations and wordnet editors' instructions (in Polish).

Relation type	definition
Synset relations	
hyponymy	gorączkowo$_1$ 'frantically' → nerwowo$_1$ 'anxiously'
value of the attribute	intensywnie$_2$ 'intensively' → intensywność$_1$ 'intensity'
gradation	brązowawo$_1$ 'in brownish colour' → brązowo$_2$ 'in brown colour'
fuzzynymy	weselnie$_1$ 'in a wedding mood' → wódka$_1$ 'vodka'
inter-register synonymy	dziwnie$_1$ 'strangely' → dziwno$_1$ 'strangely (obsolete)'
Lexical unit relations	
antonymy	apriorycznie$_1$ 'a priori' ↔ aposteriorycznie$_1$ 'a posteriori'
converseness	lepiej$_1$ 'better' ↔ gorzej$_1$ 'worse'
XPOS synonymy	gorączkowo$_1$ adv. 'frantically' → gorączkowy$_1$ adj. 'frantic'
degree	lepiej$_1$ 'better' → dobrze$_1$ 'well'
derivation	intonacyjnie$_1$ 'with regard to intonation' → intonacja$_3$ 'intonation'

Table 1: Relations in plWordNet with examples.

2.1 Synset relations

Synset relations are short-cuts for a bundle of links between lexical units belonging to two different synsets (Maziarz et al., 2013, pp. 774-775). Our test expression, then, admit pairs of lexical units belonging to synsets which are supposed to be linked by a synset relation.

We present four such tests for *hyponymy*.[5] Symbols x, y denote adverb lexical units. The awkward phrase 'does it x' is meant as "does it in a manner x", *etc.*

When we insert actual words into these tests, we can decide whether the resulting assertion is true. For example, let x and y in Listing 1 be *gorączkowo*$_1$ 'frantically' and *nerwowo*$_1$ 'anxiously.'

- Jeżeli ktoś robi coś *gorączkowo*$_1$, to robi to *nerwowo*$_1$. 'If someone does something frantically, he does it anxiously.'
- Jeżeli ktoś/coś robi coś *nerwowo*$_1$, to niekoniecznie robi to *gorączkowo*$_1$. 'If someone does something anxiously, he does not necessarily do it frantically.'

Both these statements hold for Polish: the relation **hypo**(*gorączkowo*$_1$, *nerwowo*$_1$), then, is an instance of hyponymy in plWordNet.

Listing 1: Hyponymy. Modifier of intentional verbs.

Jeżeli ktoś/coś robi coś **x**, to robi to **y**.
Jeżeli ktoś/coś robi coś **y**, to niekoniecznie robi to **x**.

```
'If someone/something does something x,
they do it y.'
'If someone/something does something y,
they do not necessarily do it x.'
```

Listing 2: Hyponymy. Modifier of unintentional verbs.

Jeżeli coś dzieje się **x**, to dzieje się **y**.
Jeżeli coś dzieje się **y**, to niekoniecznie dzieje się **x**.

```
'If something happens x, it happens y.'
'If something happens y,
it does not necessarily happen x.'
```

Listing 3: Hyponymy. Adjective modifier.

Jeżeli ktoś/coś jest **x** jakiś, to jest też **y** jakiś.
Jeżeli ktoś/coś jest **y** jakiś, to niekoniecznie jest **x** jakiś.

```
'If someone/something is x so,
they are also y so.'
'If someone/something is y so,
they are not necessarily x so.'
```

Listing 4: Hyponymy. Predicative adverb.

Jeżeli jest **x**, to jest też **y**.
Jeżeli jest **y**, to niekoniecznie jest **x**.

```
'If it is x, it is also y.'
'If it is y, it is not necessarily x.'
```

Let us now put the hyponymous pair *fiołkowo*$_1$ '$\pm$ like a violet' and *słodko*$_2$ 'sweetly' in Listing 2, and replace the generic non-volitional *dzieje się* 'it happens' with its hyponym pachnie 'it smells':

- Jeżeli coś pachnie *fiołkowo*$_2$, to pachnie *słodko*$_3$. 'If something smells like a violet, it smells sweetly.'
- Jeżeli coś pachnie *słodko*$_3$, to niekoniecznie pachnie *fiołkowo*$_2$. 'If something smells sweetly, it does not necessarily smell like a violet.'

In Listing 3, we put the hyponymous pair *bordowo*$_1$ 'maroon$_{adv}$' and *ciemnoczerwono*$_1$ 'dark-red$_{adv}$' and a specific passive participle *zabarwiony* '*-hued' to replace the generic "so".

[5] We give separate tests for the adjective modifier, the predicative position, and the modifiers of intentional and unintentional verbs; Laskowski (1998) gives an exact definition.

- Jeżeli coś jest *bordowo*$_1$ zabarwione, to jest też *ciemnoczerwono*$_1$ zabarwione. 'If something is maroon-hued, it is also dark-red-hued.'
- Jeżeli coś jest *ciemnoczerwono*$_1$ zabarwione, to niekoniecznie jest *bordowo*$_1$ zabarwione. 'If something is dark-red-hued, it is not necessarily maroon-hued.'

Finally, two hyponymous adverbs in a predicative context (to be$_{3^{rd}person}$ + adverb).[6]

- Jeżeli jest *słonecznie*$_6$, to jest też *bezchmurnie*$_4$. 'If it is sunny$_{adv}$, it is also cloudless$_{adv}$'.
- Jeżeli jest *bezchmurnie*$_4$, to niekoniecznie jest *słonecznie*$_6$. 'If it is cloudless$_{adv}$, it is not necessarily sunny$_{adv}$'.

If any of these four tests admits a given pair of lexical units, we will say they are a hyponymy pair.

The relation **value of the attribute** resembles hyponymy. It holds between an adverb, treated as a feature value and a noun, which represents certain category (attribute). For example, the attribute *intensywność*$_1$ 'intensity', has several values, among them the adverbs $intensywnie_2$ 'intesively', $fanatycznie_1$ 'fanatically' and $wydajnie_3$ 'about cough in medicine: efficiently'. Actual hyponymy and value of the attribute together form the backbone of plWordNet's adverb structure.

The **gradation** relation is applied when a series of adverbs can be arranged into a sequence according to some scale. The adverbs *brązowawo*$_1$ 'in brownish colour' and *brązowo*$_2$ 'in brown colour' represent the same attribute *hue* and could be ordered according to that attribute. Adverb sequences can be quite long. Consider adverbs of temperature: *lodowato*$_1$ 'icily', *zimno*$_5$ 'coldly', *zimnawo*$_1$ 'coldishly', *chłodno*$_6$ 'coolly', *chłodnawo*$_1$ 'coolishly', *letnio*$_1$ 'lukewarmly', *ciepło*$_1$ 'warmly', *gorąco*$_1$ 'hotly'.

Inter-register synonymy links adverbs which would be synonymous if not for minor differences in register (in usage). For example, the adverbs *dziwnie*$_1$ and *dziwno*$_1$ occupy nearly the same place in plWordNet's lexico-semantic relation net. They are related to the same lexical units, except for hyponymy (see Figure 1 at the end of section 3). They cannot be in the same synset: *dziwno*$_1$ is obsolete, so is a poor hypernym choice for contemporary vocabulary, while *dziwnie*$_1$ belongs to the general register.

[6]Unlike English, Polish allows both adjectives and adverbs in this position.

2.2 Lexical unit relations

The most prominent relation among lexical units is **cross-categorial synonymy**, which we refer to as XPOS synonymy. It binds the adjectival net with the adverbial net. Almost all plWordNet adverbs are related to their derivative bases.[3] An adverb x and its adjective base a are XPOS-synonymous if they can be replaced in the nominalisation process – see (Nagórko, 1987, p. 140) and (Jędrzejko, 1993, p. 61). Two transpositions are possible from a verb context to a nominalised phrase (denoted by the symbol $\Rightarrow$):

- krzątał się gorączkowo 'he bustled frantically' $\Rightarrow$ gorączkowa krzątanina 'frantic bustle',
- jest zimno na ulicy 'it is cold in the street' $\Rightarrow$ zimna ulica 'cold street'.

The test expressions make use of these transpositions. Let us present a test for a modifier of *intentional* verbs (Listing 5; x is an adverb, a is an adjective).

Listing 5: XPOS synonymy. Modifier of intentional verbs.

Jeżeli ktoś/coś robi coś **x**,
to jest to **a** robienie czegoś przez kogoś/coś.
Jeżeli to jest **a** robienie czegoś przez kogoś/coś,
to ktoś/coś robi to **x**.

```
'If someone/something does something x,
then it is a doing it by someone/
something.'
'If it is a doing something by someone/
something, then someone/something does
not necessarily do it x.'
```

For *gorączkowo*$_1$ and *gorączkowy*$_1$, we get the following test expressions:

- Jeżeli ktoś/coś robi coś *gorączkowo*$_1$, to jest to *gorączkowe*$_1$ robienie czegoś przez kogoś/-coś. 'If someone/something does something frantically, then it is frantic doing something by someone/something.'
- Jeżeli jest to *gorączkowe*$_1$ robienie czegoś przez kogoś/coś, to ktoś/coś robi coś *gorączkowo*$_1$. 'If it is frantic doing something by someone/something, then someone/something does something frantically.'

The tests check the truth of two hyponymy-like implications which go in opposite directions.

Since synonymy can bee seen as bi-directional hyponymy, the tests effectively investigate synonymy conditions for the two parts of speech.

Apart from XPOS-synonymy, the adverbial plWordNet has two more derivationally motivated relations: **degree** and **derivation**. The former caters for synthetic comparatives and superlatives.[7] The latter is a catch-all for other derivational relations.

Antonymy links two adverb lexical units if they satisfy the conditions in Listing 6.

Listing 6: Antonymy. Predicative context.

– Jest **x**? – Wręcz przeciwnie: jest **y** .
Jeżeli jest **x**, to nie jest **y**.
Jeżeli nie jest **x**, to niekoniecznie jest **y**.

```
- Is it x? - On the contrary: it is y.
'If it is x, then it is not y.'
'If it is not x, then it is
    not necessarily y.'
```

Semantic opposition was introduced into this test with a short dialogue, with the key word *przeciwnie* 'on the contrary, conversely' (note the predicative context):[8]

- – Jest **x**? '– Is it **x**?'
- – Wręcz przeciwnie: jest **y**. 'On the contrary: it is **y**.'

Consider the pair *słonecznie*$_6$ 'sunny$_{adv}$' and *deszczowo*$_1$ 'rainy$_{adv}$':

- – Jest *słonecznie*$_6$? – Nie, wręcz przeciwnie: jest *deszczowo*$_1$. '– Is it sunny? – On the contrary: it is rainy.'
- Jeżeli jest *słonecznie*$_6$, to nie jest *deszczowo*$_1$. 'If it is sunny, then it is not rainy.'
- Jeżeli nie jest *słonecznie*$_6$, to niekoniecznie jest *deszczowo*$_1$. 'If it is not sunny, then it is not necessarily rainy.'

[7]Degree in Polish adverbs is either synthetic (affix-*ej* for comparatives and *naj-...-ej* for superlatives) or analytic (precede with the adverb *bardzo* 'more' or *najbardziej* 'most', respectively) (Grzegorczykowa, 1998, pp. 533-534).

[8]We follow here a very interesting synonymy test (Cruse, 1997, pp. 257-258): "[N]ot all lexical items are felt to have opposites. Ask someone for the opposite of *table*, or *gold*, or *triangle*, and he will be unable to oblige. Some lexical items, it seems, are inherently non-opposable." The dialogue from our test suggests a language-game in oppositions ("[a]sk someone for the opposite of..."). This helps us throw out those lexical unit pairs which only satisfy the main condition of antonymy, *i*.e., the incompatibility implication $x \Rightarrow \sim y$ (Lyons, 1981, 154-155).

According to Lyons (1981), converseness is quite frequent among adverbs in the comparative degree whose positive degree is involved in antonymy. We found many such pairs. Listing 7 shows tests for an adjective modifier.

Listing 7: Converseness. Predicative context.

Jeżeli p robi coś **x** niż q, to q robi to **y** niż p.

```
'If p does something x than q,
  then q does it y than p.'
```

For example, the lexical units *wolno*$_6$ 'slowly' and *szybko*$_3$ 'quickly' have the comparatives *wolniej*$_1$ 'more slowly' and *szybciej*$_1$ 'more quickly'. The test becomes:

- Jeżeli p robi coś wolniej niż q, to q robi to szybciej niż p. 'If p does something more slowly than q, then q does it more quickly than p.'

3 Automatic generation of candidate adverbs

We followed six steps in the generation of new adverbs from their adjective bases. We worked all along with a copy of plWordNet, which we denote plWordNet$_c$.

1. **Derivator**. Consider every existing adjective lemma X within the domain `qualitative` in plWordNet$_c$. Using the Derivator tool (Piasecki et al., 2012) create all possible adverbial derivatives A of adjectives X housed in plWordNet$_c$. The resulting lexicon L contains adverb-adjective pairs of lemmas (A, X).

Table 2 presents the statistics of the derivation process. Since mainly qualitative adjectives form their adverbs, it is interesting that more than one-third of them have their derivatives. For example, for the adjective *czyściutki* 'pleasantly clean, clear, pure' the Derivator created its adverb derivative *czyściutko* '≈cleanly, neatly; purely', whereas for the adjective *poszkodowany* 'injured, damaged' no adverb was derived.

2. **Adverbial lexical units.** For every given qualitative adjective lexical unit x in plWordNet$_c$ representing lemma X which is present in L, create its counterpart lexical unit a representing lemma A. Omit the lexical units housed in artificial (non-lexical) synsets (Piasecki et al., 2009, p. 30). Equip every thus created adverb lexical unit with register labels and glosses copied from the corresponding adjective unit.

Lemma type	Freq.	[%]
Adj. lemmas	27,042	100.0
Qualitative Adj. lemmas	17,045	63.0
Adv. derivative lemmas, $\|L\|$	6,321	23.4

Table 2: Statistics for automatic adverb derivation by the Derivator and plWordNet$_c$. Abbreviations: Adj. – adjective, Adv. – adverb, $|L|$ – cardinality of the set L.

The rule states that whenever an adjective lexical unit x from the domain `qualitative` has an entry (A, X) in the dictionary L, we create for it its counterpart lexical unit a. For example, lemma *czyściutki* has 5 senses in plWordNet$_c$ in the domain `qualitative`, so the lemma *czyściutko* would have also 5 senses (as).

3. **Filtering rules.** Having created counterparts as for senses xs, we perform filtering based on six rules. Two of them are shown in Listings 8-9. If a rule's premise holds, we remove from plWordNet$_c$ the considered sense a_0 of a given adverb lemma A.

Listing 8: Illustration for rule #1.

$$\begin{array}{l}\mathbf{mod}(x_0, istota_1) \ \vee \\ \exists y \ [\mathbf{mod}(x_0, y) \ \wedge \ \mathbf{hypo}'(y, istota_1)] \ \vee \\ \exists y \ [\mathbf{hypo}'(x_0, y) \ \wedge \ \mathbf{mod}(y, istota_1)] \ \vee \\ \exists y, n \ [\mathbf{hypo}'(x_0, y) \ \wedge \ \mathbf{mod}(y, n) \ \wedge \ \mathbf{hypo}'(n, istota_1)]\end{array}$$

Symbols x_0, y, z in Listing 8 are lexical units, x and y are adjectives, a_0 is an adverb counterpart of adjective x_0, n is a noun. The noun $istota_1$ means 'being, causal agent, human being, spirit or animal'; $\mathbf{hypo}'(x, y)$ holds if y is a direct or indirect hypernym of x; $\mathbf{mod}(x, n)$ holds if x is a modifier of n; $\mathbf{val}(x, n)$ holds if x is a value of the attribute n.

Listing 9: Illustration for rule #4.

$$\begin{array}{l}\mathbf{val}(x_0, zachowanie_7) \ \vee \\ \exists y \ [\mathbf{hypo}'(x_0, y) \ \wedge \ \mathbf{val}(y, zachowanie_7)]\end{array}$$

Symbols in Listing 9 – see Listing 8. The noun $zachowanie_7$ means 'behaviour, manner of acting or controlling oneself'.

Rules #2 and #3 are derived from rule #1 by replacing $istota_1$ with $organizm_1$ 'living organism' and $grupa_5$ 'group of people', respectively. Rules #5 and #6 arise from rule #4 by replacing $zachowanie_7$ by *cecha osobowości*$_1$ 'character trait' and $pochodzenie_5$ 'origin, source of someone/something', respectively. The rules are based on a simple random sample of 69 *adjective* lexical units from plWordNet$_c$ (more in Section 4).

4. **Synsets.** Group all adverbial lexical units into synsets, mirroring their counterpart adjective synsets: two adverb units a_1, a_2 are in the same synset **iff** the corresponding adjective lemmas x_1, x_2 are in the same synset. An adjective lemma can also correspond to two or more adverb lemmas (each with perhaps a slightly different meaning). In such cases, all adverb lexical units $a_1, a_2, \ldots$ are considered counterparts of the same adjective lexical unit x; the register `obsolete` (Maziarz et al., 2014; Maziarz et al., 2015) is assigned to all a_k except the unit of the most frequent adverb lemma.

For example, the lemma *żmudny* 'arduous; laborious' has only one meaning in plWordNet, but two adverbial derivatives in the lexicon L: *żmudnie*, *żmudno* 'arduously; laboriously' (of which the second one is almost absent in modern Polish texts). It has also one synonym *mozolny*. Since *mozolny* has its own adverb derivative *mozolnie*, finally, we get a 3-element synset: {żmudnie$_1$, żmudno$_1$ (`obsolete`), mozolnie$_1$}.

5. **XPOS synonymy.** Add the cross-categorial (XPOS) synonymy between adverb lexical units a and the corresponding adjective lexical units x.

For the adverbs described above, the XPOS synonymy relation instances are the following:

żmudnie → *żmudny*,
żmudno → *żmudny*,
mozolnie → *mozolny*.

The last step is to copy relations from the adjective part of plWordNet$_c$.

6. **Copying relations.** Copy relations from the adjective part of plWordNet$_c$ onto the adverbial part. This step is split in two sub-steps, one for copying hyponymy chains, and another for copying various other relations.

(a) **Hypernymy/value.** If there is hyponymy between adjectives x and y, their counterpart adverbs a and b are also connected by hyponymy. There also may be "holes" in hyponymy chains, created by adjective synsets which do not have any corresponding adverb synsets (either not generated or filtered

out). Such "holes" are stepped over; see Listing 10.[9] For example, given an adjective chain $x_1 \rightarrow x_2 \rightarrow x_3$ such that only the adverbs a_1 and a_3 exist, the link $a_1 \rightarrow a_3$ is created. The relation "value of the attribute" is treated specially here; it may connect a top adjective hypernym in a chain to a noun. When copying this relation, a top adverb in a hypernymy chain will be linked to that noun if there is a hypernymy + value-of-the-attribute path from its counterpart to the noun; see Listing 11. Figure 1 is a descriptive example of this process.

(b) **Other relations.** Four other adjective-linking relations are copied onto their counterpart adverbs: gradation, inter-register synonymy, antonymy, and converseness. So, if one of these relations connects adjectives x_1, x_2, their counterparts a_1, a_2 will also be connected. Since these relations do not form chains, only immediate neighbours are considered; if one of the connected adjectives has no adverb counterpart, the relation will not be copied.

Listing 10: Illustration for hyponymy chain copying conditions.

$$\forall a,b\ \exists x,y\ \ \mathbf{hypo}'(a,b) \Leftarrow \\ \mathbf{hypo}'(x,y)\ \wedge\ \mathbf{xpos}(a,x)\ \wedge\ \mathbf{xpos}(b,y)$$

Listing 11: Illustration for value-of-the-attribute relation repair conditions.

$$\forall a,b\ \exists x,y,n\ \ \mathbf{val}(a,n) \Leftarrow \\ \mathbf{val}(x,n)\ \wedge\ \mathbf{xpos}(a,x)\ \vee \\ \mathbf{hypo}'(x,y)\ \wedge\ \mathbf{xpos}(a,x)\ \wedge\ \mathbf{val}(y,n)$$

Symbols x, y, a, b, n in Listings 10-11 are lexical units: x, y are adjectives, a and b are adverbs, n is a noun; $\mathbf{hypo}'(x,y)$ holds if y is a direct or indirect hypernym of x; $\mathbf{val}(x,n)$ holds if x is a value of the attribute n; $\mathbf{xpos}(a,x)$ holds if a is a cross-categorial synonym of x.

Figure 1 illustrates the rule with the hyponymy chain of the synset $\{\text{postrzelony}_2\}$ 'crazy'. There are 6 elements in the adjective path (on the left), including the value of the attribute relation. The Derivator did not create some derivatives, so the adverb structure (on the right) is not an exact copy of the adjective part. Luckily, in this case only derivatives forbidden in Polish (marked with "X" in the Figure) were omitted. Instances lacking relation were stepped over by pointing to the closest synset possible (*dziwnie – podobieństwo*).

[9] $\mathbf{hypo}'(\bullet, \bullet)$ stands for direct or indirect hyponymy.

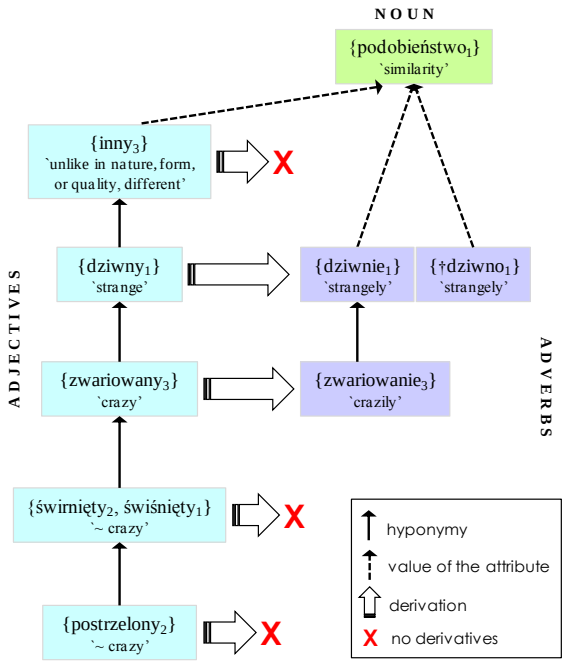


Figure 1: The hyponymy path for *postrzelony* 'crazy'. "X" marks synsets left empty by the algorithm in plWordNet$_c$.

4 Manual verification

We evaluate the procedure from section 3 in three experiments, two before copying plWordNet$_c$ onto plWordNet (S_L, S_T), and one afterwards (S_V). The former two were based on simple random samples of 69 (S_L) and 70 (S_T) *adjective* lexical units from plWordNet. The development set S_L helped write and check the filtering rules in Section 3. As a baseline B_L we chose the procedure's performance, without filtering, on the first set of 69 adjectives. The test set S_T was used to reassess the measures of efficiency. The randomly drawn adjectives were checked manually by plWordNet editors (all of them linguists with a university degree) for correspondence with adverbial lexical units .

In the S_L sample (Table 3).[10], two of 27 adverbs in plWordNet$_c$ are our procedure's "creation", and

[10] In Tables 3-5, $A+$ / $A-$ denote lexical units which are / are not proper Polish adverbs. $W+$ / $W-$ denote lexical units present / not present in plWordNet$_c$, because either the Derivator did not create them, or they were filtered by rules #1-#6 from step 3 in section 3. $P(W+)$ and $R(A+)$ are precision and recall of recognising real adverb lexical units. CI is the confidence interval.

25 of 36 existing adverbs were introduced into plWordNet$_c$. Let us calculate the precision of introducing adverbs into plWordNet $P(W_+)$ and recall of automatic recognition of adverbial lexical units $R(A_+)$, the most important measures of reliability in this case ($N(\bullet)$ is set cardinality):

$$P(W_+) = N(W_+ \cap A_+)/N(W_+) = 93\% \quad (1)$$

$$R(A_+) = N(W_+ \cap A_+)/N(A_+) = 69\% \quad (2)$$

The set $W_+ \cap A_-$ contains false positives: adverbs which do not exist in reality but were introduced by the algorithm. The set $W_- \cap A_+$ contains false negatives: adverbs which do exist in language but were omitted by the algorithm. For illustration, we present their elements.

- $W_+ \cap A_-$ = {kurczliwy$_1$ 'contractible', żeński$_3$ 'female'}
- $W_- \cap A_+$ = {redukowalny$_1$ 'reducible', jednosetowy$_1$ 'one-set [*e*.g., in tennis]', polarny$_1$ 'arctic or antarctic', ropuchowaty$_1$ 'toadlike', włókienkowaty$_1$ 'fibrillose', brutalny$_2$ 'brutal', warzywny$_3$ 'vegetable$_{Adj}$', jednopasmowy$_1$ 'single-lane', równobrzmiący$_1$ 'consonant', pilśniowaty$_1$ 'felt-like', dwupolowy$_2$ 'bi-polar'}

Precision and recall answer two questions:

- How many automatically generated lexical units are real adverb lexical units?
- How many adverb lexical units that could be generated from copying structure from adjective part of plWordNetwere indeed created?

Our procedure performed better on the S_L sample, with a statistically significant increase of precision (from 70% to 93%), and a small, not significant, decrease of recall (from 72% to 69%). The size of the adverbial base in plWordNet$_c$ was only 10% smaller after filtering the original base (see the row M in Table 3).

The results were promising, so we drew yet another sample S_T. Now precision was still high, but recall was lower, however – since we ran the very same algorithm as in S_L – the size M of adverb plWordNet$_c$ (in lexical units) did not change.

With high precision and a reasonably slight "leakage" of lexical units (reasonably high M), we finally decided to copy plWordNet$_c$ onto the live base plWordNet. The plWordNet$_c$ set consisted of

	B$_L$ ($n = 69$)		**S**$_L$ ($n = 69$)	
	W_-	W_+	W_-	W_+
A_-	22	11	31	2
A_+	10	26	11	25
M	11,402		10,190	
$P(W_+)$ 95% CI	70%* [53÷84%]		93%* [76÷99%]	
$R(A_+)$ 95% CI	72% [55÷86%]		69% [52÷84%]	

Table 3: The confusion matrix for our automatic procedure on the development set. B_L – baseline, the procedure without filtering; S_L – the development set; M is plWordNet$_c$ size, n is sample size, both in lexical units. The asterisks mark statistically significant differences between B$_L$ and S$_L$ at the confidence level 95%.

	S$_T$ ($n = 70$)	
	W_-	W_+
A_-	20	4
A_+	24	22
M	10,190	
$P(W_+)$ 95% CI	85% [65÷96%]	
$R(A_+)$ 95% CI	45% [33÷63%]	

Table 4: The confusion matrix for our automatic procedure on the test set. M is plWordNet$_c$ size, n is sample size, both in lexical units.

	S$_V$ ($n = 517$)	
	W_-	W_+
A_-	NA	86
A_+	100	331
Z	241	
$P(W_+)$ 95% CI	79% [75÷83%]	
$R(A_+)$ 95% CI	78% [72÷81%]	

Table 5: The confusion matrix for our automatic procedure on the validation set. S_V – the validation set; Z – the number of adverb lemmas in S_V, and n – sample size in lexical units. Note that the cell $W_- \cap A'_-$ is empty because we changed the interpretation of recall.

10,190 lexical units. We gave the resulting "adverbial" plWordNet to a team of 10 editors, asking them to build upon this automatically generated

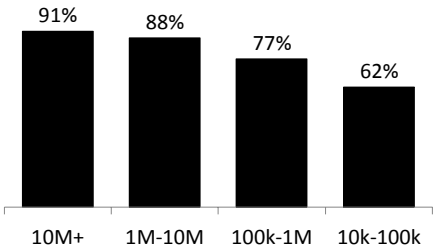


Figure 2: Coverage of lexicon built from plWordNet Corpus with regard to different frequency bins.

wordnet. Table 5 presents the results of manual verification of part of the automatically generated adverb wordnet; that is the validation set S_V. The conditions of the validation were different than in two earlier experiments S_L and S_T, in which the starting point were adjective lexical units. S_V contained only the *adverb* lemmas generated by the procedure and worked upon by the editors. In S_V, we were not interested in recall of adverbs derivable from the existing adjectives. We changed the interpretation:

- How many adverb lexical units which could have been introduced into plWordNet from generated adverb lemmas were indeed created?

Around one of four-five lexical units is not an appropriate adverb lexical unit; one of four-five existing senses of a given lemma is missing.[11]

5 Whither adverbs in plWordNet?

We have so far only considered adverbs which can be generated from adjectives in plWordNet. It stands to reason that coverage could increase if we worked instead with corpus-based frequency lists. Figure 2 presents coverage of a lexicon built from the plWordNet corpus.[12] The more frequent an adverb is, the more likely it is to appear plWordNet. Even for the least frequent adverbs, the coverage is still a high 62%.

[11]Note that this is no longer a simple random sample: editors work on packages with lists of senses of the same lemma, also synonyms and hyponyms/hypernyms of the senses. The sampling design most resembles cluster sampling. The confidence interval must be treated here as an approximation.

[12]The corpus consists of 250M tokens in the ICS PAS Corpus (Przepiórkowski, 2004); 113M tokens of news items (Weiss, 2008); ≈80M tokens in a corpus made of Polish *Wikipedia* (Wikipedia, 2010); an annotated corpus KPWr with ≈0.5M tokens (Broda et al., 2012); ≈60M tokens of shorthand notes from the Polish parliament; and ≈1.2 billion tokens collected from the Internet.

Table 6 shows that our procedure does not miss much. For example (row 3), it only omitted 1418 adverbs with frequency above 10.

	Adverb class	lemmas	%
1	in plWN, $f > 10$	3,720	42.8
2	in plWN, $f <= 10$	2,601	29.9
3	not in plWN, $f > 10$	1,418	16.3
4	multi-word adverbs, *po polsku* type, $f > 10$	958	11.0
	Total (with multi-word adverbs, a guess)	8,697 (≈9,000÷10,000)	100.0

Table 6: The estimated size of plWordNet's adverb list, based of frequencies (f) in the plWordNet corpus.

Row 4 in Table 6 refers to a productive class of multi-word adverbs such as *(mówić) po polsku, po angielsku* '(speak) Polish, English'. There also are other productive patterns, *e*.g., *(ubierać się) z polska, z niemiecka* '(dress) Polish-style, German-style', as well as non-compositional constructions, *e*.g., *z dobroci serca* 'out of the goodness of one's heart'. All such adverbial expressions must be added to plWordNet. The "po polsku" type is much more frequent than other types; we found almost 1,000 such word combinations in the corpus. Thus we estimate the number of all other multi-word adverb lexical units at yet another 1,000. We expect, all told, 9 to 10 thousand lemmas.

Clearly, the adding of adverbs to plWordNet is work in progress. Detailed instructions for the editors,[4] in keeping with our practice over the years, are meant to ensure the consistency of editorial decisions. Editors now verify, add to and complete the list of adverb lexical units, automatically generated from plWordNet's adjectives. Next, we plan to add multi-word lexical units of the *po polsku* type and of other types.

Acknowledgments

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE]. Thanks to Paweł Kędzia for help with the adverb generation algorithm. Thanks to Agnieszka Dziob and Justyna Wieczorek, the co-authors of the adverb guidelines, for the manual verification of the learning and test sets.

References

[Broda et al.2012] Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish.

[Cruse1997] Alan Cruse. 1997. *Lexical semantics*. Cambridge University Press.

[Geeraerts2010] Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press.

[Grzegorczykowa1975] Renata Grzegorczykowa. 1975. *Funkcje semantyczne i składniowe polskich przysłówków [The semantic and syntactic function of Polish adverbs]*. Ossolineum, Wrocław.

[Grzegorczykowa1998] Renata Grzegorczykowa. 1998. IV. Słowotwórstwo: Przysłówek [IV. Derivation: The adverb]. In Renata Grzegorczykowa, Roman Laskowski, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego [Grammar of contemporary Polish]*, volume 2 of *Morfologia [Morphology]*, pages 524–535. PWN, 2 edition.

[Jędrzejko1993] Ewa Jędrzejko. 1993. *Nominalizacje w systemie i w tekstach współczesnej polszczyzny [Nominalisations in language system and in texts of contemporary Polish]*. University of Silesia Press, Katowice.

[Laskowski1998] Roman Laskowski. 1998. Kategorie morfologiczne języka polskiego – charakterystyka funkcjonalna [The morphological categories of Polish – a functional characterisation]. In Renata Grzegorczykowa, Roman Laskowski, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego [Grammar of Contemporary Polish]*, volume 1 of *Morfologia [Morphology]*, pages 151–224. PWN, 2 edition.

[Lyons1981] John Lyons. 1981. *Language and Linguistics: An Introduction*. Cambridge University Press.

[Maziarz et al.2011a] Marek Maziarz, Maciej Piasecki, Stan Szpakowicz, and Joanna Rabiega-Wiśniewska. 2011a. Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition. *Cognitive Studies / Études Cognitives*, 11:161–181.

[Maziarz et al.2011b] Marek Maziarz, Maciej Piasecki, Stan Szpakowicz, Joanna Rabiega-Wiśniewska, and Bożena Hojka. 2011b. Semantic Relations Between Verbs in Polish Wordnet. *Cognitive Studies / Études Cognitives*, 11:183–200.

[Maziarz et al.2012] Marek Maziarz, Stan Szpakowicz, and Maciej Piasecki. 2012. Semantic Relations among Adjectives in Polish WordNet 2.0: A New Relation Set, Discussion and Evaluation. *Cognitive Studies / Études Cognitives*, 12:149–179.

[Maziarz et al.2013] Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796. DOI 10.1007/s10579-012-9209-9.

[Maziarz et al.2014] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. Registers in the System of Semantic Relations in plWordNet. In *Proc. 7th International Global Wordnet Conference*, pages 330–337.

[Maziarz et al.2015] Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2015. The system of register labels in plWordNet. *Cognitive Studies / Études Cognitives*, 15:in print.

[Nagórko1987] Alicja Nagórko. 1987. *Zagadnienia derywacji przymiotników [Issues on adjective derivation]*. Warsaw University Press.

[Piasecki et al.2009] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. http://www.eecs.uottawa.ca/ szpak/pub/A_Wordnet_from_the_Ground_Up.zip.

[Piasecki et al.2012] Maciej Piasecki, Radoslaw Ramocki, and Marek Maziarz. 2012. Recognition of Polish Derivational Relations Based on Supervised Learning Scheme. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 916–922, Istanbul, Turkey. European Language Resources Association (ELRA).

[Przepiórkowski2004] Adam Przepiórkowski. 2004. *The IPI PAN Corpus, Preliminary Version*. Institute of Computer Science PAS.

[Weiss2008] Dawid Weiss. 2008. Korpus Rzeczpospolitej. http://www.cs.put.poznan.pl/dweiss/rzeczpospolita. Corpus of text from the online edtion of Rzeczypospolita.

[Wikipedia2010] Wikipedia. 2010. Polish Wikipedia. https://pl.wikipedia.org, accessed in 2010.

A Language-independent Model for Introducing a New Semantic Relation Between Adjectives and Nouns in a WordNet

Miljana Mladenović
Faculty of Mathematics
University of Belgrade
ml.miljana@gmail.com

Jelena Mitrović
Faculty of Philology
University of Belgrade
jmitrovic@gmail.com

Cvetana Krstev
Faculty of Philology
University of Belgrade
cvetana@matf.bg.ac.rs

Abstract

The aim of this paper is to show a language-independent process of creating a new semantic relation between adjectives and nouns in wordnets. The existence of such a relation is expected to improve the detection of figurative language and sentiment analysis (SA). The proposed method uses an annotated corpus to explore the semantic knowledge contained in linguistic constructs performing as the rhetorical figure Simile. Based on the frequency of occurrence of similes in an annotated corpus, we propose a new relation, which connects the noun synset with the synset of an adjective representing that noun's specific attribute. We elaborate on adding this new relation in the case of the Serbian WordNet (SWN). The proposed method is evaluated by human judgement in order to determine the relevance of automatically selected relation items. The evaluation has shown that 84% of the automatically selected and the most frequent linguistic constructs, whose frequency threshold was equal to 3, were also selected by humans.

1 Introduction

In this paper, we want to demonstrate that a WordNet (WN) can be expanded by a new semantic relation between adjectives and nouns in a way that could allow for its usage in detecting figurative language and in existing methods of sentiment analysis. WN is used successfully for analysis of literal meaning of texts using SA methods (Pease et al., 2012), (Reyes and Rosso, 2012), (Rademaker et al., 2014). Resources that came out of the Princeton WordNet (PWN), such as SentiWordNet (Esuli and Sebastiani, 2006), (Baccianella et al., 2010), WordNetAffect (Strapparava and Valitutti, 2004) and others, which define the prior sentiment polarity (taken out of the context) of synsets are also being used. Still, the intensity of sentiment polarity of the lexical representation of synsets can be reduced, increased or completely changed in a given context with the usage of rhetorical figures from the group of Tropes — figures that change the meaning of words or phrases over which the figure itself is formed. These figures can be metaphor, metonymy, irony, sarcasm, oxymoron, simile, dysphemism, euphemism, hyperbole, litotes etc. (Mladenović and Mitrović, 2013). Analysing the usage of figurative language in the form of ironic similes, Hao and Veale (2010) noticed that they act similarly to valence shifters (Kennedy and Inkpen, 2006) "not", "never" and "avoid" in text, because they change the polarity of sentiment words or phrases. In general, modifiers decrease, increase or change the sentiment polarity of words or phrases. Tropes work in a similar way. By definition, irony and sarcasm change the polarity, dysphemism and hyperbole increase the existing level of sentiment expressiveness, while litotes and euphemism decrease that expressiveness. Metaphor, metonymy, oxymoron and simile have a more complex mechanism of affecting both directions of change regarding the strength and polarity of sentiment.

Automatic detection of figurative language is a new area of interest in the field of SA that can improve the existing SA methods. Reyes and Rosso (2012) showed that the precision of classification in an SA task can be improved significantly (from 54% to 89.05% max.) when predictors detecting figurative speech are involved, compared to a set of predictors that treat the text literally. Similarly, Rentoumi et al. (2010) improved the SA method of machine learning by integrating it with a rule-based method which detects the usage of figurative language, so the integrated meth-

ods achieved better precision than the baseline.

2 Related work

WordNet is a dynamic, flexible structure that can be expanded in different ways and for various purposes. In certain cases, introducing morpho-semantic relations results in solving the problems that stem from specificities of a language with rich morphology and derivation (Koeva et al., 2008). Otherwise, introducing new semantic relations can lead to the improvement of the representation of relations between synsets, e.g. Kuti et al. (2008) present a semantic relation *scalar middle* with which the antonimy relation of two descriptive adjective synsets is transformed into a triple gradable structure *lower-upper-middle*. Angioni et al. (2008) define a new relation *Commonsense* with which a literal in a synset is being connected with Wikipedia links in which it is described, while Maziarz et al. (2012) introduce a series of relations pertinent to adjectives, e.g. derivational relations *comparative* and *superlative* define gradable forms of descriptive adjectives. Derivational relation *similarity* defines a relation between an adjective and a noun such that, based on a given adjective, the structure or form of the object described by the noun can be discovered. Similarly, derivational relation *characterstic* defines a relation between an adjective and a noun where the contents or quality of an object described by the noun is known based on the adjective, e.g. based on the statement *"If someone is famous, then he is characterised by fame"* the relation *characteristic* will be set between the noun *fame* and the adjective *famous*.

The new semantic relation between nouns and adjectives in the Portuguese WordNet is described in (Marrafa et al., 2006) and (Mendes, 2006). This relation is given in the form of a pair of inverse relations *a characteristic of / has as a characteristic*. According to the authors, although the purpose of the relation is to mark significant characteristics of a noun expressed by an adjective (e.g.'{carnivorous} *is a characteristic of* {shark}'), the status of this relation in the sense of lexical knowledge is not completely clear. Authors also point out that introducing this new relation enriches a WordNet, that it can contribute to the process of determining the semantic domain of an adjective and that it can be included in reasoning applications. Veale and Hao also suggest specific enrichment of WordNet in their papers (Veale and Hao, 2008) and (Hao and Veale, 2010). As a source to be used in that enrichment, authors suggest semantic knowledge contained in language constructs of the form `as ADJ as a NOUN` which, in fact, are similes (e.g. "as free as a bird", "as busy as a bee"). In order to obtain examples of simile, the authors first extracted all antonymous pairs of adjectives in PWN and made a list of candidate adjectives. For each adjective ADJ from that list, a query in the form `as ADJ as a *` was made and sent to the Google search engine. Out of the obtained results, the first 200 snippets were kept. A collection of `as ADJ as a NOUN` constructs was made and a task of disambiguation was performed over it. In this process, one noun (*peacock*) can semantically be connected to many adjectives based on different semantic grounds. The structure, named by the authors as *frame:slot:filler*, consists of a noun (*frame*), property of the noun (*slot*) and an adjective as a value of the property (*filler*). For one noun there can be a number of instances of such structure. Authors point out that an average number of *slot:filler* constructs per one noun obtained in this particular research was 8. For instance, the noun *peacock* contains the following set of *slot:filler* values: {*Has_feather: brilliant; Has_plumage: extravagant; Has_strut: proud; Has_tail: elegant; Has_display: colorful; Has_manner: stately; Has_appearance: beautiful*}, therefore the suggested enrichment of WordNet only for the noun *peacock* leads to addition of 7 relations out of which the first one is of the form '{peacock} *Has_feather* {brilliant}'.

3 Motivation

The research described in this paper is based on the previously mentioned research results by Marrafa et al. (2006) and Mendes (2006), because we are searching for specific relations between nouns and adjectives. However, unlike the relation *has as a characteristic* which connects a number of nouns {shark, cobra, orca, predator,...} to the same adjective {carnivorous}, we consider those descriptive adjectives that are specific to a small set of nouns, or only to a single noun. In the process of generating of the new relation, we are proposing usage of the rhetorical figure simile which has a relatively high frequency of occurrence in texts written in a natural language. In that case, the re-

lation '{peludo} *is a characteristic of* {abelha}', meaning ('{furry} *is a characteristic of* {bee}'), which exists in the Portuguese WordNet, would not be an adequate example, but the new relation would be created based on the common rhetorical figure simile "as busy as a bee" in which case the relation would be '{busy} *specific of* {bee}'.

On the other hand, significant research, that the work described in this paper leans on, is depicted in papers by Veale and Hao (2008) and (2010), regarding the development of automatic methods of extracting semantic knowledge out of examples of the simile figures usage. We suggest extraction of linguistic constructs of the form `as ADJ as a NOUN` from the corpus annotated with PoS and lemmas, which means that, in contrast to the results of Google search engine, the search would be faster and more precise, because in one step, we would obtain the set of those potential figures of simile that have only nouns positioned at the end of the observed linguistic structure. Furthermore, if we do not take into account all of the attributes that are characteristic for a certain noun, but only those that are used the most in everyday language (measured by the frequency of occurrence of the corresponding figure simile in the observed corpus) we would get the possibility to describe the set of "noun-adjective" candidates for expansion of the existing structure of WordNet with one unique relation *(specificOf/specifiedBy)*. Introduction of a single relation would eliminate the risk pointed out in (Veale and Hao, 2008) that the introduction of a large number of relations expressed by the structure *slot:filler* would reduce the system's ability to recognize similar properties. In a case of one relation, for example, {frame: *Has_strut:* proud} and {frame: *Has_gait:* majestic} would be transformed into {frame: *specifiedBy:* proud} and {frame: *specifiedBy:* majestic}. Apart from that, taking into account only the most frequent ones, the described transformation would not involve all of the *slot:filler* structures of a certain noun, but only the most frequent one, which would, in the case of the noun *peacock* result in generating only one relation '{peacock} *specifiedBy* {proud}', and not all seven of them. If we introduce the frequency threshold as a parameter, its change can affect the number of *specificOf/specifiedBy* relations for the single noun synset, as well as for the total number of relations of that type.

With the suggested relation *specificOf/specifiedBy* we can determine the nature of the semantic connection between the concepts *arrow*, *light* and *rabbit*, which cannot be achieved with the existing PWN relations. Namely, the simile constructs *brz kao zec* "as fast as a rabbit", *brz kao svetlost* "as fast as light", *brz kao strela* "as fast as an arrow", obtained by querying over the Corpus of Contemporary Serbian, we can confirm that '{strela, svetlost, zec} *specifiedBy* {brz}'i.e. '{arrow, light, rabbit} *specifiedBy* {fast}'holds true.

4 Language-independent model for WordNet Expansion

The procedure of expansion with the relation *specificOf/specifiedBy* that we are proposing, will be shown on the example of expansion of the Serbian WordNet (SWN) (Krstev, 2008), but it can also be used for other wordnets. The procedure consists of the following steps:

1) From the annotated corpus of a natural language K_l extract linguistic constructs of the form `pridev kao imenica` (in the case of English `as ADJ as a NOUN`) and create the set *Sims* such that:

Sims={"as ADJ as a NOUN"}, sims$\in Sims \subset K_l$

In our case, from the Corpus of Contemporary Serbian Language[1] (Utvić, 2014) 59 concordances of the form "<as ADJ as a NOUN>" were generated, such as the following:

ri više.-Kakva je?-<**Bela kao mleko**>. *Ona traži isto crnog mrežastog šala,* <**lakog kao pero**>*, smele zelene dan od zatvorenika; lica* <**žuta kao limun**>*, radosno polete*
.....-<**White as milk**>. .....
....., <**light as a feather**>, ...
..... <**yellow as a lemon**>, ...

2) Eliminate all elements from the *Sims* set whose adjectives are not descriptive: *SimsRedycByAdj*={sims$\in Sims|ADJ\,'is\ descriptive'$} like in the following examples where the adjectives are possessive:

za taj dan. Jer reč je <**ljudska kao glad**>. *Nema za Drugog? Ljubav* <**majčinska kao vernost**>*, ljubav muško-*
..... <**human as hunger**>. .....
..... <**motherly as loyalty**>, ...

[1] `http://www.korpus.matf.bg.ac.rs/index.html/`

In our case, the result was
$|SimsRedycByAdj| = 2030$ elements.

3) From the set *SimsRedycByAdj*, eliminate all elements whose nouns are proper names, or have been replaced by acronyms (3rd example)
$SimsRedycByNoun = \{sims \in SimsRedycByAdj$
$|NOUN' is\, a\, common\, N'\}$
Like in the following examples:

Pljevlja bi bila bogata i <**bleštava kao Las**> *Vegas da bude slavna i* <**bogata kao Monika**> *Seleš. Kako zatvoru u Beogradu,* <**opštepoznatom kao CZ**>*, naći u* ..... <**glistening as Las**> *Vegas* ..... <**rich as Monika**>*, Seleš.* .... <**generally known as CZ**>, ...

In our case, the result was
$|SimsRedycByNoun| = 1059$.

4) From the set *SimsRedycByNoun* generate a subset of the most frequent elements
$SimsMostFreq = \{sims \in SimsRedycByNoun$
$|freq(sims) \geq k\}$
where *k* is the minimal frequency of occurrence `as ADJ as a NOUN` in the observed corpus K_l. In our case, for the value $k = 1$, the total number of ADJ-NOUN pairs, candidates for wordnet expansion is $|SimsMostFreq| = 1059$.

5) From the set *SimsMostFreq* create a text file Adjective_As_Noun with ADJ-NOUN pairs over which an algorithm for wordnet expansion is executed (see Algorithm).

The presented algorithm is used for sequential processing of input candidate ADJ-NOUN pairs. For each pair, it checks whether in a given wordnet there are synsets of adjectives and nouns which are lexicalized by literals of the observed *adjective* and *noun*. After that, the procedure of automatic creation of the relation *specificOf/specifiedBy* is implemented between synsets of an adjective and a noun using a restriction — both of them have to be lexicalized by only one literal whose sense is the first sense. The first sense of a literal is considered to be the sense of a word in a certain language which is defined by a relevant dictionary or a corpus as the most commonly used one. Intuition on which this restriction is based is related to minimal pairing errors in the case when there are no synonyms in the observed synsets and the sense of the literals is the first sense. In that case, the possibility of error exists only if: at least one of the synsets is not correctly complemented with synonyms and there are no correctly assigned senses, or the desired sense is not the first one and it does not exist. In this regard, since the source of errors is known in advance, it is possible to check it before applying the algorithm. On the other hand, if at least one of the synsets has more than one synonym, or has one but its sense is not the first one, the new relation is not created and *adjective-noun* pair is separated into two independent files: the file containing adjectives and all their senses from a wordnet (named *adjective_senses*) and the file containing nouns and all their senses (named *noun_senses*). These resources are later used in a web application for manual pairing of adjectives and nouns and their connection through the desired relation. Finally, pairs for which it is determined at the very beginning of the process that they do not exist in the form of literals in a given wordnet, become candidates for later regular wordnet expansion – by adding new synsets.

Algorithm

Input: Adjective_As_Noun text file
Output: 1. a pair of WordNet mutually inverse semantic relations (specificOf/specifiedBy) for each input adjective-noun pair
2. file containing adjectives and all their senses
3. file containing nouns and all their senses

```
foreach adjective-noun pair in adjective-noun pairs
if ((adjective exists in Wordnet.adjective.literals)
   and (noun exists in Wordnet.noun.literals)) {
     if ((Wordnet.senses(adjective).Count==1)
        and (Wordnet.senses(noun).Count==1)
        and (Wordnet.sense(adjective).FirstSense)
        and (Wordnet.sense(noun).FirstSense) ) {
          Create_Relation(specificOf,adjective,noun);
          Create_Relation(specifiedBy,noun,adjective);
     }
     else
       foreach (sense in Wordnet.senses(adjective)) {
         add_to_adjective_senses(adjective,sense,synsetId)}
       foreach (sense in Wordnet.senses(noun)) {
         add_to_noun_senses(noun,sense,synsetId)}
     }
}
```

Prior to the implementation of the given algorithm, we examined the SWN in order to determine its structure in terms of the previously described restrictions. SWN has more than 22,000 synsets, contains 1660 synsets of adjectives with one literal, out of which in 1452 synsets the sense

of that literal is the first sense, while the number of noun synsets with one literal, where the sense of that literal is the first sense is 15,035. By implementing the suggested algorithm, out of a total of 1059 ADJ-NOUN pairs, 69 pairs were found which are "pairs whose both members have one sense and that sense is the first sense". In SWN there are 302 ADJ-NOUN pairs in which there is more than one sense or that sense is not the first sense. The 688 pairs that are left pertain to those cases when at least one member of the ADJ-NOUN pair does not exist as a literal in SWN. Therefore, using the proposed method produces 372 candidates that can be connected in SWN by the relation *specificOf/specifiedBy* after approval.

For 302 ADJ-NOUN pairs present in SWN, but with many senses or with one sense that is not the first sense, a web page is created in the SWNE[2] application (Mladenović et al., 2014) which allows users to input adjectives, thus generating a column with synsets lexicalized by the given adjective, while inputting nouns leads to generating of the second column, with synsets lexicalized by the noun at hand. New relations can be generated by looking for appropriate synsets and senses in *adjective_senses* and *noun_senses* files as well as by chosing the desired relation from the third column.

5 Evaluation

In order to assess whether the frequency of occurrence is a valid parameter for finding ADJ-NOUN pairs which are parts of similes that are used in everyday life, we used an online survey which was carried out through Google Forms. Comparing the list (marked here as *List1*) which was automatically generated using the Corpus and filtered using steps 1-4 explained in Section 4, and ordered in a decreasing order according to pair frequency, with the list which, in fact, represents a subset of the *List1* of those pairs that were marked positively in the anonymized survey (marked as *List2*), we wanted to assess which frequency threshold value entails the results obtained in the survey.

The survey itself was conducted over the time period of 5 days, such that a total of 4 forms were published successively. Anonymous users of the social network Facebook were supposed to give an answer to each question generated on the basis of ADJ-NOUN pairs from the *List1* list with a goal of finding out whether "in everyday language we can say that someone/something is ADJ as NOUN?". The answers were `Yes` or `No` and answering all questions in a form was mandatory. The Table 1 gives an overview of the distribution of questions in each form as well as the number of participants who were involved in answering the questions.

Google form	Number of questions per form	Participants per form
1	30	46
2	42	138
3	41	150
4	41	100
Total	154	434

Table 1: Distribution of questions and participants per form.

A Phd student at the Faculty of Philology, as a linguistic expert, manually selected 154 items from *List1* for which it could be presumed with some degree of certainty that they may be used in everyday language; namely, we retrieved a lot of noisy data from the Corpus, and some items stopped carrying meaning when taken out of the context. Linguistic constructs, chosen from the given *List1*, included *čist kao apoteka* "clean as a farmacy"; *čist kao suza* "pure as a teardrop"; *hladan kao led* "cold as ice"; *lak kao pero*, "as light as a feather"; *veran kao pas* "as faithful as a dog" whereas constructs such as: *dobar kao oblik* "good as shape"; *dobar kao pisac* "good as a writer"; *poznat kao vodja* "famous as a leader" were not used as they represented occasional occurrences. As we could not predict how willing to help out the potential participants would be, we were aiming for at least 30 participants. Also, the first form had less constructs than the rest — 30 — as we wanted to test the method and to see what would be an optimal number of fields in the form. We obviously wanted to test as many constructs as possible, but had also to keep the forms interesting and easy to fill in. The rest of the forms were balanced unit-wise. The number of participants was not pre-chosen, it depended on the turnout on the particular day.

The problem with this kind of participant involvement and with posts on Facebook in general is that the novelty wears off fast and if some post is very popular today, it might not be popular at all tomorrow. The call for participation in this project did receive a lot of attention in the first few hours

[2] `http://resursi.mmiljana.com/`

after being posted on Facebook. The privacy for the post was set to Public, which meant that everyone could participate and share the link leading to the Google Forms. Due to the fact that people did share the link, and some of their friends did the same thing, we could see that the forms were being filled in quickly and that our research was getting a lot of attention. In the following three days, we posted another three forms on the same URL address (precisely because the post received a lot of attention and shares) and we were able to get enough responses in order to get valid results. On the fourth day, the novelty wore off and we were getting significantly fewer responses, which only proved our assumption that we had to move fast and to post new forms every day.

First, we measured the contribution of participants and determined the set of those participants whose results were to be taken into account as relevant, on the basis that there was no substantial difference between arithmetic means of their answers. In order to measure the participants' contribution we generated 7 subsets of questions and answers where each set had less than 30 questions (units) using four spreadsheets containing participants' answers, as it is shown in Table 2 (each Google Form, except the first one, was divided into two parts). All 7 units were converted into matrices where each row represented answers of each participant and each column represented one question in the form <adjective>as<noun>. Content of each cell of the matrix had the value 1 if the participant marked a certain expression with "Yes" and the value 0 if the participant marked that expression with "No". Rows of the matrix were compared against each other with a paired t-test in order to determine that there was no substantial difference between arithmetic means of participants' answers. From each set we selected, among all participants belonging to that set, five participants whose difference in the paired t-test was the slightest.

After that, inter-annotator (participant) agreement was evaluated using the Krippendorff α coefficient (Kalpha). When the value of α is in the [0, 1] interval, it represents the agreement level which ranges from complete disagreement, when $\alpha = 0$, to complete agreement, when $\alpha = 1$. The α measure can also have a negative value, up to -1, when two mistakes are present: mistake in sampling and mistake in systemic disagreement. Considering an acceptable level of reliability, the works of (Hayes and Krippendorff, 2007), (Lombard et al., 2002) and (Maggetti, 2013) show that agreements whose values are $\alpha \geq 0.667$ are reliable, and that agreements whose values are $\alpha \geq 0.8$ can be considered very reliable. The results we obtained using the Kalpha test over the set of 5 annotators for each of the subsets of the forms is given in Table 2. Provided that for the first two forms and a part of the third one, the value of Kalpha was such that the annotator agreement could be considered reliable, for all of the constructs in those forms, if a majority of annotators (3 or more than 3 out of 5) annotated a certain question with "Yes", that item was taken as an element of the *List2'*. Thus, we obtained 53 items in total and their distribution over form sets is given in the last column of Table 2. Furthermore, we want to draw attention to the phenomenon which we did not study in depth, which was described here in Table 2 and has to do with the decline of the Kalpha coefficient over the same questionnaire structure, related to the time period when the participants filled in the Google Forms.

Form set	**No of participants**	**No of questions**	**Kalpha value**	**No of quest. annot. with Yes**
1	5	30	$\alpha = 0.757*$	16
2a	5	21	$\alpha = 0.713*$	17
2b	5	21	$\alpha = 0.698*$	15
3a	5	21	$\alpha = 0.688*$	5
3b	5	20	$\alpha = 0.484$	
4a	5	21	$\alpha = 0.434$	
4b	5	19	$\alpha = 0.375$	
Total		154		53

Table 2: Inter-annotator agreement over Google Forms and number of items which belong to reliable forms and were annotated with "Yes".

Finally, we wanted to assess how much the change of the frequency threshold influenced the relevance of automatically selected ADJ-NOUN pairs, measured based on the results obtained through the surveys. The list *List1* has been reduced so that it contains forms 1, 2a, 2b and 3a which amounted to 93 elements, that is to say, all ADJ-NOUN pairs for which evaluation by the participants was proved relevant. That list was named *List1'*. In contrast, the list named *List2'* contained only those ADJ-NOUN pairs from the *List1'* that were marked positively. First, we wanted to set the frequency threshold

to $k = 4$, which meant that the algorithm was used to process only those pairs whose frequency of occurrence in the Corpus was $k \geq 4$. There were 23 such pairs in the list *List1'*. Out of those 23, 19 pairs were present in the list *List2'*, which meant that the participants in the survey did not recognize 4 pairs that were recognized by the algorithm. The entire statistics showing the percentage of pairs we obtained using the algorithm as well as human judgement is given in Table 3, and the graph showing the relation between human selection, as opposed to automatic selection, when the frequency threshold is being changed, is given in Figure 1.

Frequency threshold	by algorithm	by humans	humans / algorithm
$k = 1$	93	53	57%
$k = 2$	44	32	73%
$k = 3$	32	27	84%
$k = 4$	23	19	83%

Table 3: Relationship of manually and automatically selected pairs depending on the frequency threshold.

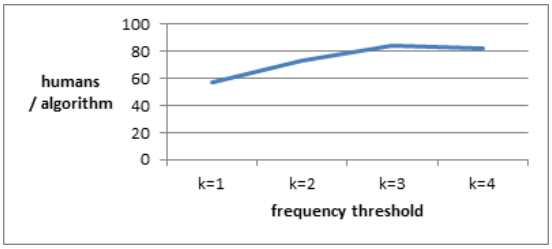


Figure 1: Relationship of selected pairs obtained with the survey method compared to the ones obtained with the method of the most frequent occurrence for different frequency thresholds.

Figure 1 shows the way in which, on a sample of 93 ADJ-NOUN pairs contained in the List2' list (Kalpha reliable), the percentage of participation of the manually selected pairs changes in the subset obtained by choosing only those pairs from the same list whose frequency was equal or higher than the set threshold, when the threshold changes. The achieved result of 84% gives us the manually measured accuracy of the Algorithm for automatic WordNet expansion with the frequency threshold of *k*=3.

6 Conclusions

In this work, we presented a general way of automatic expansion of a WordNet with the semantic relation *specificOf/specifiedBy* which was produced after extraction of semantic knowledge contained in the relation of comparison from the annotated corpus. The results of the proposed method of selection of the most frequent ADJ-NOUN pairs extracted from the described linguistic constructs `as ADJ as a NOUN` for the frequency threshold $k \geq 3$ were matched in 84% of cases with the results obtained from anonymous evaluators, on identical sets of ADJ-NOUN pairs. The Algorithm for automatic WordNet expansion can be improved in step 5) by including the Word sense disambiguation (WSD) method. That would enable literals with more than one sense to be used in automatic adding of the new relation. In future work we plan to implement WSD and to use other linguistic constructs which indicate Simile.

Using the relation *specificOf/specifiedBy* between a noun and its specific adjective, the hidden meaning of another word or a phrase can be detected, e.g. in sentences such as "My sister is like a bee" or "My sister is a bee", based on the relation *specificOf/specifiedBy* between the noun *bee* and its specific adjective *busy*, a sentiment neutral noun *sister* can have the same sentiment polarity as the adjective *busy*, i.e. positive polarity. If we say "My sister is like a lizard", based on the same principle, the same noun changes its sentiment polarity into negative polarity, considering the fact that the noun *lizard* is connected with a relation *specifiedBy* with the adjective *lazy*. In the example "My sister is as fast as a turtle" the indirect connection of the antonymous pair *fast-slow* in the construct "as fast as a turtle" indicates the existence of the rhetorical figure irony, therefore, in a given context, the noun *sister* can have a negative sentiment polarity. In our future work, we plan on analysing whether the process of sentiment classification can be improved by changing the default sentiment polarity of *n*-gram predictors, depending on the figurative context detected in the previously described way.

Acknowledgments

This research was partly supported by the Serbian Ministry of Education and Science under the grant 47003.

References

Manuela Angioni, Roberto Demontis, Massimo Deriu, and Franco Tuveri. 2008. Semanticnet: a WordNet-based Tool for the Navigation of Semantic Information. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, 21–34.

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, 2200–2204.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, 417–422.

Yanfen Hao and Tony Veale. 2010. An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes. *Journal Minds and Machines*, 20(4):635–650.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* , 1(1):77–89.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125.

Svetla Koeva, Cvetana Krstev, and Duško Vitas. 2008. Morpho-semantic Relations in WordNet. A Case Study for two Slavic Languages. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, 239–253.

Cvetana Krstev. 2008. Processing of Serbian - Automata, Texts and Electronic dictionaries. *Faculty of Philology, University of Belgrade, Belgrade*.

Judit Kuti, Károly Varasdi, Ágnes Gyarmati, and Péter Vajda. 2008. Language Independent and Language Dependent Innovations in the Hungarian WordNet. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, 254–269.

Matthew Lombard, Jennifer Snyder-Duch and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604.

Martino Maggetti. 2013. Regulation in Practice: The de facto Independence of Regulatory. *Swiss Political Science Review*, 19(1):111–113.

Palmira Marrafa, Raquel Amaro, Rui Pedro Chaves, Susana Lourosa, Catarina Martins, and Sara Mendes. 2006. WordNet.PT new directions. *Proceedings of the 3th International Global Wordnet Conference (GWC2006)*, 319–321.

Marek Maziarz, Stanisław Szpakowicz, and Maciej Piasecki. 2012. Semantic Relations among Adjectives in Polish WordNet 2.0: A New Relation Set, Discussion and Evaluation. *Cognitive Studies / Études Cognitives*, 12:149–179.

Sara Mendes. 2006. Adjectives in WordNet. *Proceedings of the 3th International Global Wordnet Conference (GWC2006)*, 225–230.

Miljana Mladenović and Jelena Mitrović. 2013. Ontology of rhetorical figures for Serbian. *LNAI, Springer*, 8082:386–393.

Miljana Mladenović, Jelena Mitrović and Cvetana Krstev. 2014. Developing and Maintaining a WordNet: Procedures and Tools. *Proceedings of the 7th International Global Wordnet Conference (GWC2014)*, 55–62.

Adam Pease, John Li, and Karen Nomorosa 2012. WordNet and SUMO for Sentiment Analysis. *Proceedings of the 6th International Global Wordnet Conference (GWC2012).*

Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Maria Real Coelho, and Maira Gatti. 2014. OpenWordNet-PT: A Project Report. *Proceedings of the 7th Global WordNet Conference (GWC2014)*, 383–390.

Vassiliki Rentoumi, Stefanos Petrakis, Manfred Klenner, George A. Vouros, and Vangelis Karkaletsis. 2010. United we stand - improving sentiment analysis by joining machine learning and rule based methods. *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010).*

Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4):754–760.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: An Affective Extension of Wordnet. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1083–1086.

Miloš Utvić. 2014. Liste učestanosti Korpusa savremenog srpskog jezika [Corpus of Contemporary Serbian – Frequency Lists]. *Naučni sastanak slavista u Vukove dane*, 241–262. Faculty of Philology, University of Belgrade, Belgrade.

Tony Veale and Yanfen Hao. 2008. Enriching WordNet with folk knowledge and stereotypes. *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, 453–461.

Identifying and Exploiting Definitions in Wordnet Bahasa

David Moeljadi, Francis Bond
Division of Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
D001@ntu.edu.sg, bond@ieee.org

Abstract

This paper describes our attempts to add Indonesian definitions to synsets in the Wordnet Bahasa (Nurril Hirfana Mohamed Noor et al., 2011; Bond et al., 2014), to extract semantic relations between lemmas and definitions for nouns and verbs, such as synonym, hyponym, hypernym and instance hypernym, and to generally improve Wordnet. The original, somewhat noisy, definitions for Indonesian came from the Asian Wordnet project (Riza et al., 2010). The basic method of extracting the relations is based on Bond et al. (2004). Before the relations can be extracted, the definitions were cleaned up and tokenized. We found that the definitions cannot be completely cleaned up because of many misspellings and bad translations. However, we could identify four semantic relations in 57.10% of noun and verb definitions. For the remaining 42.90%, we propose to add 149 new Indonesian lemmas and make some improvements to Wordnet Bahasa and Wordnet in general.

1 Introduction

A lexical database with comprehensive data about words, definitions, and examples is very useful in language research. In Princeton Wordnet, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets) which are interlinked through a number of semantic relations (Fellbaum, 1998; Fellbaum, 2005). Since its creation, many other wordnets in different languages have been built based on Princeton Wordnet (PWN) (Bond and Paik, 2012; Bond and Foster, 2013). One of them, Wordnet Bahasa, is built as a lexical database of the Malay language. At present, it consists of two language variants: Indonesian and Standard Malay. It combines data from several lexical resources: the French-English-Malay dictionary (FEM), the KAmus Melayu-Inggeris (KAMI), and wordnets for English, French and Chinese (Nurril Hirfana Mohamed Noor et al., 2011, p. 258).

We added Indonesian definitions from the Asian Wordnet project (Riza et al., 2010) to Wordnet Bahasa. To the best of our knowledge, the Asian Wordnet project is the only project that translated the English definitions of some synsets in PWN into Indonesian. However, the definitions were crowd sourced and had little quality control so not all of 14,190 definitions could be directly transferred. Many of the definitions had problems and needed to be cleaned up. The definitions for nouns and verbs which had been cleaned up were exploited to extract relations, such as synonym, hyponym, hypernym and instance hypernym, between lemmas and definitions. The method of extracting these relations was done in Bond et al. (2004) to build an ontology. We used Python (3.4, Python Software Foundation) and the Natural Language Toolkit (NLTK) (Bird et al., 2009) to process the data.

This paper is organized as follows: Section 2 describes the process of cleaning up the definitions, Section 3 explains the process of extracting hypernyms and other relations from the definitions. Section 4 presents the results and discussion and Section 5 concludes.

2 Cleaning up the definitions

As mentioned in Section 1 above, the definitions we had available were not clean. Many infelicities were found, such as misspellings, definitions using abbreviations, typos, synsets having more than one similar definitions, definitions written in English, improper use of hyphens, and lemmas written as the first word in the definitions. Each error is illustrated in the following subsections.

2.1 Correcting and deleting definitions

Words in the definitions which are not spelled correctly according to standard Indonesian, such as *dimana* "where" and *lain lain* "others", as well as typos such as *enerji* "energy" and *bagain* "part", were semi-automatically corrected. Since the typos are many and scattered throughout the file, we may have missed some. Abbreviations, most of them are prepositions, such as *dgn* "with" and *utk* "for", were also normalized to their full forms (see Table 1).

Before correction	After correction	Meaning	Number of hits
(double space)	(single space)		416
dimana	*di mana*	"where"	313
dengans	*dengan*	"with"	121
dgn	*dengan*	"with"	93
utk	*untuk*	"for"	52
kpd	*kepada*	"to"	25
pd	*pada*	"at"	23
lain lain	*lain-lain*	"others"	21
enerji	*energi*	"energy"	12
bagain	*bagian*	"part"	12
spt	*seperti*	"like"	12
dr	*dari*	"from"	10
thdp	*terhadap*	"toward"	10
sst	*sesuatu*	"something"	3

Table 1: Some examples of misspellings, abbreviations and typos, before and after the correction

Definitions which are obviously written in English or just names, were deleted (see Table 2).

Synset	Definition
03491491-n	Hanging Gardens of Babylon
09164241-n	ho chi minh city
10875910-n	George Herbert Walker Bush
11252392-n	rain in the face
13615557-n	a unit of measure for capacity officially adopted in the British Imperial System

Table 2: Some examples of deleted definitions

Some definitions had hyphens separating the words. In this case, the hyphens were deleted (see Table 3).

Synset	Definition	
14118423-n 'severe diabetes mellitus with an early onset'	*diabetes-mellitus-tergantung-insulin* "diabetes mellitus depending on insulin"	Before correction
	diabetes mellitus tergantung insulin	After correction

Table 3: An example of a definition having hyphens, before and after the correction

For definitions in which the first word is the same as the lemma with the real definition placed between brackets afterwards, the first word and the brackets were deleted (see Table 4).

Synset	Definition	
09543673-n 'an evil spirit or ghost'	*Ghoul (roh jahat atau hantu)* "Ghoul (an evil spirit or ghost)"	Before correction
	roh jahat atau hantu "an evil spirit or ghost"	After correction

Table 4: An example of a definition with lemma as the first word, before and after the correction

2.2 Choosing definitions

Some synsets have two or more different definitions as shown in Table 5. The longest one which includes other definitions, is assumed to be the correct one and automatically selected as the best definition.

Synset	Definition	
07904637-n 'gin flavored with sloes (fruit of the blackthorn)'	*buah dari semak* "fruit of the blackthorn"	Before cleaning up
	gin yang diberi rasa sloea "gin flavored with sloes"	
	gin yang diberi rasa sloea (buah dari semak) "gin flavored with sloes (fruit of the blackthorn)"	
	gin yang diberi rasa sloea (buah dari semak)	After cleaning

Table 5: An example of a synset with many parts of definition, before and after the cleaning up

However, if the definitions are all completely different and one of them was considered good based on the English and Japanese definitions, that one was chosen to be the correct one (see Table 6). This manual checking was done by the first author who has a good command of Indonesian, English, and Japanese.

If we found no satisfying definition after checking and comparing with the English and Japanese definitions, one or two of the words in the definitions were manually corrected (see Table 7).

After the cleaning up process, we made the Indonesian definitions available in the Open Multilingual Wordnet (1.2) hosted by Nanyang Technological University in Singapore (`http://compling.hss.ntu.edu.sg/omw/`). Figure 1 shows a screenshot of synset 06254371-n 'heliogram' with its Indonesian definition.

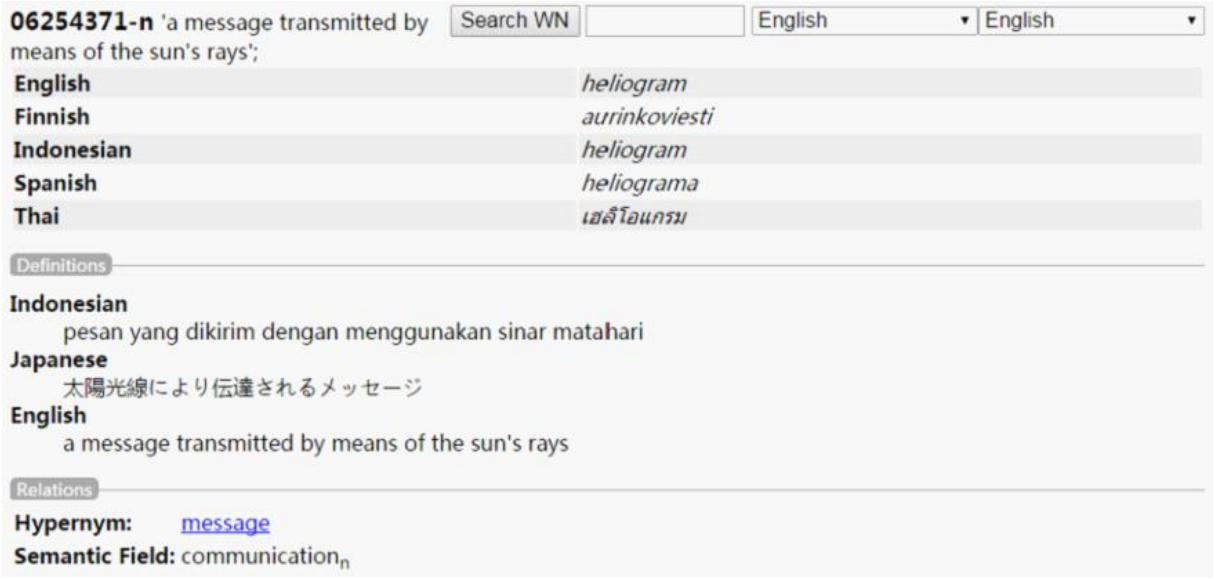


Figure 1: A screenshot of synset 06254371-n 'heliogram'

Synset	Definition	
01711910-a 'causing a sharply painful or stinging sensation'	*kedinginannya menggigit ke tulang* "the coldness bites to bones"	Before correction
	kedinginannya menusuk ke tulang "the coldness stings to bones"	
	sejuk hingga menggigit ke tulang "cool biting to bones"	
	sejuk hingga menusuk ke tulang "cool stinging to bones"	
	sejuk hingga menusuk ke tulang "cool stinging to bones"	After correction

Table 6: An example of a synset having many definitions, before and after the correction

Synset	Definition	
00731471-a 'supported by both sides'	*didukung oleh dua negara* "supported by both countries"	Before correction
	didukung oleh dua partai "supported by both parties"	
	didukung oleh dua pihak "supported by both sides"	After correction

Table 7: An example of a synset having two definitions, before and after the correction

3 Extracting relations from the definitions

Unlike Bond et al. (2004) who parsed the definition sentences using a grammar before extracting hypernyms and other relations, we simply used regular expressions. Indonesian has a strong tendency to be head-initial (Sneddon et al., 2010, pp. 160-162). In a noun phrase with an adjective, a demonstrative or a relative clause, the head noun precedes the adjective, the demonstrative or the relative clause. Typically numerals and classifiers precede the head noun (Alwi et al., 2014, pp.251-255).

Example (1) shows the Indonesian definition of synset 09500625-n 'Pegasus', the head of which is preceded by a numeral prefix *se-* "one" and a classifier *ekor* (lit. "tail") and followed by an attributive verb *bersayap* (lit. "having wings") and a prepositional phrase.

(1) *seekor **kuda** bersayap dalam mitologi Yunani*
one-CL horse winged in mythology Greece

"a winged horse in Greek mythology"

Example (2) contains a part of the Indonesian definition of synset 05316175-n 'ocular muscle'. Its head *otot-otot* "muscles" is in the plural (reduplicated) form, preceded by *satu dari* "one of" and followed by an adjective *kecil* "small".

(2) *satu dari **otot-otot** kecil pada mata...*
one of muscle-RED small at eye

"one of the small muscles of the eye"

We assume that after modifying the definitions, relations between lemmas and definitions can be extracted from the first lexical word (i.e. the head) in the definitions.

3.1 Modifying the definitions

For each definition for nouns and verbs, we removed the following words at the beginning:

(i) words which are written between brackets, such as *(Ilmu komputer)* "(Computer science)" relating to domain

(ii) numerals, such as *satu* "one", *tiga* "three", and *5* "five"

(iii) determiners, such as *setiap* "every", *sejenis* "a kind of", *semacam* "a sort of", *sembarang* "any kind of", *salah satu* "one of", *suatu* "a (for thing)", *sebuah* "a (for thing)", *seorang* "a (for person)", *seekor* "a (for animal)", *selembar* "a piece of", *sekelompok* "a group of", *beberapa* "some", *berbagai* "various", and *segala* "all"

(iv) relativizer *yang* "which"

(v) prepositions, such as *untuk* "for", *dari* "of", and *dalam* "in"

(vi) other stop words, such as *seperti* "like", *tentang* "about", *termasuk* "including", and *biasanya* "usually"

We also changed the plural (reduplicated) form of the head to its singular (non-reduplicated) form, for example *otot-otot* "muscles" was changed to *otot* "muscle" and *daun-daunan* "foliage, a cluster of leaves" was changed to *daun* "leaf". Punctuations such as slashes (/), semicolons (;), and commas (,) dividing two words were replaced as a space. After we made these changes, the first word in the definition was taken as a potential genus term.

3.2 Extracting relations

The first step was to check whether each first word of the definitions is in Wordnet or not. If it is not in Wordnet, we checked whether it is in *Kamus Besar Bahasa Indonesia* (KBBI) "The Great Dictionary of the Indonesian Language of the Language Center" or not. KBBI is published by the language institute who provides support for the standardization and propagation of Indonesian. Its third edition has been made online to public and has an official site (`http://badanbahasa.kemdikbud.go.id/kbbi/`) (Alwi et al., 2008).

The next step was to check whether the lemma synset is the same as the synset of the first word in the definition. This allows us to identify when the same word is used to define the lemma. Besides synonyms, hyponyms can also be employed to define the lemma. In order to confirm this, the lemma synset was compared to the hyponyms of the first word in the definition.

The next important step was to check whether the hypernym is used to define the lemma by comparing the hypernyms and instance hypernyms of the lemma synset with the synsets of the first word in the definition. If a lemma does not have any hypernym in Wordnet, we checked whether it has instance hypernym. Finally, lemmas having neither hypernyms nor instance hypernyms were checked by hand.

4 Results and discussion

The definition file which originally has 14,190 lines of definitions was cleaned up and 1,522 definitions (10.7%) were deleted. The remaining 12,668 definitions consist of 10,549 definitions for nouns, 1,663 definitions for adjectives, 409 definitions for verbs, and 47 definitions for adverbs. Although these definitions are considered quite clean, they may still contain small errors as mentioned in Section 2.1. Since adjectives and adverbs do not have relations such as hypernym in Wordnet, we only examined nouns and verbs. Out of 10,958 definitions for nouns and verbs, we could extract four relations from 6,257 definitions (57.10%) as shown in Table 8. The remaining 4,701 definitions (42.90%) have problems, such as words which could not be found in Wordnet and lemmas without explicit relations as shown in Table 9.

Most of the relations we extracted (95.89%) are hypernym and instance hypernym. The remaining are synonym and hyponym as shown in Table 8 for synset 00004475-n and 00029677-n. Synset 00004475-n has six Indonesian lemmas. One of these lemmas, i.e. *makhluk* "being", is used as the head of its definition and thus we regard the lemma is synonymous with the definition. Synset 00029677-n has *proses* "process" as one of its lemmas, which is the hypernym of the head of the definition *fenomena* "phenomenon".

Out of the 4,701 definitions for which we could not find the relations, most of them (83.88%) have hypernyms which are different from the first word in the definitions. We found five patterns for this problem (see Table 9):

1. The genus term is correct but Wordnet Ba-

Relation	Number of synsets	Example	
		Synset	Definition
Hypernym	5,451	00021939-n artifact	*suatu* ***objek*** *buatan manusia* "a man-made object"
Instance hypernym	549	02956500-n Capitol	***gedung*** *DPR di AS* "the government building in the United States"
Synonym	252	00004475-n organism	***makhluk*** *hidup yang dapat mengembangkan kemampuan bertindak independen* "a living thing that can develop the ability to act independently"
Hyponym	5	00029677-n process	*sebuah* ***fenomena*** *yang berkelanjutan* "a sustained phenomenon"
Total	6,257		

Table 8: Relations extracted from lemmas and definitions

Problem	Number of synsets	Example	
		Synset	Definition
No match	3,943	14350206-n myelitis	***inflamasi*** *pada syaraf tulang belakang* "inflammation of the spinal cord"
		14573846-n viremia	***kehadiran*** *suatu virus di dalam aliran darah* "the presence of a virus in the blood stream"
		13251154-n clobber	***istilah*** *informal untuk harta pribadi* "informal terms for personal possessions"
		07603411-n choc	***singkatan*** *dalam bahasa Inggris untuk coklat* "colloquial British abbreviation for chocolates"
		14364217-n sword-cut	***bekas*** *luka dari sayatan pedang* "a scar from a cut made by a sword"
		00046344-n stunt	*tidak biasa atau berbahaya* "not usual or dangerous"
Word not in Wordnet			
- Word in KBBI	252	13436063-n automatic data processing	***pemrosesan*** *data secara otomatis* "automatic data processing"
		07865105-n chili dog	***hot*** *dog dengan daging sapi diberi cabai bubuk* "a hotdog with chili con carne on it"
- Word not in KBBI	495	14099050-n visual aphasia	***ketidakmampuan*** *memahami kata-kata tertulis* "inability to perceive written words"
		09603258-n Pluto	***karakter*** *kartun anjing ciptaan Walt Disney* "a cartoon character created by Walt Disney"
		14155506-n cystic fibrosis	*disebabkan kerusakan suatu gen* "caused by defect in a single gene"
		00662589-v insure	*membagikan kawasan untuk kawalan tentara* "allot regions for soldiers"
No explicit relations	11	01773734-v grudge	*terpaksa menerima atau mengakui* "accept or admit unwillingly"
Total	4,701		

Table 9: Problems found in extracting relations

hasa does not have the right synset for the lemma. For example, synset 14350206-n 'myelitis' has 14336539-n 'inflammation' as its hypernym, which is also the first word in the English and Indonesian definitions. Wordnet Bahasa does have *inflamasi* "inflammation" but only in a different synset.

2. The semantic relation is not written explicitly in the definition. For example, synset 14573846-n 'viremia' has *kehadiran* "presence" as the first word in the English and Indonesian definitions which has nothing related with the semantic relation.

3. The genus candidate is a relational noun. For example, synset 13251154-n has *istilah* "terms" and synset 07603411-n has *singkatan* "abbreviation" as the first word in the definition. To get the real genus term requires more parsing.

4. Compounds were not extracted. For example, although the head of the definition of synset 14364217-n, was *bekas luka* "scar" (lit. "former wound"), we extracted only the first word *bekas* "former, past"

5. The definition is incomplete. For example, the Indonesian definition for synset 00046344-n lacks the head noun *usaha* "feat"

The second problem we found is that the first word in 747 definitions (15.89%) is not in Wordnet. In this case, we checked whether the word is in the Indonesian dictionary (KBBI) or not as mentioned in the previous section. We found 252 definitions having 149 unique words (the heads) which are in KBBI but not in Wordnet. Some of them are compounds as in synset 07865105-n with the definition *hot dog dengan daging sapi diberi cabai bubuk* "a hotdog with chili con carne on it". We did not distinguish compounds and thus, failed to extract *hot dog* as the head. The word *hot* does exist in KBBI as an adjective meaning 'sexually excited or exciting'.

The remaining 495 definitions have 235 unique words which are not in KBBI. We examined four patterns for this:

1. Derived words with negation are not listed as lexical items in KBBI. For example, the word *ketidakmampuan* "inability" (lit. "not able-ness") has the stem *tidak mampu* "not able" with a circumfix *ke-...-an* to nominalize. Including in this group are *ketidakadaan* "absence" (lit. "not present-ness") and *ketidaksempurnaan* "imperfection" (lit. "not perfect-ness").

2. The online KBBI data is not perfect, it does not include all Indonesian words listed in the paper dictionary. For example, the word *karakter* "character" is listed in the paper dictionary but not in the online version.

3. The Indonesian definition is incomplete. For example, the Indonesian definition for synset 14155506-n lacks the head noun *penyakit* "disease".

4. The Indonesian definition is incorrect. For example, the Indonesian definition for synset 00662589-v.

We found 11 lemmas have no explicit semantic relations with the definitions. They are all verbs: 01773734-v 'grudge', 00616857-v 'neglect', 01336635-v 'overlay', 01767949-v 'strike', 01944252-v 'hover', 02086805-v 'stampede', 02119241-v 'ignore', 02150510-v 'watch', 02413480-v 'work', 02581477-v 'prosecute', and 02673965-v 'stand out'.

5 Summary and future work

We have presented the process of cleaning up the definitions and extracting relations from the definitions. While doing the relation extraction, we spotted errors such as incompleteness and incorrectness in the definitions which we could not detect only by cleaning up the definitions. The reason why there are errors is probably because of little quality control in the translation process. In addition, we found things to be improved in Wordnet Bahasa and Wordnet in general. Based on our findings, we propose to:

1. Edit the incomplete Indonesian definitions. For example, definitions for synset 00046344-n which lacks the head noun *usaha* "feat" and 14155506-n which lacks the head noun *penyakit* "disease", as mentioned in Section 4

2. Delete the incorrect Indonesian definitions. For example, definitions for synset 00662589-v 'insure' which has the Indonesian definition *membagikan kawasan untuk kawalan tentara* "allot regions for soldiers"

3. Add 149 new lemmas from KBBI and possibly derived words with negation to Wordnet Bahasa

4. Add existing lemmas in Wordnet Bahasa to the correct synsets. For example, *inflamasi* to be added to synset 14336539-n 'inflammation'

5. Edit definitions in Wordnet to make them more informative, possibly add the hypernyms. For example, instead of having definition *jenis dari genus Soleidae* "type genus of the Soleidae" for synset 02664136-n 'Solea', we propose *jenis ikan dari genus Soleidae* "type of fish from the Soleidae genus"

6. Standardize the definitions in Wordnet, possibly make some guidelines for definitions. For example, regarding the numerals, some of them are written alphabetically, as in synset 09506337-n 'Fury' *tiga monster berambut ular...* "three snake-haired monsters...", but some of them are written in numbers, as in synset 09549416-n 'Hyades' *7 putri Atlas...* "7 daughters of Atlas...". Another problematic case is circular definitions.

For example, for synset 04658942-n ‘inhospitableness’ *memiliki sifat tidak ramah* “having an unfriendly and inhospitable disposition” and synset 04657876-n ‘unfriendliness’ “an unfriendly disposition”

Acknowledgments

Thanks to Hammam Riza who gave us permission to use the Indonesian definitions from Asian WordNet project. Thanks to Randy Sugianto and Ruli Manurung for their help. This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Hasan Alwi, Dendy Sugono, and Sri Sukesi Adiwimarta. 2008. *Kamus Besar Bahasa Indonesia Dalam Jaringan (KBBI Daring)*. 3 edition.

Hasan Alwi, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 2014. *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, 3 edition.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. 2004. Acquiring an ontology for a fundamental vocabulary. In *20th International Conference on Computational Linguistics (COLING-2004)*, pages 1319–1325, Geneva.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined wordnet bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge.

Christiane Fellbaum. 2005. WordNet and wordnets. In *Encyclopedia of language and linguistics*, pages 665–670. Elsevier, Oxford, 2 edition.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267, Singapore.

Hammam Riza, Budiono, and Chairil Hakim. 2010. Collaborative work on Indonesian WordNet through Asian WordNet (AWN). In *Proceedings of the 8th Workshop on Asian Language Resources*, pages 9–13, Beijing, China. Asian Federation for Natural Language Processing.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*. Allen & Unwin, New South Wales, 2 edition.

Semantics of body parts in African WordNet: a case of Northern Sotho

Mampaka Lydia Mojapelo

University of South Africa

Department of African Languages

mojapml@unisa.ac.za

Abstract

This paper presents a linguistic account of the lexical semantics of body parts in African WordNet, with special reference to Northern Sotho. It focuses on external human body parts synsets in Northern Sotho. The paper seeks to support the effectiveness of African WordNet as a resource for services such as in the healthcare and medical field in South Africa. It transpired from this exploration that there is either a one-to-one correspondence or some form of misalignment of lexicalisation with regard to the sample of examined synsets. The paper concludes by making suggestions on how African WordNet can deal with such semantic misalignments in order to improve its efficiency as a resource for the targeted purpose.

1 Introduction

African WordNet is a project that aims to build a lexical database for all indigenous official languages of South Africa, which will be linked to one another. It is modelled on Princeton WordNet[1] through the expand approach (Vossen, 1998). The approach was informed by experiences shared by earlier Wordnets such as BalkaNet, MultiWordNet, and other languages in the EuroWordNet, to name but a few. The expand approach takes synonym sets (synsets) from Princeton WordNet, with their relations, and convert them into the target language. The approach already lends the development of African Wordnets to the use of more than one language, that is, English and the target language concerned. African WordNet is further internally multilingual with five out of nine official African languages of South Africa that are currently part of the project. Northern Sotho (Sesotho sa Leboa)[2] is one of the languages involved. The premise in building African Wordnets is to model it on the Princeton structure while staying true to the African context.

Among the challenges that were encountered in the process of building African Wordnets was that some of the synsets extracted from Princeton for development of African WordNet did not make immediate sense for African languages and the African context, for a number of reasons. For example, among them are synsets for concepts that are geographically distant from the South African context, such as animal and plant species. This situation would result in non-lexicalised concepts. Some non-lexicalised concepts were left blank and for some it was decided that available linguistic resources would be used for coinage and borrowing. The envisaged convenience of African WordNet became clearer to the writer (a linguist, project translator or lexicographer) through other synsets of a more general nature that were easy to work with. One of the semantic domains that was considered generally applicable to any context was Anatomy, BodyPart. It was assumed that this kind of domain would have relatively fewer gaps compared to domains that are geographically or culturally more restricted. BodyPart also ranks ninth among the 50 most frequently suggested upper merged ontologies (SUMOs) in Princeton WordNet (PWN), as at 2014-03-11.

The downside of BodyPart was that the synsets extracted from Northern Sotho showed that none of the synsets done so far were aimed at the human anatomy. The SUMO_BodyPart consisted of words that were unrelated to humans, such as ‘scale’ (as in fish-scale), ‘shell’, ‘paw’, ‘feather’ and ‘wool’. Other examples to illustrate unrelatedness to humans is that the senses of the word *seatla* ‘hand’ were limited to Domain_Transport, SUMO_Device

[1] http://wordnet.princeton.edu

[2] Cf. Guthrie’s zone S30

and Domain_Factotum, SUMO_Constant Quantity, and denotation to parts of the human body did not feature. Similarly, the senses of *leoto* 'leg' were limited to Domain_Factotum, SUMO_Shape-Attribute and Domain_Zoology, SUMO_Mammal, which is a different synset from Domain_Anatomy, SUMO_BodyPart. This paper was premised on the understanding that, comparatively speaking, non-human body parts and other domains mentioned here may not demonstrate the immediate and direct societal impact of African WordNet to the extent that may be achieved with human body parts.

South Africa is a multilingual and multicultural country. According to the latest South African statistics (Statistics South Africa, 2012) on the use of home languages only 9,6% of the general population speak English as their home language (L1), while the majority speak the other ten official languages and their dialects as L1. The remainder (>90%) speak English either as a second, third or fourth language or not at all. Among this vast majority are healthcare workers, medical students and practitioners, as well as individuals and communities who should receive healthcare and medical services. Another issue is that studies incidental to most academic qualifications in South Africa are presented through the medium of English, which inevitably means that most students learn through a foreign medium. For some English schooling starts before they have duly mastered their L1. This apparent disadvantage is balanced by the foundation laid in English, which will give the student a significant headstart in his or her academic career, still with insufficient knowledge of his or her L1. L1 English speakers on the other hand are not motivated to learn other languages until they have completed their studies and happen to find themselves in an occupational environment where they have to adjust to a different language medium. It may therefore be useful to provide a multilingual platform for accessing domain lexicons on a level that is more than just a dictionary. Terminology lists and glossaries are being developed for various purposes in South Africa, including healthcare and medicine, but none of these is an African language Wordnet. African WordNet will not only provide definitions and contextual usages of words, but will be based on synsets. Synsets are sets of lexicalisations of a particular concept, and WordNet links them to other concepts through semantic relations such as hyponymy and meronymy, in the case of nouns. African WordNet will further link the languages spoken in the country to each other.

2 About the body parts lexicon in Northern Sotho

Since the available body-parts synsets in the Northern Sotho Wordnet were deemed not immediately useful for human healthcare and medicine purposes, the writer considered exploring external human body parts, which will later be followed by internal ones to complete the healthcare and medical intent. A list was drawn, verified and augmented against Northern Sotho Language Board (1988) as well as Ziervogel & Mokgokong (1975) and a paper in progress on verbs expressing physical pain. The list had Northern Sotho and English equivalents. Already when giving equivalents outside WordNet it emerged that there may be misalignment in the form of general-specific lexicalisation of senses. For example, Northern Sotho uses the same word for 'finger' and 'toe'. Unless the difference is readily apparent from the context a descriptive phrase is used for ease denotation. The question is: How big is the misalignment and how are we going to solve the problem linguistically? The sample used here is used as an index of misalignments, as well as possible solutions, for the rest of the development of the Northern Sotho Wordnet. The next step was to match the body parts on the list with English synsets.

3 Lexical entries in Northern Sotho Wordnet

In keeping with Princeton the lexical entries in African WordNet are guided by information such as part of speech (POS), domain, SUMO, definition, usage and the English ID. This paper focuses on the Northern Sotho nouns under the Domain_Anatomy, SUMO_BodyPart. According to the definition and usage provided in English as well as the ID, only body parts that are specifically human were picked out. Fellbaum (1998) contends that although the majority of lexicalised concepts are shared among languages, not every language will have words denoting concepts that are lexicalised in other languages. Therefore it is expected that some concepts may

be lexicalised in English and not in Northern Sotho, and *vice versa*. It is deemed necessary for this semantic domain to have as many lexicalised concepts as possible, given the envisaged use in the healthcare sector. The paper will also look into these semantic relations and ensure that the Northern Sotho synsets are presented in a manner that is not misconstrued.

Lexicalisation is defined as realisation of meaning in a single word or morpheme where words are already present in a language, as well as the addition of new words as new concepts enter the languages in due course. The addition of new words involves strategies of word formation such as compounding, derivation and borrowing. Another issue to lexicalisation is some level of acceptability among the speakers of a language, which will lead to general acceptability. The body-parts synsets in Northern Sotho reflect different types of lexicalisation, including addition of new words by the strategies mentioned above. There are also cases of non-lexicalisation which have yet to be resolved.

Although the expand approach has proved to be most expedient for new wordnets, lexicalisation challenges are inevitable for most of them. For example, in building the Konkani WordNet from Hindi WordNet (Walawalikar et. al 2010), which is a closely related language, some challenges were experienced. The challenges also involved the English source and they include linking errors, missing entries, definitions, concept misalignment and lexicalisation. The issue of culture-specificity is also reported as one of the causes of misalignment. In dealing with alignment in the Hebrew WordNet, which was also built on the expand approach; Ordan and Winter (2007) distinguish between contingent and systematic instances of non-equivalence. The two cases attest to the fact that lexicons of different languages mirror misalignments of both cultural and internal language structural nature.

Vincze and Almási (2014) also treat lexicalisation challenges encountered in dealing with the Hungarian WordNet. The intention of this paper is not to reinvent the wheel but to learn from others' experiences in the realisation that languages may be dissimilarly resourced, materially and structurally. Northern Sotho is a Bantu language of the Niger-Congo language family, which is agglutinating with productive morphology. Therefore one lexicalisation type or mechanism may prove to be more practical than another. For the purpose of this paper it is assumed that Northern Sotho may be differently resourced, given the object to explore how the project can try to solve extant misalignment challenges without losing the Princeton structure while remaining true to the African context, a manoeuvre requiring a certain amount of fineness.

4 Queries and results

To begin, the items on the list were queried from the English dictionary in DEBVisDic (WordNet editor and browser). Only sense 1 of SUMO_BodyPart under Domain_Anatomy was selected. The definitions, usages and synset IDs were used to obtain correct matches. General personal knowledge of Northern Sotho, as a mother tongue speaker, was complemented and verified against the Northern Sotho-English bilingual and Northern Sotho-English-Afrikaans trilingual dictionaries. The results gained from the queries confirmed some degree of misalignment between Northern Sotho and English. Clearly no comment is required on the one-to-one matches. The examples used here represent one-to-many and many-to-one mappings as well as lexicalisation gaps.

A sample of words representing 88 Northern Sotho concepts, with English equivalents, was used. The list is not exhaustive, but it is a fair representation of external human body parts. Also, not all possible connections have been indicated in the illustrations. While the initial focus was on external body parts, parts of the oral cavity were included as they are too close to the external facial body-parts and not as concealed as other internal body-parts. The English equivalents of the Northern Sotho words on the list were browsed and their IDs noted in order that their definitions and usages establish correct matches.

Queried senses in English (anatomy, human body part) were not found for the following words:

head

big hair

hair on arms and legs

protruding forehead

eye ridge

cheek

tongue

adam's apple

below the buttock (where the thigh starts)

back of hand

back

back of knee

foot

heel

When queried, the relevant senses of the words above could not be matched with the IDs found in DEBVisDic. A peculiar gap in English on human body parts relates to 'head', 'cheek', 'tongue', 'adam's apple', 'back', 'foot' and 'heel'. It is assumed that the rest of the words may be more physiologically or culturally relevant in Northern Sotho than in English. While it is still peculiar to some extent that 'back' was not found because physiologically, especially in the healthcare and medical context, the concept should have the same denotative significance in both languages, the gap was understood in the context of possible cultural dissimilarities. *Mokokotlo* 'back', as in the 'back part of the human torso', is one of the most recognisable lexical items in Northern Sotho due to what the concept represents. It is the part of the body that a baby or toddler is carried and strapped on for guaranteed safety and protection. In this context the back is culturally associated with care, nurturing, raising, acceptance and protection. The concept (and therefore the word) is culturally significant. With regard to *setšhitšhi* 'big hair' (not the same as 'long hair', which would be natural in the English lexicon) the gap in English is understood to be due to physiological difference.

Halliday et. al. (2004) explicate at length problems of cross-language mapping even for concepts that seem simple such as kinship terms. The examples of siblings and cousins between English and Australian Pitjantjatjara resonate with Northern Sotho and other Bantu languages. Therefore the issue of misalignment is not only a matter of lexical items, but of concepts as well.

The following diagrams provide reference for the current discussion. For every Northern Sotho lexical item, an English translation equivalent is provided. For combined connections, refer to appendix 1.

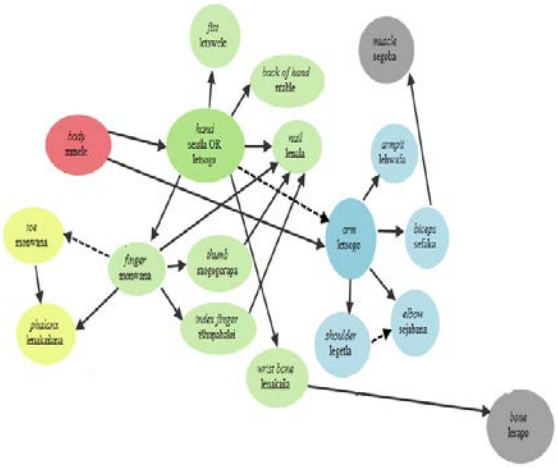


Diagram 1: Arm connections

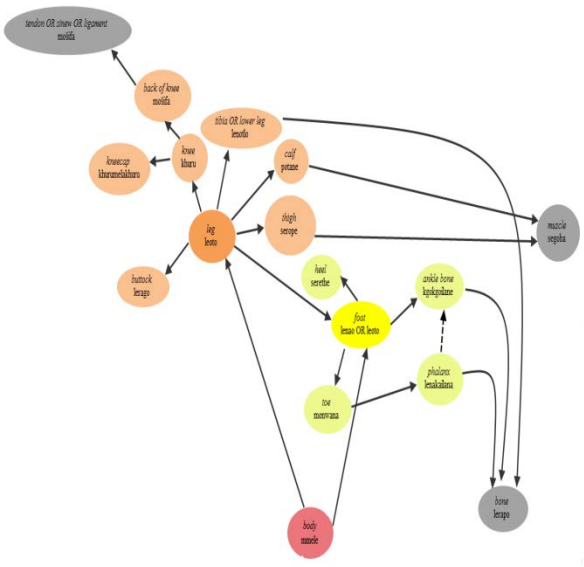


Diagram 2: Leg connections

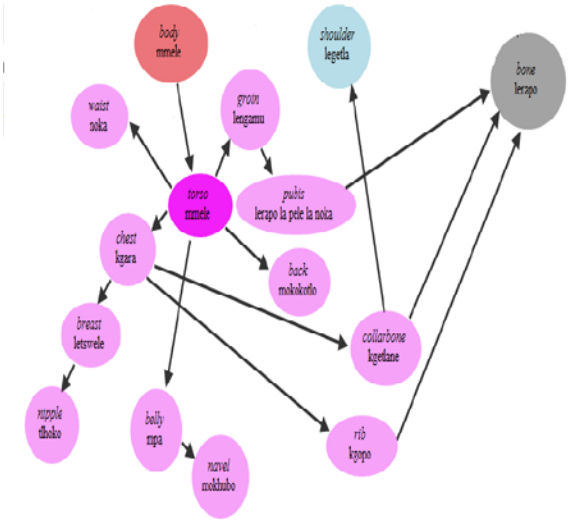


Diagram 3: Torso connections

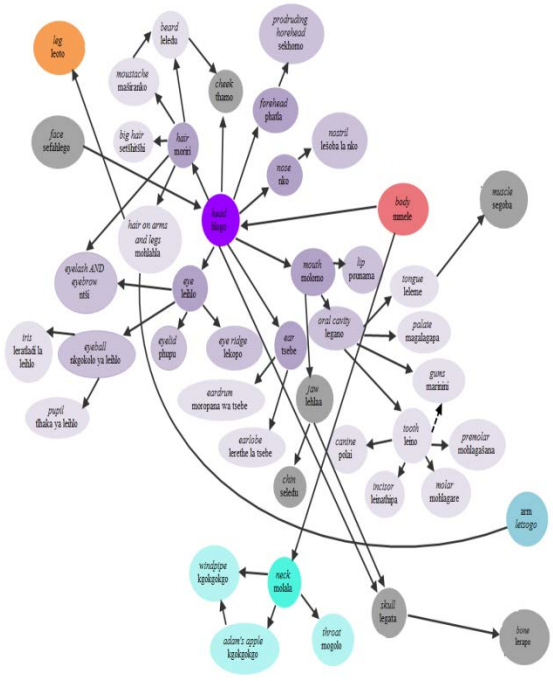


Diagram 4: Head connections

4.1 One-to-many and many-to-one

Two types of misalignment will be used for illustration here. There are cases of Northern Sotho lexicalisation of human body parts that mingle synonymy and meronymy, not in a confusing way though. In the context of WordNet words are synonymous if they express the same concept and can be interchanged in some contexts (Fellbaum 1998). Meronymy is explained by Croft and Cruse (2004) as a sense relation between meanings rather than between individual entities, that is when the meaning of one word is part of the meaning of another. The word for 'hand' in Northern Sotho is *seatla*. It expresses the same concept expressed by [POS: n ID: ENG 20-05246212-n BCS: 3], which is sense 1 of the Domain_ Anatomy, SUMO_Bodypart and defined in English as "the (prehensile) extremity of the superior limb". *Letsogo* is Northern Sotho for 'arm' [POS: n ID: ENG 20-05245410-n BCS: 3], Arm: 1, defined in English as "a human limb; technically part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb". In Northern Sotho *letsogo* refers to the whole superior limb, which includes the hand. According to the definition provided above the common usage of the English 'arm' is the same as the Northern Sotho *letsogo*, but the technical usage is not. In Northern Sotho the word *letsogo* is also used to refer to *seatla* 'hand', but the whole limb is never called *seatla*. That is, while *seatla* 'hand' is a meronym of *letsogo* 'arm', the two are also synonymous. Similarly *leoto* 'leg' [ENG20-05242579-n] is used for both 'leg' and 'foot' while a separate specific word for 'foot' is *lenao*. These examples illustrate lexicalisation that reflects the occurrence of meronymy between lexical items that are also synonymous.

Another scenario relates to the case of *monwana* for both 'toe' [ENG20-05258265-n] and 'finger' [ENG20-05247839-n], and *ntši* for 'eyebrow' [ENG20-05007503-n] and 'eyelash' [ENG20-05008887-n]. In this case Northern Sotho uses one word to express separate concepts, or concepts that are viewed as separate in English. These two examples illustrate that the words *monwana* and *ntši* are used in Northern Sotho as hypernyms. Descriptive phrases 'of the foot' and 'of the hand' are used as hyponyms of *monwana* in cases where distinction is deemed necessary. A similar descriptive strategy is not used for *ntši*; it may also be cumbersome as both 'eyebrow' and 'eyelash' belong to the eye.

4.2 Possible non-lexicalisation in English

Another concept that is lexicalised in Northern Sotho but could not be found from querying the English in DEBVisDic is *nyaraga* (Mokgokong and Ziervogel 1975), also pronounced *nyarago*. The English trees relating to 'leg' and 'buttock' were examined as the concept is understood to be either a body part below the buttock or the uppermost back part of the leg. Its absence in the two trees pointed to possible non-lexicalisation.

The following section proposes possible linguistic means of catering for the misalignment issues mentioned above in African WordNet.

5 Handling misalignments

It is necessary to provide linguistic solutions to the misalignment challenges mentioned above. Vincze and Almási (2014) suggest a number of strategies for the Hungarian lexicalisation issues, namely to shorten the tree, flatten the tree, restructure the tree and lexicalize the concepts. They are also of the opinion that the merge approach would have alleviated some of the challenges. For Konkani Walawalikar et. al (2010) suggest, among others, that the target language synsets for which there were gaps in the source language could be used to fill the gaps, thereby strengthening the HWN. Ordan and Winter (2007) detail strategies for building Hebrew synsets, which include linking Hebrew word senses to related PWN sysnsets from Hebrew to English and from English to Hebrew. Lexical gaps from both sides are acknowledged and used to preserve and link semantic information.

This paper takes a linguistic view to addressing the challenges mentioned above, which relate to lexicalisation of the concepts. The first group of Northern Sotho words which could not be matched from English seem to be a matter of misses which can be addressed if probed further. The next situation concerns *seatla* 'hand' and *lenao* 'foot' which are meronyms of *letsogo* 'arm' and *leoto* 'leg', respectively, and proved to be synonymous as well. Therefore lexical items *seatla* and *letsogo* will be in the same synset while they are meronymically related as well. The same applied to *lenao* 'foot' and *leoto* 'leg'.

The next issue concerns *monwana* 'finger' and 'toe' and *ntši* 'eyelash' and 'eyebrow'. In the language synonyms for *monwana* are provided in the form of descriptive phrases to distinguish 'finger' and 'toe'. The descriptions *wa lenao* and *wa leoto* 'of the foot'; *wa seatla* and *wa letsogo* 'of the hand' are consistent with language usage and are not expected to pose any problems. The same solution cannot work in the case of *ntši* since eyebrow and eyelash are both 'of the eye'. Northern Sotho Language Board (1988) uses compounding as a strategy to distinguish the two. While they are both *ntši* the source coined *ntšikgolo* as additional lexicalisation for 'eyebrow'. The second component of the compound *-kgolo* (*-golo*) 'big' suggests that an eyebrow is dominant. The source was produced by a standardising body (Northern Sotho Language Board) which was obviously cognisant of the gaps in terms of lexicalisation. They probably considered either the overaching position of the eyebrow in relation to the eyelashes or the perceived amount of hair in both, to come up with a suggestion that an eyebrow is the main *ntši*. Another example of compounding from the same source is *khurumelakhuru* for 'kneecap'. *-khurumela* is a verb stem which means to close or to cover. *Khuru* is 'knee'. Therefore conceptualisation points to something that covers, closes off or protects the knee. Lexicalisation strategies such as these provide promising resources for African WordNet. What remains is whether or not such lexical items will filter down to everyday usage.

The last issue relates to the apparent English non-lexicalisation of concepts that are lexicalised in Northern Sotho, and *vice versa*. *Nyaraga* 'below the buttocks' is part of the Northern Sotho lexicon whose lexicalisation could not be ascertained in English. The English equivalent is provided in Northern Sotho dictionaries as a phrase. The English lexicalisation of the Northern Sotho *ntahle* 'back of hand' could also not be ascertained. Over and above being a body part, part of a hand, *ntahle* has an added connotation relating to slapping (backhand slap). That is, slapping someone with the inner part of a hand and the outer part of a hand would be reflected by the use of different lexical items. Such words need to be added as they represent concepts that are intertwined with the idiom of the language.

An expected scenario of the expand approach where English is the source language would obviously reveal Northern Sotho non-lexicalisation of concepts that are lexicalised in English. With regard to the domain under

discussion descriptive phrases are common, for example ‘nose’ is *nko* and ‘nostril’ is *lešoba la nko*, literally ‘hole of nose’. ‘Pubis’ is *lerapo la pele la noka*, literally ‘bone of front of waist’. Another lexicalisation mechanism that is productive in Bantu languages, which was nevertheless not observed in the current sample, is derivation. Affixes are used productively to form words from different word categories. Direct borrowing is also not evident in the current sample, but it is commonly used in the lexicalisation of technological concepts and specific disease names. From this sample an example of indirect borrowing is evident in coinage that resembles the English formations such as *khurumelakhuru* above and *moropana wa tsebe* literally ‘small drum of ear’ for ‘eardrum’. Lexicalisation mechanisms that were employed for this sample hint at linguistic routes to follow in dealing with further development of human body parts.

6 Challenges

While the linguistic side of the project may prove exciting, there are challenges of an IT nature. The challenges include changes in the IT infrastructure at the hosting institutions, as well as problems with the DEBVisDic editor. Such challenges hamper the development of the wordnets, as they result in interrupted access to the server and inconsistent functionality of the editor. This becomes a challenge if one wants to browse and edit existing synsets, or add new synstes. Nonetheless, manual and semi-automatic data gathering methods are used so that when a permanent IT solution is reached there is enough linguistic data to fast-track the development of the wordnets.

7 Conclusion

The paper presented actual and possible scenarios that may pose challenges when developing the Northern Sotho Wordnet on Domain_Anatomy, SUMO_BodyPart. Human body parts are targeted in this paper due to their connection to human health care and medicine. Many speakers whose L1 is not Northern Sotho may benefit from the database as it will be linking Northern Sotho not only to English but to other South African indigenous languages as well. Not only were different types of lexical misalignment presented, but also lexicalisation mechanisms that are used in the language. While the mentioned mechanisms may be grammatically sound and fill lexicalisation gaps, the words also need to receive general acceptability to the point of being in reasonably high frequency used rather than merely existing.

It is envisaged that the proposed strategies will fill the gaps, and that inclusion of internal body parts and functions, as well as verbs of expressing physical pain will produce trees that mirror the language. It remains to be seen how far the translators in the project will go in utilising the lexicalisation strategies mentioned in this paper. To assist with acceptability and standardisation the synsets will also be shared with selected practitioners in the target field for comments.

Acknowledgement

Ms Marissa Griesel, for support with the illustrations

References

Christiane Fellbaum (Ed). 1998. *Wordnet: an electronic lexical database*. Cambridge, Mass: The MIT Press.

Dirk Ziervogel and Pothinus C. Mokgokong. 1975. *Pukuntšu ye kgolo ya Sesotho sa Leboa/ Comprehensive Northern Sotho dictionary/ Groot Noord-Sotho woordeboek.* Pretoria: J. L. Van Schaik.

M.A.K Halliday, Wolfgang Teubert, Colin Yallop and Anna Čermáková. 2004. *Lexicology and Corpus Linguistics: an introduction*. London: Continuum.

Noam Ordan and Shuly Wintner. 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation* 19(1):39-58.

Northern Sotho Language Board. 1988. *Sesotho sa Leboa Mareo le Mongwalo No. 4/ Northern Sotho Terminology and Orthography No. 4/ Noord-Sotho Terminologie en Spelreëls No. 4*. Pretoria: Government Printer.

Piek Vossen (ed.) 1998. *EuroWordNet:A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

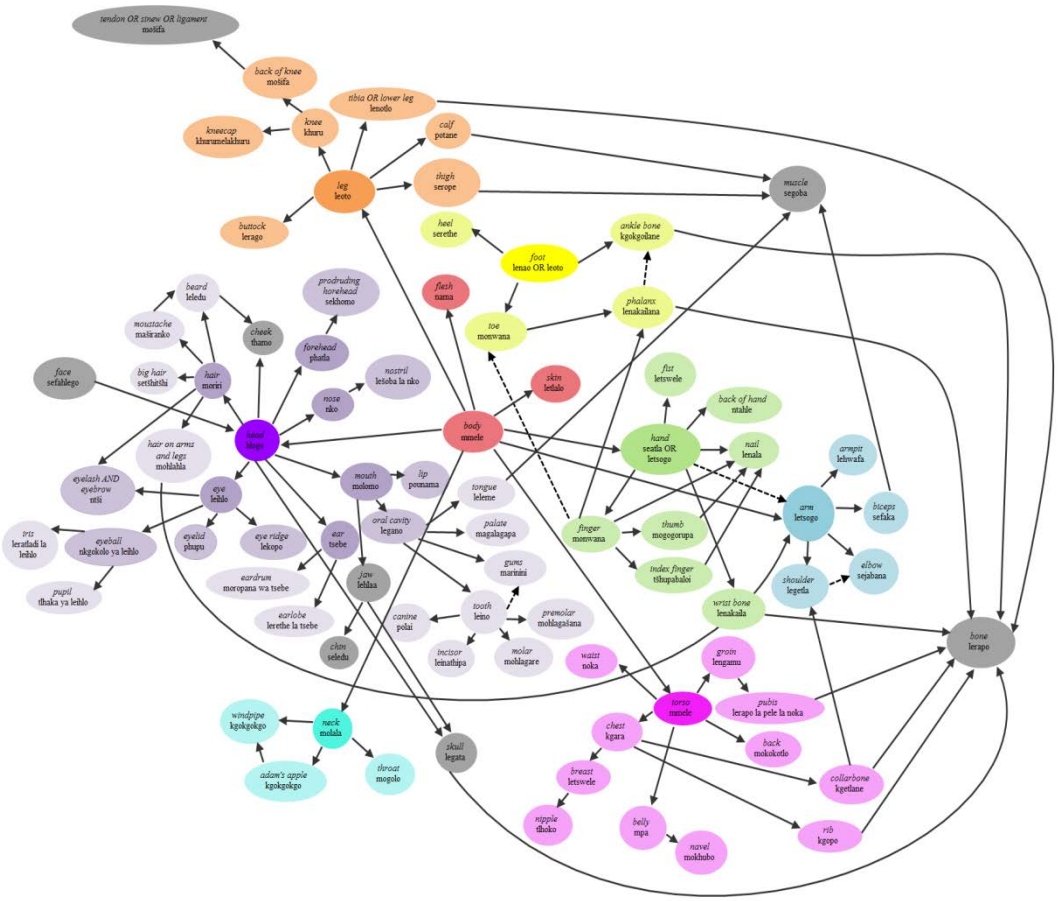
Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza and Jyoti Pawar. 2010. Experiences in Building the Konkani WordNet Using the Expansion Approach. In *Proceedings of the Fifth Global WordNet Conference, January* 2010. Mumbai, India.

Statistics South Africa http://www.stassa.gov.za accessed on 19 August 2015

Verinika Vincze and Attila Almási (2014) In *Proceedings of the Seventh Global WordNet Conference,* January 2014. University of Tartu, Estonia.

William Croft and D. Alan Cruse 2004. *Cognitive linguistics*. Cambridge: University Press.

Appendix 1: Combined connections



WME: Sense, Polarity and Affinity based Concept Resource for Medical Events

[1]Anupam Mondal **[1]Dipankar Das** **[2]Erik Cambria** **[1]Sivaji Bandyopadhyay**
[1]Department of CSE
Jadavpur University, India
[2]School of Computer Engineering
Nanyang Technological University, Singapore
[1]link.anupam@gmail.com,[1]ddas@cse,jdvu.ac.in,
[2]cambria@ntu.edu.sg,[1]sbandyopadhyay@cse.jdvu.ac.in

Abstract

In order to overcome the scarcity of medical corpora, we develop the WordNet for Medical Events (WME) for identifying such medical terms and their sense related information using a seed list. The initial WME resource contains 1654 number of medical terms. In the present task, we have reported the enhancement of WME with 6415 number of medical terms along with their conceptual features viz. gloss, semantics, polarity, sense and affinity. Several polarity lexicons viz. SentiWordNet, SenticNet, Bing Liu's subjectivity list and Taboda's adjective list were introduced with WordNet synonyms and hyponyms for expansion. The affinity feature helped us to prepare a medical ConceptNet containing the medical terms for visualization. Finally, we evaluated with respect to Adaptive Lesk Algorithm and conducted an agreement analysis for validating the expanded WME resource.

1 Introduction

In the domain of clinical text processing, sense-based information extraction is considered as a challenging task due to the unstructured nature of the corpus. The hardness in preparing structured corpora for clinical domain was found because of the less involvement of the domain experts (Smith and Fellbaum, 2004). Though several lexicons were developed and used to overcome the complexity present in the conventional NLP domain (Miller, 1995; Fellbaum, 1998).

In contrast to medical domain, the researchers introduced few number of resources e.g., Medical WordNet to overcome such problems (Burgun and Bodenreider, 2001; Bodenreider et al., 2003). The WME resource was developed along with sense-based medical information for the experts and non-expert group of people (Mondal et. al., 2015).

In the present attempt, we have expanded the WME resource with new features like semantics and affinity. The semantic feature helps to extract the relative sense-based words from the medical words and assign the type of medical words (e.g. medicine, disease etc.). The affinity feature helps to develop a medical Concept Network (ConceptNet) for visualization (Cambria et al., 2010). Started with an initial seed list of medical terms, the WordNet synonyms and hyponyms along with several polarity lexicons were employed to enrich the WME resource. The polarity lexicons viz. SentiWordNet[1], SenticNet[2], Bing Liu's subjectivity list[3] and Taboda's adjective list[4] were applied on the extracted synonyms and hyponyms for identifying the proper sense.

In next Section, we have discussed the related work associated to prepare of lexical resources for clinical domain. In Section 3, WME expansion techniques have been described along with statistics as a part of WME building. The feature selection and identification techniques were discussed under Section 4. The evaluation of the expanded WME resource and conducting agreement studies are described in Section 5. Finally, in Section 6, we conclude and mention the future scopes of the task.

2 Related Work

In the context of Bio-medical corpora, the medical terms (event) and their related information extraction can help to develop an annotation system, which is essential for representing the structured corpus (UzZaman and Allen, 2010; Hogenboom et al., 2011). The polarity, sense and concept related features are taking crucial role for preparing the structured corpus in this domain.

1 http://sentiwordnet.isti.cnr.it/
2 http://sentic.net/
3 http://www.cs.uic.edu/~liub/
4 https://www.sfu.ca/~mtaboada/research/pubs.html

Several taxonomies were designed by the researchers for understanding the medical terms and their related information for the non-experts (Tse, 2003; Zeng et al., 2003). In this concern, a research group was developed to build a medical information system using vocabulary for arbitrate the extracted information and recognize the context for the experts and non-experts (Patel et al., 2002).

Fellbaum and Smith proposed Medical WordNet (MEN) with two sub networks e.g. Medical FactNet (MFN) and Medical BeliefNet (MBN) for justifying the consumer health (Smith and Rosse, 2004). The MEN was followed the formal architecture of the Princeton WordNet (Fellbaum, 1998). The MFN guides to extract and understand the generic medical information for non-expert group whereas the MBN identifies the fraction of the beliefs about the medical phenomena (Smith and Rosse, 2004). Their primary motivation was to develop a network of medical information retrieval system with visualization effect.

The information (medical terms) extraction from the clinical corpus was treated as an ambiguous task (Pustejovsky, 1995). A group of researchers introduced the sense selection and pruning strategies for expanding the ontology of the medical domain (Toumouh et al., 2006).

WordNet of Medical Event (WME) resource was introduced as a lexical resource for identifying the medical events and their related features viz. POS, gloss, polarity and sense from the corpus (Mondal et. al., 2015). The POS signifies the lexical category of the medical events where the gloss, polarity and sense features help to provide the semantics and knowledge based information related to the medical events.

3 WME1.0 Building

The keyword extraction is essential for identifying the sense related information (e.g. "*improves"* and "*capability"* keywords provide the positive sense of the following sentence "*A supplementary component that improves capability*."). The sense-based word identification is tedious job in the domain of Bio-NLP. In this regard, in order to identify the meaning, the conventional WordNet helps to extract the word related information viz. Parts-Of-Speech (POS), synonyms, hyponyms and definition. To grasp the syntactic behavior of the medical corpus, WME1.0 resource has been prepared with medical terms and it supplies the related information, by which we can identify the syntactic and semantic behavior of the medical corpus.

The seed list of WME resource has prepared from the trial and training datasets of the SemEval-2015 Task-6.[5] The conventional WordNet and English medical dictionary were applied on the seed list for developing the initial WME resource. Primarily, the resource extracted 2479 numbers of medical events along with their attributes such as *type, span-context, sense* (*positive/negative*) from the provided datasets (e.g., *<tumor>*, *<event>*, *<An abnormal new mass of tissue that serves no purpose.>*, *<negative>*). WordNet provides the lexical information like POS, synonyms and definition of the word (medical events) (e.g., *<Abdomen>*, *<Noun>*, *<1. abdomen 2. abdominal cavity>*, *<1. "The region of the body is vertebrate between the thorax and the pelvis." 2."The cavity containing the major viscera; in mammals it is separated from the thorax by the diaphragm.>*). Meanwhile, an English Medical Dictionary identifies the POS descriptions or glosses of the words. The English Medical Dictionary was developed by H. Bateman and her group in 2007.[6] A huge amount of manual editing was carried out for the preprocessing and the preprocessed dictionary covers the 11,750 medical words in English along with POS and gloss (e.g., *<Adenoma>, <Noun>, <A benign tumor of a gland>*).

Several polarity lexicons like SentiWordNet, Taboda's adjective list etc were used for identifying the appropriate gloss of the medical events from file context, WordNet definition and dictionary gloss of the medical events. The sense-based gloss identification was considered as a task of Word Sense Disambiguation (WSD) (Basili et al., 1997). The sequential and combined WSD algorithms were applied for identifying the proper sense-based gloss of the medical terms (events) (Mondal et al., 2015).

4 WME2.0 Building

The inclusion of semantic and knowledge based features is crucial for preparing the expanded version of existing resource, WME1.0.

[5] http://alt.qcri.org/semeval2015/task6/

[6] http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+(Malestrom).pdf

The semantic, polarity, sense and affinity features have been employed as these features help to identify and extract the medical events from clinical corpus.

4.1 Feature Selection for Expansion

In order to select features, we have considered glosses and senses. The semantics and polarity features have been used for conceptual visualization (Cambria et al., 2015) and co-referencing along with the affinity relations exist among the medical events.

Gloss: It is obvious that all the words of a sentence do not always carry the concept related information (e.g., "*achievable*" is the knowledge information of the following sentence, "*The state of being achievable.*"). As the gloss identification based on concept words is crucial, in WME2.0, we have used the sequential and combined WSD approaches for extracting the proper gloss of the medical terms present in the seed lists. The extracted gloss provides the sense-based knowledge of the medical terms.

Polarity and Sense: Nowadays, opinion and sense identification is treated as an emerging task. Thus the polarity and sense features were extracted using several polarity lexicons viz. SentiWordNet, SenticNet, Bing Liu's subjective list and Taboda's adjective list. Figure 1 shows the procedure of identifying polarity and sense features of WME2.0 (e.g., <mismanage>, <-0.625>, <Negative>).

Semantic: The inclusion of semantic is to identify the similar sense-based words. In WME 2.0, the semantic of a medical term has been extracted with the help of WordNet synonyms. The example is illustrated the semantic feature of WME 2.0 (e.g., *<maltreatment>, <abuse, misuse, mismanage, overlook>*).

Affinity: The affinity feature is introduced in the present task to build a medical ConceptNet because the medical ConceptNet is essential for visualization as well as of identifying co-reference relationship. The affinity score between a pair of medical terms has been calculated by the number of similar occurrences of the semantic words. The affinity score of the medical term is measured by the following equations:

$$Affinity_{(s)} = MT_{1(s)} \cap MT_{2(s)} \quad (1)$$

$$Affinity\text{-}Score_{(s)} = Affinity_{(s)} / \sum MT_{i\,(s)}, \quad (2)$$

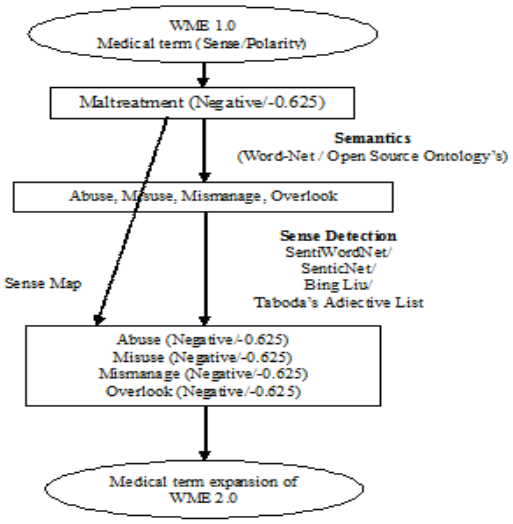


Figure 1. Sense-based technique for WME 2.0 representation

Where *i* denotes the first and second terms and $MT_{1(S)}$ and $MT_{2(S)}$ represent the semantic sets of two different medical terms. $Affinity_{(S)}$ indicates the number of common semantics of between the medical terms. $Affinity\text{-}Score_{(S)}$ is calculated with the help of $Affinity_{(S)}$ with respect to all the semantics of the medical terms. The following figure shows the medical ConceptNet along with their affinity relations.

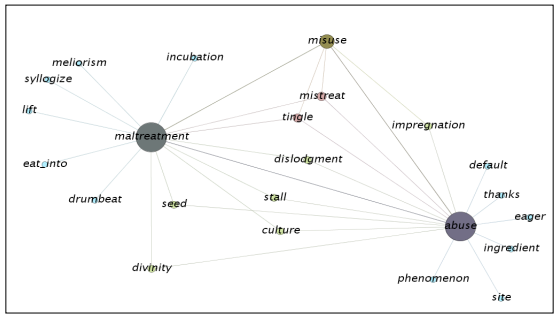


Figure 2. Partial Visualization of the Affinity score based medical ConceptNet

4.2 Statistics

We have tabulated the statistics of initial and expanded versions of WME with respect to the number of medical terms, POS and sense distributions in Table 1. The initial and expanded WME resources are termed as WME 1.0 and WME 2.0, respectively throughout the paper.

The above-mentioned statistics indicate that it is difficult to expand the WME resource with the help of word level lexical analysis (like POS distribution). The sentiment (like sense) based approaches were introduced to overcome the challenges. The detail statistics of the expanded medical terms using the above-

mentioned polarity lexicons along with a combined polarity lexicon are given in Table 2. The combined polarity lexicon has been prepared from all the above-mentioned polarity lexicons by considering the common occurrences of medical terms.

Different Basic Operation		**WME1.0**	**WME2.0**
No. of Medical terms		1654	6415
POS Distribution	Noun	1019	4219
	Verb	488	2026
	Adjective	124	111
Sense Distribution	Positive	1338	2800
	Negative	316	3615

Table 1. Comparative Statistics

Taboda's adjective list, Bing Liu's subjective list and SentiWordNet polarity lexicons were given satisfactory outputs for expanding the WME resource, where SenticNet (Cambria et al., 2016) guides us to introduce the semantic feature.

		SW	**SN**	**BL**	**TA**	**CM**
O	**S**	2938	210	1250	2509	6698
	H	4125	1136	5301	9901	19328
U	**S**	1151	196	615	1017	1592
	H	1623	698	2761	4833	6584

SW → SentiWordNet, **SN** → SenticNet, **BL** → Bing Liu's subjectivity list, **CM** → Combined Medical List, **TA** → Taboda's Adjective List
O → Original terms **U** → Unique terms
S → Synonyms **H** → Hyponyms

Table 2. Statistics based on Senses of different Polarity lexicons

5 Discussion

5.1 Evaluation

We have done the preliminary evaluation of WME 2.0 in contrast to WME1.0 with the help of sense feature. The gloss sense of the medical terms of WME 2.0 was compared with the sense extracted from the polarity lexicon, SentiWordNet. In case of clinical corpus, the SentiWordNet has a limitation of unavailability in terms of medical words. It was observed that SentiWordNet nearly covers only 40% of the medical terms of WME 2.0. On the other hand, the Lesk WSD algorithm is used to validate the senses of the medical terms of WME 2.0.

The simplified versions of Lesk algorithm primarily compares with the dictionary definition and generates the sense-based output of the term. Thus, we have found a simplified version of the Lesk algorithm which was not suitable for WME 2.0 resource due to the unavailability of dictionary definitions for most of the medical terms. To resolve it, we have applied an Adaptive Lesk algorithm for extracting the sense-based descriptions. The Adaptive Lesk algorithm not only compares the dictionary definitions but also considers the definitions of WordNet synsets.
We have evaluated the WME 2.0 using Adaptive Lesk algorithm applied to identify the proper sense-based gloss for the medical terms and represented in terms of F-Measure. F-Measure has been calculated with the help of Recall (R) and Precision (P).

$$F\text{-}Measure = 2 * [(R * P) / (R + P)] \quad (3)$$

The Precision and Recall are 82%, 62% and 57%, 29% for the WME 2.0 and Lesk algorithms, respectively. The calculated F-measure values are 71% and 38% for the WME2.0 and Lesk algorithm. The evaluation indicates that the WME 2.0 resource provides much accurate sense-based gloss information in comparison with Adaptive Lesk algorithm.

5.2 Agreement Analysis

We have conducted a manual evaluation of WME 2.0 resource for validating the expanded medical terms and their features. The agreement study is conducted by the manual annotators for the reason of unavailability of medical sense-based lexicons. The agreement analysis has been calculated by the Cohen's kappa based statistical approach.[7] The Cohen's Kappa (k) value is measured using the Proportionate (Pr(a)) and Random (Pr(e)) agreement values as follows.

$$k = [Pr(a) - Pr(e)] / [1 - Pr(e)] \quad (4)$$

Table 3 represents the agreed (Y) and non-agreed (N) medical terms and their related information for both of the annotators (denoted as A and B). The agreement score indicates a satisfactory result for WME 2.0 resource with Kappa (k) value of 0.73.

[7] https://en.wikipedia.org/wiki/Cohen's_kappa

No. of Medical Terms 6415		B	
		Y	N
A	Y	6094	51
	N	77	193

Table 3. Agreement study of WME 2.0

6 Conclusion and Future Work

The present task was initially concerned to expand the WME1.0 resource. Several polarity lexicons were used on a seed list of medical terms and their synonyms and hyponyms were also used with sense mapping for expansion. The WME 2.0 resource contains 6415 number of medical terms along with several features viz. POS, gloss, semantics, polarity, sense and affinity. The affinity feature helps us to build a medical ConceptNet for visualization. The extracted features assist to represent a system in clinical domain by which we can provide support to expert and non-expert group of people.
In future, we will attempt to enrich the WME 2.0 resource with more number of medical terms along with some concept-based features for improving the quality as well as coverage of the resource.

References

Basili. R., DellaRocca. M. and Pazienza. M. T. 1997. *Contextual word sense tuning and disambiguation.* Applied Artificial Intelligence. pp. 235-262.

Bodenreider. O., Burgun. A. and Mitchell. J. A. 2003. *Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study.* Studies in Health Technology and Informatics. pp. 379-384.

Burgun. A. and Bodenreider. O. 2001. *Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System.* In: NAACL Workshop on WordNet and Other Lexical Resources, pp. 77-82.

Cambria. E., Hussain. A., Havasi. C., Eckl. C. 2010. *SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space.* In: LNAI, vol. 6279, pp. 385–393.

Cambria. E., Fu. J., Bisio. F. and Poria. S. 2015. *AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis*. In: AAAI, pp. 508-514, Austin.

Cambria. E., Poria. S., and Schuller. B. 2016. *SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives*. In: AAAI, Phoenix.

Fellbaum. C. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge.

Hogenboom. F., Frasincar. F., Kaymak. U. and deJong. F. 2011. *An overview of event extraction from text.* In: Derive Workshop, Bonn.

Miller. G. A. 1995. *WordNet: a lexical database for English.* Comm ACM. pp. 39-41.

Mondal. A., Chaturvedi. I., Bajpai. R., Das. D., Bandyopadhyay. S. 2015. *Lexical Resource for Medical Events: A Polarity Based Approach.* IEEE 15th International Conference on Data Mining Workshops, Atlantic City.

Patel. V. L., Arocha. J. F. and Kushniruk. 2002. *A Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems.* Journal of Biomedical Informatics: 35(1). pp. 8-16.

Pustejovsky. J. 1995. *The generative lexicon.* MIT Press, Cambridge.

Smith. B. and Fellbaum. C. 2004. *Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health.* In: Coling, Geneva, pp. 31-38.

Smith. B. and Rosse. C. 2004. *The role of foundational relations in the alignment of biomedical ontologies.* In: Medinfo, San Francisco.

Toumouh. A., Lehireche. A., Widdows. D. and Malki. M. 2006. *Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus.* IEEE/ACS International Conference on Computer Systems and Applications (AICCSA).

Tse. A. Y. 2003. *Identifying and characterizing a consumer medical vocabulary.* Doctoral dissertation, College of Information Studies, University of Maryland, College Park.

UzZaman. N. and Allen. J. F. 2010. *Extracting Events and Temporal Expressions from Text.* Proceedings of IEEE International Conference on Semantic Computing.

Zeng. Q., Kogan. S., Ash. N., Greenes. R. A. and Boxwala. A. A. 2003. *Characteristics of consumer terminology for health information retrieval: A formal study of use of a health information service.* Methods of Information in Medicine.

Mapping and Generating Classifiers using an Open Chinese Ontology

Luis Morgado da Costa,[♠] **Francis Bond**[♠] **Helena Gao**[◇]
[♠]Linguistics and Multilingual Studies [◇]Chinese
Nanyang Technological University
Singapore
<luis.passos.morgado@gmail.com,bond@ieee.org,HELENAGAO@ntu.edu.sg>

Abstract

In languages such as Chinese, classifiers (CLs) play a central role in the quantification of noun-phrases. This can be a problem when generating text from input that does not specify the classifier, as in machine translation (MT) from English to Chinese. Many solutions to this problem rely on dictionaries of noun-CL pairs. However, there is no open large-scale machine-tractable dictionary of noun-CL associations. Many published resources exist, but they tend to focus on how a CL is used (e.g. what kinds of nouns can be used with it, or what features seem to be selected by each CL). In fact, since nouns are open class words, producing an exhaustive definite list of noun-CL associations is not possible, since it would quickly get out of date. Our work tries to address this problem by providing an algorithm for automatic building of a frequency based dictionary of noun-CL pairs, mapped to concepts in the Chinese Open Wordnet (Wang and Bond, 2013), an open machine-tractable dictionary for Chinese. All results will released under an open license.

1 Introduction

Classifiers (CLs) are an important part of the Chinese language. Different scholars treat this class of words very differently. Chao (1965), the traditional and authoritative native Chinese grammar, splits CLs into nine different classes. Cheng and Sybesma (1998) draw a binary distinction between *count-classifiers* and *massifiers*. Erbaugh (2002) splits CLs into three categories (*measure*, *collective* and *sortal classifiers*). Measure classifiers describe quantities (e.g. 'a bottle of', 'a mouthful of'), collective classifiers describe arrangement of objects ('a row of', 'a bunch of'), and sortal classifiers refer to a particular noun category (which can be defined, for example, by shape). Huang et al. (1997) identify four main classes, *individual classifiers*, *mass classifiers*, *kind classifiers*, and *event classifiers*. And Bond and Paik (2000) define five major types of CLs: *sortal* (which classify the kind of the noun phrase they quantify); *event* (which are used to quantify events); *mensural* (which are used to measure the amount of some property); *group* (which refer to a collection of members); and *taxonomic* (which force the noun phrase to be interpreted as a generic kind). This enumeration is far from complete, and Lai (2011) provides a detailed literature review on the most prominent views on Chinese classifiers.

Most languages make use of some of these classes (e.g. most languages have measure CLs, as in *a kilo of coffee*, or group CLs, as in *a school of fish*). What appears to be specific to some languages (e.g. Chinese, Japanese, Thai, etc.) is a class of CLs (**sortal classifiers**: **S-CL**) that depicts a selective association between quantifying morphemes and specific nouns. This association is licensed by a number of features (e.g. physical, functional, etc.) that are shared between CLs and nouns they can quantify, and these morphemes add little (but redundancy) to the semantics of noun-phrase they are quantifying.

Consider the following examples of S-CL usage in Mandarin Chinese:

(1) 两 只 狗
liǎng zhǐ gǒu
2 CL dog
"two dogs"

(2) 两 条 狗
liǎng tiáo gǒu
2 CL dog
"two dogs"

(3) 两 条 路
liǎng tiáo lù
2 CL road

"two roads"

(4) 三 台 电脑
sān tái diànnǎo
3 CL computer

"three computers"

(5) *三 只 电脑
sān zhǐ diànnǎo
3 CL computer

"three computers"

Examples (1) through (4) show how the simple act of counting in Mandarin Chinese involves pairing up nouns with specific classifiers, if incompatible nouns and classifiers are put together then the noun phrase is infelicitous, see (5).

Different S-CLs can be used to quantify the same noun, see (1) and (2), and the same type of S-CL can be used with many different nouns – so long as the semantic features are compatible between the S-CL and the noun, see (2) and (3). Extensive work on these features is provided by Gao (2010) – where more than 800 classifiers (both sortal and non-sortal) are linked in a database according to the nominal features they select, but providing only a few example nouns that can be quantified by each CL. These many-to-one selective associations are hard to keep track of, especially since they depend greatly on context, which often restricts or coerces the sense in which the noun is being used (Huang et al., 1998).

(6) 一 个 木头
yī ge mùtou
1 CL log (of wood) / blockhead

"a log / blockhead"

(7) 一 位 木头
yī wèi mùtou
1 CL blockhead

"a blockhead"

(8) 一 根 木头
yī gēn mùtou
1 CL log (of wood)

"a log"

Examples (6–8) show how the use of different CLs with ambiguous senses can help resolve this ambiguity. In (6), we can see that with the use of 个 *ge*, the most general S-CL in Mandarin Chinese, mu4tou is ambiguous because it does not restrict the noun's semantic features. With the use of 位 *wèi* (7), an honorific S-CL used almost exclusively with people, it can only be interpreted as "blockhead". And the reverse happens when using 根 *gēn* (8), a S-CL for long, slender, inanimate objects: the sense of *log (of wood)* of 木头 *mùtou* is selected.

Even though written resources concerning CLs are abundant, they are not machine tractable, and their usage is limited by copyright. Natural Language Processing (NLP) tasks depend heavily on open, machine tractable resources. Wordnets (WN) are a good example on the joint efforts to develop machine tractable dictionaries, linked in rich hierarchies. Resources like WNs play a central role in many NLP tasks (e.g. Word Sense Disambiguation, Question Answering, etc.).

Huang et al. (1998) argue that the integration between corpora and knowledge rich resources, like dictionaries, can offer good insights and generalizations on linguistic knowledge. In this paper, we follow the same line of thought by integrating both a large collection of Chinese corpora and a knowledge rich resource (the Chinese Open Wordnet: COW (Wang and Bond, 2013)). COW is a large open, machine tractable, Chinese semantic ontology, but it lacks information on noun-CL associations. We believe that enriching this resource with concept-CL links will increase the domain of it's applicability. Information about CLs could be used to generate CLs in MT tasks, or even to improve on Chinese Word Sense Disambiguation.

The remainder of this paper is structured as follows: Section 2 presents related work, followed by a description of the resources used in Section 3; Section 4 describes the algorithms applied, and Section 5 presents and discusses our results; Section 6 describes ongoing and future work; and Section 7 presents our conclusion.

2 Related Work

Mapping CLs to semantic ontologies has been attempted in the past (Sornlertlamvanich et al., 1994; Bond and Paik, 2000; Paik and Bond, 2001; Mok et al., 2012). Sornlertlamvanich et al. (1994) is the first description of leveraging hierarchical

semantic classes to generalize noun-CL pairs (in Thai). Still, their contribution was mainly theoretical, as it failed to report on the performance of their algorithm. Bond and Paik (2000) and Paik and Bond (2001) further develop these ideas to develop similar works for Japanese and Korean. In their work, CLs are assigned to semantic classes by hand, and achieve up to 81% of generation accuracy, propagating CLs down semantic classes of Goi-Taikei (Ikehara et al., 1997). Mok et al. (2012) develop a similar approach using the Japanese Wordnet (Isahara et al., 2008) and the Chinese Bilingual Wordnet (Huang et al., 2004), and report a generation score of 78.8% and 89.8% for Chinese and Japanese, respectively, on a small news corpus.

As it is common in dictionary building, all works mentioned made use of corpora to identify and extract CLs. Nevertheless, extracting noun-CL associations from corpora is not a straightforward task. Quantifier phrases are often used without a noun, resorting to anaphoric or deictic references to what is being quantified (Bond and Paik, 2000). Similarly, synecdoches also generate noise when pattern matching (Mok et al., 2012).

3 Resources

Our corpus joins data from three sources: the latest dump of the Chinese Wikipedia, the second version of Chinese Gigaword (Graff et al., 2005) and the UM-Corpus (Tian et al., 2014). This data was cleaned, sentence delimited and converted to simplified Chinese script. It was further preprocessed using the Stanford Segmentor and POS tagger (Chang et al., 2008; Tseng et al., 2005; Toutanova et al., 2003). The final version of this corpus has over 30 million sentences (950 million words). For comparison, the largest reported corpora from previous studies contained 38,000 sentences (Mok et al., 2012). In addition, we also used the latest version (2012) of the Google Ngram corpus for Chinese (Michel et al., 2011).

There are some differences between the usage of classifiers in different dialects and variations of Chinese in these different corpora, but our current goal focused on collecting generalizations. Future work could be done to single out differences across dialects and variants.

We used COW (Wang and Bond, 2013) as our lexical ontology, which shares the structure of the Princeton Wordnet (PWN) (Fellbaum, 1998). To minimize coverage issues, we enriched it with data from the Bilingual Ontological Wordnet (BOW) (Huang et al., 2004), the Southeast University Wordnet (SEW) (Xu et al., 2008), and automatically collected data from Wiktionary and CLDR, made available by the Extended OMW (Bond and Foster, 2013). The final version of this resource had information for over 261k nominal lemmas, from which over 184k were unambiguous (i.e. have only a single sense).

We filtered all CLs against a list of 204 S-CLs provided by Huang et al. (1997). Following Lai (2011), we treated both Huang's *individual classifiers* and *event classifiers* as S-CLs.

4 Our Algorithm

Our algorithm produces two CL dictionaries with frequency information: a lemma based dictionary, and a concept based dictionary, using COW's extended ontology. We tested both dictionaries with a generation task, automatically validated against a held out portion the corpus.

4.1 Extracting Classifier-Noun Pairs

Extracting CL-noun pairs is done by matching POS patterns against the training section of our corpus. To avoid, as much as possible, noise in the extracted data, we choose to take advantage of our large corpus to apply restrictive pattern variations of the basic form: (determiner or numeral) + (CL) + (noun) + (end of sentence punctuation/select conjunctions). Our patterns assure that no long dependencies exist after the CL, and try to maximally reduce the noise introduced by anaphoric, deictic or synecdochic uses of classifiers (Mok et al., 2012). Variations of this pattern were also included to cover for different segmentations produced by the preprocessing tools.

If an extracted CL matches the list of S-CLs, we include this noun-CL pair in the lemma based dictionary. The frequency with which a specific noun-CL pair is seen in the corpus is also stored, showing the strength of the association.

Extracting noun-CL pairs from the Chinese Google NGram corpus required a special treatment. We used the available 4 gram version of this corpus to match a similar pattern (and variations) to the one mentioned above: (determiner or numeral) + (CL) + (X) + (end of sentence punctuation/select conjunctions). Given we had no POS information available for the NGram corpus, we

used regular expression matching, listing common determiners, numerals, punctuation, and our list of 204 S-CLs. We did not restrict the third gram. We also transferred the frequency information provided for matched ngrams to our lemma based dictionary.

Our training set included 80% of the text portion of the corpus, from which we extracted over 435k tokens of noun-CL associations, along with the full Chinese Google NGram corpus, from which we extracted 13.5 million tokens of noun-CL associations.

This lemma based dictionary contained, for example, 59 pairs of noun-CL containing the lemma 类别 *lèibié* "category". It occurred 58 times with the CL 个 *ge*, and once with the CL 项 *xiàng*. Despite the large difference in frequencies, both CLs can be used with this lemma. Another example, where the relevance of the frequency becomes evident, is the word 养鸡场 *yǎngjīchǎng* "chicken farm", which was seen in our corpus 12 times: 6 times with the CL 个 *ge*, 3 times with the CL 家 *jiā*, twice with the CL 只 *zhǐ*, and once with the CL 座 *zuò*. Chinese native speaker judgments identified that three out of the 4 CLs identified were correct (个 *ge*, 家 *jiā* and 座 *zuò*). In addition, two other classifiers would also be possible: 间 *jiān* and 所 *suǒ*. This second example shows that while the automatic matching process is still somewhat noisy, and incomplete, the frequency information can help to filter out ungrammatical examples. When used to generate a classifier, our lemma based dictionary can use the frequency information stored for each identified CL for a particular lemma, and choose the most frequent CL. This process will likely increase the likelihood of it being a valid CL. Also, by setting a minimum frequency threshold for which noun-CLs pair would have to be seen before being added to the dictionary, we can exchange precision for coverage.

4.2 Concept Based Dictionary

The concept based dictionary is created by mapping and expanding the lemma based dictionary onto COW's expanded concept hierarchy. Since ambiguous lemmas can, in principle, use different CLs depending on their sense, we map only unambiguous lemmas (i.e. that belong to a single concept). This way, each unambiguous entry from the lemma based dictionary matching to COW contributes information to a single concept. Frequency information and possible CLs are collected for each matched sense. The resulting concept-based mapping, for each concept, is the union of CLs for each unambiguous lemma along with sum of frequencies.

Following one of the examples above, the lemma 类别 *lèibié*, was unambiguously mapped to the concept ID 05838765-n – defined as "a general concept that marks divisions or coordinations in a conceptual scheme". This concept provides two other synonyms: 范畴 *fànchóu* and 种类 *zhǒnglèi*. In the concept based dictionary, the concept ID 05838765-n will aggregate the information provided by all its unambiguous senses. This results in a frequency count of 132 for the CL 个 *ge*, and of 2 for 项 *xiàng* (both valid uses).

As has been shown in previous works, semantic ontologies should, in principle, be able to simulate the taxanomic features hierarchy that link nouns and CLs. We use this to further expand the concept based dictionary of CLs.

For each concept that didn't receive a classifier, we collect information concerning ten levels of hypernymy and hyponymy around it. If any pair of hypernym-hyponym was associated with the same CL, we assign this CL to the current concept. Since we're interested in the task of generating the best (or most common) CL, we rank CLs inside these expanded concepts by summing the frequencies of all hypernyms and hyponyms that shared the same CL. If more than one CL can be assigned this way, we do so.

Figure 1 exemplifies this expansion. While concepts A, B and C did not get classifiers directly assigned to them, they are still assigned one or more classifiers based on their place in the concept hierarchy. For every concept that didn't receive any CL information, if it has at least a hypernym and a hyponym sharing a CL (within a distance of 10 jumps), then it will inherit this CL and the sum of their frequencies. Assuming a full concept hierarchy is represented in Figure 1, concept A would inherit two classifiers, and concept B and C would inherit one each.

This expansion provides extra coverage to the concept based dictionary. But we differ from previous works in the sense that we do not blindly assign CLs down the concept hierarchy, making it depend on previously extracted information for both hypernyms and hyponyms. By following a

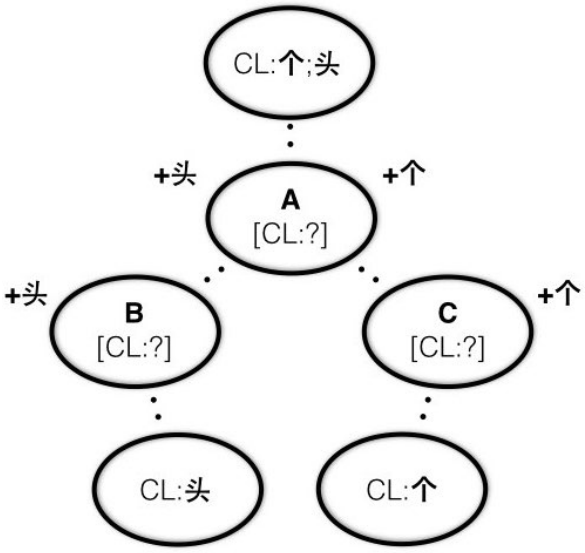


Figure 1: Classifier Expansion

stricter approach, we hope to provide results of better quality.

4.3 Automatic Evaluation

We evaluated both lemma and concept based dictionaries with two tasks: predicting the validity of and generating CLs. We used roughly 10% of held out data (dev-set), from which we extracted about 37,4k tokens of noun-CL pairs, as described in 4.1. We used this data to evaluate the prediction and generation capabilities of both dictionaries in the following ways: predicting the validity of a CL was measured by comparing every noun-CL pair extracted from the dev-set to the data contained in the dictionary for that particular lemma (i.e. if that particular classifier was already predicted by the dictionary); generation was measured by selecting the best likely classifier, based on the cumulative frequencies of noun-CL pairs in the dictionary (i.e. if the classifier seen in the example matched the most frequent classifier). This was done separately for both dictionaries.

When no other classifier had been assigned, we used 个 *ge*, the most frequent CL on the corpus, as the default classifier. And a baseline was established by assigning 个 *ge* as the only CL for every entry.

The dev-set was used to experiment with different thresholds (τ) of the minimum frequency, from one to five, for which noun-CL pairs would have to be seen in the train-set in order to be considered into the dictionaries. These different minimum frequency thresholds were compared between both tasks.

	$\tau=1$	$\tau=3$	$\tau=5$	*Test*
baseline	44.2	44.2	44.2	40.4
All lemmas				
lem-all	92.7	88.5	86.2	93.6
lem-all-mfcl	75.1	73.8	72.8	78.9
lem-all-no-info	4.7	9.2	12.1	4.1
Unamb. lemmas				
lem-unamb	93.2	88.2	85.5	94.5
wn-unamb	**95.1**	90.9	88.3	**95.9**
lem-unamb-mfcl	**77.0**	75.5	74.1	**77.9**
wn-unamb-mfcl	72.3	71.6	70.7	73.5
lem-unamb-no-info	3.4	9.5	13.6	2.8
wn-unamb-no-info	1.7	5.3	8.3	1.5
Coverage				
lemmas-w/cl	32.4k	10.4k	7.0k	
wn-concepts-w/cl	22.7k	15.0k	12.3k	

Table 1: Automatic Evaluation Results

The best performing τ was then tested in a a second held-out set of data (test-set), also containing roughly 10% of the size of the text corpus, roughly 39.9k tokens of noun-CL pairs. The test-set is used to report our final results.

The results are presented in Table 1, and are discussed in the following section.

5 Discussion and Results

In Table 1 we can start to note that the baseline, of consistently assigning 个 *ge* to every entry in the dictionary is fairly high, of roughly 40%.

In order to allow a fair comparison, since we decided that the concept based dictionary would contain only unambiguous lemmas, we only use unambiguous lemmas to compare the performance across dictionaries. All results can be compared across the different thresholds discussed in 4.3. $\tau = 1, 3$ and 5 present the results obtained in the automatic evaluation, using minimum frequencies of one, three and five, respectively.

The first three reported results report exclusively about the lemma dictionary (including both ambiguous and unambiguous lemmas). *lem-all* reports the results of the prediction task, *lem-all-mfcl* reports the results of the generation task, and *lem-all-no-info* reports the relative frequency of lemmas for which there was no previous infor-

mation in the dictionary, and which could have boosted both task's performance by falling back on the default CL $\uparrow$ *ge*.

These initial results show that it was easy to perform better than baseline, and that $\tau = 1$ achieved the best results on both predicting noun-CL pairs, and generating CLs that matched the data.

Comparing different τs shows that, even considering the over-generation reduction that imposing minimum frequencies brings (validated but not presented here), the best generation performance is achieved by not filtering the training data. And this will be consistent across the remainder of the results.

When comparing both dictionaries, we look only at unambiguous lemmas. Similar to what was explained above, *lem-unamb* and *wn-unamb* report the results of the prediction task for the lemma based and concept based dictionary, respectively. The labels *lem-unamb-mfcl* and *wn-unamb-mfcl* report the results for the generation task. And the *lem-unamb-no-info* and *wn-unamb-no-info* report about the lack of informed coverage (where backing-off to the default CL might have help the performance).

Between the lemma and the concept based dictionaries, this automatic evaluation shows that while the concept based dictionary is better at predicting if a noun-CL pair was valid, the lemma based dictionary outperforms the former in the generation task.

The final results of this automatic evaluation are shown in column *Test*, where we re-evaluated the dictionary produced by $\tau = 1$ on the test-set. *Test* shows slightly better results, perhaps because the random sample was easier than the dev-set, but the same tendencies as reported above.

Considering that the concept based dictionary should be able to provide CL information to some lemmas that have not been seen in the training data (either by expansion or by leveraging on a single lemma to provide information about synonyms), we expected the concept based dictionary to present the best results.

Many different reasons could be influencing these results, such as errors in the ontology, the fact that Chinese CLs relate better to specific senses than to concepts (i.e. different lemmas inside a concept prefer different CLs), or noise introduced by the test and dev-set (since we don't have a hand curated golden test-set). For this reason, we decided to hand validate a sample of each dictionary.

Based on a random sample of 100 concepts and 100 lemmas extracted from each dictionary, a Chinese native speaker checked if the top ranked CL (i.e. with highest frequency), that would be used to generate a CL for each of the randomly selected entries, was in fact a valid CL for that lemma or concept. This human validation showed the concept based dictionary outperforming the lemma based dictionary by a wide margin: 87% versus 76% valid guesses. This inversion of performance, when compared to the automatic evaluation, was confirmed to be mainly due to noisy data in the test-set caused by the automatic segmentation and POS tagging.

We then looked at a bigger sample of 200 lemmas and found roughly 7.5% of invalid lemmas in the lemma based dictionary. Conversely, the concept based dictionary assigns CLs by 'bags of lemmas' (i.e. synsets). This allows the noise introduced by a few senses to be attenuated by the 'bag' nature of the concept. More importantly, most of the nominal lemmas included in the extended version of COW are human validated, so the quality of the concept based dictionary was confirmed to be better – since most lemmas included in it are attested to be valid.

Comparing the size of both dictionaries in Table 1, even though the $\tau 1$ lemma based dictionary is considerably larger (32.4k compared to 22.5k entries of the concept based dictionary), we have shown that noise is a problem for the lemma based approach. Also, since the extended COW has, on average, 2.25 senses per concept, the concept based dictionary provides CL information for over 50.6k lemmas. When comparing the size of both dictionaries across τs, we can also effectively verify the potential of the expansion step possible only for the concept based dictionary. As τ increases, the size of the concept based dictionary increases relatively to the lemma based. When applied to other tasks, where noise reduction would play a more important role (which can be done by raising τ), the concept based dictionary is able to produce more informed decisions with less data.

Lastly, coverage was also tested against data from a human curated database of noun-CL associations (Gao, 2014), by replicating the automatic evaluation generation task described in 4.3. This dictionary contains information about more than

800 CLs and provides a few hand-selected examples for each CL – and hence it is not designed with the same mindset. Testing the best performing dictionaries (τ1) against the data provided for S-CLs, we achieved only 43.9% and 28.3% for prediction and generation, respectively, using the lemma based dictionary; compared to 49.8% and 22.4% using the concept based dictionary.

The same trends in prediction and generation are observed, where the concept based dictionary is able to predict better than the lemma base, but it is outperformed by the later in the generation task. Ultimately, these weak results show that even though we used a very large quantity of data, our restrictive matching patterns in conjunction with infrequent noun-CLs pairs still leaves a long tail of difficult predictions.

6 Ongoing and Future Work

Since our method is mostly language independent, we would like to replicate it with other calssifier languages for which there are open linked WN resources (such as Japanese, Indonesian and Thai). This would require access to large amounts of text segmented, POS tagged text, and adapting the matching expressions for extracting noun-CL pairs.

More training data would not only help improving overall performance on open data, by minimizing unseen data, but would also allow us to make better use of frequency threshold filters for noise reduction. Lack of training data as our biggest drawback on performance, we would like to repeat this experiment with more data – including, for example, a very large web-crawled corpus in our experiments.

In addition, we would also like to perform WSD on the training set, using UKB (Agirre and Soroa, 2009) for example. This would allow an informed mapping of ambiguous senses onto the semantic ontology and, arguably, comparable performance on generating CLs for ambiguous lemmas. We will also investigate further how to deal with words not in COW: first looking them up in the lemma dictionary, and then associating CLs to the head (character / noun) of unseen noun-phrases, as proposed in Bond and Paik (2000).

Even though this work was mainly focused on producing an external resource linked to COW, we are also investigating adding a new set of sortal classifiers concepts to COW. The absence of this class of words in COW currently prevents us from using the internal ontology structure to link nouns and classifiers. Once classifiers are represented as concepts in this lexical ontology, we will make use of this work to link nominal concepts and corresponding valid classifiers.

7 Conclusions

Our work shows that it is possible to create a high quality dictionary of noun-CLs, with generation capabilities, by extracting frequency information from large corpora. We compared both a lemma based approach and a concept based approach, and our best results report a human validated performance of 87% on generation of classifiers using a concept based dictionary. This is roughly a 9% improvement against the only other known work done on Chinese CL generation using wordnet (Mok et al., 2012).

Finally, we will merge all three data sets and, from them, produce a release of this data. We commit to make both lemma and WN mappings available under an open license, release along with the Chinese Open Wordnet at `http://compling.hss.ntu.edu.sg/cow/`.

8 Acknowledgments

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2000. Reusing an ontology to generate numeral classifiers. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, pages 90–96.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In

Proceedings of the Third Workshop on Statistical Machine Translation, StatMT '08, pages 224–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Y.R. Chao. 1965. *A Grammar of Spoken Chinese*. University of California Press.

Lisa Lai-Shen Cheng and Rint Sybesma. 1998. Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua journal of Chinese studies*, 28(3):385–412.

Mary S Erbaugh. 2002. Classifiers are for specification: Complementary functions for sortal and general classifiers in Cantonese and Mandarin. *Cahiers de linguistique-Asie orientale*, 31(1):33–69.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Helena Gao. 2010. Computational lexicography: A feature-based approach in designing an e-dictionary of Chinese classifiers. In *Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon*, pages 56–65. Coling 2010.

Helena Gao. 2014. Database design of an online e-learning tool of Chinese classifiers. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 126–137.

David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2005. *Chinese Gigaword Second Edition LDC2005T14*. Web Download. Linguistic Data Consortium.

Chu-Ren Huang, Keh-Jiann Chen, and Ching-Hsiung Lai, editors. 1997. *Mandarin Daily Dictionary of Chinese Classifiers*. Mandarin Daily Press, Taipei.

Chu-Ren Huang, Keh-jiann Chen, and Zhao-ming Gao. 1998. Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. *Quantitative and Computational Studies of Chinese Linguistics*, pages 339–352.

Chu-Ren Huang, Ru-Yng Chang, and Hshiang-Pin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of bilingual wordnet and sumo. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 825–826. European Language Resources Association (ELRA).

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Wan-chun Lai. 2011. Identifying True Classifiers in Mandarin Chinese. Master's thesis, National Chengchi University, Taiwan.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science 14 January 2011*, 331(6014):176–182.

Hazel Mok, Eshley Gao, and Francis Bond. 2012. Generating numeral classifiers in Chinese and Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 211-218.

Kyonghee Paik and Francis Bond. 2001. Multilingual generation of numeral classifiers using a common ontology. In *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*, Seoul. 141–147.

Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 556–561. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the NAACL HLT 2003 2003 - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18, Nagoya.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English wordnet. In John Domingue and Chutiporn Anutariya, editors, *The Semantic Web*, volume 5367 of *Lecture Notes in Computer Science*, pages 302–314. Springer Berlin Heidelberg.

IndoWordNet Conversion to Web Ontology Language (OWL)

Apurva Nagvenkar
DCST, Goa University
apurv.nagvenkar@gmail.com

Jyoti Pawar
DCST, Goa University
jyotidpawar@gmail.com

Pushpak Bhattacharyya
CSE, IIT Bombay
pb@cse.iitb.ac.in

Abstract

WordNet plays a significant role in Linked Open Data (LOD) cloud. It has numerous application ranging from ontology annotation to ontology mapping. IndoWordNet is a linked WordNet connecting 18 Indian language WordNets with Hindi as a source WordNet. The Hindi WordNet was initially developed by linking it to English WordNet.

In this paper, we present a data representation of IndoWordNet in Web Ontology Language (OWL). The schema of Princeton WordNet has been enhanced to support the representation of IndoWordNet. This IndoWordNet representation in OWL format is now available to link other web resources. This representation is implemented for eight Indian languages.

1 Introduction

The World Wide Web (WWW) has formed a revolution in the data availability there is no other place in the world where we can find so much of the information, but the current web structure fails to make best out of it. The user can access limitless data from the web yet, it becomes a tedious task to retrieve relevant information. Data available on the Web covers diverse structures, formats and content. It also lacks a uniform organization of scheme that would allow easy access of data and information (Candan et al., 2001). Many frameworks have been proposed to support the search engine and information access. Resource Description Framework[1](RDF), Web Ontology Language[2](OWL) is one of the framework which provides a platform for standardization and organization of data from the Web. It has been highly influenced by the web standards community.

	Noun	Verb	Adjective	Adverb	Total
Bengali	27281	2804	5815	445	36346
Gujarati	26503	2805	5828	445	35599
Hindi	29106	3306	6178	482	39072
Kashmiri	21041	2660	5365	400	29469
Konkani	23144	3000	5744	482	32370
Odia	27216	2418	5273	377	35284
Punjabi	23255	2836	5830	443	32364
Urdu	22990	2801	5786	443	34280

Table 1: POS wise statistics for Indradhanush

WordNet (Fellbaum, 1998), a lexical knowledge base system that has been adopted by the Semantic Web research community. The current essential need is to link WordNet with different resources in order to assist Natural Language Processing applications. IndoWordNet (Bhattacharyya, 2010) is an Indian community which builds WordNets for Indian languages. It is a multilingual WordNet which links WordNets of different Indian languages on a common identification number called as synset_id given to each concept (Bhattacharyya, 2010). It is constructed using the expansion model where Hindi WordNet synsets are taken as a source. The concepts provided along with the Hindi synsets are first conceived and appropriate concepts in target language are manually provided by the language experts. Figure 1 shows the statistics of Indradhanush Consortium which consist seven Indian languages belonging to Indo-Aryan family and is part of IndowordNet Consortium.

To use WordNet in Semantic Web the data model for WordNet should be extensible, interoperable and flexible. It was created as a semantic network of word meanings which at the conceptual level is a directed graph with labeled nodes and arcs (Graves and Gutierrez, 2006). Hence, OWL can be used to model WordNet since, it facilitates data manipulations and queries over the

[1] http://www.w3.org/RDF
[2] http://www.w3.org/TR/owl-features

graph structure. The main objective of this paper is to represent IndoWordNet to OWL representation.

The rest of the paper is organized as follows section 2 describes the related work. Section 3 introduces to Semantic Web Layer Cake Model. Section 4 presents the architecture of IndoWordNet OWL; section 5 gives the implementation details, followed by conclusion and future work.

2 Related Work

WordNets other than Indian languages are already available in RDF form. The work on Princeton WordNet (Assem et al., 2006) conversion to RDF/OWL was carried out by WordNet Task Force[3]. The main goal of this conversion was to represent a language in use of Semantic Web community and to provide application developers a resource. Also, the representation was done in such a way that it maintained the WordNets conceptual model.

There are other projects focusing on lexical meta-models. Lexical Markup Framework (LMF) (Francopoulo et al., 2009). IndoWordNet is already available in this format by IndoNet (Bhatt et al., 2013) which proposes modification to LMF to integrate Universal Word Dictionary (Uchida et al., 1999) and Suggested Upper Merged Ontology (SUMO) (Pease et al., 2002).

3 Semantic Web Layer Cake Model

The Semantic Web is not a separate web but a vision for the future of the Web where information is given explicit meaning which makes easier for machine to automatically process and integrate the information available on the web. OWL is a part of the growing stack of W3C recommendations related to the semantic web (McGuinness and Harmelen, 2004).

Figure 1. is the semantic web layer cake model (Hendler, 2001). This model is divided into three section:

1. Hypertext Web technologies: The bottom layer contains technologies which are used by hypertext web that includes Unicode, Universal Resource Indicator (URI), XML and XML-schema. Unicode is used to represent and manipulate text for different languages. URI represents the resources uniquely. XML provides the syntax for structured document, but does not provide any meaning to the document. XML schema restricts the structure of the document and extends XML with datatypes.

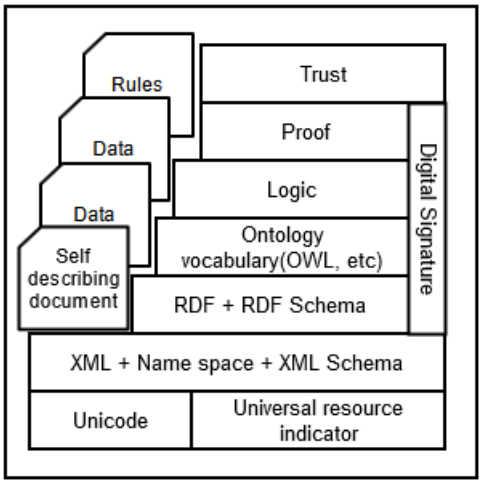


Figure 1: Semantic web layer cake model

2. Standardized Semantic Web technologies: The middle layer contains technologies which are already standardized by Semantic Web community that includes RDF, RDFS, OWL and SPARQL. RDF is a data model to represent triple, i.e. objects and relationship between them. It provides simple semantics and is represented by XML syntax. RDF schema can be viewed as an extensible, object oriented type system based on RDF (Huang and Zhou, 2007). OWL is an envelope to the RDF schema and enriches the expressibility of the RDF schema by expressing more properties like transitivity, symmetry, cardinality, etc.

3. Unrealized Semantic Web technologies: The top layer contains technologies like digital signatures, trust, proof, etc this technologies are not yet standardized by Semantic Web community and needs to be implemented in order to realize Semantic Web.

4 OWL for IndoWordNet

The architecture of the IndoWordNet OWL representation is adopted from WordNet Task Force (Assem et al., 2006). The architecture of IndoWordNet OWL contains three main classes i.e.

[3] http://www.w3.org/TR/wordnet-rdf/

Synset[4], WordSense and Word[5] .

The schema for representing IndoWordNet[6] using OWL is shown in figure 2 below.

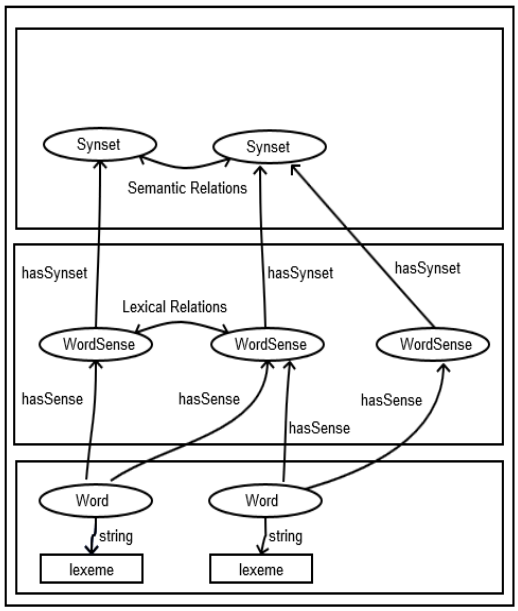


Figure 2: IndoWordNet OWL schema

The schema includes three layers, namely Concept layer, WordSense layer and Word layer which are previously described in (Huang and Zhou, 2007). Every synset has a unique concept and can have several words associated with it sharing the same concept. WordSense represents a unique sense of a word. It is also possible to represent a word with many WordSenses. IndoWordNet OWL schema handles the relations by dividing them into properties, i.e. Semantic property and Lexical property. Semantic property represents the semantic relations which are handled in concept layer, whereas lexical property represents lexical relations which are handled in WordSense layer. All the remaining types of semantic relations and lexical relations become the sub property of semantic and lexcial property. The above schema uses several predicates[7] i.e. properties.

IndoWordNet OWL schema elaborates the semantic relationship like meronymy and holonymy by classifying them into the sub properties based on their attributes[8] whereas in Princeton WordNet there is no such division.

In IndoWordNet OWL, the RDF files are organized in such a way that the management is done systematically. Unlike (Assem et al., 2006) all the RDF files are placed in one directory.

Following is the formatting of URIs for IndoWordNet:

- URI representation of a synset: `http://nlp.unigoa.ac.in/indonet/owl/hindi/v1/synset/noun/24.rdf`
- URI representation of a wordSense: `http://nlp.unigoa.ac.in/indonet/owl/hindi/v1/wordSense/ł/noun/1930.rdf`
- URI representation of a word: `http://nlp.unigoa.ac.in/indonet/owl/hindi/v1/word/ł.rdf`

5 Implementation Details

The IndoWordNet OWL is currently available in seven Indian languages. It is developed using JAVA platform, using Apache Jena[9] and IndoWordNet Application Programming Interface(API). The above architecture can be used by other Indian langauges to represent their respective wordNets in OWL format. The repository of IndoWordNet OWL is available on `http://nlp.unigoa.ac.in/indonet/owl/`.

6 Conclusion and Future Work

The heart of Semantic Web is Linked Data that provides integration and reasoning of the data on web. The representation of IndoWordNet to OWL will facilitate the semantic web community as the WordNet is strong lexical resource that has strengthened, enlarged and build up the other resources because of its taxonomy. In this paper we have presented the framework to represent the Indian wordNets in the OWL format. Currently, we have represented eight Indian language WordNets in OWL format. In future, we will like to represent the WordNets from other Indian languages in OWL format. Following are some future work to this problem.

[4] `http://nlp.unigoa.ac.in/indonet/owl/web/syn.php`

[5] `http://nlp.unigoa.ac.in/indonet/owl/web/wdSenseAndWord.php`

[6] `http://nlp.unigoa.ac.in/indonet/owl/IndoWNetSchema.rdf`

[7] `http://nlp.unigoa.ac.in/indonet/owl/web/prop.php`

[8] `http://nlp.unigoa.ac.in/indonet/owl/web/propdist.php`

[9] `https://jena.apache.org/`

Interlinking of WordNets: As the IndoWordNet is developed using ILI. The advantage of this approach is that it preserves the semantic structure, but it also has some disadvantages. The drawbacks of this approach are lexical gap and semantic gap (Fellbaum and Vossen, 2012). As a result, an effort must be made to interlink the WordNet using Common Concept Hierarchy (Bhatt et al., 2013) as a backbone to link lexicons of different languages.

Need of approach to link DBpedia: The work on linking the IndoWordNet to DBpedia should be carried out as, DBpedia is the nucleus for the web of data and most of the resources are already linked to DBpedia.

Link it to other Resources: We expect that use of the OWL representation of IndoWordNet will be used as an infrastructure to enrich and link other web resources in India.

References

[Graves and Gutierrez2006] Alvaro Graves, Claudio Gutierrez. 2006. *Data Representation for WordNet: A Case for RDF*. 3rd Global WordNet Association Conference.

[Bhatt et al.2013] Brijesh Bhatt, Lahari Poddar, Pushpak Bhattacharyya. 2013. *IndoNet: A Multilingual Lexical Knowledge Network for Indian Languages*. Association for Computational Linguistics.

[Fellbaum1998] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.*. Cambridge, MA: MIT Press.

[Fellbaum and Vossen2012] Christiane Fellbaum and Piek Vossen. 2012. *Challenges for a multilingual wordnet.*. Lang. Resour. Eval., 46(2):313326.

[McGuinness and Harmelen2004] Deborah L. McGuinness, Frank van Harmelen. 2004. *OWL Web Ontology Language*. http://www.w3.org/TR/owl-features.

[Francopoulo et al.2009] Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. *LMF for Multilingual Specialized Lexicons*. LREC Workshop on Acquiring and Representing Multilingual, Specialized Lexicons.

[Hendler2001] J. Hendler. 2001. *Agents and the Semantic Web*. IEEE Intellegent Systems.

[Candan et al.2001] K. Seluk Candan, Huan Liu, and Reshma Suvarna. 2001. *Resource description framework: metadata and its applications*. SIGKDD Explor. Newsl. 3, 1 (July 2001), 6-19.

[Kuroda et al.2010] Kow Kuroda, Francis Bond, Kentaro Torisawa. 2010. *Why Wikipedia needs to make friends with WordNet*. 5th Global WordNet Association Conference.

[Assem et al.2006] Mark van Assem, Aldo Gangemi, Guus Schreiber. 2006. *Conversion of WordNet to a standard RDF/OWL representaion*. Proceedings of LERC.

[Casado et al.2005] Maria Ruiz-Casado, Enrique Alfoneseca, Pablo Castells. 2005. *Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets*. In: Proceedings of the Atlantic Web Intelligence Conference, AWIC-2005.

[Bhattacharyya2010] Pushpak Bhattacharyya. 2010. *IndoWordNet*. Proceedings of LERC.

[Auer et al.2007] Soren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives. 2007. *DBpedia: A Nucleus for a Web of Open Data*. ISWC'07/ASWC'07 Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference.

[Huang and Zhou2007] Xiao-xi Huang, Chang-le Zhou. 2007. *An OWL-based WordNet lexical ontology*. Journal of Zhejiang Science A.

[Pease et al.2002] Adam Pease, Ian Niles and John Li 2002. *The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications*. In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web.

[Uchida et al.1999] H. Uchida, M. Zhu, and T. Della Senta 1999. *UNL- a Gift for the Millenium*. United Nations University Press, Tokyo.

A Two-Phase Approach for Building Vietnamese WordNet

Phuong-Thai Nguyen
VNU University of Engineering and Technology
thainp@vnu.edu.vn

Van-Lam Pham
VASS Institute of Linguistics
lampv.il@vass.gov.vn

Hoang-An Nguyen
Naiscorp Inc.
annh@socbay.com

Huy-Hien Vu
VNU University of Engineering and Technology
hienvuhuy@vnu.edu.vn

Ngoc-Anh Tran
Le Quy Don Technical University
anhtn69@gmail.com

Thi-Thu-Ha Truong
VASS Institute of
Lexicography and Encyclopedia
hattt.viole@vass.gov.vn

Abstract

Wordnets play an important role not only in linguistics but also in natural language processing (NLP). This paper reports major results of a project which aims to construct a wordnet for Vietnamese language. We propose a two-phase approach to the construction of Vietnamese WordNet employing available language resources and ensuring Vietnamese specific linguistic and cultural characteristics. We also give statistical results and analyses to show characteristics of the wordnet.

Length	Words	Percentage
1	6,303	15.69
2	28,416	70.72
3	2,259	5.62
4	2,784	6.93
5	419	1.04
Total	40,181	100

Table 1: Word length statistics from a popular Vietnamese dictionary, made by the Vietnam Lexicography Center (Vietlex).

1 Introduction

In order to solve various problems in NLP including information retrieval, machine translation, text classification, etc. we need language resources such as corpora and dictionaries. Wordnet is one of important resources for solving such problems. The first wordnet was created at Princeton University for English language. After that, diverse wordnets were constructed such as EuroWordNet for European languages, Asian WordNet for Asian languages, etc.

There are a number of important characteristics of the Vietnamese language that impact the construction of wordnet. Firstly, the smallest unit in the formation of Vietnamese words is the syllable. Words can have just one syllable, for example 'đẹp'$_{beautiful}$, or be a compound of two or more syllables, for example 'màu sắc'$_{color}$. As shown in Table 1, single-syllable words only cover a small proportion while two-syllable words account for the largest proportion of the whole vocabulary. Forming that vocabulary is a set of 7,729 syllables, higher than the number of single words. As in many other Asian languages such as Chinese, Japanese and Thai, there is no word delimiter in Vietnamese. The space is a syllable delimiter but not a word delimiter, so a Vietnamese sentence can often be segmented in many ways. Secondly, Vietnamese is an isolating language in which words do not change their forms according to their grammatical function in a sentence.

Constructing wordnets is a complicated task. This task involves answering questions including which approach is appropriate, how to ensure specific characteristics of the language, how to take full advantage of available resources. This paper makes an attempt to answer these fundamental questions and reports major results of a project aiming to construct a wordnet for Vietnamese language, whose database includes 30,000 synonym sets and 50,000 words with 30,000 commonly used by the Vienamese.

Figure 1 represents major steps in construction

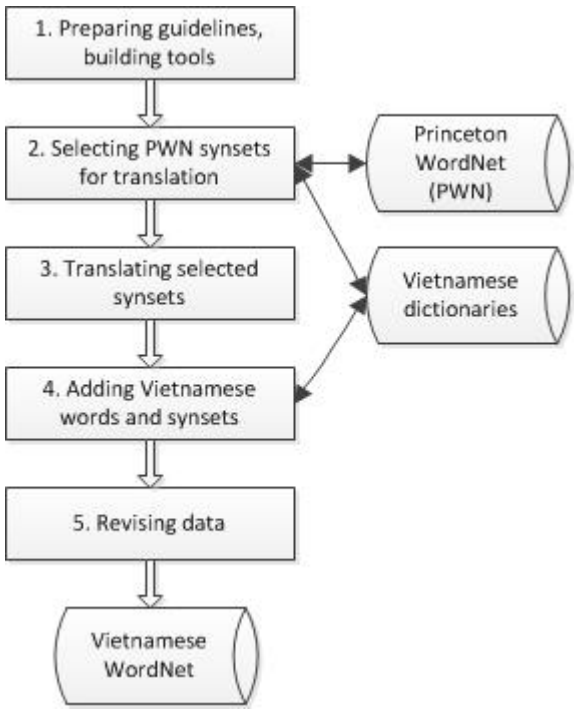


Figure 1: Steps in Vietnamese WordNet construction.

process of Vietnamese WordNet. We put these steps in two phases. Phase 1 involves steps 1-3, phase 2 involves steps 4 and 5. We exploit a number of language resources including Princeton's WordNet, a Vietnamese dictionary and an English-Vietnamese bilingual dictionary.

The class of adverbs in Vietnamese is a closed class (or a class of function words), while in English the class of adverbs is an open class (or a class of content words). Vietnamese adverbs express time (such as 'đã'$_{past}$, 'đang'$_{continuous}$), degree (such as 'rất'$_{very}$, 'hơi'$_{rather}$), and negation (such as 'không'$_{not}$). Therefore the number of adverbs in Vietnamese is much smaller than that in English. For that reason, there are only three parts of speech in Vietnamese WordNet including noun, verb, and adjective. Semantic relations in Vietnamese WordNet are similar to those in Princeton WordNet except a number of relations such as derivationally related form, participle of verb, etc.

The remaining part of this paper is organized as follows: Section 2 gives a review of several existing wordnets. Section 3 introduces our method to construct Vietnamese WordNet. Section 4 presents statistics and analyses of the wordnet being constructed. Section 5 gives a number of conclusions and future works.

2 Existing Wordnets

2.1 Princeton's WordNet

Since 1978, George Miller (Fellbaum, 1998) had researched and developed a database of words and semantic relations between words. This database was called wordnet and was considered a model of mental lexicon. Conceivably, wordnet is a large discrete graph in which nodes are synonym sets (synsets) and edges are semantic relations of synsets. A synset is a collection of synonym words of the same part of speech in which each word can be replaced by one of the others in certain contexts. For example, *car, auto, automobile, machine, motorcar* form a synset. This synset has a hyponymy relation with the synset *vehicle* because a *car* is a kind of *vehicle*.

2.2 EuroWordNet

EuroWordNet (Vossen, 2002) is a multilingual lexical database of nine European languages. Each language has its own wordnet. These component wordnets are linked via Princeton's WordNet version 1.5. More specifically, their synsets are linked to Princeton's WordNet's synsets which are equivalent or closest in meaning. EuroWordNet accepts different levels of lexicalization. For example, Princeton's WordNet contains both lexicalized and unlexicalized synsets, while Dutch WordNet contains only lexicalized ones. Component wordnets have been built by exploiting available resources such as monolingual dictionaries, bilingual dictionaries, and the Princeton's WordNet.

2.3 Asian WordNet

This project (Virach et al., 2009) aims to create wordnets for Asian languages such as Thai, Japanese, Korean, etc. Currently, there are data of 13 languages in Asian WordNet. The authors adopted a semi-automatic approach to translate Princeton's WordNet's synsets into Asian languages using bilingual dictionaries. The authors also built an online tool for editing and visualizing contents of the wordnet. By using this tool, many people can easily participate in the task of translation. They can also mod-

ify translations and can vote for the best one. In terms of wordnet design, Asian WordNet is a special case of EuroWordNet because it was built by translation approach. The major limitation of Asian WordNet is that it lacks specific concepts of Asian languages.

2.4 Laconec

This is a semantic-based multilingual dictionary available on the Internet[1]. According to the information on the website: This dictionary has been developed since 2007. The goal of Laconec is to provide multilingual lexical knowledge word lookup based on semantics. The core of the system is the large scale Princeton's Wordnet-like monolingual dictionaries linked to each other. This dictionary acknowledges Dr. Francis Bond's works (Bond and Paik, 2012) and four wordnets including English, Thai, Japanese, and Finnish.

3 A Method to Construct Vietnamese WordNet

3.1 Two Phases in Constructing Vietnamese WordNet

We construct Vietnamese Wordnet through two phases (Figure 1). In phase 1 (steps 1 to 3), we focus on translating a part of Princeton's WordNet into Vietnamese. In phase 2 (steps 4 and 5), we make use of Vietnamese resources to create the wordnet. Contents and requirements of these phases are different and separated.

The major work of phase 1 is translating a part of English Wordnet into Vietnamese. Thus, we firstly need to determine a list of English synsets to translate. Because of the significantly smaller size of our target Vietnamese wordnet, we choose to translate only a part of Princeton's WordNet. Our criteria for selecting English synsets include: (1) the lexicalization possibility in Vietnamese; (2) the connectivity of the selected part; (3) the inclusion of common base concepts.

Since the set of lexicalized concepts in English and the set of lexicalized concepts in Vietnamese are different, the data of wordnet built in phase 1 does not contain Vietnamese specific words such as *'âm dương'*$_{yin\ and\ yang}$, *'trắng ởn'*$_{white}$, *'làng xã'*$_{village}$, etc. or words relating to history, society and culture of Vietnamese such as *'truyện Kiều'*$_{a\ famous\ story\ in\ Vietnam}$, *'bánh chưng'*$_{a\ kind\ of\ cake}$, etc. Therefore in phase 2, we select coordinated compound words, reduplicative words, and subordinated compound words to add to the Vietnamese WordNet. We choose words from a popular Vietnamese dictionary, made by the Vietnam Lexicography Center (Vietlex).

3.2 Guideline Development

Editing data for wordnet is not an easy task, guideline documents are required to ensure the correctness and the consistency of data. In a wordnet, words are linked by semantic relations, therefore in the guideline document we focus on describing how to identify semantic relations especially synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy, and troponymy. We created diagnostic tests to verify relations between synsets. For instance, synonymy relation is identified on the basis of the possibility of a word being replaced by another in a specific context. This can be verified by the possibility of being mutually substitutable in sentence 'X is a *Noun*$_1$ therefore X is a *Noun*$_2$'. In addition to the tests there are a number of principles which can be used for encoding the relations, for example the Economy principle and the Compatibility principle (Fellbaum, 1998). Besides, we give guidelines as to handling Vietnamese specific linguistic and cultural characteristics. Last but not least, the guideline document contains instructions as to how to give definitions and examples, how to exploit resources such as existing dictionaries, and spelling rules.

3.3 Treatment of Vietnamese Specific Words

With regard to their structure, Vietnamese words can be divided into a number of types including single-syllable words, coordinated compound words, subordinated compound words, reduplicative words, and accidental compound words. The syllables which are not single words are bound morphemes[2], which can only be used as part of a word but not as a word on its own. The coordinated compound words (CCWs), specific to Vietnamese, are

[1] www.laconec.com

[2] They may have a meaning ('trường'$_{long}$, 'hàn'$_{cold}$) or not ('lẽo', 'nhánh')

words in which their parts– each part can be a word, single or compound words– are parallel in the sense that their meanings are similar and their order can be reversed. The meaning of a coordinated compound is often more abstract than the meanings of its parts. The proportion of this kind of words is about 10% of the number of compound words according to the statistics in the Vietlex dictionary. Reduplicative words (RWs) such as 'đất đai'$_{land}$, 'làm lụng'$_{work}$ are compounds whose parts have a phonetic relationship. This kind of words is specific to Vietnamese despite its small proportion. The identification of reduplicative words is normally deterministic and not ambiguous. Accidental compounds are non-syntactic compounds containing at least two meaningless syllables such as 'đười ươi'$_{orangutan}$, 'bù nhìn'$_{puppet}$. Subordinated compound words (SCWs) are the most problematic. A SCW can be considered as having two parts, a head and a modifier. Normally, the head goes first and then the modifiers. SCWs make up the largest proportion in the Vietnamese dictionary. Generally, discrimination between SCW and phrase is problematic because SCW's (syntactic) structure is similar to that of a phrase. This is a classical but persistent problem in Vietnamese linguistics.

The following are a number of synsets from Princeton's WordNet that were translated into Vietnamese. Words added to the synsets in phase 2 are in italics.

- (n) tree (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown): cây; *cây cối, cây cỏ (CCW)*
- (v) laugh, express joy, express mirth (produce laughter): cười; *cười đùa (CCW), cười cợt (RW)*
- (adj) strong (having strength or power greater than average or expected): mạnh, mạnh mẽ, khoẻ; *khoẻ mạnh (CCW), khoẻ khoắn (RW)*
- (adj) black (being of the achromatic color of maximum darkness): đen, màu đen, có màu đen, mun, thâm, ô, ác, mực, huyền; *đen sì, đen sì sì, đen thui, đen trũi, đen nhẻm (SCW), đen đen (RW)*

POS	Synsets	Words	Word-synset pairs
Noun	17,084	32,122	37,452
Verb	9,483	21,180	32,273
Adjective	5,846	13,590	18,289
Total	32,413	66,892	88,014

Table 2: Vietnamese wordnet statistics.

3.4 Treatment of Vietnamese Proper Names

Proper names (place name, personal name, work name, etc.) represent important information about Vietnamese history, society, culture and thought. Vietnamese WordNet contains about 4,000 such linguistic expressions. Besides, Vietnamese WordNet has to also include worldwide famous names such as Amazon, Yangtze, Bacon, Nehru, etc. However, such names occupy only a small proportion in comparison with Vietnamese ones. The following are a few examples.

- 'nhân vật'$_{character}$ > 'nhân vật kịch'$_{drama\ character}$ > 'nhân vật chèo'$_{Vietnamese\ traditional\ operetta's/character}$ > 'hề'$_{clown}$/ 'mẹ Đốp'$_{mother\ Dop}$
- 'làng'$_{village}$ > 'Đường Lâm'$_{Duong\ Lam}$/ 'Mộ Trạch'$_{Mo\ Trach}$/ 'Hành Thiện'$_{Hanh\ Thien}$
- 'dân tộc'$_{ethnic\ group}$ > 'Kinh'$_{Kinh}$/ 'Tày'$_{Tay}$/ 'Thái'$_{Thai}$
- 'bánh'$_{cake}$ > 'bánh chưng'$_{square\ glutinous\ rice\ cake}$/ 'bánh trôi'$_{floating\ cake}$/ 'bánh rán'$_{fried\ cake}$
- 'hồ'$_{lake}$ > 'Hồ Gươm'$_{Sword\ Lake}$/ 'Hồ Tây'$_{West\ Lake}$

4 Empirical Analyses of Vietnamese WordNet

4.1 Vietnamese WordNet Statistics

Table 2 shows basic statistics of Vietnamese WordNet. Nouns take the largest proportion while the number of verbs and adjectives is smaller. Like Princeton's WordNet, Vietnamese WordNet can be considered as including three subwordnets corresponding to different parts of speech. The subwordnet of nouns has a unique root 'thực thể'$_{entity}$. The

subwordnet of verbs has 255 roots. The subwordnet of adjectives has 2,201 clusters.

As shown in Table 4, there are 61,509 semantic relations, in which 34,161 between noun synsets, 18,465 between verb synsets, and 8,883 between adjective synsets. The most frequent semantic relations include hypernymy-hyponymy, synonymy, antonymy, and similar-to. Vietnamese WordNet inherits the WordNet Domains Hierarchy (Bentivogli et al., 2004) including 164 domain labels organized as a tree structure.

4.2 Synset Size Distributions

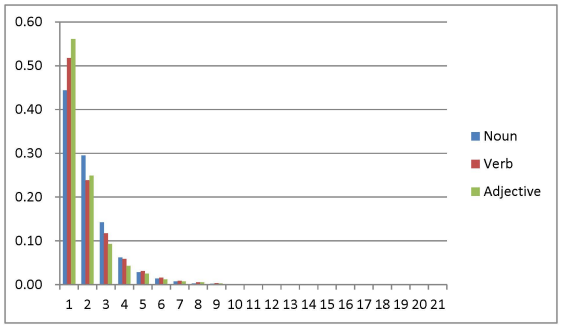


Figure 2: Synset size distributions.

Figure 2 shows synset size distributions of nouns, verbs, and adjectives. The horizontal axis represents synset size and the vertical axis represents the proportion. These distributions are not significantly different. On average each synset contains 2.42 words. When synset size increases, the corresponding proportion decreases.

4.3 Phase 2 Contributions

Table 3 represents word statistics in phase 2 of Vietnamese WordNet construction. The number of words added in this phase is 9,615. These words are specific to Vietnamese and different from words in phase 1. Besides, we also add nearly 4,000 proper nouns to Vietnamese WordNet. These nouns reflex Vietnamese anthronyms, toponyms (rivers, mountains, etc.), social events, etc.

POS	CCWs	RWs	SCWs
Noun	976	186	2,068
Verb	2,347	772	138
Adjective	1,406	1,217	505
Total	4,729	2,175	2,711

Table 3: Vietnamese WordNet statistics: phase 2.

Relation	Noun	Verb	Adjective
Antonymy	572	667	2,658
Hypernymy	15,240	8,661	
Hyponymy	15,240	8,661	
Holonymy	1,362		
Meronymy	1,362		
Entailment		307	
Cause		169	
Attribute	385		385
Similar to			5,840
Total	34,161	18,465	8,883
		61,509	

Table 4: Semantic relation statistics.

5 Conclusions

The paper has presented the most up-to-date results of the process of constructing Vietnamese WordNet. Since this project is coming to final stage, there can be slight differences between current version and the final version. We continue to revise data by lexical phenomenon or following statistical methods. Vietnamese WordNet will be published online and available for research and development purposes.

Acknowledgments

This paper has been supported by the national project number KC.01.20/11-15.

References

Luisa Bentivogli, Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of Workshop on Multilingual Linguistic Resources, COLING 2004*.

Francis Bond and Kyonghee Paik. 2012. A Survey of WordNets and Their Licenses. *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Dhanon Leenoi, Thepchai Supnithi, Wirote Aroonmanakun. 2008. Building a Gold Standard for Thai WordNet. *Proceedings of IALP*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, Chumpol Mokarat, and Hitoshi Isahara. 2009. Review on Development of Asian WordNet. *JAPIO 2009 Year Book*, Japan Patent Information Organization, Tokyo, Japan.

Piek Vossen. 2002. Wordnet, EuroWordnet and Global Wordnet. *Pub. linguistiques*, 2002/1 - Vol. VII, pages 27-38.

Piek Vossen. 2002. EuroWordNet General Document. *Online document*.

Extending the WN-Toolkit: dealing with polysemous words in the dictionary-based strategy

Antoni Oliver
Universitat Oberta de Catalunya (UOC)
Barcelona - Catalonia - (Spain)
aoliverg@uoc.edu

Abstract

In this paper we present an extension of the dictionary-based strategy for wordnet construction implemented in the WN-Toolkit. This strategy allows the extraction of information for polysemous English words if definitions and/or semantic relations are present in the dictionary. The WN-Toolkit is a freely available set of programs for the creation and expansion of wordnets using dictionary-based and parallel-corpus based strategies. In previous versions of the toolkit the dictionary-based strategy was only used for translating monosemous English variants. In the experiments we have used Omegawiki and Wiktionary and we present automatic evaluation results for 24 languages that have wordnets in the Open Multilingual Wordnet project. We have used these existing versions of the wordnet to perform an automatic evaluation.

1 Introduction

1.1 The WN-Toolkit

The WN-Toolkit[1] (Oliver, 2014) is a set of programs developed in Python for the automatic creation of wordnets following the expand model (Vossen, 1998), that is, by translation of the variants (words) associated with the Princeton WordNet synsets. The toolkit also provides some free language resources. These resources are preprocessed so they can be easily used with the toolkit.

The WN-Toolkit implements the following strategies for wordnet creation:

- Dictionary based methodology: This strategy uses bilingual dictionaries to translate the English variants associated with each synset. In previous versions of the toolkit this direct translation using dictionaries could be performed only on monosemic English, that is, variants associated to a single synset. About 82% of the English variants in the Princeton WordNet 3.0 are monosemic but frequent words tend to be polysemic. With the extension of the toolkit presented in this paper we are able to deal with polysemic English variants.

- Babelnet based strategies: BabelNet (Navigli and Ponzetto, 2010) is a semantic network and a multilingual encyclopedic dictionary with lexicographic and encyclopedic coverage of terms. In this methodology we simply extract the data from the BabelNet file to get the target wordnet. This strategy can only be applied to old versions of Babelnet, as new versions have a use restriction not allowing the creation of wordnets from its data.

- Parallel corpus based methodologies: In order to extract wordnets from a parallel corpus we need this parallel corpus to be semantically tagged with Princeton WordNet synsets in the English part. As these corpora are not easily available, we use two strategies for the automatic construction of the required corpora:

  - By machine translation of sense-tagged corpora.
  - By automatic sense-tagging of English-target language parallel corpora.

The WN-Toolkit also provides some resources, as dictionaries and preprocessed bilingual corpora.

[1] The WN-Toolkit can be freely downloaded from http://sourceforge.net/projects/wn-toolkit/

Language	Code	Synsets	Words	Senses	Core
Albanian	sqi	4,676	5,990	9,602	31%
Arabic	arb	10,165	14,595	21,751	48%
Basque	eus	29,413	26,240	48,934	71%
Bulgarian	bul	4,999	6,783	9,056	100%
Catalan	cat	45,826	46,531	70,622	81%
Chinese	cmn	42,312	61,533	79,809	100%
Croatian	hrv	23,122	29,010	47,906	100%
Danish	dan	4,476	4,468	5,859	81%
Finnish	fin	116,763	129,839	189,227	100%
French	fra	59,091	55,373	102,671	92%
Galician	glg	19,312	23,124	27,138	36%
Greek	ell	18,049	18,227	24,106	57%
Hebrew	heb	5,448	5,325	6,872	27%
Indonesian	ind	38,085	36,954	106,688	94%
Italian	ita	35,001	41,855	63,133	83%
Japanese	jpn	57,184	91,964	158,069	95%
Norwegian N.	nno	3,671	3,387	4,762	66%
Norwegian B.	nob	4,455	4,186	5,586	81%
Polish	pol	36,054	61,393	88,889	66%
Portuguese	por	43,895	54,071	74,012	84%
Slovene	slv	42,583	40,233	70,947	86%
Spanish	spa	38,512	36,681	57,764	76%
Swedish	swe	6,796	5,824	6,904	99%
Thai	tha	73,350	82,504	95,517	81%

Table 1: Statistics for the wordnets in OMW

1.2 The Open Multilingual Wordnet project

The Open Multilingual Wordnet[2] (OMW) (Bond and Paik, 2012) provides free access to several wordnets in a common format. We have performed experiments for 24 languages out of the 28 available wordnets. In table 1 we can observe some statistics about the wordnets for these languages. These wordnets have been used to perform an automatic evaluation of the results.

1.3 Omegawiki

Omegawiki[3] is a free collaborative dictionary that can be accessed through the Internet as well as downloaded as a relational database. The downloads are performed in MySQL dumps so it's easy to set up a MySQL database to have a local copy of Omegawiki. For our experiments we have downloaded all the sql dumps corresponding to the lexical data and we have created our own copy of Omegawiki. From this database we have extracted all the required data and we have filled up a new MySQL database according to the layout explained in section 2.1.

In table 2 we can observe the number of English-target language entries for Omegawiki for the languages in our experiments.

Omegawiki uses a complex set of semantic relations between its entries. It seems to be a great degree of freedom for the users to create new relations. A total number of 77 relations are found in the English Omegawiki, but only 22 of them has at least 50 occurrences. These relations can be observed in table 3.

We tried to relate the name of the relations in Omegawiki with standard relation names used in WordNet and Wiktionary (hypernym, hyponym, holonym, meronym, antonym and synonym). As holonym, meronym and antonym are already used in Omegawiki, we will try to find out the name used for hypernym, hyponym and synonym looking at examples of these relations in Wiktionary and observing if some of these examples are also present in Omegawiki. In this way we could establish the correspondence between relation codes and names in Omegawiki and standard relations names. An special case are synonyms, that are expressed as translations into the same language. In table 4 we can observe these correspondences.

In table 5 the number of definition and semantic relations in Omegawiki and Wiktionary can be observed.

[2] http://compling.hss.ntu.edu.sg/omw/

[3] http://www.omegawiki.org/

Language	Code	Omegawiki	Wiktionary
Albanian	sqi	417	4,431
Arabic	arb	3,293	17,157
Basque	eus	5,293	3,834
Bulgarian	bul	5,851	24,983
Catalan	cat	4,001	24,625
Chinese	cmn	3,368	70,553
Croatian	hrv	1,687	34,485*
Danish	dan	7,177	18,625
Finish	fin	9,654	94,193
French	fra	26,492	70,178
Galician	glg	1,636	7,832
Greek	ell	6,193	30,161
Hebrew	heb	3,447	12,452
Indonesian	ind	2,219	6,669
Italian	ita	25,083	51,098
Japanese	jpn	6,674	45,135
Norwegian N.	nno	787	5,842
Norwegian B.	nob	6,399	6,395
Polish	pol	8,280	32,486
Portuguese	por	11,858	58,925
Slovene	slv	5,102	9,036
Spanish	spa	36,139	63,512
Swedish	swe	10,271	45,016
Thai	tha	1,614	6,339

Table 2: Number of English-target language entries for each language

1.4 Wiktionary

Wiktionary[4] is also a free collaborative dictionary. This project is related with the Wikipedia and it is developed in a Mediawiki format. It can be accessed through the Internet and it can be also downloaded. The download format is an XML that includes sections in mediawiki format and for this reason it is difficult to parse.

The project Dbnary[5] (Sérasset, 2012) parses the Wiktionary content as soon as a new dump is available and provides this content in a easy to parse format.

In our first experiments we have used our own parser to extract the information for the English Wiktionary dumps but we missed a lot of information and it was very difficult and time consuming to correct the errors and expand the parser, so we started to use the results of the Dbanry project. We have used the files from Dbnary and we have stored all this information in our own database.

In table 2 we can observe the number of English-target language entries for Wiktionary for the languages in our experiments.

[4]urlhttps://www.wiktionary.org/

[5]urlhttp://kaiko.getalp.org/about-dbnary/

relation	freq.
is part of theme	16,158
parent	11,980
child	11,776
broader terms	7,299
narrower terms	5,639
is spoken in	4,692
related terms	3,717
borders on	797
is written in	633
antonym	328
official language	226
capital	209
country	192
wordt gevolgd door	178
currency	165
holonym	183
demonym	122
flows through	110
dialectal variant	78
meronym	73
flows into	68
is practiced by a	61

Table 3: Relations with at least 50 occurrences in English Omegawiki

Code OW	Relation OW	Relation S.
4	broader terms	hypernym
5	narrower terms	hyponyms
7574	antonym	antonym
375074	meronym	meronym
375078	holonym	holonym
-	translation into same language	synonym

Table 4: Conversion between Omegawiki (OW) relation codes and names and Standard (S.) relation names

2 Experimental results

2.1 MySQL database layout

We have stored all the data from Omegawiki and Wiktionary in our own MySQL database. This allows us to develop an algorithm for the construction of wordnets using this database and working in a independent way from the resource. This also allows us to add information from other sources and easily select one or more sources for the experiments. The database has the following 5 tables:

- *entry*: in this table the English word or expression, part of speech and source, along with an unique entry id are stored. The unique entry id allows us to select the information from the rest of the tables for a given entry.
- *translations*: in this table the translations for

	Omegawiki	Wiktionary
definitions	37,233	608,358
relations total	90,039	28,123
hypernyms	3,029	1,193
hyponyms	2,331	1,114
holonyms	121	92
meronyms	47	92
antonyms	171	0
synonyms	50,265	26,708

Table 5: Number of English definitions and semantic relations in the dictionaries

the target languages are stored, along with the language code and the entry id.

- *definition*: in this table the English definitions for each entry are stored.
- *tagged definition*: in order to avoid tagging each definition each time we perform an experiment we can use this table to store the tagged definition for each definition, along with the used tagger and the entry id.
- *relations*: in this table the related English words for each entry are stored along with the relation name and the entry id.

Please, note that most of the information we stored in the database is for the English language (except the translations). This is due to the fact that we plan to translate English variants from the Princeton WordNet in order to create the target wordnets.

Some indexes are create to speed up the algorithm. Most of the tables are converted into in-memory tables in order to further speed up the process of wordnet creation.

2.2 Algorithm

The wordnet extraction algorithm works as follows:

- Select all entry ids and target language words for a given target language and a given resource from the table *translations*.
  - For each entry id select the English words an pos from the table *entry*.
    - For the given English word and pos we search in the Princeton WordNet for all the synsets the word belongs to.
    - If the word belongs to one synset that means that it is monosemic and the target language word can be directly related to the given synset.
    - Otherwise, that means that it is polysemic and the disambiguation procedure is started:
      - Select all related words (hyponyms, hypernyms, holonyms, meronyms, antonyms and synonyms) along with the relation names from the table *relations*.
      - Select all the related words for all the synsets from the Princeton WordNet.
      - For each synset we count the coincident related words for each relation. A specific weight is given for each relation type.
      - Select the tagged definition from the table *tagged_definition* both for the definition coming from the dictionary as well as the Princeton WordNet definition. For each synset the coincident open class lemmata are counted and and specific weight is applied.
      - The synset with the higher score of weighted coincident relations and common open class lemmata in the definitions is attached to the target language word.

As we can see in the algorithm a set of weights has to be defined: a weight for each coincident type of relation and a weight for the number of coincident open class lemmata in the definitions. In our experiments a value of 5 has been used for all relations and a weight of 1 for coincident open class lemmata in the definitions. In section 2.4.4 a procedure for the optimization of the weights is presented.

2.3 Automatic evaluation procedure

We have used the existing wordnets in Open Multilingual Wordnet (OMW) for the 24 languages to perform an automatic evaluation. The evaluation procedure is as follows:

- Our algorithm gives us a set of synset-target language variants (SV) pairs.
  - If the extracted SV pair is also in the reference OMW, the result is evaluated as correct.
  - If the extracted SV pair is not in the reference OMW and there is other variants for the given synset in the reference OMW, the result is evaluated as incorrect.

– If the extracted SV pair is not in the reference OMW and there is no variants for the given synset in the reference OMW, the result is not evaluated.

The precision values obtained this way tend to be lower than the real values because the fact that some SV pair is not in the reference wordnet, but other variants for the same synset exist, doesn't really mean that the extracted SV pair is incorrect. May be is a valid variant for the synset, but this variant is not present in the reference wordnet.

2.4 Results

In table 6 we can observe the number of entries evaluated as correct, as incorrect and the number of entries that could not be evaluated since there is no information in the reference wordnet. The number of non-evaluated entries can give us an idea of the number of new entries we can add to the wordnet if a manual revision is performed

Lang.	Omegawiki			Wiktionary		
	C	I	N	C	I	N
sqi	58	45	249	296	207	2,263
arb	68	902	1,507	289	2,561	6,128
eus	1,339	694	881	1,192	532	635
bul	516	256	2,688	866	1,680	10,514
cat	1,671	680	554	5,697	2,915	3,881
cmn	857	526	1,344	3,640	8,140	14,775
hrv	785	274	287	2,151	7,120	4,757
dan	535	269	3,612	964	757	774
fin	3,778	2,309	18	17,551	21,325	127
fra	7,440	5,168	1,963	16,545	9,713	5,110
glg	589	134	561	1,579	498	2,328
ell	1,041	948	1,852	2,697	2,863	9,606
heb	29	575	2,018	133	1,390	5142
ind	919	484	259	1,704	1,383	758
ita	5,627	3,814	4,471	8,671	6,375	7,836
jpn	2,871	1,306	650	9,786	8,374	3,792
nno	70	17	517	326	222	2,668
nob	480	242	3,063	394	277	2,844
pol	2,348	1,310	1,434	6,133	4,402	5,817
por	4,832	1,810	474	12,892	7,741	5,410
slv	1,663	888	445	2,566	1,790	638
spa	4,088	4,567	8,525	6,179	7,155	15,274
swe	1,104	699	4,640	2,238	2,437	16,007
tha	733	464	85	1,639	1,632	330

Table 6: Figures of correct (C), incorrect (I) and nonevaluated (N) entries

In tables 7 and 8 the evaluation results are presented. For all the languages the number of extracted entries (synset-variant pairs) and the precision values (calculated in an automatic way) are presented, for several cases:

- *All no dis.*: All results, no disambiguation procedure performed.
- *All dis.*: All results, disambiguation procedure performed.
- *Non ambiguous*: Results corresponding to monosemous English variants (non ambiguous).
- *Amb. no dis.*: Results corresponding to polysemous English variants (ambiguous), no disambiguation procedure performed.
- *Amb. dis.*: Results corresponding to polysemous English variants (ambiguous), disambiguation procedure performed.

The comparison between the values with and without disambiguation procedure is interesting to observe the effectiveness of the disambiguation procedure. The results corresponding to monosemous English variants are interesting because they are the same we would obtain with the old version of the WN-Toolkit, that was not able to perform any disambiguation and was used only for monosemous English variants. They are also interested to be compared with the disambiguated results, to see if the figures are comparable.

In the tables some very low values of precision are present for languages as Arabic and Hebrew. They are due to languages specific features (as for example the writing of vowel signs than can be present or not both in the extracted variants and in the reference wordnet) that we were not able to cope with due to the our lack of knowledge of these languages. Other language-specific issues of the results will be explained in the section 2.4.3.

2.4.1 Results for Omegawiki

If we take a look at table 7 we can observe than the best overall results are obtained for Galician (precision of 81.47%) followed by Norwegian (Norsk) (precision of 80.46%). We must keep in mind that these values of precision are automatically calculated and the real values might be higher. If we concentrate on Galician we can observe than the precision of all results with no disambiguation procedure is 65.34%, so the disambiguation procedure improves the precision in 16.13 points. The precision for variants coming from monosemous English words is 83.43%, about 3 points higher than the overall values. If we concentrate on the variants coming from polysemous English words, we can see that the precision with no disambiguation is 51.16%, and it rises up to 76.85% (25.69 points) using the disambiguation algorithm.

2.4.2 Results for Wiktionary

If we now take a look at the results for Wiktionary at table 8 we can see that again the best results are

Lang.	All no dis.		All dis.		Non ambiguous		Amb. no dis.		Amb. dis.	
	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision
sqi	1,466	40.18	353	56.31	135	58.33	1,332	38.76	219	55.7
arb	9,191	4.19	2,478	7.01	1,237	9.83	7,955	3.34	1,242	4.97
eus	7,934	48.03	2,915	65.86	1,708	64.99	6,227	43.77	1,208	66.8
bul	29,183	36.66	3,461	66.84	1,862	63.09	27,322	35.66	1,600	68.46
cat	8,531	53.57	2,906	71.08	1,673	69.76	6,859	48.74	1,234	72.81
cmn-Hans	11,924	26.88	2,728	61.97	1,269	68.88	10,656	22.39	1,460	57.71
hrv	4,180	51.59	1,347	74.13	701	78.89	3,480	43.18	647	68.4
dan	11,935	48.38	4,417	66.54	2,523	58.3	9,413	46.8	1,895	70.73
fin	20,134	36.02	6,106	62.07	3,342	64.24	16,793	30.4	2,765	59.44
fra	53,499	48.3	14,572	59.01	7,850	57.48	45,650	46.75	6,723	60.74
glg	3,483	65.34	1,285	81.47	753	83.43	2,731	51.16	533	76.85
ell	12,838	34.61	3,842	52.34	2,009	52.29	10,830	31.09	1,834	52.38
heb	9,199	2.56	2,623	4.8	1,347	4.37	7,853	2.22	1,277	5.11
ind	5,589	48.06	1,663	65.5	852	64.68	4,738	44.86	812	66.33
ita	85,324	32.05	13,913	59.6	6,614	59.82	78,711	29.41	7,300	59.41
jpn	14,994	40.48	4,828	68.73	2,694	71.59	12,301	33.16	2,135	65.32
nno	1,379	59.89	605	80.46	376	80.0	1,004	55.92	230	80.7
nob	10,555	47.0	3,786	66.48	2,196	58.61	8,360	45.21	1,591	70.5
pol	16,417	41.99	5,093	64.19	2,876	64.67	13,542	35.37	2,218	63.55
por	26,301	52.52	7,117	72.75	3,761	69.16	22,541	48.36	3,357	77.10
slv	9,136	49.68	2,997	65.19	1,607	61.59	7,530	47.1	1,391	69.21
spa	68,884	31.65	17,181	47.23	8,874	41.86	60,011	30.55	8,308	51.41
swe	21,626	40.05	6,444	61.23	3,535	63.17	18,092	35.67	2,910	59.81
tha	4,065	33.08	1,283	61.24	677	59.87	3,389	27.12	607	62.72

Table 7: Results for Omegawiki

obtained for Galician (a precision of 76.02% for all the results with disambiguation). The rest of figures for this languages follows the same pattern as for Omegawiki. One important fact is that with Wiktionary we are obtaining much more results (4,406 synset-variant pairs) than with Omegawiki (1,285) as Wiktionary is a much bigger resource as can be observed in table 2

2.4.3 Comments on the results

The precision values for the experiments are very different for each languages. It can be due to several reasons, for example:

- The quality of the dictionary (Omegawiki and Wiktionary) for each language can be different, as they are collaborative dictionaries. Not only the size of the resource is important, but also the precision of the translations.
- The quality and completeness of the reference wordnet in OMW. Here again not only the size (number of synstet-variant pairs) but also the number of possible variants for the same synset are very important.

There are a lot of language-specific issues in the dictionaries and reference wordnets that must be taken into account. We already mentioned the writing of vowel signs in Arabic and Hebrew, that we could not cope with due to the lack of knowledge of these languages.

For example, if we observe the results for Bulgarian, we can see that precision for Omegawiki (66.84%) is much higher than precision for Wiktionary (34.01%). The main reason is that in Wiktionary most entries are marked with accents in vowels to express the stress (for example алкохол in Omegawiki but алкохóл in Wiktionary). This marks are not used in standard writing and so they are not used in the reference OMW wordnet. To use the Wiktionary results a simple script converting the accented characters to unaccented can be used.

For Croatian we face a double problem. Both in Omegawiki and Wiktionary some entries (but not all) are using the diacritics on vowels to express stress and intonation, but these symbols are not used in the reference OMW wordnet as they are not used in standard writing. This can also be solved with the use of a simple script. On the other hand, Wiktionary is not using a language code for Croatian (hrv) but one for Serbo-Croatian or Croatian-Bosnian-Serbian macrolanguage (hbs). Entries for this code can be Croatian words written in latin but also Serbian words written in cyrillic or latin. As in the Croatian reference OMW wordnet there are only standard Croatian

Lang.	All no dis.		All dis.		Non ambiguous		Amb. no dis.		Amb. dis.	
	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision	Entries	Precision
sqi	11,510	43.02	2,767	58.85	1,251	59.03	10,260	41.91	1,517	58.77
arb	37,540	6.75	8,980	10.14	4,431	12.3	33,110	6.21	4,550	8.79
eus	9,359	50.7	2,360	69.14	1,244	69.89	8,116	47.41	1,117	68.43
bul	59,664	25.23	13,061	34.01	5,690	34.95	53,975	24.53	7,372	33.63
cat	53,737	52.1	12,494	66.15	6,597	68.86	47,141	49.5	5,898	63.22
cmn	102,130	18.98	26,519	31.47	30	36.36	88,589	16.79	14	25.0
hrv	62,765	17.4	14,029	23.2	6,399	25.57	56,367	16.17	7,631	20.99
dan	43,052	39.95	9,469	56.01	4,866	62.01	38,187	38.4	4,604	53.65
fin	174,743	26.22	39,004	45.15	19,958	54.95	154,786	22.51	19,047	34.88
fra	119,160	53.91	31,369	63.01	17,802	66.24	101,359	51.67	13,568	58.51
glg	17,745	59.92	4,406	76.02	2,261	77.95	15,485	52.99	2,146	72.66
ell	67,014	32.3	15,168	48.51	7,408	55.4	59,607	29.85	7,761	43.35
heb	32,136	4.97	6,666	8.73	3,198	10.31	28,939	4.18	3,469	7.33
ind	17,341	41.1	3,846	55.2	1,799	54.55	15,543	39.56	2,048	55.74
ita	95,540	39.5	22,883	57.63	12,093	64.59	83,448	35.39	10,791	50.04
jpn	89,706	31.92	21,954	53.89	11,423	63.19	78,284	27.08	10,532	43.7
nno	11,670	47.37	3,217	59.49	1,751	64.94	9,920	45.66	1,467	56.95
nob	13,012	47.01	3,516	58.72	1,855	63.13	11,158	45.42	1,662	56.61
pol	69,365	36.29	16,353	58.22	8,398	65.55	60,968	30.91	7,956	49.82
por	120,069	46.11	26,044	62.48	13,486	65.64	106,584	42.82	12,559	58.84
slv	25,391	47.17	4,995	58.91	2,248	59.59	23,144	45.91	2,748	58.33
spa	114,452	38.68	28,609	46.34	15,517	46.46	98,936	37.78	13,093	46.22
swe	93,448	32.08	20,683	47.87	10,637	57.12	82,812	29.12	10,047	40.69
tha	15,660	27.77	3,602	50.11	1,784	53.62	13,877	23.85	1,819	46.56

Table 8: Results for Wiktionary

words, the values of precision for Wiktionary are lower.

So it is important that a native speaker of each language revise the obtained results in order to detect these issues and try to solve them in an automatic way.

2.4.4 Optimization of the weights

In the experiments we have used a fixed value for the weight for the different relations and common lemmata in the definitions. The extraction algorithm can give also a file with information about all the parameters. Here we can see an example for Catalan:

```
pluja àcida MONO 14517629-n
àcid POLY 14607521-n/2:1:0:0:0:0:0;
02675657-n/0:0:0:0:0:0:0
```

The first line tell us than *pluja àcida* comes from a monosemic English word having the synset 14517629-n. In the second line we can learn that *àcid* comes from a polysemic English word that is a valid variant for the synsets 14607521-n and 02675657-n. For the first synset we have two common lemmata in the definitions and one common hyponym, whereas for the second synset we don't have any information in common.

This file allow us to experiment with different weights in order to learn the best combination. In table 9 we can observe the values of overall precision for different combinations of the parameters (we have assigned one weight to the coincident lemmata in the definition and another weight for the coincident related words (the same weight for all types of relations). The values in the table are for Catalan and for Omegawiki and Wiktionary.

Def.	Rel.	Omegawiki	Wiktionary
0	1	70.01	**68.90**
1	0	70.95	66.15
1	1	71.03	66.14
1	5	**71.08**	66.15
5	1	70.98	66.14
1	10	71.06	65.90
10	1	70.98	66.14

Table 9: Precision for different combinations of the weights for Catalan

As we can observe, the best combination for Omegawiki is 1 for definition and 5 for relations. This is the combination we have used in our experiments. For Wiktionary the best combination is 0-1, that is, using no definitions and using only relations.

It would be worth to do a better analysis and to try to use some machine learning technique to find the best combination for each languages and

3 Conclusions and future work

In this paper we have presented an extension of the WN-Toolkit that allows to use the dictionary-based technique for wordnet creation for English polysemous variants, provided that the dictionary has definitions and/or relations. The algorithm have been applied to 24 languages having wordnets available in the Open Multilingual Wordnet. We have calculated values of precision in an automatic way using as reference the existing wordnets. For the experiments we have used two freely available dictionaries: Omegawiki and Wiktionary. The results demonstrate that the algorithm performs well in the task of selecting the correct translation for polysemous words.

As a future work we plan to use some machine learning technique to try to find the best combination of parameters for each language and resource. The algorithm we've presented uses a very simple strategy to find the most similar definition by comparing the number of coincident open class words. We plan to experiment with more complex strategies, as for example using a word2vec approach or similar techniques (Bjerva et al., 2014). We also plan to use other dictionaries or encyclopedia as Apertium transfer dictionaries, Wikispecies, Wikipedia, Geodata, as well as proprietary dictionaries under agreement with the copyright holders. If the dictionary has definitions and/or semantic relations the proposed disambiguation algorithm can be applied. If not, only target language variants corresponding to English monosemous variants can be extracted.

We also plan to run the algorithm for all languages in the resources, creating preliminary wordnet versions for languages not having freely available wordnet available. In this sense we would be happy to make agreement with universities or institutions in target language speaking countries to revise the results.

We want to compare and share the results with the Extended Open Multilingual Wordnet (Bond and Foster, 2013).

An lastly we want to pack the new algorithm into the WN-Toolkit and share the complete MySQL database created from the free resources. This database can be useful for wordnet creation experiment as well as for other lexicographical tasks.

Acknowledgments

This research has been partially carried out thanks to the Project SKATER (TIN2012-38584-C06-01 and TIN2012-38584-C06-06) supported by the Ministry of Economy and Competitiveness of the Spanish Government.

This research has been done during a research stay in the Vrije Universiteit in Amsterdam, thanks to a mobility grant from the Universitat Oberta de Catalunya. I would also thank Piek Vossen and his research group for welcoming me in Amsterdam.

References

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, Sofia, Bulgaria. 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japan. 64–71.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858704.

Antoni Oliver. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. In Heili Orav, Christianne Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, pages 7–15, Tartu, Estonia. Global Wordnet Association.

Gilles Sérasset. 2012. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web Journal-Special issue on Multilingual Linked Open Data*.

Piek Vossen. 1998. Introduction to eurowordnet. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer Netherlands.

A language-independent LESK based approach to Word Sense Disambiguation

Tommaso Petrolito
Filologia Letteratura e Linguistica, **University of Pisa**, Italy
tommasouni@gmail.com

Abstract

This paper describes a language-independent LESK based approach to Word Sense Disambiguation (WSD), involving also Vector Space Models applied to the Distributional Semantics Hypotesis. In particular this approach tries to solve some issues that come up with less-resourced languages. The approach also addresses the inadequacy of the Most Frequent Sense (MFS) heuristics to fit specific domain corpora.

1 Introduction

This language independent approach to WSD, even if in a very early stage of development, tries to solve two main problems.

1. Variable quality of glosses and examples (the solution would be to use glosses and examples for the aligned synsets in several languages, we will explain how).

2. Weakness of Most Frequent Sense heuristics for domain corpora (or even general corpora that, for some reasons, are not so similar to the corpus on which the frequencies were calculated), but also lack of synset annotated corpora for several non-English languages (the solution would involve Space Vector Models, we will explain how).

We use Wordnet (WN) resources (Miller et al., 1990; Miller, 1995; Fellbaum, 1998) (synsets glosses and examples) from a specific standpoint: preferring to avoid the usage of monolingual resources, even though the specific task does not involve cross-lingual WSD on aligned parallel corpora.

It has to be pointed out that the approach is being considered 'unsupervised': it does not rely on semantic annotation, although lemmatization and PoS-tagging are taken into account.

In most cases the quality of lexical resources is very variable, even though some languages have good resources, as of course English and, for instance, Italian with both MultiWordnet (Pianta et al., 2002) and ItalWordnet (Roventini et al., 2000).

An example is given with the *dog/câine* (first synset) glosses and examples[1] in Table 1. It is evident that the English synset has a richer gloss.

Assuming a WSD approach involving overlap counts, the English words *Canis*, *wolf*, *breeds* will be counted in; as for the Romanian words *Animal*, *mamifer*, *carnivor* (all IS-A relations), their English lemmas would be reached in any case in an Expanded Gloss implementation.

Anyway, *pază* and *vânătoare* ('guarding' and 'hunting') would be useful for the same task.

In the counterexample given in Table 2, the Romanian gloss is evidently richer than the English one, in particular using a WSD overlapping algorithm that is able to count on *asistenţă*, *socială*, *întreţinerea*, *bătrânilor* ('assistance', 'social', 'maintenance', 'elders') and so on.

In general, it can be noticed how variable the quality is for different corpora and for different synsets.

Anyway, usually English WN provides the best and richest set of examples for a given synset.

This variability in quality is observable also concerning the coverage of different WNs[2] (Bond and Foster, 2013).

[1] For a quick series of examples of this kind, just have a look on multilingual aligned synsets on the MultiWordnet Interface (Ranieri et al., 2004).
See http://multiwordnet.fbk.eu/online/multiwordnet.php

[2] See http://compling.hss.ntu.edu.sg/omw/ and http://globalwordnet.org/wordnets-in-the-world/ for an overview.

Synset	Lang	Gloss
dog,domestic_dog, Canis_familiaris/1	EN	a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds;
câine/1	RO	Animal mamifer carnivor domesticit, folosit pentru pază, vânătoare etc..

Table 1: Synset gloss comparison (EN:dog,domestic_dog,Canis_familiaris/1 – RO:câine/1)

Synset	Lang	Gloss
home,nursing_home, rest_home	EN	an institution where people are cared for;
azil	RO	Instituţie de asistenţă socială pentru întreţinerea bătrânilor, infirmilor, copiilor orfani etc.

Table 2: Synset gloss comparison (EN:home,nursing_home,rest_home – RO:azil)

Many LESK-inspired algorithms have been presented; see for instance Kilgarriff and Rosenzweig (2000a), Kilgarriff and Rosenzweig (2000b), Banerjee and Pedersen (2002) and Basile et al. (2014). Our approach is an adaptation that takes into account the issues about glosses and examples quality.

In particular we try to gain advantage from the usage of better resources available in other languages.

A first baseline attempt was tested here, due to time constraints: relying on English glosses and examples with non-English target corpora. For future work, a more complex adaptation will be attempted, trying to take advantage of glosses and examples in several languages at once.

This kind of approach leads to two main issues.

Either trying to use many glosses and examples from WNs in several languages or trying to use glosses and examples just from the Princeton Wordnet (PWN) working on a non-English corpus, the issue arises of how to compare the contexts of the target words with the glosses and examples of their candidate synsets.

The problem needs to be addressed, for instance, if the contexts of the target words are in a given language (not English) that is not compatible with an overlapping approach involving words from glosses and examples in many different languages or even just with the English ones.

Second, the widely used Most Frequent Sense (MFS) heuristics (Gale et al., 1992; Miller et al., 1994; Kohomban and Lee, 2005), easily implementable in English by choosing the first synset for the given lemma, cannot be used when working with other languages, as the synset ordering does not mirror sense frequency statistics.

Even working on English, MFS' usefulness varies accordingly to the similarity of the target corpus to SemCor (Mihalcea, 1998), concerning the topic(s) and the granularity of meanings.

Also this issue needs a proper solution and some help can come from Vector Space Models (VSMs) applied to the Distributional Hypothesis (DH) (Harris, 1951; Turney et al., 2010) of Semantics implementing Distributional Semantic Spaces (DSS).

2 Methodology

This approach is organized in two disambiguation steps.

The first (focused on quality) is based mainly on a kind of LESK adapted in the language-independent perspective discussed above and involves WN glosses and examples.

The second (focused on quantity) is based on VSMs and follows the assumption, coherent with the Distributional Hypothesis, that the neighbours of the target word in the Semantic Space are semantically related (in paradigmatic relations) with the target word.

Both these two disambiguation steps will be discussed in this section.

2.1 Language independent LESK algorithm

Our idea consists in counting the overlaps in couples of candidate-synset-bag-of-lemmas and

context-bag-of-lemmas. Then, the candidate synset for which the count is higher is chosen.

Let us assume that we use an Italian sentence, but we want to rely on English synsets glosses and examples (we will explain later why we would want to do that).

Let us take the Italian sentence:

Il cane abbaia spesso quando fa la guardia ai suoi giocattoli o al suo cibo
"The dog often barks when guarding its toys or food"

Given *cane* ("dog") as our target lemma and *n*, 'noun' as part of speech, the algorithm has to fulfill the following steps:

1. Find all the Italian synsets associated to the given lemma and part of speech that are aligned to the English WordNet.

2. For each candidate synset, build a 'bag of lemmas' by retrieving all content words found in the English gloss and example(s) and lemmatizing them.

3. For each sentence (in this case the current sentence containing *cane*), build a context bag of lemmas by taking English glosses and examples of the English synsets aligned to the Italian synsets of the words in the sentence (lemma and part of speech annotations are assumed to be there).

   To avoid a computational nightmare (and maybe also to avoid noise), only unambiguous lemmas and lemmas with a number of synsets less than an upper bound, previously defined, will be taken into account as sources of synset-glosses and synset-examples.

The synset for which the overlapping between the two bags is bigger is the chosen one.

With this approach we want to show that, theoretically, one can benefit from the semantic information available in different languages to help solve the ambiguity, even though the task doesn't start off as multilingual.

This means that theoretically we can disambiguate an Italian text using information from a WN in any language.

Now, let us suppose to use at once pairs of English bags (as explained above) and other pairs of bags of lemmas, built in the same way, but taken from WNs of other languages.

So we will have for each synset of *cane* a bag with lemmas from each language (separately).

Similarly, for the words in the sentence there will be a bag of lemmas for each language.

Let us take one 'monolingual' group at the time.

Each bag-of-lemmas pair (one from the candidate synset, one from the sentence words) will have an overlapping score. We can take into account all the scores, for example by summing them then choose the synset that has the higher total score.

Why should all this improve the results?

Let us suppose to include Romanian WN in these group of wordnets and try to disambiguate *cane* in the same Italian sentence seen above:

Il cane abbaia spesso quando fa la guardia ai suoi giocattoli o al suo cibo
"The dog often barks when guarding its toys or food"

We point out that `dog.n.01` and `cane.n.01` (respectively the English and Italian first synsets for the Italian lemma *cane*) have glosses and examples with no mention to 'hunting' or 'guarding', while the gloss of the Romanian synset (`câine.n.01`) refers to both.

The context word *guardia* ('guard') would be exploited much better by using the Romanian WN than by using the Italian one, even though the language of the text is Italian.

The same thing could happen with English (or any other language) texts about dogs in which 'guarding' and 'hunting' words are not exploited by a monolingual LESK approach.

This case is an evidence of how a multilingual approach, involving comparisons between the bags for the candidate synsets and for the context in several languages, could enhance overlapping counts and lead to a better synset selection.

We have provided an example showing that this approach can be applied also by building many sub-bags in distinct languages (and this was the full original idea): for each synset existing in English, Italian and Romanian (for example) a list containing the three monolingual bags can be built and the synset-scores can take into account the overlapping in all the languages (summing the overlapping scores together), taking advantage

from eventual better quality (or even just few lucky occurring keywords) in the glosses and examples in other languages.

2.1.1 Candidate synsets scoring

For the future, a more complex and representative scoring measure will be defined, maybe taking into account the good example provided by Basile et al. (2014) based on different weights for lemmas.

In the current version, due to time constraints, each synset gains a very simple score equal to the number of lemmas shared by the candidate-synset-bag and the context-synsets-bag (that is the union of the single synset-bags occurring in the sentence).

Only one specific customization is added to this naive scoring approach: unambiguous lemmas in the context have double weight (so their overlapping will be counted twice).

2.1.2 Results

Here we show a baseline experiment exploiting only English glosses and examples on an Italian target corpus.

If we set a configuration that takes context lemmas from words linked to a certain number of synsets (up to 6), this algorithm tags correctly the 36.17% of words in the Italian MultiSemCor (Pianta and Bentivogli, 2003; Bentivogli and Pianta, 2005; Bentivogli et al., 2005).

If we use it to remove the wrong synsets it works much better: removing, for each target word, synsets with score lower than max_score/2, 65% of words still have right synsets in the remaining set of synsets.

2.2 Paradigmatic relations algorithm

As for the second issue, concerning the Most Frequent heuristics, VSMs could provide a big help.

In particular, while the first disambiguation step focuses on the specificity of meanings observed in the specific contexts, a help from distributional quantities would focus on the frequencies of co-occurrences, thus providing a frequency based heuristics.

So, while the LESK based approach is context-dependent (so it will select different synsets for different usages of the same lemma in different contexts), the highest frequency heuristics would just help by pushing for the only one synset (always the same) that is the most frequent for the given lemma (independently whether observed in different contexts) in the corpus on which the frequencies have been measured.

A way to reproduce that kind of heuristics, even for languages with lack in synsets-annotated corpora[3](Petrolito and Bond, 2014) (even well resourced languages as Italian cannot provide such resources for corpora other than SemCor), could be implemented as a WSD algorithm involving a Distributional Semantic Space.

An example is provided by (McCarthy et al., 2004).

McCarthy et al. (2004) use a thesaurus, acquired from automatically parsed text, based on the method of Lin (1998), in order to find the predominant sense of a target word.

This thesaurus provides, through distributional similarity scores, the nearest neighbours to each target word. Then they use the WordNet similarity package (Patwardhan and Pedersen, 2003) to obtain semantic similarity measures to weight the contribution that each neighbour gives to the various senses of the target word.

Here we do something similar, but we specifically exploit paradigmatic relations.

1. In the DSS, neighbour words with high cosine similarity share the same contexts and are therefore supposed to be in paradigmatic relation.

2. Also through WordNet we can infer words in paradigmatic relation with the target word, such as hypernyms, hyponyms, cohyponyms, synonyms and antonyms.

Also this method consists of measuring the overlapping between bags of lemmas, as for the algorithm described previously.

The first bag contains the *N* (chosen arbitrarily) neighbours of the target lemma in the Semantic Space.

Both the target lemma and the neighbours are combinations lemma-PoS (the lemma and PoS information is assumed to be there).

The second bag contains lemmas taken through an exploration of the paradigmatic relations in the WN ontology for the candidate synset.

So for the candidate synset the following will be taken: lemmas, antonyms of lemmas, hyper-

[3]See `http://globalwordnet.org/wordnet-annotated-corpora/`

nyms lemmas, hyponyms lemmas, cohyponyms (hyponyms of the hypernyms) lemmas.

Actually, also in this bag (as for the one of the neighbours) instead of simple lemmas, we have combinations of lemma-PoS (in this case the information is obviously provided because we are taking lemmas from WN synsets).

Then the synset with maximum score among the overlapping values between the Semantic Space 'neighbourhood' and the paradigmatic-relations-bag is selected.

2.2.1 Candidate synsets scoring

Also in this case the score is a simple count of the intersection between the two bags.

2.2.2 Results

Also this algorithm on its own is not achieving good performances, only a 34.5% of correct annotations on the Italian SemCor.

3 Data Set

All the experiments have been done on the Italian MultiSemCor (Pianta and Bentivogli, 2003; Bentivogli and Pianta, 2005; Bentivogli et al., 2005) corpus, already sense-tagged.

SemCor is the perfect data set for this task, as it is the first case of corpus annotated with WN synsets and it is available in various languages (English, Italian, Romanian and Japanese). The Italian MultiSemCor contains 14,144 sentences and 261,283 tokens, 119,802 of which are annotated with senses.

The availability in a good number of languages makes MultiSemCor a good resource to try this language-independent approach.

Also the NTU-Multilingual Corpus (NTU-MC) (Tan and Bond, 2011) could be a perfect resource for this kind of experiments.

NTU-MC is a corpus designed to be multilingual from the start. It contains parallel text in eight languages: English, Mandarin Chinese, Japanese, Indonesian, Korean, Arabic, Vietnamese and Thai.

4 Implementation and Evaluation

The two algorithms have been implemented as Python scripts importing the NLTK (Bird, 2006) WN Interface and the Gensim (Řehůřek and Sojka, 2010) `word2vec` (Mikolov et al., 2013) library.

At first, the two algorithms were implemented separately, achieving the results discussed in Subsection 2.1.2 and Subsection 2.2.2, then the two algorithms have been implemented together in sequence.

The first algorithm has been used for a first disambiguation step excluding candidate synsets with scores lower than the 50% of the maximum, then the second algorithm has been applied taking into account only the remaining candidate synsets (provided by the first step of disambiguation) instead of considering all the possible synsets.

When the candidate with higher score in the paradigmatic relations algorithm differs from the one with higher score in the LESK based one, the two scores are normalized in a minimum-maximum 0-1, range and the candidate synset with the highest average is chosen.

The results have improved a lot achieving an encouraging result: 38.67% of the content words have been correctly annotated, with a maximum number of 6 synsets for the context words.

5 Future Work

There is reason to hope that some further attempts based on the approach described in this paper will lead to significant improvements in language-independent WSD.

A first improvement will be exploiting the disambiguated glosses at least for English, as most of the English glosses are disambiguated.

A second improvement will be the extension of the LESK based algorithm with other languages; considered that many glosses are translations of English, we should focus on Merge WNs (Dutch, Polish, etc) in particular.

To do that it will be useful to extend NLTK multilingual support: the `.definition()` and `.examples()` methods of WN synsets would be much more useful for tasks like this by exploiting a `lang` attribute.

A third improvement will be a further development of scores definitions and a complete testing of parameters like: for the first algorithm, the lower bound for the candidate synsets to be saved and passed to the second step of disambiguation and the upper bound for the number of synsets of the context words; for the second algorithm, the number of neighbours or even try to include the approach defined by McCarthy et al. (2004).

References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING*, pages 1591–1600.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Luisa Bentivogli, Emanuele Pianta, and Marcello Ranieri. 2005. Multisemcor: an english italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop*, volume 90.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439.

Zellig S Harris. 1951. Methods in structural linguistics.

Adam Kilgarriff and Joseph Rosenzweig. 2000a. English senseval: Report and results. In *LREC*.

Adam Kilgarriff and Joseph Rosenzweig. 2000b. Framework and results for english senseval. *Computers and the Humanities*, 34(1-2):15–48.

Upali S Kohomban and Wee Sun Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics.

Rada Mihalcea. 1998. Semcor semantically tagged corpus. *Unpublished manuscript*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244.

G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *In Proceedings of the Human Language Technology Workshop*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Siddharth Patwardhan and Ted Pedersen. 2003. The cpan wordnet:: similarity package.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*.

Emanuele Pianta and Luisa Bentivogli. 2003. Translation as annotation. In *Proceedings of the AI* IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy*, pages 40–48. Citeseer.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the first international conference on global WordNet*, volume 152, pages 55–63.

Marcello Ranieri, Emanuele Pianta, and Luisa Bentivogli. 2004. Browsing multilingual information with the multisemcor web interface. In *Proceedings of the LREC 2004 Satellite Workshop on The Amazing Utility of Parallel and Comparable Corpora*, pages 38–41. Citeseer.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. Italwordnet: a large semantic database for italian. In *LREC*.

Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). In *PACLIC*, pages 362–371. Citeseer.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

plWordNet in Word Sense Disambiguation

Maciej Piasecki, Paweł Kędzia and Marlena Orlińska
Wrocław University of Technology, Poland
{Maciej.Piasecki,Pawel.Kedzia,Marlena.Orlinska}@pwr.edu.pl

Abstract

The paper explores the application of plWordNet, a very large wordnet of Polish, in weakly supervised Word Sense Disambiguation (WSD). Because plWordNet provides only partial descriptions by glosses and usage examples, and does not include sense-disambiguated glosses, PageRank-based WSD methods perform slightly worse than for English. However, we show that the use of weights for the relation types and the order in which lexical units have been added for sense re-ranking can significantly improve WSD precision. The evaluation was done on two Polish corpora (*KPWr* and *Składnica*) including manual WSD. We discuss the fundamental difference in the construction of both corpora and very different test results.

1 Introduction

Large wordnets are often treated as sense inventories that describe and enumerate word senses. If we want to process texts at the level of wordnet senses, a very useful operation, we first must map text words to those senses, i.e. to perform Word Sense Disambiguation (henceforth WSD). This is only trivial for monosemous words. WSD methods built upon supervised Machine Learning achieve good accuracy but are intrinsically impractical in their dependence on corpora that have been manually disambiguated with respect to word senses. Needless to say, such corpora are very laborious to annotate.

Weakly supervised WSD methods that use a wordnet as the basic knowledge source, but do not depend on a manually annotated corpus, can fully utilise wordnet senses, i.e. they can in theory assign any sense stored in a wordnet to words in text. So, in spite of their lower precision they seem to be noteworthy as a potentially practical solution. Most wordnet-based weakly supervised WSD methods are based on the idea of spreading activation in the wordnet graph, where the initial activation comes from the words in a textual context.

Several methods based on this general scheme were proposed. A short overview is presented in Section 2. Most such methods were developed and tested on Princeton WordNet (PWN) (Fellbaum, 1998) that is slightly different than plWordNet (Piasecki et al., 2009, Maziarz et al., 2013a), currently the world largest wordnet. First attempts to transfer the methods with good performance on PWN to plWordNet (Kędzia et al., 2015) were encouraging; the performance is relatively close to the performance of the supervised methods observed for Polish on limited test sets (Baś et al., 2008, Młodzki and Przepiórkowski, 2009). In addition to the differences between both wordnets, PWN has been enriched with various other resources in order to obtain better performance of unsupervised WSD. First of all, additional links between synsets were created on the basis of the manually disambiguated SemCore corpus (Miller et al., 1993). Such links have contributed significantly to the increase of WSD performance. There is no Polish corpus similar to SemCore.

The goal of the work presented here is to explore the structure and specific properties of plWordNet in order to improve the precision of the WSD methods based on the spreading activation in the wordnet graph, here the plWordNet graph.

In the rest of the paper, first we will briefly overview the existing wordnet-based unsupervised WSD methods, including their known applications to plWordNet. Next, the plWordNet model will be discussed and compared with PWN from the perspective of utilising different features in WSD method. On this basis, several possible versions of unsupervised WSD will be introduced. Finally,

we will present data sets used in the evaluation and the results achieved for different settings used in WSD methods. Based on the results, we will analyse the the specific properties of plWordNet and its development process and its influence on wordnet-based unsupervised WSD methods for Polish.

2 Wordnet-based WSD

Unsupervised WSD methods (Pantel, 2003) use corpora to induce word senses and tune mechanisms for assignment of the induced senses to words. However, it is difficult to map the induced word senses to the wordnet. Weakly supervised WSD that are based on a wordnet as the knowledge base work directly on wordnet synsets and do not depend on manually disambiguated corpus.

Lesk's algorithm (Lesk, 1986) can be applied to textual definitions constructed on the basis on of synsets, e.g. from glosses, examples and synset members. The definitions are next compared with the occurrence contexts of words. Different similarity measures can be applied. The main problems are limited lenghts of the constructed definitions and high computational complexity, because many word sets must be compared.

Weakly supervised wordnet-based WSD algorithms assume that if we map words senses pertaining to a text fragment onto the wordnet graph, we can expect that the "hits" are located in short distances (in terms of paths) from each other in the wordnet graph. Moreover, we can use a kind of spreading activation algorithm in order to move this information along the wordnet graph, analyse the "hot" areas and identify word sense, i.e. lexical units (LUs),[1] located in them or close to them. Those LUs should be the most likely senses for words in the text. There are several parameters to set in this general scheme: the initial activation (text words vs LUs), spreading algorithm (topology and relations) and identification of association between "hot" areas and LUs to be chosen. Various methods propose a range of decisions.

Weakly supervised WSD methods are mostly based on the PageRank algorithm (Page et al., 1999) for spreading. Mihalcea et al. (2004) proposed application of the original PageRank to WSD called Static PageRank.

Page Rank algorithm (henceforth PR) is an iterative method for ranking nodes in the graph G. In WSD the nodes in G represent synsets and the edges of G correspond to wordnet relations (between synsets and in other case between synsets and between LUs). The spreading is done iteratively in the following way:

$$\mathbf{P}^{(\mathbf{new})} = cM\mathbf{P}^{(\mathbf{old})} + (1-c)\mathbf{v} \qquad (1)$$

$M_{N\times N}$ ins the adjacency matrix of the wordnet graph with N nodes (synsets), where $m_{ij} = \frac{1}{d_i}$ if the edge from the node s_i to s_j exists, 0 otherwise; d_i is degree of the node s_i (representing the synset i); where c is the damping factor; $\mathbf{v_{N\times 1}}$ is the vector of the initial scores for nodes and $\mathbf{P_{N\times 1}}$ is a vector of node scores updated in every iteration. In Static PageRank (SPR) all values in $\mathbf{v}$ are equal $1/N$.

Agirre and Soroa (2009), Agirre et al. (2014) proposed a modified version called Personalised PageRank (PPR) in which the values in $\mathbf{v}$, called personalised vector, depends on the text context of the disambiguated word. The non-zero score values are assigned to those nodes which are contextually supported. In PPR all words from the context are disambiguated at once. The $\mathbf{v}$ values are equal to:

$$\mathbf{v}[i] = \frac{1}{\frac{CS}{NS(i)}}, \quad i = 1, 2, ..., N \qquad (2)$$

where CS is the number of different lemmas in the context, $NS(i)$ – the number of synsets sharing the same context lemma with the synset i.

Agirre and Soroa (2009), Stevenson et al. (2012) proposed a modified version of PPR called Personalised PageRank Word-to-Word (PPR_W2W), in which a word to be disambiguated is excluded from the occurrence contexts, i.e. all synsets of this word have initial scores in $\mathbf{v}$ set to zero. Thus, PPR_W2W cannot be run once for all ambiguous words in the context. The vector $\mathbf{v}$ must be initialised individually for each ambiguous word in the context – this is a disadvantage of PPR_W2W. A potential advantage is the removal of the effect of mutual amplification of the closely connected senses of the word being disambiguated. The best results (measured in recall) are obtained on the *Senseval-2* dataset for a graph built from WordNet 1.7 and eXtended WordNet (Harabagiu et al., 1999). For nouns the best results are obtained using PPR (recall 71.1%), for verbs and adjectives with PPR_W2W recall was between 38.9% and 58.3%. For adverbs SPR achieved the best result of 70.8%. The best result

[1] See Section 3 for more on LUs.

for nouns, 71.9%, was achieved by PPR_W2W on the basis of the combination of WordNet 3.0 with disambiguated glosses.

In (Kędzia et al., 2014), SPR algorithm for Polish was based on plWordNet 2.1. The graph consisted of synsets linked by edges representing a selected subset of the synset relations. The precision on nouns (43%) and verbs (28%) was low in comparison to the works for English. The algorithm was evaluated on the *KPWr* corpus of Polish discussed in Section 5. In the second version, a Measure of Semantic Relatedness was utilised to add links to plWordNet. The measure had been extracted automatically from a large corpus of 1.8 billion words. However, there was no improvement: the precision for nouns was 37% and 27% for verbs. Nevertheless, we observed that even a WSD method of limited precision can be helpful in improving the performance of text clustering.

Next we adapted several algorithms: SPR, PPR and PPR_W2W – to Polish resources Kędzia et al. (2015). plWordNet 2.2 was used with all synset relations for the edges. Due to the lack of word-sense disambiguation of glosses, no additional synset links could be added. The achieved precision (on *KPWr*) was in the range 42.79%-50.73% for nouns and 29.79%-32.94% for verbs. PPR_W2W produced the best results. We also tested different variants of combining plWordNet with the *Suggested Upper Merged Ontology* (SUMO) (Pease, 2011) on the basis of the mapping constructed in (Kędzia and Piasecki, 2014). All three PR-based algorithm were evaluated. A slight improvement of the precision for nouns up to 50.89% for PPR_W2W could be observed when the two joined graphs were treated as one large graph.

3 plWordNet properties

plWordNet is a very large wordnet built independently from PWN and expresses several unique features. Word senses are represented in plWordNet as *lexical units* (LUs), i.e. pairs: lemma[2] plus sense identifier. LUs are the basic building blocks of plWordNet, but one LU belongs to exactly one synset. plWordNet includes about 40 main types of lexico-semantic relation. Half of them links synsets, the rest directly link LUs (Piasecki et al., 2009, Maziarz et al., 2012, 2013a, Piasecki et al., 2013). Many relations, e.g. meronymy, have subtypes, so the total number of lexico-semantic relations in plWordNet 2.3 exceeds 90.

The detailed description of the model underlying plWordNet can be found in (Maziarz et al., 2013b), below we present only a concise overview due to the space limit. LUs that share a set of constitutive lexico-semantic relations are grouped into *synsets* that are considered to consists of *near synonyms*. Synset relations are notational abbreviations for the relations shared between LUs from the linked synsets. The relations are the basic means of describing word senses. Different types of relations express different semantic associations, and provide different semantic information. This properties can be explored in WSD to improve the use of knowledge during spreading activation in the graph.

plWordNet provides as well some additional means of semantic description: *stylistic registers*, *glosses* and *use examples*. Stylistic registers signal pragmatic constraints on the use of LUs. However, such subtle differences are difficult to explore in WSD methods, so we have not done it. Glosses in plWordNet are comments to the LUs (not to synsets like in PWN) provided for a human reader in order to explain the motivation behind the given word sense and clarify its difference from other senses of the same lemma. Glosses are short descriptions but they are not proper lexicographic definitions and are much less elaborated from the point of view of their application in Lesk's algorithm (Lesk, 1986). Glosses are intended to be secondary and additional to the lexico-semantic relations that are the primary tool for the description of the lexical meanings in plWordNet, e.g. the genus information is expressed by hypernymy and should not be provided in a gloss. As such they have been added only to a subset of LUs. In addition to glosses, LU can be described by one or more use examples. They are also focused on human readers, but they can be used in WSD as an additional source of information. There have been not attempts so far to disambiguate word senses in the plWordNet glosses and examples.

plWordNet has been automatically mapped onto SUMO with high precision. The extended graph, plWordNet plus SUMO, has been already used in WSD with positive signals, discussed in Section 2.

plWordNet LUs are not clustered into semantic

[2]A lemma is a basic morphological form representing a group of word forms that have the same meaning but differ in the values of the morphological categories.

domains, but only into PWN-like, i.e. domains that correspond to the lexicographer files introduced in early stages of PWN development (Fellbaum, 1998). They do not seem to provide important knowledge for WSD.

Finally, there is no information about the frequency or salience of LUs, e.g. in comparison to other LUs of the same lemma. Numerical identifiers of LUs and the order of synsets in the plWordNet database mostly originate from the order in which editors introduced them into the database.

4 Exploring plWordNet in WSD

Taking as a starting point the work of Kędzia et al. (2015) and the observations in the previous section, we explored several ways of using the knowledge present in plWordNet to improve WSD performance.

4.1 Glosses and Examples

As the number of glosses and examples has been increased in the version 2.3 of plWordNet[3] we can apply Lesk's algorithm in a straightforward way – further on called basic Lesk's:

1. For a word w to be disambiguated, we select all synsets s_i that include LUs with lemma identical to the lemma of w.

2. Description sets $D(s_i)$ encompass all lemmas that are included in glosses and examples describing LUs from s_i, as well lemmas from s_i.

3. For each occurrence of w a context set $C(w)$ is collected, such that it contains all lemmas from the fixed size context of the w occurrence.

4. s_i such that the set $D(s_i)$ that have the maximal intersection with $C(w)$ is selected as the sense of the given occurrence of w.

The results obtained with the basic Lesk's algorithm are presented in Table 5.

4.2 Structural Description

In all experiments presented in (Kędzia et al., 2015) the wordnet graph was treated as a direct but uniform graph, i.e. every relation link was represented in the same way independent of the relation type. In order to increase the density of the graph LU relations were mapped on the synset level, i.e. if there was a link between LUs, then a link between their synsets was added. However, different relations represent different types of semantic association and provide different descriptions for the elements (synsets or LUs) they are attached to. On the basis of preliminary experiments, we assumed that synset relations and LU relations convey information of different importance for WSD and we assigned different weights to both types of links: $w_{LU} = 0.3$ for LU relations and $w_S = 0.7$ for synset relation[4]. The assigned weights can be next used in the spreading activation algorithm.

4.3 Sense order

In the case of highly polysemous words, some word senses located close to each other in the word graph are difficult to be distinguished. However, for practical applications, sometimes there is no need to differentiate such closely related word senses. So, we also tested partial WSD in which the top-ranked LUs within the range of $k = 30\%$ of the maximal score from the WSD algorithm were selected as a joint result. In a natural way, this relaxation of the task resulted in significantly improved precision.

It is well known that the most frequent sense baseline is difficult to be beaten by WSD. This is due to the mostly skewed distribution of word senses, in which one or few senses dominate among occurrences. Having LUs ordered according to their frequency in plWordNet, we could use this information to boots WSD performance. However, both Polish corpora annotated with word senses are much too small to provide such data. Regardless, LUs are numbered in plWordNet according to the order in which they have been added for the given lemma. The detailed guidelines for plWordNet editors say nothing about the order in which LUs should be defined[5], and our null hypothesis was that this would be almost a random factor from the point of WSD, i.e. the use of this information should not have any positive effect on the WSD performance. Nevertheless, we suspected that the null hypothesis does not match the

[3] However, most glosses take the form of short comments that are several words long.

[4] The highest weight of 1.0 was implicitly assigned to the synonymy relation that was not present in the graph structure but was expressed by synsets. The synsets collected activations from the occurrence of their members in the contexts of disambiguation.

[5] In fact it would be very difficult to define this in guidelines in a way resulting in consistent decisions of editors.

data and that the order of LUs identifiers is not accidental. We assumed that LUs with the highest identifiers represent the most salient senses of lemmas. Thus, selecting them should bring us closer to selecting the most frequent sense.

The relatively good results, presented in Section 5, seem to be in favour of rejecting the null hypothesis. They give some insights into the work of plWordNet editors, see Section 5.2.

5 Results and evaluation

Evaluation was based on applying the analysed algorithms to a corpus with manually disambiguated LUs (word senses). As a main criterion for evaluation we used the precision, calculated by comparing the LUs assigned by annotators and the algorithms, see Equation (3):

$$Pr = \frac{t}{t+f} \tag{3}$$

- t: the number of correctly disambiguated instances,
- f: the number of incorrectly disambiguated instances.

5.1 Experimental settings

Two corpora including disambiguated assignment of LUs to words were used during the evaluation. They have different character and were built by two independent teams but both are based on plWordNet, so that seems to be an interesting opportunity for evaluation.

The *KPWr* corpus (Corpus of the Wrocław University of Technology) (Broda et al., 2012), available under the Creative Commons license,[6] contains 1,127 documents (≈250,000 tokens) divided into 11 thematic categories. *KPWr* has been manually annotated and disambiguated at several levels: morpho-syntactic, syntactic relations, semantic relations, Named Entities. The documents are also described with manually assigned keywords and meta-information, like genre, author, etc.

In the case of 88 different lemmas, all their occurrences have been manually described with LUs from plWordNet by two annotators plus a super-annotator, who was responsible for solving conflicts. In the case of all lemmas annotated, their descriptions in plWordNet have been verified according to the defined set of LUs and the information provided for them, i.e. relation links, glosses and usage examples. In the case of lacking LUs (missing word senses), they have been added. If for some LU of one of the 88 lemmas there was no usage examples in *KPWr* or the number was very small, *KPWr* was expanded with some new texts. The WSD part of *KPWr* has been built in two stages, and in the second stage all previous annotations have been verified.

The WSD lemma set includes 58 different nouns and 30 verbs, see the statistics in Table 1. The lemmas were not selected randomly, but were chosen by linguists in such a way that all the lemmas are polysemous and represent different types of homonymy and polysemy. Moreover they vary according to numbers of possible lexical meanings, i.e. possible LUs. From the very beginning this set of WSD annotations was meant to be a gold standard for the evaluation of WSD methods.

	Total	**Nouns**	**Verbs**
Tagged words	88	58	30
Tagged instances	6048	3846	2202

Table 1: Statistic of WSD annotations in *KPWr*.

For 58 nouns and 30 verbs, the average number of word senses per word are 5.98 and 7.50 respectively. The standard deviation is 4.30 for nouns and 3.96 for verbs. The median of number of senses for the nouns is 5; 4 nouns have the number of senses equal to the median. 28 nouns have more senses than the median, and 26 have fewer. The median number of senses for the verbs is 6; 5 verbs have a number of senses equal to the median. 12 verbs have fewer senses than the median, and 13 have more. Thus, the annotated words are quite diversified and challenging for WSD.

Składnica (Hajnicz, 2014a), a treebank of Polish, is the second test set used during the evaluation. It includes 20,000 sentences among which more than 8,200 have manually assigned parse trees. For all these sentences, nouns, verbs and adjectives occurring in them have been manually mapped to LUs from plWordNet 1.6 (Hajnicz, 2014b). Proper Names included in them have been marked and semantically classified. Lemmas or word senses not found in plWordNet have been marked. *Składnica* includes sentences randomly selected from the open part of NKPJ (National Corpus of Polish) (Przepiórkowski et al., 2009). All sentences are described by identifiers and links to the original paragraphs, so it is possible to use

[6] http://nlp.pwr.edu.pl/kpwr

the whole paragraphs as contexts for WSD. *Składnica* differs significantly from *KPWr* with respect to words disambiguated with word senses: the selection was made at the level of sentences, so in the case of most lemmas only selected senses are covered. In *KPWr* all senses of every selected word are represented. Moreover, the *KPWr* builders paid attention to acquiring as many usage examples as possible for every senses, including those that are infrequent.

	Total	**MN**	**PN**	**MV**	**PV**
Tag. words	6309	1717	2424	684	1484
Tag. instances	15342	3560	6610	1307	3865

Table 2: Statistics of WSD annotations in *Składnica*.

WSD annotations in *Składnica* has been provided not only for polysemous words, but also for monosemous – in Table 2 the column *MN* contains statistics for monosemous nouns, *PN* for polysemous nouns, *MV* for monosemous verbs, *PV* polysemous verbs.

5.2 Results

5.2.1 Baseline PageRank approaches

As a baseline, we repeated experiments from (Kędzia et al., 2015) using plWordNet 2.2 as originally, but also version 2.3 as a basis for the WSD algorithm. All tests were performed on *KPWr* ; the results are shown in Table 3. The columns grouped under the label *PPR* include results achieved by the application of the *Personalized PageRank* algorithm, while the joint label *Static* signals the application of *Static PageRank*. The description of the tested combinations (algorithm parameters and the wordnet version) could make the table too large, so the combinations have been encoded as follows:

C1 the results achieved on plWordNet 2.2,

	PPR			**Static**		
	V	**N**	**All**	**V**	**N**	**All**
C1	28.64	47.25	40.45	28.14	43	37.57
C2	33.70	50.23	44.58	34.11	44.17	40.73
C3	29.57	48.06	37.57	29.79	42.79	38.05
C4	32.61	52.22	45.52	32.19	44.63	40.38

Table 3: Comparison of disambiguation precision using PLWN 2.2 and PLWN 2.3 evaluated on *KPWr*

	KPWr			***Składnica***		
	V	**N**	**All**	**V**	**N**	**All**
C5	34.11	44.17	40.73	47.08	57.37	53.37
C6	33.70	50.23	44.58	42.05	54.15	49.44
C7	32.19	44.63	40.38	47.00	57.97	53.70
C8	32.61	52.22	45.52	41.99	55.40	50.17

Table 4: Precision of disambiguation achieved on *KPWr* and *Składnica*.

C2 as above, but for plWordNet 2.3,

C3 and C4 the results achieved on plWordNet versions 2.2 and 2.3, respectively, merged with the SUMO ontology; in both only nodes belonging to plWordNet are initialised (i.e. receive non-zero values in the initial vector).

In Table 3 we can observe that the increasing size of plWordNet affects positively the precision when the same configuration of the algorithm is applied. This effect can be caused by the increasing number of text words covered by the wordnet that results in the increasing number of initially activated nodes in the PR graph. Moreover, in plWordNet 2.3 the number of adjectives and relation links between adjectives and nouns have been increased significantly. Thus cross-categorial connections have been improved, facilitating the activation flow in PR-based algorithms.

Next, we performed similar tests but using both data sets, i.e. *KPWr* and *Składnica*. Once again algorithms and parameters from (Kędzia et al., 2015) were applied, but this time we concentrated only on plWordNet 2.3. This resulted in better precision in the experiments presented above. Table 4 contains the results achieved for the following configuration of the algorithms:

C5 *Static* algorithm, only plWordNet 2.3 synset graph used,

C6 *PPR* algorithm, only plWordNet 2.3 synsets,

C7 *Static* algorithm, plWordNet 2.3 synset graph merged with *SUMO* ontology, but only nodes from plWordNet are initialised,

C8 *PPR* algorithm, as above, plWordNet 2.3 synset graph merged with *SUMO* ontology, but only nodes from plWordNet are initialised for disambiguation.

Results on *Składnica* are higher and close to the results obtained for English. The precision is

	KPWr			Składnica		
	V	**N**	**All**	**V**	**N**	**All**
Lesk	16.80	18.80	18.12	39.34	38.56	38.87

Table 5: Simple Lesk algorithm run on *KPWr* and *Składnica*

	KPWr			*Składnica*		
	V	**N**	**All**	**V**	**N**	**All**
C8	32.61	**52.22**	45.52	49.02	64.02	58.48
C9	**42.66**	47.91	**46.12**	**47.51**	**61.67**	**56.16**

Table 6: Static PageRank WSD algorithm based on the weighted plWordNet graph (**C9**) in comparison to the PPR algorithm.

clearly boosted by the monosemous words, while monosemous words are not annotated *KPWr*. However this influence is too small to be the only reason for the difference, e.g. in Tab. 6 in the case of *Składnica* only polysemous words were evaluated, i.e. for polysemous and monosemous words the precision of **C9** is: 69.08% for nouns, 53.86% for verbs and 63.46% for all. The higher precision on *Składnica* can be also caused by the different way of selecting words for WSD annotation. In *Składnica* they come from the running text and we can expect some bias towards most frequent LUs (word senses), while the authors of *KPWr* tried to cover in WSD annotation all LUs for the selected lemmas, so less frequent LUs received more occurrences than we could expect in a text sample. Tests on *KPWr* illustrate the ability of the algorithms to distinguish between all possible senses, while tests on *Składnica* are a better picture of average precision we can expect in practical applications (especially when monosemous words are included in the result).

5.2.2 Glosses and Examples

The results of the simple Lesk's algorithm based on plWordNet 2.3 run on both corpora are presented in Tab. 5, where the precision is given for verbs and nouns in percentage points. This algorithm can be treated as the second baselines. The results illustrate the amount of disambiguating information included in the textual descriptions of plWordNet. They are much lower than obtained by PageRank-based algorithms, that explore the rich structure of plWordNet relations

5.2.3 Structural Description

Tab. 6 presents a comparison of the best baseline configuration for *KPWr*, namely **C8** with the approach using the information about the relation types called **C9**. In **C9** *Static* algorithm based on plWordNet 2.3 was used, but synset relations were assigned weights equal to 0.7 and LU relations weights equal to 0.3. Moreover, the top-scoring LUs within the range of 10% from the best score (according to the WSD algorithm) are re-ranked according to their order (i.e. their identifiers) in the plWordNet database. The re-ranking is limited to those cases in which the values from WSD are very close and the differences can be insignificant.

On *KPWr*, the use of weighting gave improvement only for verbs. Verbs have a higher ratio of LU relations in comparison to synset relations than nouns, so this supports the intuition that synset relations provide more information for WSD. However, a more in-depth analysis of different weights for different relations is needed. Such an optimisation would need larger training-testing WSD data sets. The situation was completely different in tests on *Składnica* – here in all cases a significant improvement can be observed. It seems that the higher weights for synset relations and synonymy (the weight 1.0) favour the most frequent senses.

	KPWr			*Składnica*		
	V	**N**	**All**	**V**	**N**	**All**
C10	38.57	43.20	41.62	48.77	**61.74**	**56.69**
C11	39.76	39.30	39.46	**49.28**	61.12	56.51

Table 7: PageRank-based WSD algorithms supported by re-ranking based on the synset order in plWordNet.

5.2.4 Sense order

Finally, we tested the use of the order of adding LUs to plWordNet for a given lemma as an additional source of knowledge for WSD algorithms. In all cases this knowledge was used for post-re-ranking. Two configurations were tested:

C10 *Static* algorithm, plWordNet 2.3 synset graph only, WSD results post-processed by re-ranking of the top highest scored LUs within the range of $k = 30\%$ of the maximal score, the re-ranking is based on LUs numbers in plWordNet.

C11 Similar to **C10**, but re-ranking is limited to $k = 40\%$ of the maximal score.

The results obtained with the help of **C10** and **C11** are presented in Tab. 7. In comparison to the

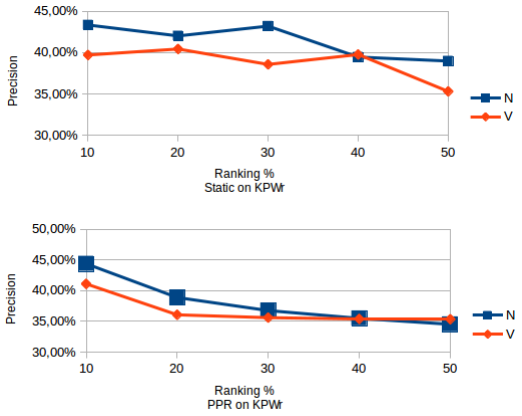


Figure 1: Influence of ranking % on precision evaluated on *KPWr* with Static and PPR.

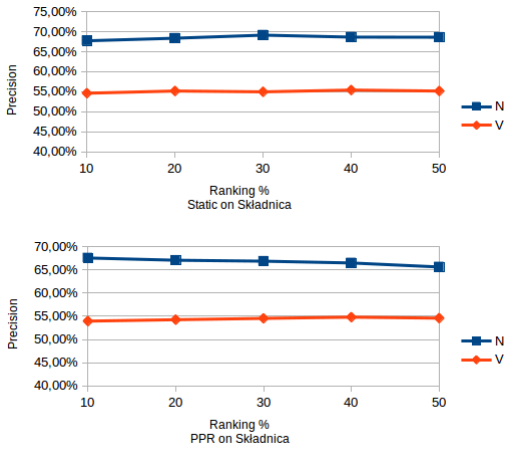


Figure 2: Influence of ranking % on precision evaluated on *Składnica* with Static and PPR.

baselines shown in Tab. 4, we can notice that re-ranking brought significant improvement in tests on *Składnica* for both configurations. The situation is different for *KPWr*. *KPWr* includes more occurrences of less frequent senses, while *Składnica* has a bias towards more frequent senses as built on randomly selected sentences. This difference supports our assumptions that LU numbers in plWordNet are correlated with their frequency in corpora. This correlation is next transferred to re-ranking. This observation is important for practical applications. Thus, we guess that the wordnet editors share some notion of the word sense saliency or their frequency. For a new lemma being edited, they seem to add to the plWordNet its more prominent and more frequent senses first. plWordNet 1.6 noun synsets were automatically ordered according to the estimated frequency of the word senses they represent (McCarthy et al., 2004, 2007). However, this method is of limited accuracy and all synsets added later (a large number, the majority) were not ordered in this way.

In Tab. 1 and 2 the analysis of the relation between the re-ranking threshold and precision is presented. In the case of *KPWr* the best results were obtained for the 10% re-ranking threshold. However, in the case of *Składnica* the highest results are concentrated around the threshold 30% and decrease beyond it, so scores produced by the WSD algorithm are at least useful in selecting the most likely LUs for a given word.

6 Conclusions

Weakly supervised WSD methods based on plWordNet have slightly lower precision in tests on Polish WSD corpora than similar PWN-based methods. However, plWordNet does not provide glosses for all LUs and the existing glosses are not disambiguated. Instead we looked into utilisation of other features. We showed that except glosses and examples, we can explore relation types by weighting them for the needs of WSD and the order in which LUs have been added to plWordNet. Both resulted in the increased precision of WSD on one of the test corpora – the one that seems to be closer to the practical applications. While the positive influence of the relations weights on PageRank-based WSD algorithm had been expected, the positive influence of the LUs adding order is a surprise, as the wordnet editors were not asked to use any specific order in introducing new LUs into plWordNet. Thus they have to share some idea of the salience or frequency of the individual LUs for the given lemma. This effect may not be visible when we analyse lists of LUs of individual lemmas, but it seems to be the most probable explanation for the results WSD algorithms using this order as a knowledge source. In future work we plan to develop more sophisticated system of weights assigned to relations for WSD and to work on combining different knowledge sources in one complex WSD algorithm.

Acknowledgment

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

References

Eneko Agirre and Aitor Soroa. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609070.

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. Random walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84, 2014.

Dominik Baś, Bartosz Broda, and Maciej Piasecki. Towards Word Sense Disambiguation of Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, pages 65–71, 2008. URL http://www.proceedings2008.imcsit.org/pliks/162.pdf.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. KPWr: Towards a free corpus of Polish. In *Proceedings of LREC'12*, Istanbul, Turkey, 2012. ELRA.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. ISBN 026206197X.

Elżbieta Hajnicz. The procedure of lexico-semantic annotation of Składnica treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014a. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Elżbieta Hajnicz. Lexico-semantic annotation of *składnica* treebank by means of PLWN lexical units. In *Proceedings of the 7th International WordNet Conference*, pages 23–31, 2014b.

Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. WordNet 2 - a morphologically and semantically enhanced resource. In *SIGLEX99: Standardizing Lexical Resources*, pages 1–8, 1999. URL http://www.aclweb.org/anthology/W99-0501.

Paweł Kędzia and Maciej Piasecki. Ruled-based, interlingual motivated mapping of plWordNet onto SUMO ontology. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Paweł Kędzia, Maciej Piasecki, Jan Kocoń, and Agnieszka Indyka-Piasecka. Distributionally extended network-based Word Sense Disambiguation in semantic clustering of Polish texts. *IERI Procedia*, 10(Complete):38–44, 2014. doi: 10.1016/j.ieri.2014.09.073.

Paweł Kędzia, Maciej Piasecki, and Marlena J. Orlińska. Word sense disambiguation based on large scale Polish CLARIN heterogeneous lexical resources. *Cognitive Studies*, 14(To appear), 2015.

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings SIGDOC '86 Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan, January 2012.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stanisław Szpakowicz. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*, pages 443–452, 2013a. URL http://aclweb.org/anthology/R/R13/R13-1058.pdf.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. The chicken-and-egg problem in wordnet design: Synonymy, synsets and constitutive relations. *Language Resources and Eval-*

uation, 47(3):769–796, 2013b. doi: 10.1007/s10579-012-9209-9.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1218955.1218991. URL http://dx.doi.org/10.3115/1218955.1218991.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Unsupervised acquisition of predominant word senses. *Comput. Linguist.*, 33(4):553–590, December 2007. ISSN 0891-2017. doi: 10.1162/coli.2007.33.4.553. URL http://dx.doi.org/10.1162/coli.2007.33.4.553.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. PageRank on semantic networks, with application to Word Sense Disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220517. URL http://dx.doi.org/10.3115/1220355.1220517.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7. doi: 10.3115/1075671.1075742. URL http://dx.doi.org/10.3115/1075671.1075742.

Rafał Młodzki and Adam Przepiórkowski. The WSD development environment. In Zygmunt Vetulani, editor, *LTC*, volume 6562 of *Lecture Notes in Computer Science*, pages 224–233. Springer, 2009. ISBN 978-3-642-20094-6. URL http://dblp.uni-trier.de/db/conf/ltconf/ltconf2009.html#MlodzkiP09.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web, 1999.

Patrick A. Pantel. *Clustering by Committee*. PhD thesis, University of Alberta Edmonton, Alta., Canada, 2003.

Adam Pease. *Ontology: A Practical Guide*. 2011.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground up*. Oficyna Wydawnicza Politechniki Wroclawskiej, 2009.

Maciej Piasecki, Stan Szpakowicz, Christiane Fellbaum, and Bolette Sandford Pedersen. Introduction to the special issue: On wordnets and relations. *Language Resources and Evaluation*, 47(3):757–767, 2013. ISSN 1574-020X. doi: 10.1007/s10579-013-9247-y. URL http://dx.doi.org/10.1007/s10579-013-9247-y.

Adam Przepiórkowski, Rafał Górski, Barbara Lewandowska-Tomaszczyk, and Marek Łaziński. Narodowy Korpus Języka Polskiego, 2009.

Mark Stevenson, Eneko Agirre, and Aitor Soroa. Exploiting domain information for Word Sense Disambiguation of medical documents. *JAMIA*, 19(2):235–240, 2012. URL http://dblp.uni-trier.de/db/journals/jamia/jamia19.html#StevensonAS12.

plWordNet 3.0 – Almost There

Maciej Piasecki[A], **Stan Szpakowicz**[B], **Marek Maziarz**[A], **Ewa Rudnicka**[A]
[A] G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Technology, Wrocław, Poland
[B] School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada
[A] maciej.piasecki@pwr.wroc.pl, mawroc@gmail.com,ewa.rudnicka78@gmail.com
[B] szpak@eecs.uottawa.ca

Abstract

It took us nearly ten years to get from no wordnet for Polish to the largest wordnet ever built. We started small but quickly learned to dream big. Now we are about to release plWordNet 3.0-emo – complete with sentiment and emotions annotated – and a domestic version of Princeton WordNet, larger than WordNet 3.1 by nearly ten thousand newly added words. The paper retraces the road we travelled and talks a little about the future.

1 Wordnet makers' ambition

A respectable wordnet ought to be a fair model of the lexical-semantic system of the language it represents; a nearly comprehensive model is a dream worth pursuing. A wordnet linked to other wordnets, and to world knowledge, is a dream come true. This paper tells the story of plWordNet, a resource for Polish built over a decade of concentrated effort. Our wordnet is well published, but we are reaching a really large milestone, so we want take a bird's eye view of that decade.

We began cautiously. Our starting point in 2005 was a list of 10,000 most frequent lemmas in the IPI PAN Corpus of Polish, a mere quarter billion words from not quite balanced sources (Przepiórkowski, 2004). More than 30 person-years later, we are but a small step from completing the work on plWordNet 3.0-emo. With 177,003 lemmas, 255,733 lexical units, 193,286 synsets and more than 550,000 instances of relations, it is – in numbers – the largest wordnet created to date. Practically all its elements are in place, and the rollout is imminent. We think that it is an opportunity to present a synthetic picture of the whole endeavour.

The paper first recalls the initial fundamental assumptions, which have held astonishingly well, even if they had, inevitably, to be adjusted as our wordnet grew. We discuss the central lessons learned, and present the structure and statistics of plWordNet 3.0-emo. Finally, there is an overview of applications, and plans for the future.

2 Assumptions

We based the development of plWordNet on several unique assumptions, formulated *a priori*. They have been discussed at length in previous publications, notably (Piasecki et al., 2009; Maziarz et al., 2013c), so we will only recapitulate them briefly just to ease into the further discussion.

First and foremost, we believe that lexico-semantic systems of different languages differ in deep – and interesting – ways. That is why plWordNet, meant as *a precise description of the Polish lexical system*, had to be built in a way that avoided widespread influence of the material and structure of other wordnets. We were aware of the high cost of not simply translating Princeton WordNet, the only resource large enough for our ambition, but it felt most important to be faithful to the complex reality of our language.[1]

When the project began, there were no public-domain and no open-licence large electronic lexico-semantic resource for Polish.[2] We opted for a *corpus-based wordnet development process*. A very large corpus, the main knowledge source, is supplemented by a variety of linguistic substitution tests, mono-lingual dictionaries and other semantic language resources, encyclopædias, discussions among linguists, and the wordnet editors' linguistic and lexicographic intuition.

[1] In retrospect, this decision has been borne out by the scale of differences between plWordNet and WordNet when we got deep enough into the mapping between the two.

[2] There are scarcely any such resources even now (Vetulani et al., 2009; Miłkowski, 2007), unless one counts plWordNet ☺.

Corpus-based work, unaided by specialised software, would necessarily be rather slow. We assumed large-scale software support for *semi-automatic wordnet construction*, predicated on the availability of support tools for editing. Such tools were designed and built (and then perfected over the years) in parallel with fully manual construction of a small wordnet core to serve as the springboard for further expansion. This ensured much reduced workload for the editors and improved exploration of the corpus data. In many cases, the editor needs only to conform the support system's suggestions.[3]

It soon became clear that there were significant problems with making the usual synset definition operational, and with the consistency of the editors' decisions. We chose a smaller-grain basic element for plWordNet: the *lexical unit*.[4] A synset was then defined indirectly as a set of lexical units which share a number of *constitutive lexico-semantic relations and features* (Maziarz et al., 2013c). Relations between synsets are a notational abbreviation for the shared relations between lexical units grouped into those synsets. Constitutive relations, which define the structure of the wordnet, are complemented by relations which only link lexical units. Three categories of constitutive features are lexical registers, semantic classes of verbs and adjectives, and aspect. In this model, synonymy is also a derived concept: constitutive relation- and feature-sharing lexical units grouped into a synset are understood to be synonymous.

Finally, in the construction of plWordNet we tried to follow the principle of *a minimal commitment*, that is to say, to keep the number of assumptions small, to make plWordNet transparent to linguistics theories of meaning, and to shape it in a close relation to language data.

3 Lessons learned

3.1 Tools and organisation of work

Ten years of continuous wordnet development gave us a lot of practical experience which confirmed the initial assumptions.

The building of plWordNet was what can be termed a *corpus-based wordnet development process*. It starts with the lemmatisation of a large corpus and the extraction of the lemma frequency ranking. A top sublist of *new lemmas*, those not yet included in plWordNet, is selected for the given iteration of wordnet expansion. Typically, 6000-9000 new lemmas selected for an iteration meant 3-6 months of work. Each iteration processed lemmas in the same part of speech. We tried to "sanitise" every list by removing obvious non-words (mostly proper names), but serious cleaning would double the workload: it requires searching corpora and identifying potential senses.

Several tools examine the corpus to extract *knowledge sources* which help merge a new batch of lemmas with what is already in plWordNet: a Measure of Semantic Relatedness (MSR) and lists of lemma pairs potentially linked by hypernymy. The LexCSD system (Broda and Piasecki, 2011) extracts usage examples for the new lemmas. The extracted MSR was next used to cluster lemmas into semantically motivated groups we call *packages*, each package assigned to one editor. A package is clearly homogenous; usually, 2-3 domains are most prominent (lemmas were grouped by dominating senses), so the editor can stay focused. The acquired knowledge sources were input to the WordnetWeaver system (Piasecki et al., 2009) which, for each new lemma, automatically suggests the number and location in the network of lexical units. The suggestions are visually presented in the wordnet editing system WordnetLoom (Piasecki et al., 2010).

The plWordNet team consists of rank-and-file editors and coordinators.[5] Before tackling lemmas in any of four parts of speech, we prepared guidelines with detailed relation definitions and substitution tests. A coordinator entered the definitions and tests into WordnetLoom, and trained the editors. The coordinator assigns lemmas to editors in batches, performs selective verification, answers questions, refines the guidelines, and monitors the pace and progress of the editors' work.

For frequent lemmas, the editor uses supporting tools in a specified order of importance: WordnetWeaver suggestions; corpus browsers; usage

[3] Software support has also greatly assisted in the mapping between plWordNet and Princeton WordNet. Likewise, a mapping to knowledge resources, notably to ontologies, had to be built semi-automatically from scratch.

[4] A lexical unit is understood here as a triple: (lemma, part of speech, sense identifier). A lemma is the basic morphological form of a word. Each lexical unit represents a unique word sense.

[5] At the height of plWordNet development, several coordinators supervised a small group of editors each. Separate teams work on plWordNet-to-WordNet mapping, and on sentiment annotation. All this allows cross-checking: the teams exchange information about likely errors.

examples generated automatically by LexCSD and the induced senses they represent; lists of highly related lemmas according to MSR; existing electronic dictionaries, lexicons, encyclopaedias; and, last but not least, the linguistic intuition of the editor and the team. The importance of WordnetWeaver and MSR dropped for lower-frequency lemmas. In the case of nouns editors tend to use dictionaries as the main source, but still remember the other sources. Adjectives and adverbs are much less richly described in the existing dictionaries, so LexCSD examples and corpus browsers became the primary tools.

Before adding any relation instance to the wordnet, WordnetLoom presents the appropriate substitution test with the variable slots filled by the lexical units of the two synsets. The instantiated substitution test reminds about the constraints included in the relation definition, likely improving the consistency of the editors' definitions. Similarly, consistency increases with the use of the same supporting tools in the same order.

3.2 The role of corpora

Corpus-based development is surely slower and more costly than the merge method based on the previously existing lexical resources, but it is the only method which allows going beyond the existing dictionaries, often closely related. Corpus-based development also promotes a wordnet's better coverage of lemmas described and lexical units, assuming that the procedure recapped above is carefully followed. Obviously, a lot depends on the type of corpus. We aimed at building a comprehensive wordnet, so we tried to acquire or collect as large a corpus as possible. We made a practical assumption that the larger the corpus and the more diverse its text sources, the more balanced and representative the corpus becomes.

The development of plWordNet 1.0 relied on the IPI PAN Corpus (IPIC) (Przepiórkowski, 2004), ca. 260 million tokens, the first publicly available large corpus of Polish.[6] IPIC represents a range of genres, biased towards parliamentary documents and scientific literature. That is why we put much effort into collecting corpora and texts, and combining them with IPIC.

The work on plWordNet 2.1 built upon a plWordNet corpus of 1.8 billion tokens, recently expanded to almost 4 billion tokens. This merged corpus encompasses IPIC, the corpus of text from the newspaper *Rzeczpospolita* (Weiss, 2008) and Polish Wikipedia; it is complemented by texts collected from the Internet, filtered according to the percentage of unrecognised words by Morfeusz (Woliński, 2006), with duplicates removed with respect to the whole corpus.

Finally, plWordNet 3.0 describes all lemmas with 30+ occurrences in 1.8 billion words, as well as a significant number of those less frequent.[7] At the final stage of work on plWordNet 3.0, we plan to add missing lemmas with the frequency 30+ from the 4-billion-token corpus.

3.3 The underlying model

The strategy of making the lexical unit the basic building block helped us formulate definitions of relations, and substitution tests for those relations, so they refer primarily to language data and the distribution of lemmas in use examples. We could also refer to the linguistic tradition in defining lexico-semantic relations better matching the background of our editors. We are convinced that the use of elaborate relation definitions, substitution tests and the procedure of lexicographic work have improved the mutual understanding of the plWordNet model among the members of the linguistic team, as well as the consistency of editing decisions across the pool of editors.

The model of plWordNet, based on the sharing of constitutive relations and features, allowed us to write up and implement an operational definition of the synset. Still, specific leaves deep in the wordnet hypernymy tree often could not be easily separated into different synsets without referring to some notion of synonymy (or – more important in practice – to the absence of synonymy). We "pinned it down" as a combination of two parallel hyponymy relations. We think that the need for synonymy in wordnet editors' everyday work can be reduced in the future as the list of relations grows. That was what happened with verbs, adjectives and adverbs, for which we introduced, *e.g.,* several cross-categorial constitutive relations.

3.4 The progress of work

We deliberately avoided putting non-lexical elements in plWordNet, a lexical resource *par ex-*

[6] Oddly, it is even now the only freely available corpus of Polish. It is a pity that the newer and larger National Corpus of Polish (Przepiórkowski et al., 2012) is not all in the public domain (http://nkjp.pl/).

[7] Editors were free to add any existing lemma, after checking corpora (Przepiórkowski et al., 2012) and the Internet.

cellence. For example, we only included proper names from which frequent lexical units are derived; other proper names are kept in a separate large lexicon mapped onto plWordNet. We have also developed an elaborate procedure for assessing the lexicality of multiword expressions. We made an exception for "artificial" (non-lexical) synsets first proposed for GermaNet (Hamp and Feldweg, 1997). They usually make a wordnet's hypernymy structure more readable for humans. The added artificial nodes also help editors maintain the hypernymy structure. Consequently, a significant number of *artificial lexical units* (language expressions) have been placed in singleton synsets. Such synsets and lexical units, clearly marked, can be removed or made transparent, if needed. They are not treated as part of the lexical system described by the wordnet.

The WordnetWeaver system implements a complex frequency-based method of wordnet expansion[8] (Piasecki et al., 2013). The method worked fine in the first phase of plWordNet development, for frequent lemmas, mostly nouns. With the move to less frequent lemmas, the importance of WordnetWeaver waned. Its Measure of Semantic Relatedness (MSR), an essential knowledge source, proves useful for lemmas occurring 200+ times (an observed empirical rule); below 100 occurrences, it begins to produce many accidental associations. The thresholds are even higher for verbs, if the description of their occurrences is not based on the output of a reliable parser.

While we abandoned WordnetWeaver for less frequent lemmas, several of its components remain in use. Most important, even if the MSR's quality decreases, it helps automated semantic clustering of lemmas in aid of assigning work to individual editors. Semantically motivated packages for this purpose, even if imperfect, handily beat such schemas as alphabetic order. Also, the LexCSD system automatically extracts use examples meant to represent various senses of a new lemma. LexCSD clusters all occurrences of the lemma, and tries first to identify occurrence groups representing different senses, and then to find the most prominent example in each group.

Examples extracted by LexCSD are also presented in WordnetLoom. Such examples have become the first knowledge source which plWordNet editors consult when they work on adjectives and adverbs. Existing Polish dictionaries neglect both categories, so we rely on corpus-derived examples. Lexico-syntactic patterns used for the extraction of lemma pairs potentially linked by a given relation also apply to less frequent words; the practice shows, however, that they are also less frequent in language expressions matching the patterns. Automated methods were very helpful in expanding derivational relations in plWordNet (Piasecki et al., 2012a; Piasecki et al., 2012b). Regardless of which automatic method was used, the results were always verified by human editors and revised if necessary.

The manual mapping of plWordNet onto Princeton WordNet has incurred a high labour cost, even though we deliberately stayed away – for now – from the opposite direction (Rudnicka et al., 2012). We built an automated system to suggest inter-lingual links (Kędzia et al., 2013). Its precision is acceptable, but too low to let the results stand without intervention. We have also introduced several inter-lingual relations (Rudnicka et al., 2012) in order to cope with non-trivial differences between the two wordnets. All that investment was worth the price. The bilingual resource we now have is unique in scale (two largest wordnets, over 150,000 interlingual links between synsets) and nature (two wordnets based on slightly different models). The mapping opens many interesting paths for further exploration.

Early on, we assumed tacitly that glosses were not part of the relational model of language which our wordnet represented. We still think that it is better first to invest in building a larger gloss-free wordnet than to construct a much smaller but more lexicographically complete resource.[9] A wordnet describes the meaning of a lexical unit *via* its network of lexico-semantic relations. Inevitably, though, as plWordNet gained popularity (through its Web page and mobile application), we soon noted that glosses help non-specialist users understand the meaning of wordnet entries. It is a technicality, perhaps, but glosses also help wordnet editors see clearly the editing decisions made by other members of the team: glosses serve as a form of control information. Similarly, use examples help, and appear more important for Natural Language Engineering applications of plWordNet.

[8] automated, but subject to editors' final approval

[9] Come to think of it, glosses in Princeton WordNet were an afterthought, too. ☺

4 The structure of plWordNet 3.0-emo

Maziarz et al. (2013a) presented plWordNet 2.1. In most ways, plWordNet 3.0 is just better and larger, as planned two years ago (Maziarz et al., 2014). In comparison to version 2.1:

- noun and adjective sub-databases have grown very substantially – see the statistics in Section 5; the verb, already a large list, have been only amended;
- the set of adjective relations has been revised, while only minor changes were introduced for nouns and verbs;
- a new adverb sub-database has been constructed from scratch with the help of a semi-automatic method based on exploring derivational relations and mapping between adjective relations and adverb relations;
- an elaborate procedural definition of Multiword Lexical Units was designed (Maziarz et al., 2015), together with a work procedure supported by the semi-automatic system for collocation extraction and their further editing as potential candidates;
- the plWordNet-to-WordNet mapping has been very significantly expanded to adjectives, with coverage vastly increased to 151,200 interlingual links of various types (38,471 I-synonymy links);
- the constructed bilingual mapping was used to build a rule-based automated procedure of mapping plWordNet to SUMO (Pease, 2011; Kędzia and Piasecki, 2014).

4.1 Mapping to WordNet

To this planned development, we added two derived resources. While mapping onto Princeton WordNet, we observed that the most frequent inter-lingual relation is I-hyponymy (over twice more frequent than I-synonymy). That is to say, there were no counterparts in WordNet 3.1 for many specific lexical units in plWordNet. The cause: differences in coverage between both wordnets rather than any major differences in lexicalisation between Polish and English (Maziarz et al., 2013a), even though we dutifully checked English dictionaries and corpora for direct translations. Now, I-hyponymy is more vague – gives us less useful information for language processing – than I-synonymy. That is why we decided to add material to WordNet 3.1. The result is a resource we call enWordNet 0.1, included in the plWordNet distribution as a large bilingual system. It has been built by adding to WordNet 3.1 about 8,000 new noun lemmas (9,000 noun lexical units).[10]

We aimed to improve the mapping of plWordNet (by adding to WordNet the missing corresponding entries), and then to replace I-hyponymy with I-synonymy as much as possible. This could be done simply by translating plWordNet synsets into English and putting the translations in enWordNet,[11] but we resisted that temptation.

We decided to let I-hyponymy guide expansion. The lemmas of all plWordNet 'leaf' synsets linked by I-hyponymy to WordNet synsets were automatically translated by a large cascade dictionary. The translations were then filtered by the existing WordNet lemmas and divided into three groups, lemmas for which the dictionary found: (i) equivalents whose lemmas were absent from WordNet; (ii) no equivalents; (iii) equivalents whose lemmas were already present in WordNet. Editors started with the first group, carefully verifying the suggestions with corpora, especially BNC (BNC, 2007) and ukWaC (Ferraresi et al., 2008), and all available resources. For the second group, they tried to find equivalents on their own (in all available resources). Finally, they investigated the third group, checking the existing mapping relations. Whenever editors started work with a particular WordNet 'nest', they were encouraged to look for its possible extensions on their own, not just limit themselves to the cascade dictionary suggestions.

We began with nouns. That segment of Princeton WordNet figures in applications more often than other parts of speech. Also, our experience with developing plWordNet suggested that adding to the nouns in WordNet would be relatively easy. We used the same set of relations as in Princeton WordNet but, following the plWordNet practice, the relations have been specified by definitions and substitution tests in the WordnetLoom editing system. The editor team consisted of graduates of English philology and native speakers.

In the first phase, we used bilingual dictionaries to select from the list those lemmas which appeared to be missing translation equivalents for plWordNet synsets lacking I-synonymy. Even so, the processing of the selected lemmas was in-

[10] The estimated target size is 10,000 new nouns.

[11] That would mean applying the transfer method in an "unorthodox" direction. One normally translates English synsets into whatever language one is building a wordnet for.

dependent of their potential Polish counterparts. Only after new lexical units had been added to enWordNet would the interlingual mapping be modified or expanded. For each English lemma, the editors identified its senses by searching for use examples in the corpora. We allowed into enWordNet only lexical units with 5+ occurrences, supported by examples.

In the second phase, we used the rest of the lemma list extracted from the corpora going through the lemmas of decreasing frequency.

4.2 Sentiment and emotions

Section 6 shows how plWordNet has become an important resource for language engineering applications in Polish. A notable exception were applications in sentiment analysis, despite their growing importance among research and commercial systems. That is why we decided to annotate manually a substantial part of plWordNet with sentiment polarity, basic emotions and fundamental values (Zaśko-Zielińska et al., 2015). The suffix "-emo" in the name of this plWordNet version signals the presence of this annotation. All in all, 19,625 noun lexical units and 11,573 adjective lexical units received two manual annotations. The team consisted of linguists and psychologists, whose coordinator was tasked with breaking ties. Each lexical unit was annotated with:

- its sentiment polarity (positive, negative, ambiguous) and its intensity (strong, weak);
- basic emotions associated with it: joy, trust, fear, surprise, sadness, disgust, anger, anticipation (Plutchik, 1980);
- fundamental human values associated with it: *użyteczność* 'utility', *dobro drugiego człowieka* 'another's good', *prawda* 'truth', *wiedza* 'knowledge', *piękno* 'beauty', *szczęście* 'happiness' (all of them positive), *nieużyteczność* 'futility', *krzywda* 'harm', *niewiedza* 'ignorance', *błąd* 'error', *brzydota* 'ugliness', *nieszczęście* 'misfortune' (all negative) (Puzynina, 1992).

The annotation of nouns encompassed those hypernymy sub-hierarchies which we expected to include lexical units with non-neutral sentiment polarity. Those were the sub-hierarchies for affect, feelings and emotions, nouns describing people, features of people and animals, artificial lexical unit *events rated negatively*, *evaluated as negative* and the sub-hierarchy of entertainment. The adjectival part of plWordNet was in major expansion during that time, so we only annotated the parts for which the expansion had been completed.

It is worth emphasizing that the amount of manual annotation is several times higher than in other wordnets annotated with sentiment. This pilot study can be a good starting point for semi-automated annotation of the whole plWordNet.

POS	synsets	lemmas	LUs	avs
N-PWN	82,115	117,798	146,347	1.78
N-enWN	88,381	125,819	155,437	1.76
N-plWN	123,985	126,746	167,243	1.35
V-PWN	13,767	11,529	25,047	1.81
V-enWN	13,789	11,540	25,061	1.82
V-plWN	21,669	17,398	31,841	1.47
A-PWN	18,156	21,785	30,004	1.65
A-enWN	18,185	21,808	30,072	1.65
A-plWN	39,204	27,041	45,899	1.17
Adv-PWN	3,625	4,475	5,592	1.54
Adv-enWN*	3,625	4,475	5,592	1.54
Adv-plWN	8,080	5,719	10,416	1.29
GermaNet	101,371	119,231	131,814	–
PWN	117,659	155,593	206,978	1.74
enWN	124,266	164,032	216,623	1.73
plWN	**193,286**	**177,003**	**255,733**	1.32

Table 1: The count by part of speech (PoS) of Noun/Verb/Adjective synsets, lemmas and lexical units (LUs), and average synset size (avs), in PWN 3.1 (PWN), enWordNet 0.1 (enWN), plWordNet 3.0 (plWN) and GermaNet 10.0 (www.sfs.uni-tuebingen.de/GermaNet/).

*This part of WordNet remains to be extended.

5 Statistics

Wordnets are treated as basic lexical resources, so their sizes matter a lot for potential applications. See Table 1 for the general statistics in plWordNet 3.0-beta-emo and a comparison with the other very large wordnets. We note that plWordNet has been consistently expanded in all parts of speech (PoS). The ratio between the size of plWordNet and Princeton WordNet is roughly the same for all PoS. The development of enWordNet has been intentionally concentrated on nouns.

Moreover, plWordNet has become larger than all modern dictionaries of general Polish in terms of the entries included: 130k (Zgółkowa, 1994 2005), 125k [180k lexical units] (Doroszewski, 1963 1969), 100k [150k lexical units] (Dubisz,

2004), 45k [100k lexical units] (Bańko, 2000). One of the main reasons is that those dictionaries do not contain many specialised words and senses from science, technology, culture and so on. Such material, however, is appropriate for a wordnet due to its applications in processing of texts of many genres coming from different sources, including the Internet. We could also observe that lemma lists added to plWordNet (based on the corpus) included quite a few words that are now frequent, but not described in those dictionaries.

The largest ever Polish dictionary, from the early 1900s, has 280k entries (Karłowicz et al., 1900 1927; Piotrowski, 2003, p. 604) and is still much larger than plWordNet, but it contains many archaic words, perhaps useful in the processing of texts from specialised domains. The achieved size of plWordNet has already exceeded the target size estimated for it considering a corpus of 1.8 billion words (Maziarz et al., 2014).

Lexico-semantic relations are the primary means of description of lexical meanings represented in a wordnet by synsets. The average number of relation links per synset, which is called *relation density*, tells us about the average amount of information provided by the wordnet for a single lexical meaning. Table 2 compares the relation density in Princeton WordNet and plWordNet for different parts of speech (obligatory inverse relations have been excluded from the count).[12] The relation density is higher in plWordNet for all parts of speech. We can name two reasons for this difference: smaller synsets in plWordNet on average, see Table 1, and the assumed way of defining synsets by the constitutive relations – more relations are needed to distinguish different synsets (*i.e.,* lexical meanings). However, plWordNet has a rich set of relations (more than 40 main types and 90 sub-types). Some of them have originated from the derivational relations. That can also increase the relation density.

If a wordnet is treated as a reference source, we expect to find in it most of the lemmas from the processed text. The complete coverage is not possible, but the higher it is, the more information a wordnet provides for the analysed text. Table 3 compares the coverage of Princeton WordNet and plWordNet for two corpora of a comparable size. From both corpora, two lemma frequency lists were extracted. Both corpora were first morphosyntactically tagged and only lemmas of the parts of speech described in wordnets were taken into account. For Polish, we worked with the plWordNet corpus (version 7) of ≈1.8 billion words from several available corpora (see section 3.2), supplemented by texts collected from the Internet. As an English corpus, we took texts from the English Wikipedia, ≈1.2 billion words, a size similar to that of the plWordNet corpus.[13]

POS	Princeton WordNet	plWordNet
nouns	2.5	3.17
verbs	3.32	3.95
adjectives	3.05	3.20
adverb	0.88	4.53

Table 2: Synset relation density in Princeton WordNet 3.1 and in plWordNet 2.0 by part of speech.

FRC	≥1000	≥500	≥200	≥100	≥50
PWN	0.383	0.280	0.170	0.107	0.064
plWN	0.732	0.644	0.515	0.416	0.327

Table 3: Percentage of Princeton WordNet noun lemmas in *Wikipedia.en* and plWordNet (plWN) lemmas in the plWordNet corpus. FRC is lemma frequency in the reference corpus.

The coverage is much higher for plWordNet, but the corpora differ. Many more specialised and rare words appear in English Wikipedia than in the Polish corpus. Even so, the statistics bode well for plWordNet's potential in applications. The coverage for the most frequent words (≥ 1000) is not 100% because the list includes many proper names and misspelled words recognised by the tagger as common words. In comparison with plWordNet 2.1 (Maziarz et al., 2013b), the coverage of less frequent words increased significantly, because the development of plWordNet moves towards the bottom of the frequency ranking list.

The average polysemy – the ratio of lexical units to lemmas – is higher in plWordNet than in WordNet both for nouns (1.32 *vs* 1.24) and adjectives (1.71 *vs* 1.38). The difference is lower than in

[12] The relation structures differ among the parts of speech, so we do not show relation density for the whole wordnets.

[13] We used the plWordNet corpus to build the wordnet and to evaluate it. This may suggest a biased comparison. WordNet is evaluated on a corpus unrelated to its development, so only a qualitative comparison is warranted. Regardless, both wordnets more willingly absorb frequent than infrequent lemmas (Maziarz et al., 2013b).

plWordNet 2.1: we added more specific monosemous lemmas as a result of the focus given to lexical units and the tendency to describe exhaustively all existing lexical units for a given lemma. For verbs we have 1.83 *vs* 2.17, maybe because of aspect and rich derivation in Polish verbs.

The comparison of hypernymy path lengths did not change much from plWordNet 2.1 (Maziarz et al., 2013a). WordNet's much longer paths are caused by the elaborate topmost part of its hypernymy hierarchy; plWordNet has ≈100 linguistically motivated hypernymy roots.[14]

6 Applications

Wordnet-building costs a lot of public money, so as a rule the effect should be free for the public use. This good rule, grounded in Princeton WordNet's practice, is central for languages other than English, still less resourced. The availability of plWordNet on the WordNet-style open licence has stimulated, over the years, many interesting applications in linguistic research, language resources and tools, scientific applications, commercial applications and education.

The plWordNet Web page and Web service have had tens of thousands of visitors, and hundreds of thousands of searches. There are over 100 citations and over 700 users, individual and institutional, who optionally registered when downloading the plWordNet source files. Most of the registered users described the intended use of plWordNet, and a rich tapestry it is. The limited space only allows us to single out a handful in citations.

First of all, plWordNet has been applied in linguistic research: valency frame description and automated verb classification; verb analysis for semantic annotation in a corpus of referential gestures; contrastive/comparative studies, *etc*.

Increasingly often, plWordNet is treated as a large monolingual and bilingual dictionary, *e.g.,* in text verification during editing or as a source of meta-data for publications. Miłkowski (2010) included plWordNet among the dictionaries in a proofreading tool and as a knowledge source for an open Polish-English dictionary, which many translators and translation companies say they use. Open Multilingual Wordnet (Bond, 2013) now includes plWordNet. It is referred to in several other projects on wordnets and semantic lexicons (Pedersen et al., 2009; Lindén and Carlson, 2010; Borin and Forsberg, 2010; Mititelu, 2012; Zafar et al., 2012; Šojat et al., 2012). Practical machine translation systems use plWordNet. We are aware of applications in measuring translation quality and building the MT component embedded in an application supporting English teaching to children.

There are more research and commercial projects, both under way and announced by plWordNet users. They include ontology building and linking, information retrieval, question answering, text mining, semantic analysis, terminology extraction, word sense disambiguation (WSD), text classification, sentiment analysis and opinion mining, automatic text summarisation, speech recognition, or even the practice of aphasia treatment.

7 The lexicographer's work is never done

When in 2012 we established the target size of plWordNet 3.0, we were convinced that we would go to limits of the Polish lexical system. We now see that – even if major paths have been explored – we are discovering numerous smaller paths going deeper into the system.

The Polish side of plWordNet could have more relation links per synset. The constitutive relations do not differentiate all hypernymy leaves yet. There are cross-categorial relations, more numerous than in many other wordnets, but still not enough for WSD or semantic analysis. The connection to the valency lexicon could be tighter. The description of verb derivation (as highly productive in Polish as in other Slavic languages) needs much more work, and so do some relations, *e.g.,* meronymy. More information useful for WSD could be introduced, *e.g.,* further glosses or links to external sources like Wikipedia. Finally, for applications in translation (manual and machine-based) we must not only complete the mapping to WordNet, but also go inside synsets, *i.e.,* map lexical units. We are fortunate to have so much more intriguing work to do.

Acknowledgment

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

[14] They do not have hypernyms according to the definitions assumed in plWordNet.

References

[Bańko2000] Mirosław Bańko, editor. 2000. *Inny słownik języka polskiego PWN [Another dictionary of Polish]*, volume 1-2. Polish Scientific Publishers PWN.

[BNC2007] BNC. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

[Bond2013] Francis Bond. 2013. Open Multilingual Wordnet. http://compling.hss.ntu.edu.sg/omw/.

[Borin and Forsberg2010] Lars Borin and Markus Forsberg. 2010. From the People's Synonym Dictionary to fuzzy synsets – first step. In *Proc. LREC 2010*.

[Broda and Piasecki2011] Bartosz Broda and Maciej Piasecki. 2011. Evaluating LexCSD in a large scale experiment. *Control and Cybernetics*, 40(2):419–436.

[Calzolari et al.2012] Nicoletta Calzolari et al., editor. 2012. *Proc. Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association.

[Doroszewski1963 1969] Witold Doroszewski, editor. 1963–1969. *Słownik języka polskiego [A dictionary of the Polish language]*. Państwowe Wydawnictwo Naukowe.

[Dubisz2004] Stanisław Dubisz, editor. 2004. *Uniwersalny słownik języka polskiego [A universal dictionary of Polish], electronic version 1.0*. Polish Scientific Publishers PWN.

[Ferraresi et al.2008] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proc. 4th Web as Corpus Workshop (WAC-4)*, pages 47–54.

[Hamp and Feldweg1997] Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proc. ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Madrid.

[Isahara and Kanzaki2012] Hitoshi Isahara and Kyoko Kanzaki, editors. 2012. *Advances in Natural Language Processing: Proc. 8th International Conference on NLP, JapTAL*, volume 7614 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag.

[Karłowicz et al.1900 1927] Jan Karłowicz, Adam Antoni Kryński, and Władysław Niedźwiedzki, editors. 1900-1927. *Słownik języka polskiego [A dictionary of the Polish language]*. Nakładem prenumeratorów i Kasy im. Józefa Mianowskiego [Funded by subscribers and Józef Mianowski Fund], Warsaw.

[Kędzia et al.2013] Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka, and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies*. to appear.

[Kędzia and Piasecki2014] Paweł Kędzia and Maciej Piasecki. 2014. Ruled-based, interlingual motivated mapping of plwordnet onto sumo ontology. In Nicoletta Calzolari et al., editor, *Proc. Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association.

[Lindén and Carlson2010] Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNorcdica*, 17.

[Maziarz et al.2013a] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013a. Beyond the Transfer-and-Merge Wordnet Construction: plWordNet and a Comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Proc. International Conference on Recent Advances in Natural Language Processing*. Incoma Ltd.

[Maziarz et al.2013b] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013b. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In *Proc. International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452. INCOMA Ltd. Shoumen, BULGARIA.

[Maziarz et al.2013c] Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013c. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

[Maziarz et al.2014] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proc. Seventh Global Wordnet Conference*, pages 304–312.

[Maziarz et al.2015] Marek Maziarz, Stan Szpakowicz, and Maciej Piasecki. 2015. A procedural definition of multi-word lexical units. In *Proc. RANLP 2015*, page to appear.

[Miłkowski2010] Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software – Practice and Experience*.

[Mititelu2012] Verginica Barbu Mititelu. 2012. Adding Morpho-semantic Relations to the Romanian Wordnet. In *Proc. LREC 2012*.

[Miłkowski2007] Marcin Miłkowski. 2007. Open Thesaurus - polski Thesaurus. http://www.synomix.pl/.

[Pease2011] Adam Pease. 2011. *Ontology - A Practical Guide*. Articulate Software Press.

[Pedersen et al.2009] Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*.

[Piasecki et al.2009] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. http://www.eecs.uottawa.ca/ szpak/pub/A_Wordnet_from_the_Ground_Up.zip.

[Piasecki et al.2010] Maciej Piasecki, Michał Marcińczuk, Adam Musiał, Radosław Ramocki, and Marek Maziarz. 2010. WordnetLoom: a Graph-based Visual Wordnet Development Framework. In *Proc. Int. Multiconf. on Computer Science and Information Technology – IMCSIT 2010, Wisła, Poland, October 2010*, pages 469–476.

[Piasecki et al.2012a] Maciej Piasecki, Radosław Ramocki, and Marek Maziarz. 2012a. Automated Generation of Derivative Relations in the Wordnet Expansion Perspective. In *Proc. 6th Global Wordnet Conference*.

[Piasecki et al.2012b] Maciej Piasecki, Radosław Ramocki, and Paweł Minda. 2012b. Corpus-based semantic filtering in discovering derivational relations. In Allan Ramsay and Gennady Agre, editors, *Proc. 15th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, volume 7557 of *Lecture Notes in Computer Science*, pages 14–42. Springer.

[Piasecki et al.2013] Maciej Piasecki, Radosław Ramocki, and Michal Kaliński. 2013. Information spreading in expanding wordnet hypernymy structure. In *Proc. International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 553–561. INCOMA Ltd. Shoumen, BULGARIA.

[Piotrowski2003] Tadeusz Piotrowski, 2003. *Współczesny język polski [Contemporary Polish], edited by Jerzy Bartmiński*, chapter Słowniki języka polskiego [Dictionaries of Polish]. Marie Curie-Sklodowska University Press.

[Plutchik1980] Robert Plutchik. 1980. *EMOTION: A Psychoevolutionary Synthesis*. Harper & Row.

[Przepiórkowski et al.2012] Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [in Polish]*. Wydawnictwo Naukowe PWN. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

[Przepiórkowski2004] Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences.

[Puzynina1992] Jadwiga Puzynina. 1992. *Język wartości [The language of values]*. Scientific Publishers PWN.

[Rudnicka et al.2012] Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A Strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proc. COLING 2012, posters*, pages 1039–1048.

[Šojat et al.2012] Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, 0(1):111–142.

[Vetulani et al.2009] Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Jacek Marciniak, Paweł Konieczka, and Przemysław Rzepecki. 2009. An Algorithm for Building Lexical Semantic Network and Its Application to PolNet – Polish WordNet Project. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society, Third Language and Technology Conf., Poznań, Revised Selected Papers*, LNCS 5603, pages 369–381. Springer.

[Weiss2008] Dawid Weiss. 2008. Korpus Rzeczpospolitej [Corpus of text from the online edition of "Rzeczpospolita"]. http://www.cs.put.poznan.pl/dweiss/rzeczpospolita.

[Woliński2006] Marcin Woliński. 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag.

[Zafar et al.2012] Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, and Asad Mustafa. 2012. Developing Urdu WordNet Using the Merge Approach. In *Proc. Conference on Language and Technology*, pages 55–59.

[Zaśko-Zielińska et al.2015] Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A Large Wordnet-based Sentiment Lexicon for Polish. In *Proc. RANLP 2015*, page to appear.

[Zgółkowa1994 2005] Halina Zgółkowa, editor. 1994–2005. *Praktyczny słownik współczesnej polszczyzny [A practical dictionary of contemporary Polish]*. Wydawnictwo Kurpisz.

Open Dutch WordNet

Marten Postma
VU Amsterdam
Amsterdam, The Netherlands
m.c.postma@vu.nl

Emiel van Miltenburg
VU Amsterdam
Amsterdam, The Netherlands
emiel.van.miltenburg@vu.nl

Roxane Segers
VU Amsterdam
Amsterdam, The Netherlands
roxane.segers@gmail.com

Anneleen Schoen
VU Amsterdam
Amsterdam, The Netherlands
a.m.schoen@vu.nl

Piek Vossen
VU Amsterdam
Amsterdam, The Netherlands
piek.vossen@vu.nl

Abstract

We describe Open Dutch WordNet, which has been derived from the Cornetto database, the Princeton WordNet and open source resources. We exploited existing equivalence relations between Cornetto synsets and WordNet synsets in order to move the open source content from Cornetto into WordNet synsets. Currently, Open Dutch Wordnet contains 117,914 synsets, of which 51,588 synsets contain at least one Dutch synonym, which leaves 66,326 synsets still to obtain a Dutch synonym. The average polysemy is 1.5. The resource is currently delivered in XML under the CC BY-SA 4.0 license[1] and it has been linked to the Global Wordnet Grid. In order to use the resource, we refer to: https://github.com/MartenPostma/OpenDutchWordnet.

1 Introduction

The main goal of this project is to convert the Dutch lexical semantic database Cornetto version 2.0 (Vossen et al., 2013) into an open source version. Cornetto is currently not distributed as open source, because a large portion of the database originates from the commercial publisher Van Dale.[2] The main task of this project is hence to replace the proprietary content of the database with open source content. In order to create Open Dutch WordNet, we used all the synsets and relations from WordNet 3.0 (Fellbaum, 1998) as our basis. We then exploited existing equivalence relations between Cornetto synsets and WordNet synsets in order to replace WordNet synonyms by Dutch synonyms. We further added new concepts that were not matched through hyperonym relations to the WordNet hierarchy. Any new and manually-created semantic relation from Cornetto was added to the database as well. We limited the synonyms, concepts and relations to those on which there are no copy-right claims. In addition, the inter-language links in various external resources were used to add synonyms to the resource. The result is an open source wordnet that combines the merge and expand method described in (Vossen, 1999).

The resource is currently delivered in XML under the CC BY-SA 4.0 license.[3] In order to inspect and improve the resource, a Python module has been created. This module can be found at : https://github.com/MartenPostma/OpenDutchWordnet.

The outline of this paper is as follows. We start with the motivation to create Open Dutch WordNet in section 2, followed by the methodology to create the resource in section 3. An overview of the main components will be provided in section 4. Finally, we discuss the process of making the resource and plans to improve the resource in section 5.

2 Background and motivation

The first version of the Dutch WordNet was developed within the EuroWordNet project starting from a database developed by Van Dale publisher. This database already contained synset-like structures and lexical semantic relations that could be used to efficiently derive a wordnet structure. Licenses were agreed for commercial and research usage. The Dutch WordNet and the Referentie Bestand Nederlands (RBN) (Van der Vliet, 2007) were combined in the Cornetto project (Vossen et al., 2013). RBN has detailed information on

[1] https://creativecommons.org/licenses/by-sa/4.0/

[2] http://www.vandale.nl/

[3] https://creativecommons.org/licenses/by-sa/4.0/

morpho-syntactic, semantic and pragmatic properties of lexical units, with a focus on the combinatorics. The Cornetto database thus provides the semantic organization of a wordnet and the details on each synonym in a synset as can be found in lexical unit based lexicons. An important characteristic of Cornetto is that it has been developed independently from Princeton WordNet (PWN). The synsets in Cornetto were then mapped to synsets in PWN following a merge approach (Vossen, 1999). First, all possible equivalence relations were created between synonyms in synsets using bilingual dictionaries, after which the mappings were ranked on the basis of shared properties, e.g. hyperonyms and hyponyms already linked manually, similar domain labels, and synset membership of multiple translations (Vossen et al., 2008). The Van Dale publisher however decided to stop all collaborations with the research community. This motivated us to develop Open Dutch WordNet, for which we wanted to keep as much as possible the concepts and word meanings that are defined independently of PWN. This implies that we cannot simply follow an expand approach to translate English synonyms in PWN to Dutch words but we need to also match PWN synsets to RBN lexical units.

Figure 1 introduces the main components of the Dutch lexical semantic database Cornetto.

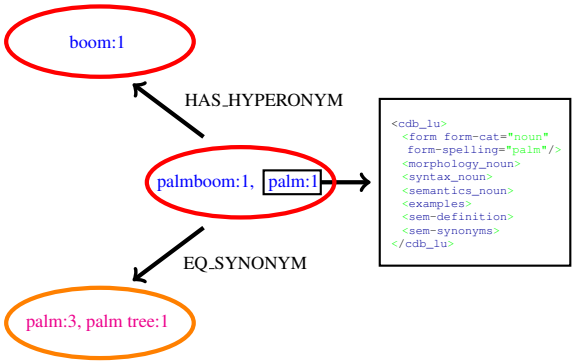


Figure 1: The most important components of Cornetto are visualized. The ellipses in red are examples of Cornetto synsets, which contain Lexical Units (LU). Each LU can contain rich information about its morphology, syntax and semantics. Cornetto synsets can have Internal Semantic Relations (ISRs) to other Cornetto synsets (e.g. HAS_HYPERONYM), but also Equivalence Semantic Relations (ESRs) to PWN synsets (e.g. EQ_SYNONYM).

Figure 1 visualizes the most important components of Cornetto. Cornetto synsets, or Cornetto sets of synonyms, are shown in red. The synonyms inside the Cornetto synsets are called Lexical Units (LU), because they can contain rich information about its morphology, syntax and semantics, especially if these LU's originate from RBN. Synonyms that originate from the Van Dale database only have part-of-speech information. Cornetto synsets can have Internal Semantic Relations (ISRs) to other Cornetto synsets (e.g. HAS_HYPERONYM), but also Equivalence Semantic Relations (ESRs) to PWN synsets (e.g. EQ_SYNONYM). ESRs are mainly used to define synonymy or near synonymy between Cornetto synsets and PWN synsets. Most ISR relations originate from the Van Dale database. A small set of relations were added manually in the various projects. All synonyms and relations have provenance tags which enables us to trace data from Van Dale and data that can transferred to the Open Dutch WordNet.

Table 1 presents the provenance statistics for the most important components of the database:

Component	Van Dale	RBN	Cornetto
LU	60	57	1.5
S	70	1	0
ISR	77	0	33
ESR	0	0	82

Table 1: The provenance information for Lexical Units (LU), Synsets (S), Internal Semantic Relations (ISR), and Equivalence Semantic Relations (ESR) is shown for each of the three sources: Van Dale, RBN, and Cornetto (if the source is Cornetto, this means that the data was created manually in the Cornetto project and does not originate from Van Dale).

Table 1 clearly shows that a large part of the LU's, synsets, and ISRs originate from Van Dale. The removal of this licensed content creates large gaps in the resource. The main goal is hence to use open source resources to replace the licensed content with open source content as much as possible. One of the most promising components to transfer information from Cornetto into Open Dutch WordNet are the ESRs that were created semi-

automatically during the EuroWordNet and Cornetto project and are 100% open source.

3 Methodology

We used the following procedure to create Open Dutch WordNet.

We use English WordNet3.0 (PWN) (Miller, 1995; Fellbaum, 1998) as our basis for the concept structure. This means that we copied the PWN synsets and relations to ODWN and ignored all synsets and relations from Van Dale. The next step is to transfer the LU's from RBN to the PWN-based synsets.

Before copying these LU's we improved the quality of the ESRs. We defined a set of ESRs that are either likely to be more difficult or that play an important role in the transfer. This subset was checked manually and was also used as training to filter the remaining ESRs using a decision tree algorithm. This process is described in subsection 3.1.

Subsequently, we make use of the ESRs between Cornetto synsets and WordNet synsets to copy the LU's that do not originate from Van Dale from a Cornetto synset into a WordNet synset, which is described in subsection 3.2.

The transfer still leaves us with many synsets from PWN without a Dutch LU. We therefore use open source resources to translate the WordNet synonyms into Dutch, which is described in subsections 3.3 and 3.4, respectively. This results on the one hand in more synsets to have Dutch synonyms but also in further evidence for transferred synonyms to be correct because of evidence through other sources.

Finally, we manually checked 8,257 Dutch synonyms, which is described in subsection 3.5.

3.1 Revision of equivalence relations

Firstly, we manually filtered the ESRs, from which we focused on the synonymy relations. Each ESR links a Cornetto synset to a WordNet synset with a certain relation type. The mapping of an ESR is one of many to many. We considered three main aspects of Cornetto synsets in deciding whether to manually check an ESR: the synset depth, the number of children, and the number of ESRs. We decided to manually check the deepest and shallowest synsets because these relations got little attention in previous projects. In addition, we checked the synsets with most children because they play an important role in a wordnet. Finally, the Cornetto synsets with most ESRs were checked because we suspect that the equivalence relation is complex and likely to contain many wrong mappings. Four students manually checked 12,966 of the total 82,285 ESRs, of which 6,575 were removed.

The manually revised relations were used to train a pruned C4.5 decision tree algorithm (Quinlan, 1993; Hall et al., 2009) that was used to filter the remaining ESRs. An ESR consists of an equivalence relation between a Cornetto synset and a WordNet synset. We used properties of the Cornetto synset and the WordNet synset as well as of the synset relation itself as features.

1. the number of equivalence relations in which a Cornetto synset and a Wordnet synset are present.

2. the depth of the Cornetto synset and the Wordnet synsets. The difference of the depth is also used.

3. Because a Cornetto synset can be present in multiple ESRs to WordNet synsets and vice versa, we average the semantic similarity scores (using the Leacock & Chodorow similarity measure (Leacock and Chodorow, 1998)) of of all combinations of these ESRs.

Interestingly enough, the features in which Cornetto properties were used yielded the best results. This might be caused by the fact that the relations were also generated using Cornetto. The filtering of the ESRs using the decision tree algorithm resulted in an additional removal of 32,258 ESRs.

3.2 Cornetto synonyms

When there exists an ESR between a Cornetto synset and a WordNet synset and the relation type is either EQ_SYNONYM or EQ_NEAR_SYNONYM, all LU's that do not originate from Van Dale are inserted into the WordNet synset. Using figure 1 as an example, the LU's *palmboom:1* and *palm:1* would replace *palm tree:1* and *palm:3*. If the ESR was checked manually, the provenance tag is **cdb2.2_Manual**. If the ESR was checked using the decision tree algorithm, the provenance tag is **cdb2.2_Auto**. The provenance tag **cdb2.2_None** is given to all other strategies that were used to add LU's to

Open Dutch WordNet. One of the most dominant strategies of this class is when a LU in a Cornetto synset does not have a direct ESR (no ESR or one of EQ_HAS_HYPERONYM) to a WordNet Synset but the parent of the Cornetto synset does have an ESR to a WordNet synset. In that case a new synset (not represented in WordNet) is created as a hyponym of the target of the ESR of the hyperonym. Finally, the ESRs are used to insert Cornetto synset relations into Open Dutch WordNet that do not originate from Van Dale but were created manually in one of the projects.

3.3 External resources

Using various external open source resources such as Wiktionary (Foundation, 2014b), Omegawiki [4], and Google (Google, 2014), Oliver (2014) translated both monosemous and polysemous lemmas into Dutch for the part of speeches noun, verb, and adjective. For the monosemous lemmas, the English lemmas are simply translated into Dutch. For the polysemous lemmas, the gloss overlap between examples in an external resource and the possible WordNet synsets for a lemma are used to determine the correct synset for a lemma. We used a similar procedure to add synonyms from Wikipedia (Wikipedia, 2014; Foundation, 2014a).

3.4 Adjectives extended

We created a mapping for two kinds of adjectives: monosemous adjectives, that have only one sense in WordNet, and 'slightly polysemous adjectives' that have exactly one adjectival sense and one nominal sense. Adjectives of the latter kind are typically nationalities (*Cameroonian*), religious denominations (*Buddhist*), and words like *purebred*. To create the mapping, we translated the English word forms using Google Translate and Bing Translate. We also use the word alignments from the OPUS project (Tiedemann, 2012). These resources provide us with Dutch candidate word forms that should correspond to the original WordNet synonyms in synsets. We then checked for each word form how many senses are associated with them in RBN. If there is only one (and the word is indeed an adjective), we conclude that this Dutch sense corresponds with the original WordNet synset.

One problem with the translation-based approach is that Dutch adjectives are sometimes inflected with the suffix *-e*. For example, the English *ontological* is automatically translated by Google to *ontologische*. In RBN, all word forms are stored without the inflectional ending, which means that the translation does not match the lemma. To solve this issue, in the cases where we could not find a direct match, we applied an automatic stemming rule to remove the suffix and tried to find a match using the stem.

[4] http://www.omegawiki.org/

3.5 Manual editing

Finally, we checked the resulting Dutch wordnet manually. We focused on two main editing tasks. Firstly, we inspected all synsets that had 10 or more synonyms since excessive synsets may contain false synonyms. In addition, because one Cornetto synset could have multiple ESRs, it occurred that the same sense was copied into multiple WordNet synsets. This may lead to excessive polysemy. The second task therefore consisted of indicating which WordNet synset was the correct synset for a sense that occurred in more than one WordNet synset. In total, 8,257 LU's were checked in this phase.

4 Overview and statistics

In this section, we provide an overview of Open Dutch Wordnet in terms of general statistics, the format it is delivered in, evaluation, and a Python module which allows to interact with the resource.

Open Dutch Wordnet contains 117,914 synsets, of which the majority are noun synsets: 98,049. There are 18,782 verb synsets and 1,083 adjectival synsets. 51,588 synsets contain at least one Dutch synonym, which leaves 66,326 synsets still to obtain a synonym. The resource contains 92,295 synonyms, of which 75,173 are nouns, 15,979 are verbs, and 1,143 are adjectives. The average polysemy is 1.5. 19,996 relations were added to the WordNet hierarchy.

4.1 Format

Open Dutch WordNet is stored in a type of XML called Global WordNet Grid LMF (`https://github.com/globalwordnet/schemas`), which is an adaptation of WordnetLMF (Vossen et al., 2012). The XML contains two main elements: LexicalEntry and Synset. LexicalEntry elements contain information about a specific synonym, whereas Synset elements contain information about synsets. A simplified example

of a LexicalEntry element can be found in figure 2:

```
<LexicalEntry id="ondernemer-n-1"
                partOfSpeech="noun">
      <Lemma writtenForm="ondernemer"/>
      <Sense
       id="r_n-25922"
       senseId="1"
       definition="iemand met eigen bedrijf"
       synset="eng-30-10060352-n"
       provenance="cdb2.2_Auto+wiktionary+google"
       annotator="">
</LexicalEntry>
```

Figure 2: A simplified example of a LexicalEntry element is shown.

In figure 2, an example of a LexicalEntry element is shown. The attributes **id** and **partOfSpeech** of the LexicalEntry element indicate the identifier and the part of speech, respectively. In this example, the identifier is *ondernemer-n-1*, which refers to the first noun sense of the Dutch translation of *entrepreneur* in the sense of "someone who organizes a business venture and assumes the risk for it". The attribute **writtenForm** of the element Lemma indicates the lemma. Following the structure of Cornetto, the LexicalEntry structure represents a lexical unit and not a form unit. The motivation for this is that form properties can differ from one meaning to another for a lemma. The same form can thus appear in multiple LexicalEntry elements.

Finally, the Sense element contains five attributes:

1. **senseId** refers to the synonym sense number.

2. **id** stores the synonym sense identifier. If the identifier starts with *r*, the synonym originates from RBN. In this case, more information about the synonym can be found in RBN. In all other cases, this is not available.

3. **definition** presents the definition for the sense.

4. **synset** points to the synset to which this synonym belongs.

5. Concatenated by '+', the attribute **provenance** shows which resources proposed this particular synonym for this particular synset.

6. the attribute **annotator** shows the name of an annotator and marks that the synonym has been checked manually. The default value is an empty string. Currently, 6,370 LexicalEntry elements have been checked manually.

The LexicalEntry used in Figure 2 belonged to the synset "eng-30-10060352-n". Figure 3 presents a simplified example of that Synset element.

```
<Synset id="eng-30-10060352-n"
                  ili="i89775">
      <Definitions>
         <Definition
         gloss="iemand met eigen bedrijf"
         language="nl"
         provenance="odwn"/>

         <Definition
         gloss="someone who organizes
         a business venture and
         assumes the risk for it"
         language="en"
         provenance="pwn"/>

      <SynsetRelations>
         <SynsetRelation
         provenance="pwn"
         relType="has_hyperonym"
         target="eng-30-09882716-n"/>

         <SynsetRelation
         provenance="odwn"
         relType="role_agent"
         target="eng-30-01651293-v"/>
         ....
         </SynsetRelations>
</Synset>
```

Figure 3: A simplified example of a Synset element is shown.

In figure 3, a simplified example is shown of a Synset element. The Synset attributes **id** and **ili** provide information about the synset identifier and the interlingual index identifier, respectively: `http://data.lider-project.eu/ili`.

The elements Definitions/Definition provide information about the **gloss**, **language**, and **provenance** of the definitions. Finally, the element SynsetRelations/SynsetRelation stores the information about the relations between synsets. Again the **provenance** attribute is used to mark whether the relation originates from PWN or from Cornetto.

4.2 Analysis Lexical Entries

Open Dutch WordNet contains 92,295 synonyms, originating from various resources. Table 2 presents information about the number of synonyms from each resource:

Table 2 presents the number of synonyms proposed by each resource. Note that the same synonym can be proposed by multiple resources, which is why the sum of all numbers is higher than

Provenance	instances	% of all LE
cdb2.2_Auto	32806	35.5
cdb2.2_None	19073	20.7
wiktionary	17968	19.5
cdb2.2_Manual	13075	14.2
omegawiki	12589	13.6
google	8374	9.1
opus	612	0.7
bing	506	0.5
wikipedia	375	0.4

Table 2: The number of synonyms from each resource is shown. In addition, the second column indicates what percentage this number is relative to all synonyms in Open Dutch Wordnet.

the total number of synonyms. The vast majority of synonyms originate from the ESRs (prefixed by *cdb2.2*) between Cornetto synsets and WordNet synsets.

In order to evaluate the quality of each resource for the creation of Open Dutch Wordnet, we randomly evaluated 50 monosemous and polysemous instances. The results can be found in table 3:

Provenance	m	p
Google	0.84	NA
Wiktionary	0.86	0.68
Wikipedia	0.88	0.62
Omegawiki	0.90	0.86
Cdb2.2_Manual	0.88	0.74
Cdb2.2_Auto	0.80	0.80
Cdb2.2_None	0.96	0.78

Table 3: The evaluation results of randomly selected 50 monosemous (m) and polysemous (p) instances per resources is shown.

Table 3 shows that the overall precision of the resource is high as far as the quality of a synonym that bears a certain provenance is concerned. What it does not show, is a fair comparison of the quality of each resource, because not exactly the same strategy was used to extract information from each resource. For example, only monosemous words were used from the output from Google. Overall, we observe that 87% of the proposed monosemous synonyms were correct in the evaluation, whereas this was 76% for the polysemous synonyms. The most valuable external resource for Open Dutch WordNet seems to be Omegawiki, which is not only present in 13.6% of the LexicalEntry elements, but also performed well in the evaluation. For comparison, Sevens (Sevens et al., 2014) performed an independent evaluation of the equivalence relations in Cornetto and reported precision of 52.18% for a sample based on all synsets and 88.94% for a subset that was likely to have manually created links. Although it is difficult to compare both samples for evaluation, the precision for Open Dutch Wordnet is thus very much in line with the precision of Cornetto as reported by them.

4.3 Depth Distribution

66,326 synsets in Open Dutch Wordnet still lack a synonym. We were interested in knowing in which part of the hierarchy these synsets were located. Breadth-first search was used to calculate synset depth. Figure 4 presents the distribution of synsets with and without synonyms per depth layer.

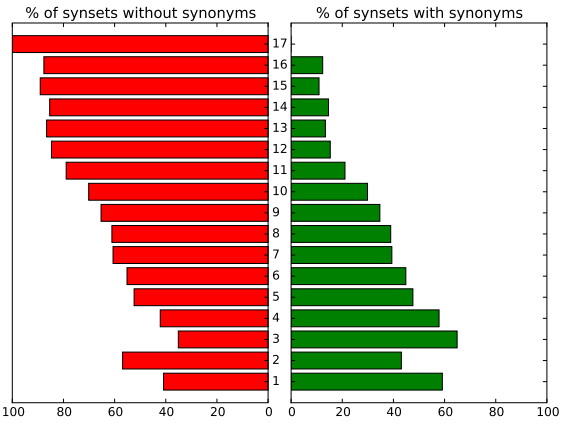


Figure 4: For each depth layer in Open Dutch WordNet, which ranges from the top level 1 to the most deepest layer 17, the percentage of synsets in that layer with and without synonyms is shown.

Figure 4 presents the distribution of synsets with and without synonyms per depth layer. In general, we observe that the top layers have relatively few synsets without synonyms, whereas the opposite is true for the deeper layers. It is likely that these lower level synsets can be filled easily if bilingual resources extend their coverage. These words usually have a single meaning and only one translation.

Also the opposite situation occurs that we added new synsets to the hierarchy that are not in WordNet. These synsets appear to be spread

over all levels of the hierarchy. It is more difficult to resolve these cases since searching for possible matches in WordNet that could have been missed can only partially be supported through e.g. gloss-comparison but in the end needs to be verified manually. To support this process, we visualized these concepts in the hierarchy. An example can be found in Figure 5.

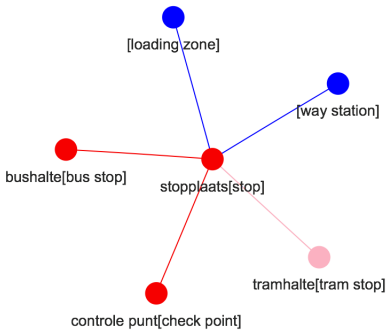


Figure 5: In this visualisation, pink nodes are new concepts, red nodes are WordNet synsets with Dutch synonyms and blue nodes are WordNet synsets without Dutch synonyms.

Figure 5 presents an example of a new concept that has been added to the hierarchy. We added the concept of *tramhalte* (tram stop) as a hyponym of the concept 'stop'. In general, we observed that we mostly added concepts that are represented in Dutch by compounds, such as *polderlandschap* (flat, barren landscape).

4.4 Python module

A Python module has been created to use Open Dutch WordNet. The module can be found at `https://github.com/MartenPostma/OpenDutchWordnet`. It is designed in Python 3.4. The module allows the user to inspect the LexicalEntry and Synset elements and to gather general statistics about the resource. Finally, it is possible to edit the resource using this module.

5 Discussion and future work

In this section, we discuss the process of creating Open Dutch WordNet as well as future work to further improve the resource.

A part of Open Dutch WordNet consists of synonyms that originate from the inter-language links in external resources such as Omegawiki, Wiktionary, and Wikipedia. It is interesting to observe that we obtained mostly noun synonyms

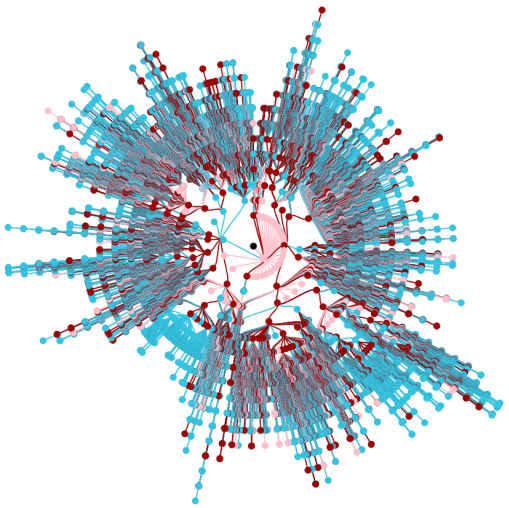


Figure 6: This figure visualizes the noun hyperonym hierarchie in ODWN. The black center node represents the top noun node ('entitiy'). In this visualisation, pink nodes are new concepts, red nodes are WordNet synsets with Dutch synonyms and blue nodes are WordNet synsets without Dutch synonyms.

from these resources. There are two main reasons why this is the case. Firstly, nouns simply have more entries in these resources. In addition, it is obviously more difficult to disambiguate verbs than nouns. In order to get a better understanding of where we added Dutch noun synonyms, we visualized the noun hyperonym hierarchy, which can be found in Figure 6.

In Figure 6, the noun hyperonym hierarchy is visualized, focusing on which synsets contain a Dutch synonym. The lower left side shows a large blue spot, which means that no Dutch synonyms are located in that part of the hierarchy. We identified the synset *genus* ('taxonomic group containing one or more species)' as the main hyperonym of this part. In addition, we observe pink nodes around the top node, which we identified as religious terms such as *Heer* (Lord), and *Jaweh* (Jaweh).

In order to improve the resource, we strive to both improve the quality and quantity of the resource. The quality will be improved by manually inspecting the synsets ranging from 5 to 10 synonyms. The quantity will be improved by adding synonyms in the deeper parts of the resource. This can be done by using more or improved public bilingual resources, both English-Dutch but also

by combining more languages, or by using parallel corpora. In addition, we plan to assess the most important parts of the hierarchy. This involves the top nodes of the hierarchies and the base concepts. Errors in these synsets are likely to propagate to other synsets in lower parts of the hierarchy. Finally, the relations imported from Cornetto are now added to the PWN relations. As a result, we obtained 115,077 hyperonym relations from PWN and 19,996 hyperonym relations from Cornetto. Additional hyperonym relations result in tangled hierarchies with more complex semantics. Whereas PWN has 559 top nodes for verbs, ODWN has 154 tops. The reduction of the tops is due to the additional relations that were created in Cornetto to provide more structure to the verb hierarchy. In Cornetto, there are only two top nodes for the verb hierarchy.

Open Dutch WordNet currently contains a limited amount of monosemous adjectives. We hope to be able to map the polysemous adjective synsets to PWN synsets by translating the Dutch glosses and by making use of the synset relations in Cornetto and Princeton WordNet. Because Dutch is very close to German, another possibility is to map the Cornetto synsets to GermaNet (Hamp and Feldweg, 1997) and make use of the rich set of synset relations that it provides.

Finally, the current format of the resource is XML. We would also like to make the resource available in RDF (Klyne and Carroll, 2006).

6 Conclusion

We described Open Dutch WordNet, which is derived from the Cornetto database, Princeton WordNet and various external resources. We exploited existing equivalence relations between Cornetto synsets and WordNet synsets in order to replace WordNet synonyms by Dutch synonyms. In addition, the inter-language links in various external resources such as Wiktionary and Omegawiki were used to add synonyms to the resource. In addition, we manually evaluated each resource and manually edited the most problematic synsets. The Princeton-based hierarchy was also extended with manually created relations came from Cornetto.

Open Dutch Wordnet contains 92,295 synonyms, which are located in 51,588 synsets. There are 75,173 nouns, 15,979 verbs, and 1,143 adjectives. In total, the resource consists of 117,914 synsets, which leave 66,326 synsets still to obtain a synonym. The average polysemy is 1.5.

The resource is currently delivered in XML under the CC BY-SA 4.0 license.[5] In order to use and improve the resource, a Python module has been created. This module can be found at: `https://github.com/MartenPostma/OpenDutchWordnet`.

Acknowledgments

This project has been co-funded by the Nederlandse Taalunie (`http://taalunie.org/`). In addition, we thank Anne Broekhuis, Anja Stoop, Marjolein Klaassen, and Amber Witsenburg for their work on evaluating the ESRs manually. Moreover, we thank Isa Maks (`https://www.linkedin.com/pub/isa-maks/24/b47/`) and Hennie van der Vliet (`https://www.linkedin.com/pub/hennie-van-der-vliet/0/869/512`) for their valuable input. Finally, we would like to thank Adam Rambousek (`http://www.muni.cz/fi/people/60380`) for his help in creating and updating the DebVisDic editor.

References

Christiane Fellbaum. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Wikimedia Foundation. 2014a. Wikipedia. `http://en.wikipedia.org/`.

Wikimedia Foundation. 2014b. Wiktionary. `http://en.wiktionary.org/`.

Google. 2014. Google translate. `https://translate.google.nl/`.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Graham Klyne and Jeremy J Carroll. 2006. Resource description framework (rdf): Concepts and abstract syntax.

[5] `https://creativecommons.org/licenses/by-sa/4.0/`

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

George A. Miller. 1995. Wordnet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Antoni Oliver. 2014. Wn-toolkit: Automatic generation of wordnets following the expand model. *Proceedings of the 7th Global WordNetConference, Tartu, Estonia*.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Leen Sevens, Vincent Vandeghinste, and Frank Van Eynde. 2014. Improving the precision of synset links between cornetto and princeton wordnet. *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing,Coling 2014, Dublin, Ireland*, pages 120–126.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Hennie Van der Vliet. 2007. The Referentiebestand Nederlands as a multi-purpose lexical database. *International Journal of Lexicography*, 20(3):239–257.

P Vossen, I Maks, R Segers, and H Vliet. 2008. van der, zutphen, h. van,(2008). the cornetto database: the architecture and alignment issues. In *Proceedings of the Fourth International GlobalWordNet Conference-GWC 2008*, pages 22–25.

Piek Vossen, Claudia Soria, and Monica Monachini. 2012. Wordnet-lmf: a standard representation for multilingual wordnets. In G. Francopoulo, editor, *LMF: Lexical Markup Framework, theory and practice*, pages 51–66. Hermes, Lavoisier, ISTE.

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: a Combinatorial Lexical Semantic Database for Dutch. In Jan Odijk Peter Spyns, editor, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 165–184. Springer.

Piek Vossen. 1999. Eurowordnet: General document. version 3 final. *University of Amsterdam. EuroWordNet LE2-4003, LE4-8328*.

Wikipedia. 2014. Plagiarism — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350`.

Verifying Integrity Constraints of a RDF-based WordNet

Fabricio Chalub
IBM Research
fchalub@br.ibm.com

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Abstract

This paper presents our first attempt at verifying integrity constraints of our openWordnet-PT against the ontology for Wordnets encoding. Our wordnet is distributed in Resource Description Format (RDF) and we want to guarantee not only the syntax correctness but also its semantics soundness.

1 Introduction

Lexical databases are organized knowledge bases of information about words. These resources typically include information about the possible meanings of words, relations between these meanings, definitions and phrases that exemplify their use and maybe some numeric grades of confidence in the information provided. The Princeton English Wordnet (Fellbaum, 1998), is probably the most popular model of a lexical knowledge base. Our main goal is to provide good quality lexical resources for Portuguese, making use, as much as possible, of the effort already spent creating similar resources for English. Thus we are working towards a Portuguese wordnet, based on the Princeton model (de Paiva et al., 2012).

In a previous paper (Real et al., 2015) we reported the new web interface[1] for searching, browsing and collaborating on the improvement of OpenWordnet-PT. Correcting and improving linguistic data is a hard task, as the guidelines for what to aim for are not set in stone nor really known in advance. While the WordNet model has been paradigmatic in modern computational lexicography, this model is not without its failings and shortcomings, as far as specific tasks are concerned. Also it is easy and somewhat satisfying to provide copious quantitative descriptions of numbers of synsets, for different parts-of-speech, of triples associated to these synsets and of intersections with different subsets of Wordnet, etc. However, the whole community dedicated to creating wordnets in other languages, the Global WordNet Association[2], has not come up with criteria for semantic evaluation of these resources nor has it produced, so far, ways of comparing their relative quality or accuracy. Thus qualitative assessment of a new wordnet seems, presently, a matter of judgment and art, more than a commonly agreed practice.

Believing that this qualitative assessment is important, and so far rather elusive, we propose that having many eyes over the resource, with the ability to shape it in the directions wanted, is a main advantage. This notion of volunteer curated content, as first and foremost exemplified by Wikipedia, needs adaptation to work for lexical resources.

Our openWordnet-PT was distributed since its beginning in RDF, following the Semantic Web standards proposed by Tim Berners-Lee (Berners-Lee, 1998). Nevertheless, so far, although we make available not only the data but also its model definition in OWL[3], we have not addressed the task to confront the data with its model to guarantee that data is compliance with the defined model. This is the main contribution of this paper.

2 OpenWordnet-PT

The OpenWordnet-PT (Rademaker et al., 2014), abbreviated as OpenWN-PT, is a wordnet originally developed as a projection of the Universal WordNet (UWN) x(de Melo and Weikum, 2009). Its long term goal is to serve as the main lexicon for a system of natural language processing focused on logical reasoning, based on representation of knowledge, using an ontology, such as SUMO (Pease and Fellbaum, 2010).

[1] http://wnpt.brlcloud.com/wn/

[2] http://globalwordnet.org/

[3] https://github.com/own-pt/openWordnet-PT

OpenWN-PT has been constantly improved through *linguistically motivated* additions and removals, either manually or by making use of large corpora. This is also the case for the lexicon of nominalizations, called NomLex-PT, that is integrated to the OpenWN-PT (Freitas et al., 2014). One of the features of both resources is to try to incorporate different kinds of quality data already produced and made available for the Portuguese language, independent of which variant of Portuguese one considers.

The philosophy of OpenWN-PT is to maintain a close connection with Princeton's wordnet since this minimizes the impact of lexicographical decisions on the separation or grouping of senses in a given synset. Such disambiguation decisions are inherently arbitrary (Kilgarriff, 1997), thus the multilingual alignment gives us a pragmatic and practical solution. It is practical because Princeton WordNet remains the most used lexical resource in the world. It is also pragmatic, since those decisions will be more useful, if they are similar to what other wordnets say. Of course this does not mean that all decisions will be sorted out for us. As part of our processing is automated and error-prone, we strive to remove the biggest mistakes created by automation, using linguistic skills and tools. In this endeavor we are much helped by the linked data philosophy and implementation, as keeping the alignment between synsets is facilitated by looking at the synsets in several different languages in parallel. For this we make use of the Open Multilingual WordNet's interface (Bond and Foster, 2013) through links from our interface.

This lexical enrichment process of OpenWN-PT reported in employs three language strategies: (1) translation; (2) corpus extraction; and (3) dictionaries. The interested reader will find more details in (Rademaker et al., 2014; Real et al., 2015). The essential fact is that given the constant release of new versions of our openWN-PT, we must ensure the quality of the data that we make available. By quality here we mean not only the data content but its encoding consistency.

3 OpenWordnet-PT in RDF

As reported in (Rademaker et al., 2014), since its beginning OpenWN-PT is distributed using the Resource Description Format (RDF) (Cyganiak and Wood, 2003). We have being following the increasingly popular way of addressing the issue of interoperability by relying on Linked Data and Semantic Web standards such as RDF and OWL (Hitzler et al., 2012), which have led to the emergence of a number of Linked Data projects for lexical resources (de Melo and Weikum, 2008; Chiarcos et al., 2012). The adoption of such standards not only allows us to publish both the data model and the actual data in the same format, they also provide for instant compatibility with a vast range of existing data processing tools and storage systems, triple stores, providing query interfaces based on the SPARQL standard (Harris and Seaborne, 2013).

To encode any data in RDF, one needs to decide which classes and properties (vocabulary) will be used. The adoption of already defined vocabularies helps on the data interoperability since these makes data easily integrate with other resources.

We chose to use the vocabulary for wordnets encoding proposed by (van Assem et al., 2006) which is based on Princeton Wordnet 2.0. Their work includes (1) a mapping of WordNet 2.0 concepts and data model to RDF/OWL; (2) conversion scripts from the WordNet 2.0 Prolog distribution to RDF/OWL files; and (3) the actual WordNet 2.0 data. The suggested representation stayed as close to the original source as possible, that is, it reflects the original WordNet data model without interpretation. The WordNet schema proposed by (van Assem et al., 2006) has three main classes: *Synset*, *WordSense* and *Word*. The first two classes have subclasses for each lexical group present in WordNet. Each instance of Synset, WordSense and Word has its own URI.

Since (van Assem et al., 2006) is based on Princeton Wordnet 2.0, its use required few adaptations. Our first decision was to adapt the WordNet 2.0 vocabulary to version 3.0, having our own URIs for all entities (classes and properties). We converted the WordNet 3.0 data to RDF in such a way that OpenWN-PT is an extension of WordNet 3.0, with its instances, connected to Princeton instances through *owl:sameAs* relations. That is, for each Princeton WordNet synset, we created an equivalent synset in OpenWN-PT synset, with no additional synsets or relations so far. Given that OpenWN-PT's RDF is only useful together with an RDF version of Princeton WordNet and we wanted to ensure that all information in the WordNet 3.0 distribution was transformed to RDF, we wrote our own script to translate the Princeton

WordNet 3.0 data files to RDF so they can be distributed alongside OpenWN-PT.[4].

For the URI schema, we adopted a similar approach of (van Assem et al., 2006) of pattern for the URIs by classes. Moreover, we created the domain `https://w3id.org/own-pt/` under our control as suggested by the Linked Data principles. In Table 1, under the namespace [1] we have the classes and properties of our vocabulary (TBox), adapted from (van Assem et al., 2006). The namespace [2] holds the instances of our openWordnet-PT and [3] holds the Princeton instances. Our Nomlex-PT (Freitas et al., 2014) data also has its vocabulary and data namespace, respectively, [4] and [5].

1	`https://w3id.org/own-pt/wn30/schema/`
2	`https://w3id.org/own-pt/wn30-pt/instances/`
3	`https://w3id.org/own-pt/wn30-en/instances/`
4	`https://w3id.org/own-pt/nomlex/schema/`
5	`https://w3id.org/own-pt/nomlex/instances/`

Table 1: the used URIs

4 Consistency check of OWL and Integrity Constraints in RDF

The Web Ontology Language (OWL) [5] is a family of knowledge representation languages for authoring ontologies (or Knowledge bases) composed by OWL Lite, OWL DL and OWL Full. The OWL languages are built upon the W3C standard RDF and characterized by formal semantics. OWL Lite and OWL DL semantics are based on Description logics (DLs) (Baader, 2003). DL are a family of logics that are decidable fragments of first-order logic with attractive and well-understood computational properties.

A DL knowledge base is comprised by two components, TBox and ABox. The TBox contains intensional knowledge in the form of a terminology and is built through declarations of the general properties of concepts[6]. The ABox contains extensional knowledge, also called assertional knowledge. The knowledge that is specific to the individuals of the domain of discourse. Intensional knowledge is usually thought not to change and extensional knowledge is usually thought to be contingent, and therefore subject to occasional or even constant change.

[4] `https://github.com/own-pt/wordnet2rdf`

[5] `http://www.w3.org/OWL/`

[6] In this paper the TBox is sometimes called the vocabulary.

Given an ontology encoded in OWL (Lite or DL) one can use DL reasoners for different tasks such as: concepts consistency checking, query answering, classification, etc. In particular, classification amounts to placing a new concept expression in the proper place in a taxonomic hierarchy of concepts, it can be accomplished by verifying the subsumption relation between each defined concept in the hierarchy and the new concept expression. Validating an ontology means to guarantee that all concepts are satisfiable, that is, the concepts definition do not contain contradictions.

The basic reasoning task in an ABox is instance checking, which verifies whether a given individual is an instance of (or belongs to) a specified concept. Although other reasoning services are usually employed, they can be defined in terms of instance checking. Among them we find knowledge base consistency, which amounts to verifying whether every concept in the knowledge base admits at least one individual; realization, which finds the most specific concept an individual object is an instance of; and retrieval, which finds the individuals in the knowledge base that are instances of a given concept (query answering).

In some use cases, we need a method to validating the RDF data regarding a given model. In this case, OWL users intend OWL axioms to be interpreted as constraints on RDF data (Pérez-Urbina et al., 2012). For that, one has to define a semantics for OWL based on the Closed World Assumption and a weak variant of the Unique Name Assumption (Baader, 2003). OWL default semantics adopts the Open World Assumption (OWA) and does not adopt the Unique Name Assumption (UNA). These design choices make it very difficult to treat these axioms as ICs. On the one hand, due to OWA, a statement must not be inferred to be false on the basis of failures to prove it; therefore, the fact that a piece of information has not been specified does not mean that such information does not exist. On the other hand, the absence of UNA allows two different constants to refer to the same individual.

In the next section, we present some preliminary experiments with TBox and ABox consistency check and integrity constraints (IC) validation in our RDF/OWL data, reporting our experience with most well-know freely available tools. Nevertheless, it is important to emphasize the capabilities that semantic web technologies that ex-

ceed the currently mainstream technologies.

Most research groups that are still using XML for lexical resources distribution would argue that XML Schema (Fallside and Walmsley, 2004) can ensure some constraints that we verify in the next section. Relational database users would argue that SQL is an already mature and declarative query language. We argue that OWL/RDF brings much more expressivity allowing much more robust and semantics aware verification with queries such as:

```
select ?w ?ws1 ?ws2
{
  ?ss1 wn30:containsWordSense ?ws1 .
  ?ws1 wn30:word ?w .
  ?ss2 wn30:containsWordSense ?ws2 .
  ?ws2 wn30:word ?w .
  ?ss1 wn30:hyponymOf* ?ss2 .
}
```

In the SPARQL query above, we are asking for words that occur repeated in the same branch of the hierarchy of synsets formed by the `wn30:hyponymOf` transitive closure.

5 Validating OpenWN-PT

We were interested in checking our RDF and OWL files against a wide variety of errors, both minor and major and to increase our coverage we opted to use a variety of reasoners.

We started with Protégé [7], which is an ontology editor that among other features has interface with two well-know DL reasoners: FaCT++ (Tsarkov and Horrocks, 2006) and HermiT (Shearer et al., 2008). Starting in version 4, Protégé also gives us the opportunity to search for explanations that caused an inconsistency (Horridge et al., 2008). Racer (Haarslev et al., 2012) and Pellet (Sirin et al., 2007) are reasoners that have this feature built-in.

In order to verify OWN-PT files we needed to combine all files in `https://github.com/own-pt/openWordnet-PT` and the Simple Knowledge Organization System (SKOS) [8] ontology file. There are a number of tools available for this, we chose *RDF*$_{pro}$ (Corcoglioniti et al., 2015), which was the fastest in our benchmarks.

The errors found can be categorized in three different classes: datatype errors, domain and range errors, structural errors.

[7] `http://protege.stanford.edu/`

[8] `http://www.w3.org/TR/skos-reference/`

5.1 Datatype errors

Errors such as missing datatype declarations and wrongly typed literals were found by both Hermit and Pellet. Hermit identified the following missing classes:

```
wn30:AdjectiveWordSense
    rdfs:subClassOf wn30:WordSense .

wn30:VerbWordSense
    rdfs:subClassOf wn30:WordSense .
```

And the following verification fails due to incorrectly typed literals:

```
Literal value "00113726" does not
  belong to datatype nonNegativeInteger

Literal value "104" does not belong
  to datatype nonNegativeInteger
```

These errors were caused by the fact that `wn30:synsetId` and `wn30:tagCount` are defined as properties of synsets and word senses that are non-negative integers, but they were incorrectly stored without the type qualifier, for example: the literal in `synset-13363970-n synsetId "13363970"` should have been specified as `"13363970"^^xsd:nonNegativeInteger`.

Pellet Lint, like lint tools for programming languages, aims to detect possibly incorrect constructions that generally indicate bugs. For brevity we omit the prefix `https://w3id.org/own-pt/` from the individuals below.

```
[Untyped classes]
wn30/schema/BaseConcept
nomlex/schema/Nominalization
wn30/schema/CoreConcept
[...]
```

```
[Untyped datatype properties]
wn30/schema/senseKey
wn30/schema/syntacticMarker
wn30/schema/lexicographerFile
[...]
```

```
[Untyped individuals]
wn30-en/instances/wordsense-01362387-a-2
wn30-en/instances/wordsense-01362387-a-1
wn30-en/instances/wordsense-01722140-a-1
[...]
```

What Pellet Lint calls an untyped class is an object of a triple involving `rdf:type`, but it was never formally defined as an OWL class. The same idea applies to untyped properties: these are never formally defined as an OWL property,

and lacks any information about its domain and range. Untyped individuals also are used as objects, but never participate in triples as a subject, which seems like a mistake on some previous data import task. These likely need to be removed.

5.2 Domain and range errors

Moving beyond these initial type checks, we used initially Protégé with the FaCT++ reasoner to match our triple store statements against the OWL definition. The ontology was found to be inconsistent, with the following explanation:

```
Explanation for: Thing SubClassOf Nothing
classifiedByRegion Domain Synset
current_account classifiedByRegion Britain
current_account Type WordSense
Synset DisjointWith WordSense
```

We now give a detailed analysis of this explanation; we'll omit such details from the other inconsistencies found later on this section. The relation `wn30:classifiedByRegion` was created from the `;r` pointer symbol in Princeton WordNet data distribution, documented in `wninput(5wn)`.[9] In the explanation above, `current_account` is the label of `wordsense-13363970-n-3` and `Britain` the label of `wordsense-08860123-n-4`. These two subjects are related via the following triple:

```
wordsense-13363970-n-3 classifiedByRegion
   wordsense-08860123-n-4
```

This triple was generated from the following line in original Princeton `data.noun` file (formatted for clarity):

```
13363970 21 n 03
  checking_account 0 chequing_account 0
  current_account 1 004
   @ 13359690 n 0000
  ;r 08860123 n 0304
  ;r 08820121 n 0201
  ;r 09044862 n 0101
  | a bank account against which the
    depositor can draw checks that are
    payable on demand
```

Notice that the triple in the explanation above is a relationship between two word senses, while our definition of the `wn30:classifiedByRegion` property is as follows:

```
wn30:classifiedByRegion
  a rdf:Property, owl:ObjectProperty ;
  rdfs:domain wn30:Synset ;
  rdfs:range wn30:NounSynset ;
  rdfs:subPropertyOf wn30:classifiedBy .
```

In other words, it is a property whose domain contains synsets and its range contains all noun synsets. This is contradicted by the example, where the `rdfs:domain` and `rdfs:range` restrictions were violated.

To fix the inconsistency, we need to understand the source of the error: is the problem in our translation from the Wordnet file to RDF, the OWL definition of `wn30:classifiedByRegion`, or an issue in Wordnet itself? In the excerpt from `data.noun` above, all three domain/region pointers are between word senses, which was preserved in the translation to RDF. Looking at the other entries there, we find that `chequing_account` and `Canada` and `checking_account` and `United_States` are also word senses labels that are related by `wn30:classifiedByRegion`. This indicates a desire to differentiate between the different lexical forms and their regions of usage, which can be seen as a form of lexical relationship. This indicates an issue with the formalization of the relation `wn30:classifiedByRegion`. Going back to the original definition in `wninput(5wn)` we find the following (emphasis ours):

> The following pointer types are *usually* used to indicate lexical relations: Antonym, Pertainym, Participle, Also See, Derivationally Related. The remaining pointer types are *generally* used to represent semantic relations.

While generally a domain/region pointer is a semantic relationship, our examples show that this is not always the case. Also, by using words such as 'generally' and 'usually' the informal description above accommodates such cases. This leads us to think that `wn30:classifiedByRegion` is both a semantic and a lexical relation, unlike our formal definition states.

We can query for the statistics of the `wn30:classifiedByRegion` domain in our endpoint.[10] The SPARQL query below selects all individuals that are involved in `wn30:classifiedByRegion` relations, their

[9] `http://goo.gl/AbkdaZ`

[10] `http://goo.gl/ptPw6S`

types, and counts the number of individual by type.

```
select ?t (count(?t) as ?ct)
{ ?s wn30:classifiedByRegion ?o ;
    a ?t
} group by ?t
```

The majority of the subjects – over 1200 – are synsets, but there are 15 word senses as well, meaning that `wn30:classifiedByRegion` is definitely not strictly a semantic relation. To fix this issue, the definition needed to be changed so that the domain and range contains both synsets and word senses. This is done using the `owl:unionOf` operator, which represents set union.

```
wn30:classifiedByRegion
  a rdf:Property, owl:ObjectProperty ;
  rdfs:subPropertyOf wn30:classifiedBy ;
  rdfs:range [ a owl:Class ;
    owl:unionOf (wn30:NounWordSense
                 wn30:NounSynset)] ;
  rdfs:domain [ a owl:Class ;
    owl:unionOf (wn30:WordSense wn30:Synset)] .
```

We found similar problems with the properties `wn30:classifiedByUsage`, `wn30:classifiedByTopic`, and `wn30:frame`. We selected the latter since it highlights one of the issues that we find while performing formal verifications, which is the complexity of the proofs/explanations. This is the explanation found for the issue:

```
synset-01345109-v hypernymOf
    synset-01220528-v
 VerbWordSense subClassOf WordSense
 frame domain VerbWordSense
 synset-01220528-v frame
    "Somebody ----s something"
 hypernymOf range Synset
 Synset disjointWith WordSense
```

While this example can be understood, it definitely could be made simpler. For instance, `synset-01220528-v` found to be of type 'synset' due to the fact that it is the object of a triple containing the predicate `wn30:hypernymOf` combined with that fact that the range of this predicate is the set of all synsets. A more concise way is to realize that `synset-01220528-v` is a verb synset and that verb synsets are a subset of synsets. In any case, interpreting the explanation, we see that `wn30:frame` is being used as a relation whose domain contains a synset, but its definition prohibits this. We can query our triple store for the *de facto* domains of `wn30:frame` via a SPARQL query similar to the one used for `wn30:classifiedByRegion`. We again omit the results for brevity, but there are both word senses and synsets in the domain of this relation. Checking the definition of `wn30:frame` in `wninput(5wn)` we find that its original formal definition is too restrictive as it allows frames to exist between both synsets *and* word senses.

After fixing those, only a couple of issues remained:

```
NounSynset SubClassOf Synset
hemolysis Type WordSense
holonymOf Domain NounSynset
adjectivePertainsTo SubPropertyOf meronymOf
meronymOf SubPropertyOf inverse(holonymOf)
Synset DisjointWith WordSense
haemolytic adjectivePertainsTo hemolysis
```

While `wn30:adjectivePertainsTo` is a relation between word senses, it was marked as a subproperty of `wn30:meronymOf`, which is a relationship between synsets. It was also marked as the inverse of `wn30:holonymOf`, which is also a semantic relation. Both restrictions are, of course, incorrect and were removed.

The final issues were investigated using the Pellet reasoner. This allows us to verify our work and also experiment with the different implementations of the explanations for inconsistencies.

```
Axiom: Thing subClassOf Nothing

inSynset range Synset
VerbWordSense subClassOf WordSense
synset-00105023-a containsWordSense
  wordsense-00105023-a-2
synset-00105023-a seeAlso synset-00885415-a
AdjectiveWordSense subClassOf WordSense
seeAlso domain AdjectiveWordSense
              or VerbWordSense
inSynset inverseOf containsWordSense
Synset disjointWith WordSense
```

Here, `wn30:seeAlso` usually indicates lexical relations, but the explanation shows relationship between two synsets.

5.3 Structural errors

Our last example show cases yet another trap that should be avoided when designing ontologies, which is to assume that once it is consistent, there is nothing else to do. In our case, our modifications so far lead us to a consistent ontology, but unfortunately that doesn't mean that there weren't any issues left. In fact, there were two extremely serious errors in our RDF distribution that were not caught by the analyses so far and were

found accidentally through a cursory look: during one of our post-processing jobs we mistakenly implemented a blank node renaming algorithm and ended up having two invalid situations: (a) two or more words associated to a single word sense subject; (b) two or more lexical forms associated to a single word subject.

After fixing our ontology to give the proper restrictions on word senses, words, and lexical forms, Pellet was able to identify the issues. The following excerpt describes a single word sense (`wordsense-01860795-v-2`) with two words associated ('deixar', 'parar').

```
wordsense-01860795-v-2 type WordSense
word-deixar lexicalForm "deixar"@pt
word-parar lexicalForm "parar"@pt
wordsense-01860795-v-2 word word-deixar
Word subClassOf lexicalForm exactly 1
wordsense-01860795-v-2 word word-parar
word-deixar type Word
word-parar type Word
WordSense subClassOf word exactly 1 Word
```

The last tool that we tested was Stardog [11]. Stardog is the only reasoner and database system that supports ICV. Under the ICV semantics, the axioms below from the *wn30:WordSense* class were taken as constraints rather than terminology definitions. In other words, if Stardog finds an instance of the class *wn30:WordSense* connected to more than one instance of *wn30:Word*, it will raise an exception instead of infer that the two different *wn30:Word* instances should be the same.

```
wn30:WordSense
    a rdfs:Class, owl:Class ;
    rdfs:subClassOf [
     a owl:Restriction ;
     owl:onProperty wn30:inSynset ;
     owl:qualifiedCardinality
       "1"^^xsd:nonNegativeInteger ;
      owl:onClass wn30:Synset ], [
     a owl:Restriction ;
     owl:onProperty wn30:word;
     owl:qualifiedCardinality
      "1"^^xsd:nonNegativeInteger ;
     owl:onClass wn30:Word ] .
```

Unfortunately, in all tests that we run, Stardog hung without producing any output, even when we executed it with few axioms of our ontology. We hope to investigate the problem in a future report.

6 Conclusion

The use of different systems, with different functionalities, give us more confidence in our validations. Unfortunately, it required considerable effort to prepare data in different formats and interpret the results. Racer and RDFUnit did not give us meaningful results. We could not use Stardog at all. We will continue to try them, though, as we believe the diversity of tools and techniques are beneficial to the coverage of potential problems.

Performance is still an issue. Some of these experiment took hours to complete, in a relatively simple ontology. It looks like most of DL reasoners are not prepared to handle large ABoxes.

Most DL reasoners are based on some variation of tableaux or other refutation based procedure (Baader, 2003). Prove by refutation does not preserve information and tableaux proofs usually have exponential size. In the future, we hope to implement a proof-theoretical based reasoner for DL based on (Rademaker, 2012).

It is also worthy to mention that the tools that we tested do not always have an user-friendly interface, making adoption for people outside the area difficult.

Reasoning with closed world assumption for ICV is a future work given the problems that we faced with Stardog. Finally, DL Learning (Lehmann, 2009) is another possible interesting technique to explorer, it would allow us to extract the minimum required TBox for a given ABox.

[11] `http://www.stardog.com`.

References

[Baader2003] Franz Baader. 2003. *The description logic handbook: theory, implementation, and applications*. Cambridge university press.

[Berners-Lee1998] Tim Berners-Lee. 1998. Semantic web road map. Technical report, W3C, September.

[Bond and Foster2013] Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Chiarcos et al.2012] Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked data in linguistics: Representing and connecting language data and language metadata*. Springer.

[Corcoglioniti et al.2015] Francesco Corcoglioniti, Marco Rospocher, Michele Mostarda, and Marco Amadori. 2015. Processing billions of rdf triples on a single machine using streaming and sorting. In *ACM SAC 2015 Proceedings*.

[Cyganiak and Wood2003] Richard Cyganiak and David Wood. 2003. RDF 1.1 concepts and abstract syntax. Technical Report Draft 23 July 2013, W3C.

[de Melo and Weikum2008] Gerard de Melo and Gerhard Weikum. 2008. Language as a foundation of the Semantic Web. In *Proc. of ISWC 2008*, volume 401.

[de Melo and Weikum2009] Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

[de Paiva et al.2012] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian wordnet for reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

[Fallside and Walmsley2004] David C. Fallside and Priscilla Walmsley. 2004. Xml schema part 0: primer second edition. Technical Report W3C Recommendation 28 October 2004, W3C.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

[Freitas et al.2014] Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. 2014. Extending a lexicon of portuguese nominalizations with data from corpora. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos, Brazil, oct. Springer.

[Haarslev et al.2012] Volker Haarslev, Kay Hidde, Ralf Möller, and Michael Wessel. 2012. The RacerPro knowledge representation and reasoning system. *Semantic Web Journal*, 3(3):267–277.

[Harris and Seaborne2013] Steve Harris and Andy Seaborne. 2013. SPARQL 1.1 query language. Technical Report W3C Recommendation 21 March 2013, W3C.

[Hitzler et al.2012] Pascal Hitzler, Markus Krotzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. 2012. OWL 2 web ontology language primer. Technical Report W3C Rec 11 Dec 2012, W3C.

[Horridge et al.2008] Matthew Horridge, Bijan Parsia, and Ulrike Sattler. 2008. Explanation of OWL entailments in protégé 4. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008*.

[Kilgarriff1997] Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

[Lehmann2009] Jens Lehmann. 2009. DL-Learner: learning concepts in description logics. *Journal of Machine Learning Research (JMLR)*, 10:2639–2642.

[Pease and Fellbaum2010] Adam Pease and Christiane Fellbaum. 2010. Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 2, pages 25–35. Cambridge University Press.

[Pérez-Urbina et al.2012] Héctor Pérez-Urbina, Evren Sirin, and Kendall Clark. 2012. Validating rdf with owl integrity constraints. Technical report, Clark & Parsia, LLC.

[Rademaker et al.2014] Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Maria Real Coelho, and Maira Gatti. 2014. Openwordnet-pt: A project report. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan.

[Rademaker2012] Alexandre Rademaker. 2012. *A Proof Theory for Description Logics*. Springer-Briefs in Computer Science. Springer.

[Real et al.2015] Livy Real, Fabricio Chalub, Valeria dePaiva, Claudia Freitas, and Alexandre Rademaker. 2015. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 20–29, Beijing, China, July. Association for Computational Linguistics.

[Shearer et al.2008] R. Shearer, B. Motik, and I. Horrocks. 2008. HermiT: a highly efficient OWL reasoner. In *Proceedings of the Fifth International Workshop on OWL: Experiences and Directions (OWLED)*.

[Sirin et al.2007] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. 2007. Pellet: A practical OWL-DL reasoner. *Web Semant.*, 5(2):51–53, June.

[Tsarkov and Horrocks2006] Dmitry Tsarkov and Ian Horrocks. 2006. FaCT++ description logic reasoner: System description. In *Proceedings of the Third International Joint Conference on Automated Reasoning*, IJCAR'06, pages 292–297, Berlin, Heidelberg. Springer-Verlag.

[van Assem et al.2006] Mark van Assem, Aldo Gangemi, and Guus Schreiber. 2006. RDF/OWL representation of WordNet. Technical Report W3C Working Draft 19 June 2006, W3C.

DEBVisDic: Instant Wordnet Building

Adam Rambousek and **Aleš Horák**
Natural Language Processing Centre
Faculty of Informatics
Masaryk University
Brno, Czech Republic
{rambousek,hales}@fi.muni.cz

Abstract

The semantic network editor DEBVisDic has been used by different development teams to create more than 20 national wordnets. The editor was recently redeveloped as a multi-platform web-based application for general semantic networks editing. One of the main advantages, when compared to the previous implementation, lies in the fact that no client-side installation is needed now. Following the successful first phase in building the Open Dutch Wordnet, DEBVisDic was extended with features that allow users to easily create, edit, and share a new (usually national) wordnet without the need of any complicated configuration or advanced technical skills.

The DEBVisDic editor provides advanced features for wordnet browsing, editing, and visualization. Apart from the user-friendly web-based application, DEBVisDic also provides an API interface to integrate the semantic network data into external applications.

1 Introduction

The original wordnet, Princeton WordNet (PWN), is one of the most popular lexical semantic resources in the NLP field (Fellbaum, 1998). The publication of PWN was followed by the multilingual EU projects EuroWordNet 1 and 2 (1998–1999) (Vossen, 1998) and the Balkanet project (2001–2004) (Christodoulakis, 2004) in which wordnets for 13 languages were developed (English, Dutch, Italian, Spanish, French, German, Czech, Estonian, Bulgarian, Greek, Romanian, Serbian and Turkish). In the course of this work, new software tools for browsing and editing wordnets were designed and implemented. Within the EuroWordNet project the Polaris (and Periscope) tools were implemented and used (Louw, 1998).

For Balkanet project, a new browser and editor VisDic was built at the NLP Laboratory at the Faculty of Informatics, Masaryk University (FI MU) (Horák and Smrž, 2003), since the development of the Polaris tool was closed by 1999.

In comparison with the previous tools, VisDic exploits the XML data format thus making the wordnet-like databases more standard and exchangeable. Not only that, thanks to the XML data format used and to its dictionary specific configurability, VisDic can serve for developing various types of dictionaries, i.e. monolingual, translational, thesauri and multilingually interlinked wordnet-like databases. The experience with the VisDic tool during the Balkanet project has been positive (Horák and Smrž, 2004) and it was used as the main tool with which all Balkanet wordnets were developed.

VisDic, however, has its disadvantages, particularly it was designed for a single user off-line use, and team coordination was really difficult.

2 The DEB platform and the DEBVisDic editor

Based on the experience with VisDic, the team at the NLP Centre FI MU has designed and implemented a universal dictionary writing system that can be exploited in various lexicographic applications to build large lexical databases. The system has been named Dictionary Editor and Browser (further DEB platform) (Horák and Rambousek, 2007) and it has been used in many lexicographic projects so far, among others for the development of the Czech Lexical Database (Rangelova and Králík, 2007), or currently running projects of Pattern Dictionary of English Verbs (Hanks, 2004), and Family names in UK (Hanks et al., 2011).

The DEB platform is based on the client-server architecture, which brings along a lot of benefits.

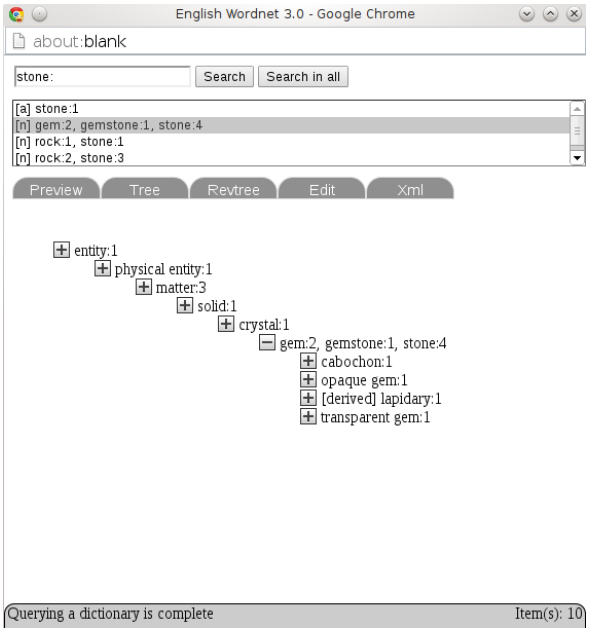


Figure 1: Example of a hypero-hyponymic tree in DEBVisDic.

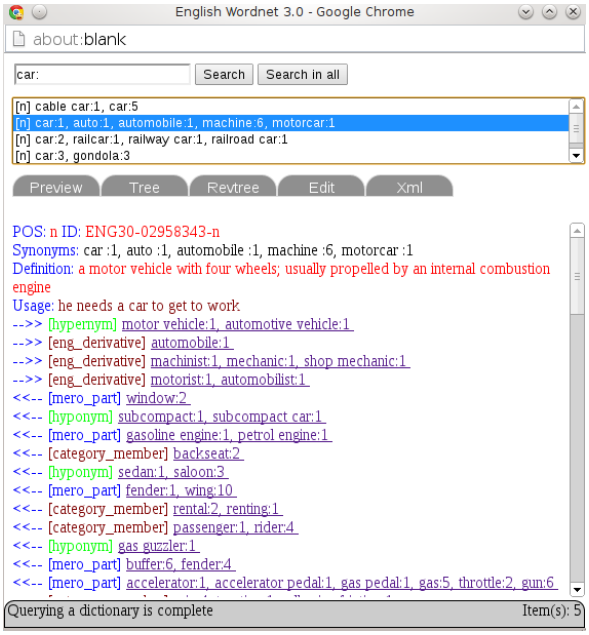


Figure 2: Example of a synset preview.

All the data are stored on the server side and a considerable part of the client-side functionality is also implemented on the server, thus the client application can be very lightweight.

This approach provides very good tools for team cooperation: all data modifications are immediately seen by all involved users. The server also provides well arranged authentication and authorization functions.

The general design of the DEB platform allows to adapt it also for building wordnet-like databases. For this purpose, the original VisDic tool was completely re-implemented on top of the DEB platform, as the current DEBVisDic editor (Horák et al., 2006).

The first version of the DEBVisDic editor was designed as a client application for the corresponding DEB server module, and was implemented using the flexible Mozilla Development Platform (XUL based applications (Boswell et al., 2002)), which was at that time the best option to design and build really cross-platform GUI applications, utilizing open standards.

However, any applications based on the Mozilla Development Platform are limited to the use via Mozilla-based browsers (mainly Firefox) only, while users prefer many different web-browsers. Since the development of DEBVisDic, the Firefox browser has introduced several major changes to the XUL application interface, thus limiting DEBVisDic to be used only in specific versions of the Firefox browser. As a result, the editor would need major changes to work with recent Firefox versions.

Fortunately, the current standards for web-based applications support many new features, which are implemented and supported by all the major web browsers. Considering all the options, we decided to re-implement DEBVisDic editor as a general web application, not limited to a single web browser and without the need to install specific browser extensions.

3 DEBVisDic 2

Thanks to the client-server architecture of the DEB platform, no changes were needed on the server side. Only the client side application had to be reimplemented, reusing the existing DEB interface. The main feature requests within the design of the new version were to keep all the DEBVisDic features and to provide an application working in all major web browsers.

As in the previous XUL version, *DEBVisDic 2* (Rambousek and Hrušo, 2013) aims primarily at wordnet-type semantic network browsing and editing, but supports different types of dictionaries. The application consists of a main window with settings and separate windows for each dictionary that the user needs to edit or browser. Single dictionary window includes a list of entries (synsets) and a set of tabbed panels with sev-

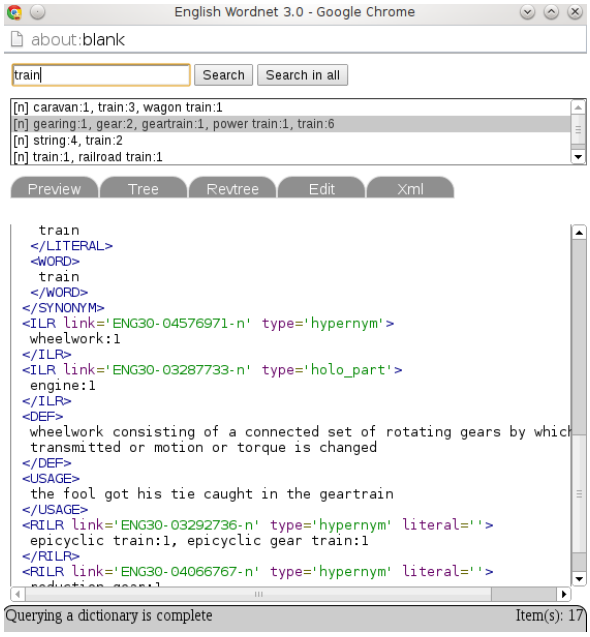


Figure 3: Example of the synset XML representation.

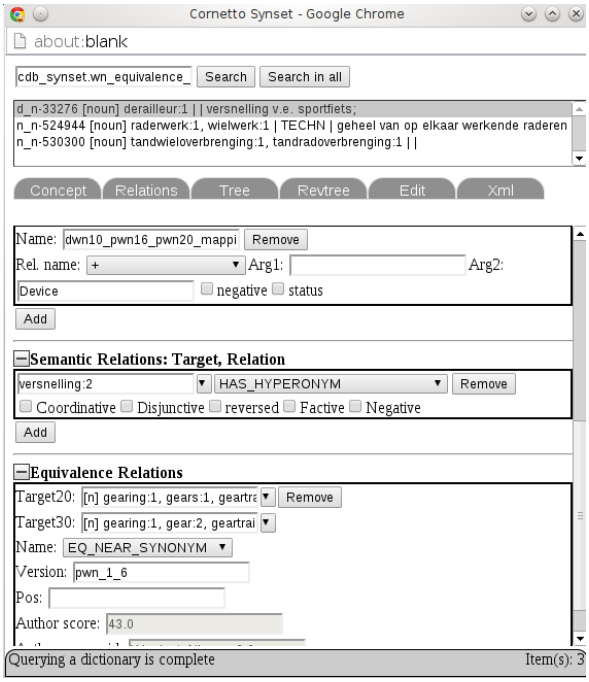


Figure 4: Example of the editing form.

eral types of view on the selected entry: a basic preview, the XML representation, the entry position withing the hypero-hyponymic tree, and an editing form. A context pop-up menu (right-click menu) provides functions for displaying and creating inter-dictionary links (e.g. to show the selected synset in another national wordnet or to display all synsets using the selected ontology term).

DEBVisDic 2 utilizes the Model-View-Controller (MVC) architecture and its design follows the MVC principle. Current open standards are used in the application: HTML and CSS for data presentation (view), and JavaScript for application logic scripting (model, controller). The application itself is modular, with separate core shared by all the dictionaries, and a plugin with specific functionality for each dictionary type.

As the implementation of web-related standards (mainly JavaScript) may vary in different browsers, several frameworks and libraries provide a unified environment on top of the browser interface. After reviewing several frameworks, we have decided to use the jQuery library (jQuery Foundation, 2015), which is a versatile JavaScript library for basic document and data manipulation without adding unnecessary features, thus staying lightweight and not slowing down the application.

One of the most challenging features was the implementation of the context menu functions, because of large differences in its implementation within the main web browsers. In the end, we were able to implement the context menu to keep the same behaviour as in DEBVisDic with the help of the jQuery contextMenu plugin.[1] Syntax highlighted view (pretty printing) of an entry in the XML format is provided by the Prettify plugin[2] (see Figure 3).

Apart from complete reimplementation of the DEBVisDic tool, the new version comes with several new features, especially concentrating on team work and complex dictionary editing. For example, *saving user settings* (opened dictionaries and window positions, with the possibility to store more information) *on the server*, thus allowing the user to switch browsers and computers and continue in the work.

Another major new feature is the implementation of *generalized links and relations* between dictionaries. It is possible to use any part of the entry structure (XML-based query) to build inter-dictionary search queries. For instance, selecting all lexical units in a synset, automatically view details of an ontology term for the selected synset, or all synonymic or near synonymic synsets between two wordnet languages.

[1] http://medialize.github.io/jQuery-contextMenu/
[2] http://google-code-prettify.googlecode.com/

Figure 5: User access management.

4 Building new wordnet

To support fast preparation of national wordnets, the DEBVisDic editor was extended with features similar to the DEBWrite application (Rambousek and Horák, 2015). The new application allows any user of the DEBVisDic server to create a new wordnet, without any complicated configuration or a need of advanced technical skills. Right after the straightforward set-up of a new wordnet, users may edit the wordnet data in the DEBVisDic application.

The DEBVisDic application supports copying of synsets from other wordnets (e.g. PWN). Editors have two options: either copy the original synset and translate it to the target language, or to create a new synset and link it to the "pivot" wordnet.[3]

In the case when the wordnet data are prepared in advance, e.g. in another editor or via "manual" editorial work, it is possible to import a XML file to DEBVisDic. The application also supports an export in the DEBVisDic XML file format.

One of the major assets of the DEBVisDic application lies in its support of team cooperation on the wordnet editing process. DEBVisDic classifies authorized users into one of three possible user roles: a manager, an editor, or a reader (see Figure 5 for an example of user access management).

- The user who created the wordnet is the *manager*. Managers may alter any settings. They may grant access to other users specifying their role. The manager may also decide to make synsets publicly available, which means that no password is needed to browse the semantic network (this *free access* might be regarded as a fourth user role in the dictionary access management).
- An *editor* may edit synsets before they are set to be published.
- *Readers* may browse and navigate through the published synsets with advanced search capabilities.

5 Data sharing

For sharing the resulting databases and its inclusion to wordnet repositories, e.g. Open Multilingual Wordnet (Bond and Foster, 2013), it is possible to export the wordnet data in several XML formats:

- *DEBVisDic XML*, used as an export/import format in several wordnet editors,
- *Wordnet-LMF* (Soria et al., 2009), developed during the KYOTO project as an extension of the Lexical Markup Framework data model,
- *Lemon RDF* (McCrae et al., 2011), ontology model adopted by the Princeton Wordnet.

To enhance the possibility to share and reuse wordnet resources, DEBVisDic provides a published API (Application Programming Interface) through which the DEBVisDic server functionality can be integrated in external applications, e.g. Visual Browser (Nevěřilová, 2007) for wordnet visualization. Features accessible by the API include complex searching for synsets, extracting various synset data, getting information on the semantic network graph, synset batch editing, or synset translation between several interconnected languages.

6 Conclusions

We have introduced a new web-based version of one of the most widespread wordnet editor, the DEBVisDic application. Besides the easy-to-distribute client application, DEBVisDic now

[3] This can be either an *interlingual index* (Vossen, 1998) or, which is the most common option today, the Princeton WordNet directly.

comes with an extension that allows fast and simple wordnet building. The application is currently in public testing, available at `http://deb.fi.muni.cz/debvisdic`. Future versions will incorporate features based on user feedback, and new options to link separate wordnets together or to shared ontologies. We believe this application will help with creation of new national wordnets, especially for sparsely resourced languages.

In the future, we plan to enhance the DEB-VisDic editor features, both in user experience (e.g. redesign the graphical interface) and format checks to ensure the synset structure follows specified rules.

Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the national COST-CZ project LD15066.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362. The Association for Computer Linguistics.

David Boswell, Brian King, Ian Oeschger, Pete Collins, and Eric Murphy. 2002. *Creating Applications with Mozilla*. O'Reilly and Associates, Inc., Sebastopol, California.

D. Christodoulakis. 2004. *Balkanet Final Report*. University of Patras, DBLAB. No. IST-2000-29388.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Patrick Hanks, Richard Coates, and Peter McClure. 2011. Methods for Studying the Origins and History of Family Names in Britain. In *Facts and Findings on Personal Names: Some European Examples*, pages 37–58, Uppsala. Acta Academiae Regiae Scientiarum Upsaliensis.

Patrick Hanks. 2004. Corpus Pattern Analysis. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France. Universite de Bretagne-Sud.

Aleš Horák and Adam Rambousek. 2007. DEB Platform Deployment – Current Applications. In *RASLAN 2007: Recent Advances in Slavonic Natural Language Processing*, pages 3–11, Brno, Czech Republic. Masaryk University.

Aleš Horák and Pavel Smrž. 2003. VisDic – wordnet browsing and editing tool. In *Proceedings of the Second International WordNet Conference – GWC 2004*, pages 136–141, Brno, Czech Republic. `http://nlp.fi.muni.cz/projekty/visdic/`.

Aleš Horák and Pavel Smrž. 2004. New features of wordnet editor VisDic. In *Romanian Journal of Information Science and Technology*, volume 7, pages 1–13.

Aleš Horák, Karel Pala, Adam Rambousek, and Martin Povolný. 2006. DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In *Proceedings of the Third International WordNet Conference - GWC 2006*, pages 325–328, Jeju, South Korea. Masaryk University, Brno.

jQuery Foundation. 2015. jQuery. `http://jquery.org`.

M. Louw. 1998. Polaris user's guide. Technical report, Belgium.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan, editors, *The Semantic Web: Research and Applications*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer Berlin Heidelberg.

Z. Nevěřilová. 2007. The Visual Browser Project. `http://nlp.fi.muni.cz/projects/visualbrowser`.

Adam Rambousek and Aleš Horák. 2015. DEBWrite: Free Customizable Web-based Dictionary Writing System. In Iztok Kosem, Miloš Jakubíček, Jelena Kallas, and Simon Krek, editors, *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 443–451. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Adam Rambousek and Tomáš Hrušo. 2013. Web Application for Semantic Network Editing. In *RASLAN 2013: Seventh Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 13–19, Brno, Czech Republic. Tribun EU.

Albena Rangelova and Jan Králík. 2007. Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century. In *Proceedings of the Computer Treatment of Slavic and East European Languages 2007*, pages 209–217, Bratislava, Slovakia.

C. Soria, M. Monachini, and P. Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of IWIC2009*, New York. ACM Press.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.

Samāsa-Kartā: An Online Tool for Producing Compound Words using IndoWordNet

Hanumant Redkar
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India
hanumantredkar@gmail.com

Nilesh Joshi
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India
joshinilesh60@gmail.com

Sandhya Singh
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India
sandhya.singh@gmail.com

Irawati Kulkarni
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India
irawatikulkarni@gmail.com

Malhar Kulkarni
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India
malharku@gmail.com

Pushpak Bhattacharyya
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India
pushpakbh@gmail.com

Abstract

Samāsa or compounds are a regular feature of Indian Languages. They are also found in other languages like German, Italian, French, Russian, Spanish, *etc.* Compound word is constructed from two or more words to form a single word. The meaning of this word is derived from each of the individual words of the compound. To develop a system to generate, identify and interpret compounds, is an important task in Natural Language Processing. This paper introduces a web based tool – *Samāsa-Kartā* for producing compound words. Here, the focus is on Sanskrit language due to its richness in usage of compounds; however, this approach can be applied to any Indian language as well as other languages. IndoWordNet is used as a resource for words to be compounded. The motivation behind creating compound words is to create, to improve the vocabulary, to reduce sense ambiguity, *etc.* in order to enrich the WordNet. The *Samāsa-Kartā* can be used for various applications *viz.*, compound categorization, sandhi creation, morphological analysis, paraphrasing, synset creation, *etc.*

1 Introduction

Word compounding is an essential feature of any language. In literature, there are various definitions of the compound word[1]. A compound word is a lexeme that consists of more than one stem. An English compound is a word composed of more than one free morpheme. However, in Sanskrit, a compound, also known as समास (*samāsa*) is defined as पृथगर्थानामेकार्थीभावः समासः (*pṛthagarthā nāmekārthībhāvaḥ samāsaḥ*, placing together two or more words so as to express a composite sense, which is a compound composition)[2]. Example, शिवपत्नी (*śivapatnī*, wife of *śiva* and a benevolent aspect of *devī*) is a *samāsa* or a compound formed from two words शिव (*śiva*, a major divinity in the later Hindu pantheon) and पत्नी (*patnī*, a married woman) which are formed from paraphrase शिवस्य पत्नी (*śivasya patnī*, wife of *śiva* and a benevolent aspect of *devī*). Sanskrit language has high usage of compounds in literature and is rich in producing

[1] http://grammar.ccc.commnet.edu/grammar/compounds.htm
[2] http://lukashevichus.info/knigi/abhyankar_shukla_sans_gram_dic.pdf

compound words. *Pāṇini*, the most referred Sanskrit grammarian, mentioned various types of *samāsa* and compounding system stated in the form of 110 *sutras* (rules) in his grammar book *Aṣṭādhyāyī* (Mishra, 2010).

1.1 Types of *Samāsa* in Sanskrit

In Sanskrit, there are four major types of *Samāsa*:

- **अव्ययीभाव** (***Avyayībhāva***) - In *avyayībhāva samāsa*, first member has primacy[3] (पूर्वपदार्थप्रधान, *pūrva-padārtha-pradhāna*). Here, the first member of this type of nominal compounds is indeclinable, to which another word is added so that the newly formed compound also becomes indeclinable (*i.e.*, अव्यय, *avaya*). Example, यथाशक्ति (*yathāśakti*, in accordance with one's strength).
- **तत्पुरुष** (***Tatpuruṣa***) - In *tatpuruṣa samāsa*, second member has primacy (उत्तरपदार्थप्रधान, *uttara-padārtha-pradhāna*) and the first component is in a case relationship with another. Example, सन्ध्याकालः (*sandhyākālaḥ*, evening time).
- **द्वन्द्व** (***Dvandva***) – In *dvandva samāsa*, both members have primacy (उभयपदार्थप्रधान, *ubhaya-padārtha-pradhāna*). Here, the members are usually noun stems, connected in sense with 'and'. Example, रामलक्ष्मणभरतशत्रुघ्नाः (*rāmalakṣmaṇabharata śatrughnāḥ*, Ram and Laxman and Bharat and Shatrughn).
- **बहुव्रीहि** (***Bahuvrīhi***) - In *bahuvrīhi samāsa*, both members refers to a thing which in itself is not part of the compound (अन्यपदार्थप्रधान, *anya-padārtha-pradhāna*). Example, गजाननः (*gajānanaḥ*, one whose face is that of an elephant).

1.2 IndoWordNet as a Resource

WordNet is a lexical resource composed of synsets and their semantic and lexical relations. Synsets are sets of synonyms or synonymous words (Miller et al., 1990). IndoWordNet[4] is a linked structure of WordNets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families (Bhattacharyya, 2010).

In this paper, we have taken Sanskrit WordNet[5] as a resource. Sanskrit is an Indo-Aryan language and is one of the ancient languages. It has vast literature and a rich tradition of creating lexica (Kulkarni et al., 2010(a)). The roots of most of the languages in the Indo European family in India can be traced to Sanskrit (Kulkarni et al., 2010(b)). Also, as stated in the article 351 of the constitution of India, the need arises for coining new words when the new object or an action related to it becomes part of the language and gets lexicalized[6]. The grammatical features of Sanskrit are prescribed for use and compounding is an important feature of Sanskrit.

The paper is organized as follows: Section 2 introduces *Samāsa-Kartā* and its components in detail. Section 3 lists the salient features of *Samāsa-Kartā*. Section 4 gives the limitation of *Samāsa-Kartā*. Section 5 describes the related work. Finally, we conclude the paper with the mention of scope and enhancements to this tool and its usefulness in the entire WordNet community.

2 *Samāsa-Kartā*: The Compound Word Producer

2.1 What is *Samāsa-Kartā*?

The *Samāsa-Kartā*[7], also known as Compound Word Producer is an online tool developed to produce compound words. The produced words are formed using rule based system which takes two words from IndoWordNet database (Prabhu et. al, 2012) with the help of IndoWordNet APIs (Prabhugaonkar et. al, 2012). The new word which is produced, is another word, which falls under any of the four types of *samāsas* mentioned above.

There are two types of users for this tool – the lexicographer and the validator. The basic job of lexicographer is to enter words, generate compound words and temporarily add these compound words to the synset in WordNet database. The main task of validator is to validate if the compound words are properly produced and added to the WordNet database.

Samāsa-Kartā basically produces compounds between Noun-Noun (NN-NN), Noun-Adjective

[3] primacy – the fact of being pre-eminent or most important.
[4] http://www.cfilt.iitb.ac.in/indowordnet/
[5] http://www.cfilt.iitb.ac.in/wordnet/webswn/wn.php
[6] http://www.constitution.org/cons/india/p17351.html
[7] http://www.cfilt.iitb.ac.in/wordnet/samaaskarta/

(NN-JJ), Noun-Verb (NN-VM), Adjective-Noun (JJ-NN), Adverb-Noun (RB-NN) pairs. However, it does not deal with the word combinations such as Noun-Adverb (NN-RB), Verb-Verb (VM-VM) and Verb-Noun (VM-NN) as they cannot be compounded.

2.2 Components of *Samāsa-Kartā*

Samāsa-Kartā, the tool, has multiple components which follows the pipeline architecture. Figure 1 shows the block diagram and figure 2 shows the basic interface of the *Samāsa-Kartā*.

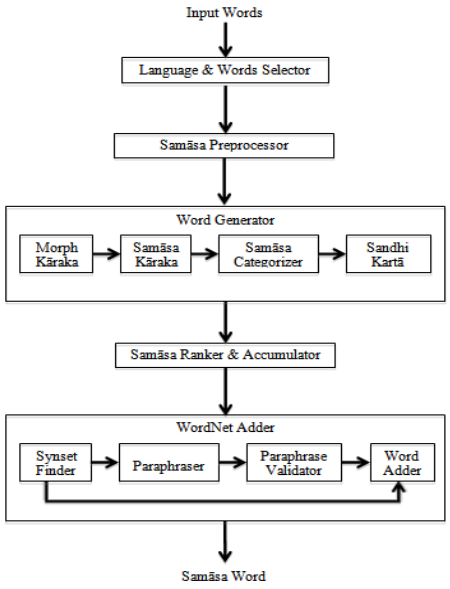


Figure 1. Block diagram of *Samāsa-Kartā*

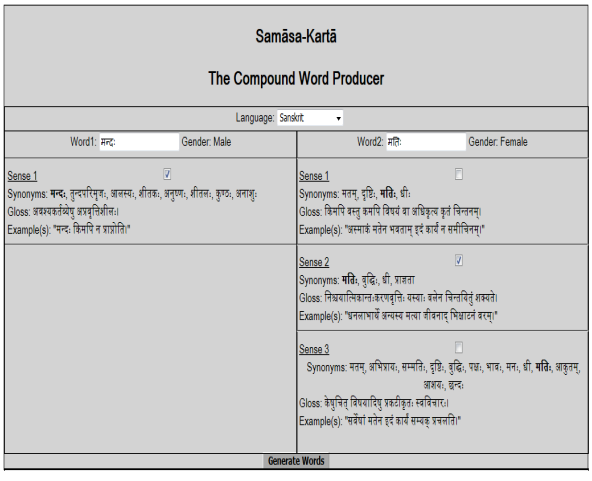


Figure 2. Interface of *Samāsa-Kartā*

The components of *Samāsa-Kartā* are described in detail as follows:

2.2.1 Language and Words Selector

In this module, user selects input language, in our case, Sanskrit and the words are taken from IndoWordNet database to form a compound. Here, the lexicographer types-in any character in the selected input language and all the words in database starting with typed character appear in the drop down list. Once words are selected, their corresponding synset information is displayed accordingly.

For example, if a lexicographer inputs two words, मन्दः (*mandaḥ*, disinclined to work or exertion) as first word and मतिः (*matiḥ*, knowledge and intellectual ability) as second word; we get the following synset information:

For the word मन्दः (*mandaḥ*)

Sense 1

Synonyms: मन्दः (*mandaḥ*), तुन्दपरिमृजः (*tundaparimṛjaḥ*), आलस्यः (*ālasyaḥ*), शीतकः (*śītakaḥ*), अनुष्णः (*anuṣṇaḥ*), शीतलः (*śītalaḥ*), कुण्ठः (*kuṇṭhaḥ*), अनाशुः (*anāśuḥ*)
Gloss: अवश्यकर्तव्येषु अप्रवृत्तिशीलः। (*avaśyakartavyeṣu apravṛttiśīlaḥ*)
Example(s): "मन्दः किमपि न प्राप्नोति।" (*mandaḥ kimapi na prāpnoti*)

For the word मतिः

Sense 1

Synonyms: मतम् (*matam*), दृष्टिः (*dṛṣṭiḥ*), मतिः (*matiḥ*), धीः (*dhīḥ*)
Gloss: किमपि वस्तु कमपि विषयं वा अधिकृत्य कृतं चिन्तनम्। (*kimapi vastu kamapi viṣayaṃ vā adhikṛtya kṛtaṃ cintanam*)
Example(s): "अस्माकं मतेन भवताम् इदं कार्यं न समीचिनम्।" (*asmākaṃ matena bhavatām idaṃ kāryaṃ na samīcinam*)

Sense 2

Synonyms: मतिः (*matiḥ*), बुद्धिः (*buddhiḥ*), धीः (*dhīḥ*), प्राज्ञता (*prājñatā*)
Gloss: निश्चयात्मिकान्तःकरणवृत्तिः यस्याः बलेन चिन्तयितुं शक्यते। (*niścayātmikāntaḥkaraṇa-vṛttiḥ* yasyāḥ balena cintayituṃ śakyate)
Example(s): "धनलाभार्थे अन्यस्य मत्या जीवनाद् भिक्षाटनं वरम्।" (*dhanalābhārthe anyasya matyā jīvanād bhikṣāṭanaṃ varam*)

Sense 3

Synonyms: मतम् (*matam*), अभिप्रायः (*abhiprāyaḥ*), सम्मतिः (*sammatiḥ*), दृष्टिः (*dṛṣṭiḥ*), बुद्धिः (*buddhiḥ*), पक्षः (*pakṣaḥ*), भावः (*bhāvaḥ*), मनः (*manaḥ*), धी (*dhī*), मतिः (*matiḥ*), आकुतम् (*ākutam*), आशयः (*āśayaḥ*), छन्दः (*chandaḥ*)

Gloss: केषुचित् विषयादिषु प्रकटीकृतः स्वविचारः। (*keṣucit viṣayādiṣu prakaṭīkṛtaḥ svavicāraḥ*)

Example(s): "सर्वेषां मतेन इदं कार्यं सम्यक् प्रचलति।" (*sarveṣāṃ matena idaṃ kāryaṃ samyak pracalati*)

Here, the lexicographer chooses sense 1 of the word मन्दः (*mandaḥ*) and sense 2 of the word मतिः (*matiḥ*) to form a compound word मन्दमति (*mandamati*, lacking intelligence). He/she also has freedom to select/deselect the synonymous words of these selected synsets. Also, in this module, the proper care has been taken to avoid words which cannot form *samāsa*. There are some words having specific case endings which cannot be compounded, *e.g.,* a word यथा (*yathā*, in which manner) can be compounded; however its synonyms यत्प्रकारेण (*yatprakāreṇa*), येन_प्रकारेण (*yena_prakāreṇa*) cannot be compounded, as they are specific case ending adverbs.

After selecting the appropriate synset and its synonyms, lexicographer finally proceeds to generate *samāsas* or compound words. The compound words are then processed using the following modules.

2.2.2 *Samāsa* Preprocessor

The *Samāsa* Preprocessor performs a check whether the input words are valid to form a *samāsa* or not. Here, it will check part-of-speech (POS) of each input word and validates if the combinations of POS like NN-NN, NN-JJ, JJ-NN, RB-NN, *etc.* can be formed.

2.2.3 Word Generator

The Word Generator internally processes each input word by using *Morph-Kāraka, Samāsa-Kāraka, Samāsa* Categorizer and *Sandhi-Kartā* to form a compound word. The details of these sub modules are as follows:

Morph Kāraka

Morph-Kāraka or Morph Analyzer is executed once the *Samāsa* Preprocessor provides it the validated input words. In this module, each input word is taken and converted to its root form by applying standard morphological rules. This is required, as in Sanskrit WordNet, all nouns are stored in nominative singular form. In order to make compound of these words, we need to bring these nouns to their root form. Table 1 illustrates some of the words processed through *Morph-Kāraka*.

स्वरान्त-शब्दाः (*svarānta-śabdāḥ*) (vowel-ending words)		**व्यञ्जनान्त-शब्दाः** (*vyañjanānta-śabdāḥ*) (consonant-ending words)	
अकारान्त (*akārānta*)	मन्दः → मन्द (*mandaḥ → manda*)	**चकारान्त** (*cakārānta*)	वाक् → वाच् (*vāk → vāc*)
आकारान्त (*ākārānta*)	विद्या → विद्या (*vidyā → vidyā*)	**जकारान्त** (*jakārānta*)	भिषक् → भिषज् (*bhiṣak → bhiṣaj*)
इकारान्त (*ikārānta*)	मतिः → मति (*matiḥ → mati*)	**तकारान्त** (*takārānta*)	भगवान् → भगवत् (*bhagavān → bhagavat*)
ईकारान्त (*īkārānta*)	नदी → नदी (*nadī → nadī*)	**दकारान्त** (*dakārānta*)	शरद् → शरद् (*śarad → śarad*)
उकारान्त (*ukārānta*)	भानुः → भानु (*bhānuḥ → bhānu*)	**नकारान्त** (*nakārānta*)	आत्मा → आत्मन् (*ātmā → ātman*)
ऋकारान्त (*ṛkārānta*)	माता → मातृ (*mātā → mātṛ*)	**सकारान्त** (*sakārānta*)	तेजः → तेजस् (*tejaḥ → tejas*)

Table 1. Words processed through *Morph-Kāraka*

Once the morphological analysis is done on input words, they are given to *Samāsa-Kāraka* for further processing.

Samāsa-Kāraka

The *Samāsa-Kāraka* takes the processed words from *Morph-Kāraka* and applies standard *samāsa* rules based on grammar. The *Samāsa-Kāraka* works at the semantic as well as syntactic level. At semantic level, meanings of the words are considered from the gloss to form the compounded word. At syntactic level, the inflections are appended/not appended to the morphed words. The processed words along with its *Samāsa* type are passed to the *Samāsa* Categorizer as an input.

For example,

1) आत्मन् + शक्ति (*ātman + śakti*) – Here, *Samāsa-Kāraka* identifies that both the words आत्मन् (*ātman*) and शक्ति (*śakti*) follows 2.2.8 rule षष्ठी (*ṣaṣṭhī*) *of Pāṇinian* grammar. Hence, आत्मन् (*ātman*) is eligible to form *Samāsa* with the

शक्ति (*śakti*). However, the rule number 8.2.7 नलोपः प्रातिपदिकान्तस्य (*nalopaḥ prātipadikān tasya*) *of Pāṇinian* grammar says that the न् (*n*) should be removed from the word आत्मन् (*ātman*). Hence, words आत्म (*ātma*) and शक्ति (*śakti*) is sent to *Samāsa* Categorizer for further processing.

2) देव + ईश (*deva* + *īśa*) – Here, there is no infection, hence these words are directly passed to the *Samāsa* Categorizer.

Samāsa Categorizer

Samāsa Categorizer identifies category of a *samāsa* like *Avyayībhāva, Tatpuruṣa, Dvandva* & *Bahuvrīhi* as per the *samāsa* rules. Further, it identifies its sub categories. It generates paraphrased information using gloss of input words. This paraphrased information is stored here, which is further used in the WordNet Adder for paraphrasing of compound words.

Sandhi-Kartā

Sandhi-Kartā or Sandhi Joiner helps in joining two words together which are passed through *Samāsa* Categorizer. The words are joined together by following sandhi rules of the language into consideration. The *Sandhi-Kartā* performs on all the combinations of the selected synset words and produces list of joined words. All these joined words are given to the *Samāsa* Ranker & Accumulator module.

Some of the examples of *Sandhi-Kartā* usage for words in Sanskrit are illustrated in table 2.

2.2.4 *Samāsa* Ranker and Accumulator

In this module, all the combinations of words are ranked and accumulated together as per the most frequent usage of words in the original WordNet synsets. Here, *Samāsa* Ranker Algorithm is used to rank the accumulated *samāsas.* Once the ranking and accumulating of words are done, the *samāsas* will be passed through the WordNet Adder module where its validity is checked and added to the WordNet.

2.2.5 WordNet Adder

WordNet Adder is a semi-automatic process where newly formed *samāsas* are passed through sequence of steps before adding to the synset in the WordNet.

स्वरान्त-शब्दाः (*svarānta-śabdāḥ*) (vowel-ending words)	
अकारान्त (*akārānta*) (words ending with a)	देव + ईश → देवेश (*deva + īśa → deveśa*)
आकारान्त (*ākārānta*) (words ending with ā)	विद्या + आलय → विद्यालय (*vidyā + ālaya → vidyālaya*)
इकारान्त (*ikārānta*) (words ending with i)	प्रति + उत्तर → प्रत्युत्तर (*prati + uttara → pratyuttara*)
ईकारान्त (*īkārānta*) (words ending with ī)	नदी + ईश → नदीश (*nadī + īśa → nadīśa*)
उकारान्त (*ukārānta*) (words ending with u)	भानु + उदय → भानूदय (*bhānu + udaya → bhānūdaya*)
ऋकारान्त (*ṛkārānta*) (words ending with ṛ)	मातृ + ऋण → मातॄण (*mātṛ + ṛṇa → mātṝṇa*)
व्यञ्जनान्त-शब्दाः (*vyañjanānta-śabdāḥ*) (consonant-ending words)	
तकारान्त (*takārānta*) (words ending with ta)	भगवत् + गीता → भगवद्गीता (*bhagavat + gītā → bhagavadgītā*)
दकारान्त (*dakārānta*) (words ending with da)	शरद् + हविष् → शरद्धविष् (*śarad + haviṣ → śaraddhaviṣ*)
नकारान्त (*nakārānta*) (words ending with na)	आत्मन् + शक्ति → आत्मशक्ति (*ātman + śakti → ātmaśakti*)
सकारान्त (*sakārānta*) (words ending with sa)	मनस् + रथ → मनोरथ (*manas + ratha → manoratha*)

Table 2. Words processed through *Sandhi-Kartā*

Following are the sub modules of WordNet Adder.

Synset Finder

Here, the lexicographer checks if the intended synset already exists in the WordNet. If it exists then the words are directly appended to the intended synset's vocabulary. If the synset does not exist, then it passes through the ***Paraphraser to create gloss of the compound word*** which will help in creating new synset**.**

Paraphraser

The Paraphraser automatically generates most likely gloss of the intended synset on the basis of input words. This gloss or a concept definition of a

synset is given to Paraphrase Validator for further processing.

Paraphrase Validator

Here, the lexicographer checks if the paraphrased gloss is properly generated. If not, it is created / edited manually by using the three principles of synset creation, *viz.*, principle of minimality, coverage and replaceability (Bhattacharyya, 2010). This is given to the Word Adder module.

Word Adder

The lexicographer finally fills-in other synset information like examples, gender, *etc.* and adds to the WordNet using an online synset creation tool - *Synskarta* (Redkar et al., 2014). The resultant *Samāsas* will either be the member of an existing synset or it can be a new synset altogether.

3 Salient Features of *Samāsa-Kartā*

Some of the salient features of *Samāsa-Kartā* are as follows:

- *Samāsa* or compounds are created on the flow.
- *Samāsa* in WordNet helps in identifying meaning or concept of a compound occurring in the literature.
- *Samāsa-Kartā* helps in enriching the standard of the language and to simplify the case-ending words in language under consideration.
- It assists in developing vocabulary, which in turn, helps in improving the word count in a language.
- It helps in automatic generation of paraphrases.
- It helps in compound type identification.
- The compound words produced can be helpful to understand the multi-words.

4 Limitation of *Samāsa-Kartā*

Some of the limitations of *Samāsa-Kartā* are:

- Used only for words in WordNet.
- Possibility of over generation of compounds.
- In Sanskrit, verbs are in its root form; hence word pairs such as VM-VM and RB-VM are not implemented.
- The word combination NN-RB is not possible as adverbs cannot come as a second word in the compound.

5 Related Work

In past, many researchers have worked on compound words, more particularly for Sanskrit Language. To understand the need of the tool presented here, a study is done on different kinds of tools available for usage. Some tools and work which were reviewed in the domain of compound word and its related fields are presented here.

Kumar et al. (2010) presented a Sanskrit compound processor tool, which automatically segments and identifies the type of a compound using the manually annotated data. To understand the compound; their approach involved segmentation, constituency parsing, compound type identification and paraphrasing. This tool can identify the type of compound and suggest its component's root word.

Jha et.al. (2009) proposed an Inflectional Morphology Analyzer for Sanskrit that identifies and analyzes the inflected noun-forms and verb-forms in any sandhi-free text. The tool checks and labels each word as three basic POS categories - *subanta*, *tiṅanta*, and *avyaya*. It is based on a reverse *pāṇinian* approach to analyze *tiṅanta* verb forms into their verbal base and verbal affixes. The methodology used to create database tables to store various morphological components of Sanskrit verb forms is based on the well defined and structured process of Sanskrit morphology described by *Pāṇini* in his *Aṣṭādhyāyī*. This analyzer also includes the analysis of derived verb roots.

Gupta et al. (2009) proposed a Rule Based Algorithm for *Sandhi-Vichedā* of compound Hindi words where one letter (whether single or conjoined) is broken to form two words. Part of the broken letter remains as the last letter of the first word and later part of the broken word forms the first letter of the next letter. A *Sandhi-Vichedā* module breaks the compound word in a sentence into constituent words, which enables to understand the meaning of the words better. This work aids in learning about the language grammar in an easy way.

Satuluri et al. (2013) studied the generation of Sanskrit compounds and rewrote the grammar as a combination of phrase structure rules and the regular grammar. It listed various semantic features as constraints governing the formation of compounds in Sanskrit. The rules taken from *Pāṇini* for compound formation are classified into two sets – the ones which designate a technical term to the input

string or a part thereof termed as *saṃjñāsūtra,* and the others which transform the input string into another termed as *vidhisūtra.* Also, the various semantic information needed by the compound formation rules is stated through ontological approach.

Sanskrit being a highly inflected language in nature, each of its word is inflected. If the words are not used in the correct case-endings, it may lead to a different meaning altogether, giving different context. To simplify the usage of these case-endings, compound words are used. Also, if these compound words are added to the WordNet, it may help in identifying meaning of a compound occurring in the literature.

Hence, we have developed a web based tool called *Samāsa-Kartā*. The approach used in this tool follows the rule-based system which takes two words from IndoWordNet database as an input and produces a compound word. This resultant compound word or *samāsa* can be included as a synset member along with its paraphrase as a gloss in the WordNet.

Work done by Kumar et al. is about identification of compound word, whereas, our tool deals with creation of new compound words. Jha et al. created morphological analyzer; similarly, we have implemented *Morph Kartā* which is created, specifically for WordNet words. Gupta et al., created Sandhi Splitter, however, we have created Sandhi Joiner, which is also specific for *samāsa* of WordNet words. Hence, *Samāsa-Kartā* can be considered as a complete tool of producing compound words related to words in IndoWordNet database.

6 Conclusion

Samāsa is a significant part of most of the languages which is used to express meaning using less number of words. The tool *Samāsa-Kartā*, discussed in this paper, is an attempt to improve upon the richness and coverage of a language using a semi-automated approach. It takes words from IndoWordNet and creates *Samāsa* or compound word(s). *Samāsa-Kartā* uses rule based system to form the compounds by passing through various rules of grammar at each sub module. This tool is able to create new compound words along with its paraphrase which can be added to the WordNet.

7 Future Scope and Enhancements

In future, the tool can be extended to other Indian languages belonging to Indo-Aryan, Dravidian and Sino-Tibetan families *viz.*, Hindi, Marathi, Gujarati, Bengali, Konkani, Kannada, *etc*. It can also be extended to other non-Indian languages like English, German, Italian, *etc.* This tool can have additional features such as non-WordNet words. This will be useful in the light of development of improving the vocabulary of the language, thus enhancing the richness of the language. Some of the major modules of this tool such as *Morph Kāraka*, S*andhi Kartā*, *Samāsa* Categorizer, Paraphraser, etc. can be made available independently.

Acknowledgments

We graciously thank all the members of CFILT[8] lab, IIT Bombay for providing necessary help and guidance needed for the development of *Samāsa-Kartā.* Further, we sincerely thank the members IndoWordNet and Global WordNet community.

References

Amba Kulkarni, Soma Paul, Malhar Kulkarni, Anil Kumar , Nitesh Surtani : Semantic Processing of Compounds in Indian Languages, Proceedings of COLING 2012, Mumbai, December 2012.

Anil Kumar, Vipul Mittal and Amba Kulkarni: *Sanskrit Compound Processor*, Sanskrit Computational Linguistics - 4th International Symposium, New Delhi, India, 2010.

George Miller, R., Fellbaum, C., Gross, D., Miller, K. J. 1990. *Introduction to wordnet: An on-line lexical database.* International journal of lexicography, OUP. (pp. 3.4: 235-244).

Girish Nath Jha, Muktanand Agrawal, Subash, Sudhir K. Mishra, Diwakar Mani, Diwakar Mishra, Manji Bhadra, Surjit K. Singh, *Inflectional Morphology Analyzer for Sanskrit*, Sanskrit computational linguistics. Springer Berlin Heidelberg, 2009.

Hanumant Redkar, Jai Paranjape, Nilesh Joshi, Irawati Kulkarni, Malhar Kulkarni, and Pushpak Bhattacharyya. 2014. *Introduction to Synskarta: An Online Interface for Synset Creation with Special Reference to Sanskrit.* ICON 2014, Goa, India.

[8] http://www.cfilt.iitb.ac.in/

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda and Pushpak Bhattacharyya. 2010(a). *Introducing Sanskrit Wordnet*. In Principles, Construction and Application of Multilingual WordNets, Proceedings of the 5th GWC, edited by Pushpak Bhattacharyya, Christiane Fellbaum and Piek Vossen, Narosa Publishing House, New Delhi, 2010, pp 257 – 294.

Malhar Kulkarni, Irawati Kulkarni, Chaitali Dangarikar and Pushpak Bhattacharyya. 2010(b). *Gloss in Sanskrit Wordnet*. In Proceedings of Sanskrit Computational Linguistics. Jha. G. Berlin: Springer-Verlag / Heidelberg. pp 190-197.

Neha R. Prabhugaonkar, Apurva S. Nagvenkar, and Ramdas N. Karmali. 2012. *IndoWordNet Application Programming Interfaces.* In 24th International Conference on Computational Linguistics (COLING 2012), p. 237.

Pavankumar Satuluri, Amba Kulkarni, *Generation of Sanskrit Compounds*, Proceedings of ICON, 2013.

Priyanka Gupta, Vishal Goyal. 2009. *Implementation of Rule Based Algorithm for Sandhi-Vicheda of Compound Hindi Words*. JCSI International Journal of Computer Science Issues, Vol. 3, 2009.

Pushpak Bhattacharyya. 2010. *IndoWordNet*. In the Proceedings of Lexical Resources Engineering Conference (LREC), Malta.

Ramashankar Mishra. 2010. *अष्टाध्यायीसूत्रपाठः*. Motilal Banarasidas publishers pvt. ltd, New Delhi (ISBN 978-81-208-2748-6).

Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, Neha Prabhugaonkar, Apurva Nagvenkar, Ramdas, Karmali. 2012. *An Efficient Database Design for IndoWordNet Development Using Hybrid Approach*. COLING 2012, Mumbai, India. p 229.

Arabic WordNet: New Content and New Applications

Yassir Regragui
Mohammadia School of Engineers, Mohammed V University of Rabat, Morocco
yasserregragui@research.emi.ac.ma

Lahsen Abouenour
Mohammadia School of Engineers, Mohammed V University of Rabat, Morocco
abouenour@yahoo.fr

Fettoum Krieche
Letter Faculty of Sais-Fez, Sidi Mohammed Ben Abdellah University of Fez, Morocco
fettoum.krieche@gmail.com

Karim Bouzoubaa
Mohammadia School of Engineers, Mohammed V University of Rabat, Morocco
Karim.bouzoubaa@emi.ac.ma

Paolo Rosso
NLE Lab, PRHLT Research Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract

The Arabic WordNet project has provided the Arabic Natural Language Processing (NLP) community with the first WordNet-compliant resource. It allowed new possibilities in terms of building sophisticated NLP applications related to this Semitic language. In this paper, we present the new content added to this resource, using semi-automatic techniques, and validated by Arabic native-speaker lexicographers. We also present how this content helps in the implementation of new Arabic NLP applications, especially for Question Answering (QA) systems. The obtained results show the contribution of the added content. The resource, fully transformed into the standard Lexical Markup Framework (LMF), is made available for the community.

1 Introduction

WordNets are important as lexical resources containing not only words of the targeted language but also synsets and semantic relations between them such as synonymy, meronymy and antonymy.

Synsets are groups of words that each can substitute others in a sentence without changing its general meaning. Therefore, in Natural Language Processing (NLP), various applications used this information, especially Query Expansion (QE), Information Retrieval (IR) and Question Answering (QA) systems.

Thus, the development of new WordNets targeting new languages and dialects and/or enriching existing ones, witnessed regular experiences and research works.

Arabic, as a Semitic language spoken by around 300 million people worldwide, is concerned by these experiences as well as the use of Arabic WordNet (AWN) (Felbaum 1998; Elkateb et al., 2006) in recent NLP applications such as information retrieval (Abbache et al., 2014) E-learning of Arabic (Karkar et al., 2015), semantic-based applications (Bouhriz et al., 2015), conceptual search (Al-Zoghby and Shaalan 2015), etc.

After the first release of AWN, There were many attempts to enrich its content (Al khalifa and Rodriguez 2009; Rodriguez et al., 2008a; 2008b). Nevertheless, the gap between AWN and the Arabic language as well as other similar WNs remained one of the limitations for its use. Also, some particularities of the Arabic language, including Broken Plurals (BP), has not been sufficiently addressed. Indeed, the morphological analysis of BP is not an easy task in NLP since they are irregular forms of plurals, and cannot be identified using patterns. Making these BP forms in a resource such as AWN is much helpful for the developers of NLP applications.

In previous research (Abouenour et al., 2013), we presented a new content that has been added

to the AWN in order to cover more words and synsets and, therefore, enhance the usability of this resource for Arabic NLP applications.

This paper keeps on the track of this previous research by presenting the last experiments conducted using the new content of AWN and the different refinements brought by manual validation made by lexicographers. This paper also presents the transformation of the new content of AWN into the Lexical Markup Framework (LMF) format in order to make this resource available for the community in the context of the Open Multilingual WordNet project (Bond and Kyonghee, 2012), providing free access to WordNets in several languages in a common format.

The paper is organized as follows: Section 2 recalls some existing works around the AWN enrichment and its use in NLP applications. Section 3 draws a synthesis of the main techniques that we used to enrich AWN. Section 4 presents the conducted experiments to show the usability of the new content as well as to validate this content. Section 5 addresses the transformation of the enriched AWN into the LMF format. Finally, Section 6 provides a conclusion of this research and highlights the future works.

2 Related works

The AWN project followed the development of WordNets for other languages, including Euro WordNet (Vossen 1998; 1999) by focusing, first, on the most common concepts and word-senses in PWN 2.0 (Fellbaum 1998). The first release was available on 2007[1] (Black et al., 2006; Elkateb et al., 2006).

The first AWN release contains 9,698 synsets, corresponding to 21,813 Modern Standard Arabic (MSA) words, and 6 different relation types (hyponymy, meronymy, instance, etc.). A later version of AWN, is also available and contains 11,269 synsets, corresponding to 23,841 words, including new Named Entities (Rodriguez et al., 2008).[2] This content is smaller than the PWN 2.0 and much smaller than what is expected by Arabic NLP applications.

Indeed, although various research experiences used AWN in many Arabic NLP applications, the common limitation reported in these experiences was the shortcomings of this resource in terms of the coverage of the Arabic language.

In a previous work (Abouenour et al., 2013), we made a comparison between the size of AWN and a dictionary for MSA on one side, and between AWN and English and Spanish WNs on another side. This comparison allowed us to measure the gap to be filled in to improve the quality and usability of this resource.

The previously mentioned experiences in using AWN (Elghamry 2008; El Amine 2009; Baldwin et al.,, 2010; Sharaf 2009; Benajiba et al.,, 2008; 2009; Kreaa et al., 2014; Suhad et al., 2015) show the importance of the target community once this resource is enriched with a content that suits the size of MSA and the expectations by Arabic NLP applications.

In this direction, it is worth mentioning that the experiences reported in other languages (English, Spanish, etc.) show the opportunity to use WordNets in more sophisticated and complex applications, such as humour detection (Reyes et al., 2010).

Nevertheless, AWN remains one of the few and important resources that can support the development of Arabic NLP applications regarding the following findings that we reported in (Abouenour et al., 2013):

- The current AWN considers the most common concepts and word-senses in PWN 2.0 so that its use in a cross-language context is possible.
- Similarly to other wordnets, AWN is connected to SUMO (Suggested Upper Merged Ontology) (Niles and Pease, 2001; Niles and Pease, 2003; Black et al.,, 2006). A significant number of AWN synsets was, indeed, linked to their corresponding concepts in SUMO. Statistics show that 6,556 synsets in AWN (65.56% of the synsets) are linked to 659 concepts in SUMO (65.9% out of 1000 concepts). Definitions that are provided by SUMO and its related domain-specific ontologies can be of great interest, complementing the information contained in AWN (SUMO also covers the Arabic culture domain).

[1] http://www.globalwordnet.org/AWN/

[2] In our work, we referred to the content of the first release.

Despite the above advantages of the AWN project, there were just a few attempts to enrich its content. These attempts relied on existing tools and resources. Del Gratta and Nahli (2014) proposed an enhancement of Arabic WordNet content using the PWN and AraMorph bilingual dictionary as bilingual resource.[3] This attempt resulted in adding new words and synonyms.

There was also another attempt to build a WordNet for an Arabic dialect based on the content of AWN and a parallel dictionary (Cavalli-Sforza et al., 2013).

Boudabous et al., (2013) proposed a linguistic method based on two steps to enrich AWN. The first step defines morpho-lexical patterns and the second step enriches semantic relations using these patterns. The Wikipedia resource is also used in both steps.

In comparison with those attempts, our approach is particular in that: (i) it uses techniques and resources with higher confidence, (ii) it is followed by a significant validation by lexicographers, and (iii) the usability of the enriched AWN resource is proven through experiments in the context of real-world application, i.e., Arabic QA.

3 New Content for AWN

3.1 Techniques used for AWN enrichment

The AWN lexical resource and its semantic relations showed the ability to support QE, QA and other applications (Abouenour et al., 2014; Abouenour et al., 2010), giving rise to the improvement of performance in comparison to the baseline systems, respectively. Nevertheless, this resource has many coverage shortcomings that we emphasized through the theoretical and experience-based perspectives. These shortcomings affect the usability of this resource and have been the reasons behind its limited use in Arabic NLP applications. To tackle this problem, we proposed in a previous research (Abouenour et al., 2013) an enrichement of AWN by targeting three types of content needed by Arabic QA as observed in the experience-based analysis:

- Instances or NEs enrichment: since our aim was to answer questions from the Web, we were interested in integrating YAGO (Suchanek et al., 2007) entities and relations in AWN after their automatic translation and validation. This kind of dynamic information is widely used in questions and other texts (Abouenour et al., 2013);
- Verbs and nouns enrichment: the coverage of these main Common Linguistic Categories is poor in AWN with respect to the Arabic lexicon and the coverage registered in experiments for TREC[4] and CLEF[5] nouns and verbs. The proposed enrichment consists in: (i) extending the list of verb senses in AWN using the translation of both English VerbNet (Kipper-Schuler 2006) and Unified Verb Index (UVI)[6] by means of three heuristic rules already used in the EuroWordNet project and (ii) refining the hyponymy relation among AWN noun synsets using a technique based on pattern discovery and Maximal Frequent Sequences (MFS) (García-Hernández 2007) over Web snippets and starting from a list of AWN synsets seeds (Abouenour et al., 2013).
- Broken plurals enrichment: BP is among the forms of plural that are widely and specifically used in Arabic. The analyzed questions showed that the enrichment of AWN forms in terms of BP is important to apply the QE process for a higher number of questions in real-world applications, especially QA (Abouenour et al., 2013).

The content to be added in AWN was generated from semi-automatic techniques, i.e. automatic translation and MFS, and using external resources such as YAGO, Arabic VerbNet, UVI and Web snippets. Therefore, a manual validation by lexicographers was required to guarantee a high confidence content in the enriched AWN. More details about how these techniques and resources were used can be found in (Abouenour et al., 2013).

In the next sub section, we present this validation.

3.2 Content validation

[3] http://www.nongnu.org/aramorph/english/dictionaries.html

[4] Text REtrieval Conference, http://trec.nist.gov/data/qa.html

[5] Cross Language Evaluation Forum, http://www.clef-campaign.org

[6] http://verbs.colorado.edu/verb-index/index.php

The manual validation focused on the new entries related to nouns, verbs and BPs. This validation involved 3 lexicographers that are also Arabic native speakers.

Each lexicographer has to validate the BP part and a specific part of nouns and verbs. For each entry, the decision to make is three-fold: (i) a given word is correct or not, (ii) the word can be member of the given synset or not, and (iii) a synset has the given relation (synonymy or hyponymy) with the given synset or not. In the latter case, the lexicographers can propose the right relation if it exists between both synsets. If the given relation is correct but not obvious, the lexicographer can mention this by "Lenient synonymy" or "Lenient hyponymy" tag. For instance, the synset "وافق - waAfaqa_v1AR" (to agree) has been assigned a new member which is the word "أقر" (to adopt), while the lexicographers classified this as "Lenient Synonymy". Figures 1, 2 and 3 show the results of this manual validation.

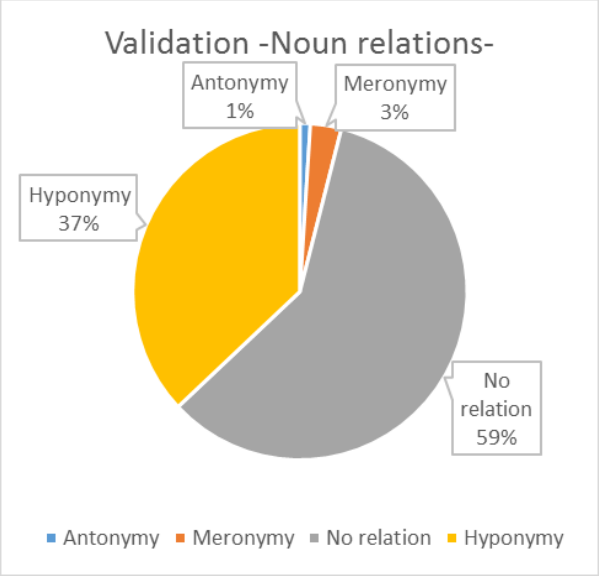


Figure 1. Manual validation of nouns

From Figure 1, 37% of the new proposed hyponymy relations between noun synsets were right. From the remaining cases, 4% were classified under the Meronymy (3%) and Antonymy (1%) relations. Hence, the overall successful relations generated by the MFS-based techniques represent 41% of the proposed ones.

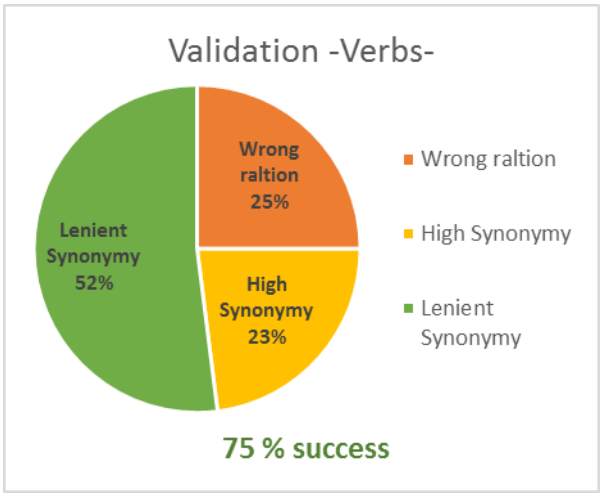


Figure 2. Manual validation of verbs

As for the verbs validation, the percentage of successful new proposed verb lemmas is roughly 75%. This percentage can be detailed as follows: (i) 23% of the new verbs can highly be regrouped into the given synset, and (ii) 52% have lenient synonymy with the given synset members.

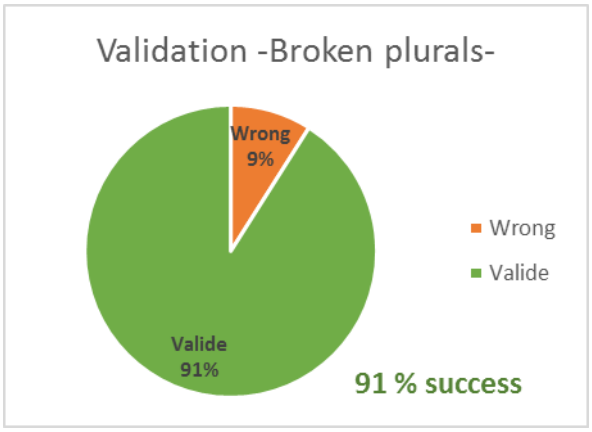


Figure 3. Manual validation of BP forms

Finally, the BP forms registered the highest percentage of success (91%), which means that the related external resource has higher precision.

4 New experiments using the enriched AWN

In (Abouenour et al., 2014; Abouenour et al., 2013; Abouenour et al., 2012), experiments were conducted in the field of Arabic QA systems using the AWN that we enriched following the techniques described above.

Let us recall that the experimental process relied on a three-level approach for Arabic QA as follows:

- Keyword-based level generating related terms from the enriched AWN;
- Structure-based level filtering passages and weighting those where question keywords and their related terms extracted from AWN appear in higher n-gram density;
- Semantic-based level comparing Conceptual Graphs (CG) of questions and candidate passages on the basis of their semantic similarity. CGs are built using AWN in addition to Arabic VerbNet.

For example, if the user question is " أين تتكون "اللويحات المتسببة في مرض الزهايمر؟" (Where do plaques causing the Alzheimer's disease are made up?). The keyword-based level generates related terms of "اللويحات", "المتسببة", "مرض" and "الزهايمر" from AWN, the structure-based level ranks the resulting passages on the basis of the N-gram density, i.e., a passage is highly ranked when these terms and their related terms appear in it and form a high density. The semantic-based level uses the AWN and the Arabic VerbNet resources to represent the Conceptual Graph (CG) (Sowa 1983) of the resulting ranked passages as well as the question itself in order to compare the semantic similarity between both CGs (Abouenour et al., 2014). The idea of the conducted experiments is to process the three levels with the original version of AWN and thereafter with the extended version, and compare performance before and after AWN content improvement.

The effectiveness of the new content added in AWN was proven by means of different experiments previously reported in (Abouenour et al., 2013; 2014):

- **Experiment #1**: The test-set of 2,264 CLEF and TREC questions (1999-2008) shows an improvement of the keyword-based and structure-based levels using the enriched AWN according to the considered QA measures: accuracy (+53%), Mean Reciprocal Rank (+45%), answered questions (+55%) and C@1 (+50%).
- **Experiment #2**: The test-set of the 2013 QA4MRE question set (Sutcliffe et al., 2013) is composed of 284 question classified into 4 topics, namely "Aids", "Climate change", "Music and Society" and "Alzheimer". This test-set shows the ability of the enriched AWN to support the semantic-based level at an acceptable extent (56% of the questions were represented in CG while this percentage is 61% for the candidate passages thanks to the new content of AWN).

Thus, the Experiment #1 shows that the QA performance based on the keyword-based and structure-based levels is better when using the enriched AWN. In order to show at which extent the performance of the Arabic QA process can be improved with the new AWN, we had to include the semantic-based level. The Experiment #2 shows that with the new AWN content, it is possible to semantically represent a higher percentage of questions and passages.

In this paper, we reconducted Experiment #1 using the 2,264 CLEF and TREC questions (1999-2008), considering also the semantic-based level. This is to assess the ability of the enriched AWN to improve the three levels of the Arabic QA approach regardless the considered test-set and passage collection (the QA4MRE test-set has local collection, the CLEF and TREC are assigned a Web collection).

The obtained results show a significant improvement in terms of C@1. Indeed, the three-level approach reaches a 0.51 C@1 which is higher than the 0.21 C@1 registered by the participating Arabic QA systems in QA4MRE Track 2012, including the IDRAAQ system (Abouenour 2012) that we built.

According to results of the manual validation of the new content as well as the promising results obtained with this content in a challenging task like Arabic QA, we decided to make available this content for the community of Arabic NLP.

The next section shows the steps followed to achieve this goal.,

5 The enriched AWN in LMF

5.1 Presentation of the LMF

LMF is the ISO standard for NLP lexicons and Machine Readable Dictionaries (MRD). The ISO code number for LMF is ISO-24613:2008. LMF has been developed under the aegis of TC37/SC4

by Gil Francopoulo and Monte George as editors and with Nicoletta Calzolari as convenor (Francopoulo et al., 2007).

The main goals of LMF are to use the lexical resources, manage and exchange the data among these resources
, and finally, provide a common model for the creation.

Types of individual instantiations of LMF can include monolingual as in our case, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons, for both simple and complex lexicons, for both written and spoken lexical representations. The descriptions range from morphology, syntax, and computational semantics to computer-assisted translation.

5.2 Transforming the raw extended AWN to LMF

As mentioned above LMF contains several packages such as syntax, morphology, semantics, MRD, and Multilingual notations. In our approach, we used the semantics and morphology packages.

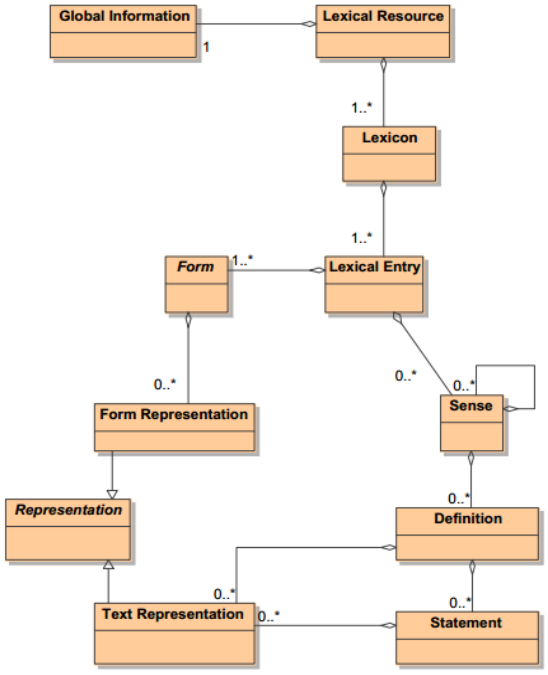


Figure 4. LMF core package**.**

After examining the content of the extended AWN and LMF, we made the following mapping to process the transformation:

AWN extended	LMF
Item	Sense
Word	Lexical Entry
Links	Synset relations
Forms (including BP and roots)	Word Form

Table 1. Mapping of the correspondence between LMF and the Element in the extended AWN.

Some attributes from the Element of the AWN needed to be transformed into element in the LMF to respect the DTD.

AWN extended attribute	LMF
Item id	Synset
Word value	Lemma
Form value	Written form

Table 2. List of attributes that became elements in LMF

```
<LexicalEntry id="sanap__5">
<Lemma partOfSpeech="n" writtenForm=" سَنَة
مَدَنِيَّة"/>
<Sense id="sanap__5_sanap__n3AR" syn-
set="sanap__n3AR"/>
</LexicalEntry>
<LexicalEntry id="sanap__6">
<Lemma partOfSpeech="n" writtenForm=" سَنَة
شَمْسِيَّة"/>
<Sense id="sanap__6_sanap__n3AR" syn-
set="sanap__n3AR"/>
</LexicalEntry>
….
<Synset baseConcept="3" id="sanap__n3AR">
<SynsetRelations>
<SynsetRelation relType="hyponym" tar-
gets="tAriyx_n1AR"/>
<SynsetRelation relType="hypernym" tar-
gets="sanap_n1AR"/>
</SynsetRelations>
</Synset>
```

Figure 5. Sample of AWN relation represented in LMF

5.3 Statistics

The process of transformation was done by a Java code using DOM API. After the transformation, we got 1036 lexical entries without part of speech (POS), this error occurred because some words did not have a corresponding item to get the POS from. To address this issue we followed a semi-automatic approach by using the SAFAR API (Jaafar and Bouzoubaa 2015; Souteh and Bouzoubaa 2011; Sidrine et al., 2010). The use of SAFAR is due to its integration of the most known Morphological Analyzers

(MA) and preprocessing tools, which simplify their use in a complementary way.

This process results in 1,012 POS from the integrated MAs, from which 24 POS were manually identified as wrong and were manually corrected.

At the end of our process, we obtained a document in LMF format totalizing 56,164 lexical entries (words grouped into synsets), 17,498 word forms and 41,136 synset relations.

6 Conclusion and Future Works

As a conclusion, the AWN project is important for Arabic NLP as witnessed by various attempts and research having used this resource. However, its content needs much extension and refinement. Our research tries to fill in a part of the gap registered between the current coverage of AWN and the expected one.

We presented in this paper the last experiments and manual validation of the AWN enrichment proposed in previous research (Abouenour et al., 2013). This enrichment was based on semi-automatic techniques and used external resources, thus the added content required refinement brought by lexicographers.

Also, the experiments conducted in the context of this paper shows a significant improvement of Arabic QA after using the enriched content of AWN. The new experiments together with the results previously presented in (Abouenour et al., 2013; 2014) show that regardless the considered test-set of questions, the new AWN content allowed an improvement of Arabic QA performance through various measures, especially the C@1 (from 0.21 to 0.51 after using the enriched AWN as a lexical resource in the keyword-based level and as a support resource for the semantic-based level).

Thus, the new release of AWN, manually validated by lexicographers and experimentally tested in the context of Arabic QA, is now available for the community in its LMF format[7]. The process of transformation into this format was described in Section 5 of this paper.

As future works, we can mention the requirement to add new relation types such as meronymy and antonymy that currently are slightly present in AWN. In addition, new techniques and resources could be investigated for this enrichment.

[7] The AWN v2 can be downloaded from the Open Multilingual WordNet project Web site. The resource is available at: http://compling.hss.ntu.edu.sg/omw/

Acknowledgement

The research of the second author was carried out in the framework of the grant provided by the Council for the Development of Social Science Research in Africa (CODESRIA) Ref. SGRT. 38/T13.

We would like to thank Dr Francis Bond, Nanyang Technological University Singapore, for his collaboration to make the new AWN v2 online for the community.

We would like to thank the lexicographers Dr Hakima Khammar, Faculty of Letters and Human Sciences, and Dr Rachida Tajmout, Mohammadia School of Engineers, for their contribution to validate the content of AWN v2.

References

Abbache, A., Barigou, F., Belkredim, F. Z., & Belalem, G. (2014). The Use of Arabic WordNet in Arabic Information Retrieval., International Journal of Information Retrieval Research (IJIRR), 4(3), 54-65.

Abouenour, L., Bouzoubaa, K., Rosso, P. (2010). An evaluated semantic QE and structure-based approach for enhancing Arabic Q/A. In the Special Issue on Advances in Arabic Language Processing for the IEEE International Journal on Information and Communication Technologies (IJICT), ISSN: 0973-5836, Serial Publications, June 2010.

Abouenour, L., Bouzoubaa, K., Rosso, P. (2012). IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval., CLEF'2012 (Online Working Notes/Labs/Workshop).

Abouenour, L., Bouzoubaa, K., Rosso, P. (2013). On the Evaluation and Improvement of Arabic WordNet Coverage and Usability. In: Languages Resources and Evaluation, vol. 47, issue 3, pp. 891-917.

Abouenour, L., Nasri, M, Bouzoubaa, K., Kabbaj, A., Rosso, P. (2014). Construction of an ontology for intelligent Arabic QA systems leveraging the Conceptual Graphs representation. Journal of Intelligent and Fuzzy Systems.

Al Khalifa, M., & Rodríguez, H. (2009). Automatically extending NE coverage of Arabic WordNet using Wikipedia. In Proceedings of the 3rd international conference on Arabic language processing CITALA'09, May, Rabat, Morocco.

Al-Zoghby, A. M., & Shaalan, K. (2015). Conceptual Search for Arabic Web Content. In Computational Linguistics and Intelligent Text Processing (pp. 405-416). Springer International Publishing.

Baldwin, T., Pool, P., & Colowick, S. M. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. In Proceedings of Coling 2010, Demonstration Volume, (pp. 37–40), Beijing.

Benajiba, Y., & Rosso, P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. In: Proc. Workshop on HLT & NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 26-31.

Benajiba, Y., Diab M., & Rosso, P. (2009). Arabic Named Entity Recognition: A Feature-Driven Study. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. Special Issue on Processing Morphologically Rich Languages, pp. 926-934. 2009.

Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Introducing the Arabic WordNet project. In Proceedings of the third international WordNet conference. Sojka, Choi: Fellbaum & Vossen (eds).

Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M.,Vossen, P., Pease, A., and Fellbaum, C., (2006). Introducing the Arabic WordNet Project. In Proceedings of the Third International WordNet Conference, Fellbaum and Vossen (eds).

Boudabous, M.M., Chaâben, N., Khedher, N., Hadrich Belguith, L., Sadat, F. (2013). Arabic WordNet semantic relations enrichment through morpho-lexical patterns, The First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), Sharjah, UAE, February 12–14.

Bouhriz, N., Benabbou, F., & Benlahmer, H. (2015). Text Concepts Extraction based on Arabic WordNet and Formal Concept Analysis International Journal of Computer Applications (0975 – 8887) Volume 111 – No 16, February.

Cavalli-Sforza, V., Saddiki, H., Bouzoubaa, K., Abouenour, L., Maamouri, M., & Goshey, E. (2013). Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources. In Computer systems and applications (aiccsa), 2013 acs international conference on (pp. 1–8).

Clark, P., & Fellbaum, C., & Hobbs, J. (2008). Using and extending WordNet to support question-answering. In: Proceedings of the Fourth Global WordNet Conference, University of Szeged, Hungary, pp. 111–119. COLING, pages 42.488.

Del Gratta, R.; Nahli, O., Enhancing Arabic WordNet with the use of Princeton WordNet and a bilingual dictionary, in Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in , vol., no., pp.278-284, 20-22 Oct. 2014.

El Amine, M. A. (2009). Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. In Proceedings of the 2nd conférence internationale sur l'informatique et ses applications (CIIA'09), May 3-4, Saida, Algeria.

Elghamry, K. (2008). Using the Web in building a corpus-based hypernymy-hyponymy lexicon with hierarchical structure for Arabic. Faculty of computers and information (pp. 157-165).

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., & Al khalifa, M. (2006). Arabic WordNet and the challenges of Arabic. In Proceedings of Arabic NLP/MT conference, London, U.K.

Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C., (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M.,Vossen, P., Pease, A., and Fellbaum, C., (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Fellbaum, C. (ed.), WordNet: An electronic lexical database. Cambridge, MA: MIT Press, 1998.

Francis Bond and Kyonghee Paik (2012). A survey of wordnets and their licenses In Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue. 64–71.

Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M., Soria C. 2007 Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. GLDV (Gesellschaft für linguistische Datenverarbeitung), Tubingen

García-Blasco, S., Danger, R., & Rosso, P. (2010). Drug-Drug interaction detection: A new approach based on maximal frequent sequences. Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN, 45, 263-266.

García-Hernández, R. A. (2007). Algoritmos para el descubrimiento de patrones secuenciales maximales. Ph.D. thesis, INAOE. September, Mexico.

García-Hernández, R. A. (2007). Algoritmos para el descubrimiento de patrones secuenciales maximales. Ph.D. thesis, INAOE. September, Mexico.

García-Hernández, R. A., Martínez Trinidad, J. F., & Carrasco-ochoa, J. A. (2010). Finding maximal sequential patterns in text document collections and single documents. Informatica, 34(1), 93-101.

Jaafar, Y., & Bouzoubaa, K. Arabic Natural Language Processing from Software Engineering to Complex Pipelines Cicling Cairo, Egypt 4/ 2015.

Karkar, A., Alja'am, J. M., Eid, M., & Sleptchenko, A. (2015). E-LEARNING MOBILE APPLICATION FOR ARABIC LEARNERS. Journal of Educational & Instructional Studies in the World, 5(2).

Kipper-Schuler, K. (2006). VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D Thesis.

Kreaa, A., Ahmad S Ahmad and Kassem Kabalan (2014). Arabic Words Stemming approach using Arabic WordNet. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.6, November 2014.

Mousser, J. A Large Coverage Verb Lexicon For Arabic. In: Proceedings of the 7th conference on International Language Resources and Evaluation (LREC) (2010), Valetta, Malta.

Mousser, J. Classifying Arabic Verbs Using Sibling Classes. In: Proceeding of the International Conference on Computational Semantics (IWCS) (2011), Oxford, UK.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In Proceedings of FOIS-2 (pp. 2–9), Ogunquit, Maine.

Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the 2003 international conference on information and knowledge engineering, Las Vegas, Nevada.

Reyes A., Rosso P., Buscaldi D. (2010). Finding Humour in the Blogosphere: The Role of WordNet Resources. In: Proc. 5th Global WordNet Int. Conf., GWN-2010, Bombay, India, January 31-February 4.

Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., & Martí, A. (2008a). Arabic WordNet: Semi-automatic extensions using Bayesian Inference. In Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008, May, Marrakech, Morocco.

Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Mart., M., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., & Fellbaum, C. (2008b). Arabic WordNet: Current state and future extensions. In Proceedings of the fourth global WordNet conference, January 22-25, Szeged, Hungary.

Rodríguez, R., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M.A., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., and Fellbaum, C., (2008). Arabic WordNet: Current State and Future Extensions.Proceedings of The Fourth Global WordNet

Sharaf, A. M. (2009). The Qur'an annotation for text mining. First year transfer report. School of Computing, Leeds University. December.

Sidrine, S., Y. Souteh, K. Bouzoubaa and T. Loukili, SAFAR: vers une Plateforme Ouverte pour le Traitement Automatique de la Langue Arabe. In: Proceeding of the 6th Conference of Intelligent Systems: Theory and Applications SITA (2010), May, Rabat, Morocco.

Souteh, Y., & Bouzoubaa, K. SAFAR platform and its morphological layer, In Proceeding of the Eleventh Conference on Language Engineering ESOLEC'2011, Cairo, Egypt, 14/ 12/ 2011.

Sowa J. F. (1984). Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Company.

Sowa, F. Conceptual Structures: Information Processing in Mind and Machine, 1984, Addison-Wesley Company.

Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In Proc. of the 16th WWW, pp. 697-706 (2007).

Suhad A. Yousif, Venus W. Samawi, Islam Elkabani and Rached Zantout (2015). Enhancement of Arabic Text Classification Using Semantic Relations of Arabic WordNet. Journal of Computer Science, Volume 11, Issue 3, Pages 498-509.

Sutcliffe, R., A. Peñas, E. Hovy, P. Forner, A. Rodrigo and C. Forascu (2013). Overview of QA4MRE Main Task, CLEF.

Vossen P. (ed). EuroWordNet, A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, 1999, The Netherlands.

Vossen, P. (ed). (1998). EuroWordNet, a multilingual database with lexical semantic networks. Kluwer Academic Publishers, The Netherlands.

Hydra for Web: A Browser for Easy Access to Wordnets

Borislav Rizov
Institute for Bulgarian Language
Bulgarian Academy of Sciences
body@dcl.bas.bg

Tsvetana Dimitrova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
cvetana@dcl.bas.bg

Abstract

This paper presents a web interface for wordnets named *Hydra for Web* which is built on top of Hydra – an open source tool for wordnet development – by means of modern web technologies. It is a Single Page Application with simple but powerful and convenient GUI. It has two modes for visualisation of the language correspondences of searched (and found) wordnet synsets – single and parallel modes. Hydra for web is available at: http://dcl.bas.bg/bulnet/.

1 Introduction

As the wordnets of the world are growing in number, implementations, applications, and the complexity of the information encoded in their relational format, wordnets data need tools for flexible but also readily accessible and easy to comprehend visualisation. Further, the tools used for creation of wordnets and visualisation of the lexical-semantic information also have to consider the relational character of the wordnet data in order to give the users access to most of the rich data without further complications and without much hidden information, especially the information concerning the relations between the synonym sets and concepts these synonym sets encode.

In the last decades, a number of web interfaces for browsing wordnet databases have been developed, with Wordvis, Mexidex, etc. among the most often used. Additionally, many web tools (mainly dictionaries) which use wordnet (especially the English wordnet) as a database for definitions and information about synonyms are available (e.g., Bee Dictionary; LookWAYUp; a2zDefined; cozyenglish, among others). Although based on wordnet, these dictionaries do not provide access to all the information about the relational organisation of the data which is one of the most valuable information in the wordnet.

There are popular user interfaces that visualise wordnet relationships as graphs. Wordvis (Vercruysse and Kuiper, 2013), for instance, do not support a parallel view of two or more wordnet language databases. Besides, being based on modern visual technological solutions, WordVis still prevents the whole needed information to be readily accessible, especially for wordnet developers. There are also tools that support parallel view as graphs such as Visual Browser (Neverilova, 2005) that can process wordnet synsets from a DEB server storage to convert them into RDF notation for visualisation (Horak et al., 2008). In the DEB platform environment, all the wordnets are stored on a DEBVisDic server; the client application supports a core module and individual modules for wordnets, so different data structure, workflow, external sources, etc. can be defined for each wordnet. The DEBVisDic was used as a basis for several multilingual projects including the Global Wordnet Grid (Horak et al., 2008). he web interface is very complicated though it is really useful for wordnet developers and for tasks involving heavy linking between wordnets, ontologies, and other lexical and semantic resources.

2 User interface and functionalities

The Hydra for web tool [1] is a web interface GUI implementation for wordnet that uses as backend the freely accessible open source modal logic tool for wordnet development Hydra (Rizov, 2008; Rizov, 2014).[2] The interface presented in this section is dependable on most of the functionalities of the

[1] Hydra for web can be checked at http://dcl.bas.bg/bulnet/.

[2] Hydra is freely available at http://dcl.bas.bg/en/hydra.html and through the META-SHARE repository at the Institute for Bulgarian Language: http://metashare.ibl.bas.bg/repository/search/.

Hydra which uses a convenient relational model to present and manage linguistic resources with relational structure.

Hydra for web is designed as Single Page Application that supports two modes – a Single Wordnet mode and Parallel Wordnets mode. Currently, it allows users to make queries into a wordnet database containing the Princeton wordnet (PWN) 3.0 (Fellbaum, 1999), the Bulgarian wordnet (BulNet) 3.0 (Koeva et al., 2004), and the Romanian wordnet (RoWN) (Tufis et al., 2013); the SentiWordnet data is in process of deployment. Thus, the web tool allows for searching into the databases of different language wordnets with a single query.

Hydra for webs interface is currently available in English, Bulgarian, and Romanian. The names of the relations and other elements were manually translated into Bulgarian (the part-of-speech – pos – and language markers – en, bg, ro – are still kept in English), and (partly) in Romanian. However, as the interface supports internationalisation, it is possible for other languages to be used.

The window has a top panel for switching the wordnets to be viewed which currently allows for the options of a Single mode – only selected synset is visualized, and the three pairs of wordnets in the Parallel wordnets mode, namely BulNet vs. PWN, BulNet vs. RoWN, and RoWN vs. PWN. However, the tool has functionalities that can further allow users to select and query any wordnet in the database.

The search panel is also present in both Single and Parallel wordnets modes. It allows searching for an exact match of a word string – a single word such as **[dance]** as shown on Fig. 1) or a multiword unit, e.g., **[barn dance]** – see Fig. 2.

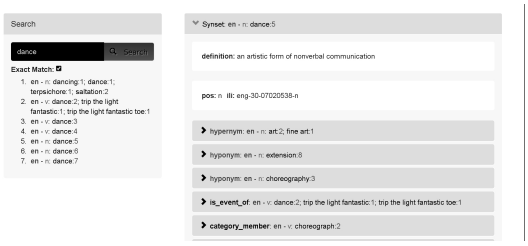


Figure 1: Hydra for web – exact match search

The non-exact match search returns any synset where the searched word string is found, as shown on Fig. 2 where the search for **[dance]** returns 24 different synsets from the Princeton word-

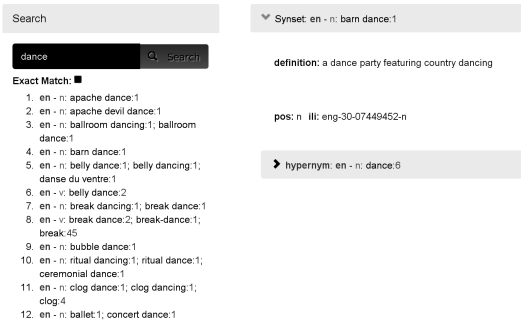


Figure 2: Hydra for web – non-exact match search

net database, among them two-word units such as **[apache dance:1]**, three-word units such as **[apache devil dance:1]**, and a hyphenated word string such as **[counter-dance:1]**, etc.

To limit the results shown, the search respects word (string) boundaries, i.e., the user can search only for whole words but not parts of the words (e.g., **[dance]** but not **[danc]** as this would return more than one hundred results – an option that is otherwise available in the Hydra software for the purposes of wordnet development. This also means that searching for the string **[dance]** will not return **[dancer]** or **[dancing]** although this word string is only part of the derived word.

2.1 Single wordnet mode

The layout in a single wordnet mode consists of two panels, namely a search panel to the left, and the synset view panel of the selected word to the right of the screen (as shown on Fig. 3 for the Princeton wordnet).

When searching for a word string in a Single wordnet mode, the search returns the synsets that contain the searched literals in all the languages in the database. The right panel displays the synset selected (e.g., the search for **[canis]** on Fig.3 returns all synsets with **[canis]** in English, Bulgarian, and Romanian wordnets).

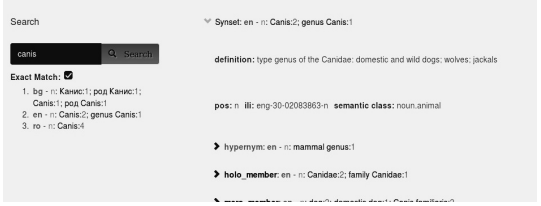


Figure 3: Hydra for web – single mode

2.2 Parallel wordnets mode

The parallel wordnets mode consists of three panels, with the second and the third panel visualising the parallel wordnets – see on Fig. 4. The two wordnet panels show the correspondences of the synset in the selected language. In this way, the user can search for a word in English, e.g., **[dog]** and with the selection of the synset **[dog:2, domestic dog:1, Canis familiaris:2]**, she can access the parallel synsets in the Bulgarian wordnet (BulNet) **[kuche:1, Canis familiaris:1]**, and in the Romanian wordnet (RoWN) – **[caine:1]**, as shown on 4. This option is very useful for fast checking the translation equivalents.

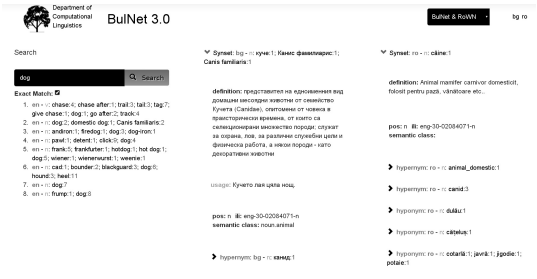


Figure 4: Hydra for web – parallel wordnets mode

On a small width (mobile), the responsive layout orders the panels successively – the search panel, then the synset views (see Fig. 5).

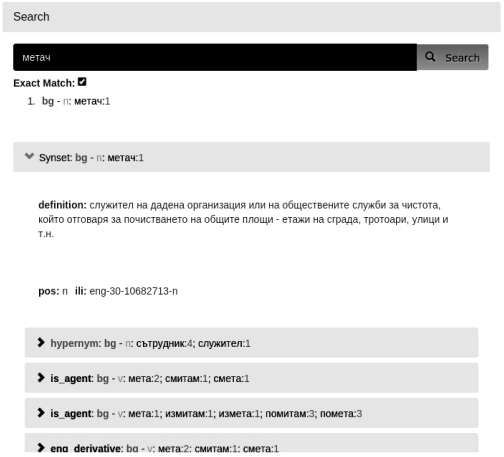


Figure 5: Hydra for web – small width (mobile) version

2.3 Synset visualisation

The elements of the synset structure are visualised in a predefined order, as shown on Fig. 6.

The literals in a synset are shown first - such as **[sweep:4, broom:2]**. The definition comes second

Figure 6: Hydra for web – ordering of synset elements

as shown by *'sweep with a broom or as if with a broom'*.

Relations that are most closely connected to the meaning of the synset are below the definition (these are usage examples in the extralinguistic relations USAGE: *'Sweep the crumbs off the table'* and *'Sweep under the bed'* on Fig. 6). The part-of-speech (pos) and the interlanguage index (ILI) are next, with hypernym(s) and hyponym(s) following, and all other relations – verb_group, is_agent, etc. (SNOTE coming at the bottom).

The relations are visually distinguished by their colour (in addition to their ordering in the synset structure) – this can be seen on the web. The part-of-speech and usage links are given in orange, the hypernyms are in blue, the hyponyms – in green, while the other relations are coloured in white.

The information in the relations are processed according to the synsets ILI. Thus, the current synset and the synset with the same ILI are marked with the arrow bullet turning red. The resulting visualisation in Hydra for web is shown on Fig. 7 where the verb synset **[sweep:4, broom:2]** is linked to the noun synset **[sweeper:3]** via the relations *is_agent* and *eng_derivative* that are both marked by the same arrow bullet – light-coloured on the Figure (on the web, the arrow turns red). The same is true for the same noun synset **[sweeper:3]** and the verb synset **[sweep:5]**. The same notification appears on the synsets in the parallel wordnet – the Bulgarian wordnet on this Figure (and if these synsets are available in the parallel wordnet).

The visualisation is recursive in a sense that every relation that leads to a synset (hypernym, holo_part, etc.) is expandable in the same way as the root one. The data like pos, ILI, etc. are available immediately, while the relations are loaded by means of AJAX query, but without blocking the

UI.

Figure 7: Hydra for web – selection of elements

3 Implementation

Hydra for web is implemented by means of modern web technologies and libraries. Its source code is relatively small, straightforward and it is easy to maintain and extend *Hydra for web* with new features.

Hydra for web is built with Node.js[3] and Express[4]. It is a single page application and uses one of the most popular HTML, CSS and JS frameworks – Bootstrap[5].

Hydra for web is themed in Slate from Bootswatch[6]. Bootstrap makes easy the GUI to be responsive, and so it is mobile friendly.

For the html rendering, the very clean and elegant JADE template engine[7] is used.

Many of the tasks in the GUI are solved in the client with the use of Knockout.js[8] framework. It uses declarative bindings, dependency tracking and provides automatic UI refresh.

The wordnet data retrieval is made by means of the Wordnet Service. The retrieval uses AJAX and is completely asynchronous (non-blocking).

3.1 Wordnet Service for wordnets

Wordnet service is a RESTful web service written in Python and Twisted[9]. The service uses the Hydra API to extract the information from the wordnet database.

The services API provides requests for searching and extracting the objects from the database (synsets, literals, and texts). It is also useful for retrieving the neighbours of a particular wordnet object by all the relations (hypernyms, hyponyms, antonyms, etc.) and its correspondent synsets in the other languages.

3.2 Hydra library

Hydra is implemented in Python, using the platform independent GUI library Tkinter. The data is managed by a MySQL server. The program allows users to query any number of wordnets simultaneously. Individual wordnets can be synchronized, allowing simultaneous visualisation of the equivalent synsets in different languages.

The program allows concurrent access by multiple users. The changes in the database are available to all users right after they are made and this option is very useful for simultaneously working wordnet developers. The important thing in our case is that it provides API for wordnet data extraction and manipulation which is at the heart of the Wordnet Service for Wordnet.

4 Applications

The most obvious application of Hydra for web is for queries into different wordnets, as well as for viewing parallel wordnet resources. Such parallel data can be used for comparative lexical and other linguistic studies. It highlights the links between words and concepts.

Hydra for web is a convenient tool for using wordnet from every place, computer, phone or other device with internet connection.

One obvious application, alongside the wordnet databases behind it, is as a multilingual dictionary.

Searching and return of single words and multiword units may help in building certain models for text identification and categorisation, word sense disambiguation, etc.

The list of results (single words and multiword units) returned contains also information about other (synonym) words and the part-of-speech of the resulting words.

[3]Node.js is a JavaScript runtime: https://nodejs.org/

[4]Web application framework for Node.js http://expressjs.com/

[5]http://getbootstrap.com/

[6]https://bootswatch.com/

[7]http://jade-lang.com/

[8]http://knockoutjs.com/

[9]https://twistedmatrix.com/

References

Christiane Fellbaum. 1999. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ales Horak, Karel Pala, and Adam Rambousek. 2008. The Global Wordnet Grid Software Design. *Proceedings of the Fourth Global Wordnet Conference*, Szeged, Hungary, 194–199.

Svetla Koeva, Tinko Tinchev, and Stoyan Mihov. 2004. Bulgarian Wordnet – Structure and Validation. *Romanian Journal of Information Science and Technology*, 7(1–2):61–78.

Zuzana Neverilova. 2005. Visual Browser: A Tool for Visualising Ontologies. *Proceedings of I-KNOW'05*, Graz, Austria 453–461.

Borislav Rizov. 2008. Hydra: A Modal Logic Tool for Wordnet Development, Validation and Exploration. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, ELRA.

Borislav Rizov. 2014. Hydra: A Software System for Wordnet. *Proceedings of the Seventh Global Wordnet Conference*, Tartu, Estonia, 142–147.

Dan Tufis, Verginica Barbu Mititelu, Dan Stefanescu, and Radu Ion. 2013. The Romanian Wordnet in a Nutshell. *Language Resources and Evaluation*, December 2013, 47(4): 1305–1314.

Steven Vercruysse and Martin Kuiper. 2013. WordVis: JavaScript and Animation to Visualize the WordNet Relational Dictionary. *Advances in Intelligent Systems and Computing*, 179: 137-145.

Towards a methodology for filtering out gaps and mismatches across wordnets: the case of noun synsets in plWordNet and Princeton WordNet

Ewa Rudnicka
Wrocław University of Technology, Poland
ewa.rudnicka@pwr.edu.pl

Wojciech Witkowski
University of Wrocław, Poland
wojciech.witkowski@uwr.edu.pl

Łukasz Grabowski
Opole University, Poland
lukasz@uni.opole.pl

Abstract

This paper presents the results of large-scale noun synset mapping between plWordNet, the wordnet of Polish, and Princeton WordNet, the wordnet of English, which have shown high predominance of inter-lingual hyponymy relation over inter-synonymy relation. Two main sources of such effect are identified in the paper: differences in the methodologies of construction of plWN and PWN and cross-linguistic differences in lexicalization of concepts and grammatical categories between English and Polish. Next, we propose a typology of specific gaps and mismatches across wordnets and a rule-based system of filters developed specifically to scan all *I(inter-lingual)-hyponymy* links between plWN and PWN. The proposed system, it should be stressed, also enables one to pinpoint the frequencies of the identified gaps and mismatches.

1 Introduction

Since the development of the first wordnet, that is, Princeton WordNet (henceforth PWN, cf. Fellbaum, 1998), a number of wordnets for the multitude of languages have followed. Their construction was usually based on either of the two major approaches: the *merge* approach assuming manual wordnet creation on the basis of language data collected from dictionaries (e.g. Hindi Wordnet cf. Narayan et al 2001) and the *expansion* approach taking the content and structure of one of the existing wordnets as input for translation to another language (e.g. IndoWordnet, Bhattacharyya, 2010). Some wordnets were also built by means of the 'mixed', *transfer-and-merge* (also called *merge-expand*) method joining the previously mentioned approaches (cf. EuroWordNet, Vossen, 2002; Romanian Wordnet, Cristea et al, 2004; Open Multilingual WordNet, Bond and Foster, 2013). Thus, the process of their construction was often intertwined with the process of their linking to PWN, which served as the 'input' wordnet. The obvious advantage of the *expansion* and, partly, *transfer-and-merge* method is time and cost effectiveness, yet it looses on capturing the actual structure and content of the lexical system of a language in question. One of the few wordnets created independently of PWN is plWordNet, a wordnet of Polish language (henceforth plWN), built manually with the help of a unique method of extracting information on lexico-semantic relations from large text corpora (cf. Piasecki et al., 2009; Maziarz et al., 2014). Although much more time-consuming and expensive, such method of construction yields a resource which more closely reflects a lexical system of a language. The noun part of plWN has been already linked to PWN using a set of 7 inter-lingual relations (modelled on by those used in EuroWordNet, cf. Vossen, 2002). All of them were introduced manually by a team of bilingual lexicographers working in accordance with a detailed, three-stage mapping procedure (cf. Rudnicka et al., 2012). Already the first effects of this mapping process have showed a variety of contrasts in the structure and content of plWN and PWN. Some of them could be traced to different concept and (partly) grammatical categories' lexicalization between English and Polish; other resulted from different con-

struction methods of plWN and PWN. The structure of Princeton WordNet was motivated by the results of psycholinguistic studies, while its content was largely based on individual lexicographers' choices and the data obtained from monolingual dictionaries.

In the paper, we present the results of a final stage of plWN to PWN noun synsets mapping and a proposal of a rule-based system of filters that enables one to identify the sources and measure the degree of gaps and mismatches between the two wordnets. The paper is organised as follows: in Section 2 the manual mapping strategy is described and the statistics of inter-lingual relations are given, in Section 3 different types of gaps and mismatches revealed in the course of mapping are discussed, in Section 4 a procedure for filtering out gaps and mismatches across wordnets is presented, in Section 5 the results of filtering are presented, while in Section 6 the conclusions are given.

2 Mapping results

Mapping between plWN and PWN was carried out by a team of trained and supervised bilingual lexicographers working in accordance with a detailed mapping procedure (cf. Rudnicka et al., 2012). The mapping was performed on the level of synsets (as in the case of all world wordnets) and consisted in linking plWN and PWN synsets corresponding in meaning and position in wordnet structure by means of one of 7 hierarchically ordered inter-lingual relations, such as S*ynonymy, Partial synonymy, Inter-register synonymy, Hyponymy, Hypernymy, Holonymy* and *Meronymy*. The mapping procedure consisted of three major steps: recognising the sense of the source synset, searching for the most corresponding target synset and selecting an inter-lingual relation to be established. In their work, lexicographers consulted the whole variety of available dictionaries and encyclopedias. Also, they were supported by a custom-designed system of automatic prompts, based on the relaxation labeling algorithm paired with a filtering by a large cascade dictionary (cf. Kędzia et al., 2013).

So far, the process of mapping has been conducted for noun and adjective synsets. The noun part is almost finished, the work on adjective part is still in progress. Therefore, in this paper we focus on the results of noun mapping. In Table 1, we compare basic numbers for plWN and PWN, while in Table 2 the counts of the established I(nter-lingual) relations are given.

Data analyzed	**plWN**	**PWN**	**plWN - Nouns**	**PWN - Nouns**
no. of synsets	198029	109505	123709	87695
no. of lexical units	269347	190049	166938	154385
no. of lemmas	182374	151162	126482	124879

Table 1: plWN 2.3. and PWN 3.1. general statistics[1]

I-relation	**Instances all**	**Instances nouns**
Synonymy	37191	33613
Hyponymy	85338	67680
Meronymy	6428	6428
Partial synonymy	5166	3767
Hypernymy	4142	4077
Holonymy	3025	3025
TOTAL	**141290**	**118770**

Table 2: plWN 2.3. to PWN 3.1. mapping statistics: instances of I-relations

One may plausibly argue that the most striking feature of the obtained results is the frequency of *I(inter-lingual)-hyponymy* links, which is two times higher than the frequency of the 'highest priority' *I(nter-lingual)-synonymy* links. Such results definitely point to a number of discrepancies between the content and structure of the two wordnets. Some sources of those discrepancies were already identified (Rudnicka et al., 2012): they encompass those due to the differences between lexical systems of English and Polish and those relating to different construction methods of the two wordnets under scrutiny. Still, the paper presents and discusses the results of only the very first stage of the mapping process. As shown in Table 2, the tendency of the double predominance of *I-hyponymy* over *I-synonymy* has prevailed and there arises the need to explain the reasons behind it.

[1]The data given in Table 1 and Table 2 are taken from the official plWordNet's website: http://plwordnet.pwr.wroc.pl/wordnet/stats

3 Gaps and mismatches across wordnets

In this section, we will discuss, first, the contrasts resulting from different construction methods of plWN and PWN and, second, various types of gaps and mismatches that may occur between lexical systems of natural languages (on the example of English and Polish).

Hence, the main research problem addressed in this paper refers to identification of any gaps and mismatches between linguistic data stored in two electronic lexical databases, that is, PWN and plWN. In general terms, the language-pair specific gaps and mismatches, which will be described in greater detail later in this paper, result from the following factors: 1) differences in structures of PWN and plWN; 2) differences in methodologies used to compile PWN and plWN; 3) specificity of mapping procedure applied to plWN and PWN; 4) systemic differences between English and Polish lexicon, morphology and syntax (e.g. varying degrees of lexicalization; differences in encoding of grammatical categories (e.g. gender); varying degrees of morphological productivity of affixal derivation); 5) cross-cultural differences between English and Polish. These differences, as applicable to the lexical data stored in plWN and PWN, are discussed in greater detail in the following sections 3.1 and 3.2.

3.1 Structural and methodological differences between plWN and PWN

In their analysis of the first stage of mapping of noun synsets, Rudnicka et al. (2012) identify two main sources of the observed predominance of I-hyponymy over I-synonymy links: these include contrasts in wordnet structure resulting from the application of different construction methods for each wordnet as well as morpho-lexico-semantic gaps and mismatches attributable to cross-linguistic differences between English and Polish. The former ones include the use of *Hyponymy and* vs. *Hyponymy or*, the use of different intra-lingual relations (*Hyponymy* vs. *Meronymy*) to capture the same conceptual dependencies and the occasional placement of mass/count, singular/plural and hyponym/hypernym lexical units in the same synset on the PWN side. The latter ones consist of greater degree of lexicalization of such grammatical categories in Polish as gender, diminutiveness and augmentativeness.

In the present study, the results of the final stage of mapping of noun synsets are analysed with an eye to other sources of the predominance of *I-hyponymy* relation over *I-synonymy* relation. Since mapping was carried from plWN to PWN side, we have searched for peculiarities of plWN's structure in order to develop a methodology that would lead to the creation of a large number of synsets lacking direct equivalents on PWN side. Three such groups of synsets were identified: gerund forms, multi-word expressions and forms belonging to marked registers. The latter ones will be discussed in the subsequent section, since most of them belong to the category of the so-called referential gaps (cf. Svensen 2009, also called 'cultural mismatches' cf. Bond et al. 2014). In plWN, there is a number of gerund forms under the category of noun. This is motivated by their ability to function as both verb participles and nouns. The creators of PWN did not adopt a similar strategy, hence there are not that many "-ing forms" in PWN noun synsets. Consequently, there could not be many *I-synonymy* links established in this category. The creators of plWN originally introduced many multi-word expressions and only recently a complex procedure for identifying multi-word lexical units has been applied (cf. Maziarz et al., 2015). The structural and methodological differences between PWN and plWN are summarized in Table 3 below:

plWN	**PWN**
hyponymy *and* {musical 1} - 'musical' hypo > {film 1}'film' {musical 2} - 'musical' hypo > {przedstawienie 7}- 'play'	**hyponymy *or*:** {musical 1} hypo > {movie 1}, hypo > {film 2}(a play or film whose action and dialogue is interspersed with singing and dancing)
use of different intra-lingual relations (hyponymy vs. meronymy) to capture the same conceptual dependencies	
{naszyjnik 1}[necklace] - mero-> {biżuteria 1} [jewellery]	{bracelet 2} hyponymy > {jewellery 1}
mass and count nouns in the same synset	
{mebel 1} (piece of furniture),{umeblowanie 2}(furniture)	{furniture 1, piece of furniture 1} 'furnishings that make a room or other area ready for occupancy'
singular and plural in the same synset	
{pieróg 2} 'small	{dumpling 1, dumplings

boiled ball of dough with various stuffing'	1} 'small balls or strips of boiled or steamed dough
gerunds in plWN	
{kopanie 2} 'the act of kicking'	----------
plWN multi-word synsets	
{eskadra bobmowa 1} 'bomber squadron'	----------
{eskadra niszczycieli 1} 'destroyer squadron'	----------

Table 3: Structural and methodological differences between plWN and PWN handled by *I-hyponymy* relation

3.2 Morpho-lexical mismatches ad lexico-semantic gaps

As already mentioned in the previous section, the second important source of the high frequency of *I-hyponymy* links between plWordNet and Princeton WordNet identified by Rudnicka et al. (2012) are the differences between lexicalisation (and structuralisation) of concepts and grammatical categories between English and Polish. The latter ones are called **morpho-(syntactic) mismatches** by (Bond et al., 2014: 252). They result from systemic differences between languages; in practice this means varying degrees of lexicalization of certain grammatical categories, such as number or gender (e.g. Pol. *kuzyn/kuzynka* vs. Eng. *cousin*; Pol. *Amerykanka* vs. Eng. *American girl*). In other words, certain concepts are "lexicalized through words with different morpho-syntactic properties across languages" (ibid.). [This resembles what Catford (1965/1978) refers to as category shifts in the context of translation]. Such differences may also result from varying degrees of morphological productivity of derivational morphemes, notably in the case of diminutives (e.g. Pol. *samochód / samochodzik* vs. Eng. *car*), augmentatives (Pol. *dom/domisko* vs. Eng. *house*). Due to its productivity, we expect a high number of such cases in plWN to PWN mapping. Also, their recognition should not pose major problems, since they can be identified by intralingual plWN morpho-lexical relation links holding between lexical units, such as *Żeńskość* - 'Feminine gender', *Diminutywność* - **'**Diminutiveness' and *Augmentatywność* - **'**Augmentativeness' (cf. Maziarz et al., 2012).

The more challenging part are the differences arising from different lexicalisation of concepts. These are widely discussed in the literature. Cvilikaite (2006: 129) argues that the so-called *lexical gaps* should be identified on the level of individual meanings of lexical items. The reason for that is that translators are usually interested in context-specific individual meanings of lexical items rather than semantic structures of lexemes, often polysemous ones (ibid.). In fact, lexical gaps occur when a given concept is not lexicalized in a given language (Cvilikaite, 2006) or when it is it is expressed with a lexical unit in one language and with a free combination of words in another language (Bentivogli, Pianta and Pianesi, 2000; Hutchins & Somers, 1992). In specialist literature, one may find a number of typologies of lexical gaps and mismatches between data stored in bilingual dictionaries or multilingual wordnets (e.g. Svensen, 2009; Bond et al., 2014); also, specialist literature on translation studies and linguistic typology addresses the problem of incompatibility of lexicons of different languages (e.g. Talmy, 2000). In this paper, we aim to synthesize the aforementioned typologies in order to capture lexical gaps and mismatches between linguistic data, more specifically, between nouns stored in PWN and plWN.

The first group are **referential gaps** (Svensen 2009: 271), which roughly correspond to what Bond et al. (2014: 252) subsume under an umbrella label of 'cultural concepts'. These include culture-specific concepts that are lexicalized in one language and not lexicalized in another. Such concepts are tied to the history, customs, traditions making up the cultural heritage of a given linguistic community. For example, concepts such as *szmalcownik* **'**a person who extorted money from Jews under threat of denouncing on them; a word used in the period of German occupation of Poland during World War II' or *noc Kupały* or *kupała* **'**summer solstice celebrated on the night of 23/24 June, the shortest night during entire year' are cultural concepts specific to or deeply rooted in the Polish culture and hence not lexicalized in English. In a similar vein, names of national holidays, institutions, administrative functions and units, historical names, etc. fall into this category.

The next group are the so-called 'pragmatic lexicalizations' (Bond et al., 2014: 252), which correspond to what Svensen (2009: 273) refers to as **lexical gaps**. In short, these include concepts that are familiar in many cultures yet they are not lexicalized in all of them (Bond et al., 2014: 252). Because such concepts are known across cultures, they reveal differences in lexicalization

of their conceptual structure e.g. Eng. *uncle* vs. Pol. *stryj/wuj*; Pol. *palec* vs. Eng. *finger/toe*. The last group of gaps resulting from cross-cultural differences are the so-called **differences in perspective** (Bond et al., 2014: 252) or **standpoint gaps**, that is, the differences resulting from structuring conceptual reality from various perspectives or standpoints (who does what to whom and how) e.g. Eng. *married* vs. Pol. *żonaty/mężatka*; Eng. *house/home* vs. Pol. *dom*; Eng. *bring/take* vs. Pol. *przynieść*. Table 4 summarizes the gaps and mismatches discussed in the foregoing.

plWN	PWN
Differences arising from productive morphological derivation	
Diminutives: {samochód 1} 'a car', {samochodzik 2} 'a small car'	**Diminutives:** {car 1}
Augmentatives: {dom 1} 'a house', {domisko 1} 'a large house'	**Augmentatives:** {house 1}
Referential gaps/Cultural concepts	
{szmalcownik 1} 'blackmailer'	----------
{noc Kupały 1} 'summer solstice celebration'	----------
Lexical gaps/Pragmatic lexicalization	
{stryj 1} 'father's brother', {wuj 1} 'mother's brother'	{uncle 1} 'the brother of your father or mother'
{palec 1} 'digit of a hand or foot'	{finger 1}, {toe 1}
{kończyna górna 1} 'upper limb', {kończyna dolna 1} 'lower limb'	{limb 1} 'one of the jointed appendages of an animal used for locomotion or grasping'
Differences in perspective/ Standpoint gaps	
{żonaty 1} 'married man', {mężatka 1} 'married woman'	{married 1} 'a person who is married'
Morpho-lexical mismatches: grammatical gender lexicalization	
{kuzyn 1} 'male child of your uncle or aunt'	{cousin 1} 'the child of your aunt or uncle'
{kuzynka 1} 'female child of your uncle or aunt'	

Table 4: Taxonomy of gaps and mismatches between plWN and PWN

4 Methodology: a procedure for filtering out gaps and mismatches

In this section, we propose a rule-based system of filters designed for the recognition of the different types of gaps and mismatches that may occur in wordnet mapping. Based on the typology of gaps and mismatches described in Section 3, the system scans all *I-hyponymy* links from plWN to PWN side. Its ultimate aim is to filter out, first, contrasts resulting from different construction methods of plWN and PWN, second, all and any systematic mismatches resulting from different lexicalization patterns of grammatical categories, third, cultural gaps. Ultimately, the system aims to produce the set of proper lexical gaps. The system's implementation is conducted in a number of steps presented in greater detail below.

Step 1. I-hyponymy

- select all plWN **noun** synsets that have I-hyponymy relation to PWN synsets.
- Create a list of plWN - PWN noun synset pairs.

Step 2. From the list obtained in [1] filter out:

- all plWN gerund forms. Do this by filtering out those synsets whose L(exical)U(nit)s have *Synonimia międzyparadygmatyczna V-N (Cross-paradigm Verb-Noun synonymy)* relation
- all plWN synsets that belong to [sys(tematics)] domain
- all plWN synsets built from LUs denoting proper names or LUs derived from proper names. Do this by removing all plWN synsets which have Typ/*Egzemplarz (Type / Instance)* relation.
- all plWN multi-word synsets which are not tagged as multi-word (fixed) phrases in plWN
- (keep on a separate list) all Princeton WordNet synsets that are built in the following manner: {LU1 (lemma1)}, {LU2 (lemma1+s)} or {LU1 (lemma1)}, {LU2 (lemma1+ing)})

Step 3. Filtering out morpho-lexico-syntactic mismatches. From the set remaining after completion of [Step 2], sort out all plWN synsets that include lexical units which have specific intra-lingual lexical unit relations (such as (1st) *żeńskość (feminine form),* (2nd) *diminutywność (diminutiveness) & augumentatywność (augumentativeness)*) to other plWN LUs. For each filtering stage save the list of filtered out results.

Step 4. From the set remaining after [Step 3] has been carried out, (tests for filtering out cultural gaps) - remove PWN synsets with relation *Topic/Domain*[2] (keep on a separate list)

Step 5. Filter out Polish domain specific synsets - From the list of synsets remaining after the implementation of [Step 4], sort out synsets containing LUs belonging strictly to Polish language domain. The target are those synsets whose LUs have the following register markers[3]: ##K: pot., ##K: posp., ##K: wulg., ##K: daw., ##K: środ. or ##K: reg. marked.

5 Results

The results are summarized in Table 5. The filtering procedure resulted in removing out only 44.83% i.e. 30679 synsets out of the overall 67680 plWN synsets mapped onto PWN synsets by means of *I-hyponymy* relation. The biggest percentage of those constitute gerund forms and proper names (21%). The former ones, together with multi-word synsets[4] (5.39%) (both removed in Filter 2), are the effect of plWN's methods of construction. The next groups in line are diminutives and augmentatives (5.32%) and feminine forms (3.73%) (removed in Filter 3) which reflect the specificity of Polish morphology. Another group are PWN synsets that have the intra-lingual *Topic-Domain* relation (5.79%), removed in Filter 4 aimed at removing mainly culture-dependent concepts found in PWN. The last and the least numerous group are plWN synsets including lexical units marked for register (3.4%), also aimed at removing culture-specific concepts.

With respect to the data presented in Table 5, it should also be noted that the remaining number of 37001 synsets is too large for any manual analysis and hence it needs to be treated with caution; the said number is primarily influenced by the size differences between plWN and PWN. Accordingly, in order to minimize the effect of database size, the results of filtering were divided into three groups defined in terms of dictionary and wordnet coverage. The results of this division are presented in Table 6.

F	Details	no. of synsets removed	% removed
2	gerunds, pr. names, [sys] domain	14478	21%
	plural number errors	155	0.2%
	multi-word synsets (but not multi-word units)	3649	5.39%
3	diminutives and augmentatives	3606	5.32%
	feminine form	2526	3.73%
4	topic / domain relation	3921	5.79%
5	[posp] - everyday common	91	0.13%
	[pot] - everyday non-standard	1137	1.68%
	[reg] - regional variants	173	0.2%
	[srod] - social group specific	0	0%
	[wulg] - vulgar	9	0.01%
	[daw] - archaisms	934	1.38%
TOTAL REMOVED		30679	44.83%
REMAINING - candi-		37001	----------

[2]In [Filter 3] all plWN synsets that hold *I-hyponymy* relation to PWN synsets with *Topic/Domain* relation within PWN are removed. That may seem a 'drastic' move, yet we aimed at removing all potential cultural gaps. Thus, the number of synsets removed by [Filter 3] - 3921 - should be treated with caution as it is overestimated, since it also includes synsets that lexicalize concepts common to both Polish and English.

[3]The abbreviations used to mark relevant registers are explained in Table 5.

[4]What is meant by a multi-word synset is a synset whose LUs are built of more than one word but are not treated as multi-word units in the sense of Maziarz et al. (2015) e.g. {eskadra niszczycieli 1} - 'fighter sqaudron', where multi-word units are defined as those composed of a sequence of words that cannot be separated from each other and occur in a fixed order, e.g. {chlorek amonu 1} - 'ammoniun chloride'.

dates for actual lexical gaps		

Table 5: Filtering procedure results

Synset type	Instances
[Group 1] - synsets whose lemma are not present in Princeton WordNet and whose equivalent was found in a 'cascade dictionary'	9420
[Group 2] - synsets whose lemma are not present in Princeton WordNet and whose equivalent was not found in a 'cascade dictionary'	18567
[Group 3] - synsets whose lemma are present in Princeton WordNet and whose equivalent was found in a 'cascade dictionary' but which are not related via I-synonymy Pol-Eng or I-partial synonymy Pol-Eng relation	9014

Table 6: Group division of possible candidates for lexical gaps in plWN and PWN comparison

The data in Table 6 show that the number of candidates for actual lexical gaps can be lowered by 9420 Group 1 synsets, a decision that yields 27581 possible candidates. The resulting number, however, is still too large for a manual analysis. However, with due caution it can be lowered by the reduction of cases in Group 2, where, given large enough language resources, a substantial number of English equivalents can be found.

6 Discussion and conclusions

This study constituted an attempt at identifying any gaps and mismatches between lexical data, specifically, nouns, stored in two inter-liked electronic databases, that is, plWN, the wordnet of Polish, and PWN, the Princeton wordnet of English. The results confirmed our initial hypotheses that the gaps and mismatches result from wordnet-specific structural and methodological differences, specificity of the interlingual mapping procedure, systemic differences between Polish and English as well as cross-cultural differences. In order to identify any gaps and mismatches across two aforementioned wordnets, a custom-designed filtering procedure was developed and described in this paper. The results of filtering procedure revealed groups of synsets (mainly gerunds and multi-word synsets) whose existence and high numbers are the effect of the assumed methodology of construction of plWN. Next, the procedure revealed groups of synsets such as diminutives, augmentatives and feminine forms that reflect the specificity of the morphology of the Polish language and, finally, PWN synsets holding the relation *Topic-Domain* and plWN synsets marked for different registers that attest to the presence of culture-dependent concepts were filtered out/identified.

The approach presented in this study has a number of limitations which need to be addressed in future research. First, the results revealed up to 28000 synsets that are required to be manually analyzed, the process that is bound to be time-consuming and labour-intensive. Second, the procedure described in this paper does not allow, in its current form, checking the filtering results against larger lexical resources (e.g. larger than the cascade dictionary used in this study) where more potential equivalents for Polish lemmas could be found. To this end, it is possible to use additional resources such as Polish-English parallel corpora (e.g. PARALELA[5], a large collection of Polish-English parallel texts); this could provide an improvement in terms of filtering out the results. A small-scale manually conducted experiment aimed at identification of equivalents in corpora and Internet resources has revealed that the number of lemmas present in plWN and, at the same time, not found in the cascade dictionary used in the filtering procedure can be lowered by approximately 37% (see Rudnicka and Witkowski, 2015). Finally, it should be stressed that at the current stage there are no means of filtering out exactly those Polish synsets whose potential equivalents could be multi-word units with compositional meaning in English. Removal of all plWN multi-word synsets with no special tag (cf. Maziarz et al., 2015 for separate treatment of multi-word units in plWN) as a part of [Filter 2] appears to be a significantly imprecise tool with respect to compositionality of meaning, i.e. this operation removed all multi-word synsets in one fell swoop, regardless of the internal semantic dependencies of the words in a multi-word unit. To resolve this problem, it is possible to target the relevant plWN multi-word synsets by identifying instances in which the synsets at hand have *Hyponymy* and *Meronymy: element* relation links to

[5]http://paralela.clarin-pl.eu

synsets whose LUs have the same bases as the LUs in question.

As for the future, the procedure described in this study and its results may come in useful for exploration of the different types of equivalence relations obtained between lexical data stored in plWN and PWN. This could enable one to turn the study results into actionable knowledge useful for lexicographers and translators, among others.

Acknowledgement

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

References

Arleta Adamska-Salaciak. 2014. Bilingual Lexicography: Translation Dictionaries. *International Handbook of Modern Lexis and Lexicography.* Springer-Verlag, Berlin Heildelberg.

Luisa Bentivogli, Emanuele Pianta and Fabio Pianesi. 2000. Coping with lexical gaps when building aligned multilingual wordnets. [In:] *Proceedings of LREC 2000*, Athens, Greece, pp. 993-997.

Pushpak Bhattacharyya. 2010. IndoWordNet. Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.

Francis Bond. 2005. Translating the Untranslatable. A Solution to the Problem of Generating English Determiners. *CSLI Studies in Computational Linguistics*.

Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Adam Pease and Piek Vossen 2014. A Multilingual Lexico-Semantic Database and Ontology. [In:] *Towards the Multilingual Semantic Web* Paul Buitelaar and Philipp Cimiano (eds), Springer pp 243–258 (Publisher's page). http://compling.hss.ntu.edu.sg/who/bond/pdf/2014-msw-omw.pdf

Jurgita Cvilikaite. 2006. Lexical Gaps. Resolution by functionally complete units of translation. *Darbai ir Dienos*, 45, 127-142. donelaitis.vdu.lt/publikacijos/dd45_cvilikaite.pdf

Christiane Fellbaum (ed). 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusets.

W. John Hutchins and Harold L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press, London.

Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies,* 13: 123-142.

Dipak Kumar Narayan, Debasri Chakrabarty, Prabhakar Pande, Pushpak Bhattacharyya. 2001. *An Experience in Building the Indo WordNet - a WordNet for Hindi*. 1st International Conference on Global WordNet (GWC 02), Mysore, India.

Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. Approaching *plWordNet* 2.0. [In:] *Proceedings of the 6th Global Wordnet Conference*, Matsue. 189-196.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the Transfer-and-Merge WordNet Construction: plWordNet and a Comparison with WordNet. [In:] *Proceedings of RANLP*, Hissar.

Marek Maziarz, Stan Szapkowicz and Maciej Paisecki. 2015. A Procedural Definition of Multiword Lexical Units. [In:] *Proceedings of RANLP, Hissar.*

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda 2009. *A Wordnet from the Ground Up.* Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki and Stan Szpakowicz. 2012. A Strategy of Mapping Polish WordNet onto Princeton WordNet. [In:] Proceedings of COLING 2012. ACL.

Ewa Rudnicka and Wojciech Witkowski. 2015. Towards the Methodology for Extending Princeton WordNet. *Cognitive Studies 15*.

Bo Svensen. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making.* Cambridge University Press, Cambridge.

Leonard Talmy. 2000. *Toward a Cognitive Semantics. Typology and Process in Concept Structuring*. MIT Press, Cambridge, Massachusetts.

Piek Vossen (ed.). 2002. *EuroWordNet General Document,* Version 3 (final) URL: http://www.hum.uva.nl./~ewnAlfred.

Folktale similarity based on ontological abstraction

Marijn Schraagen
Digital Humanities Lab, Utrecht University, The Netherlands
M.P.Schraagen@uu.nl

Abstract

This paper presents a method to compute similarity of folktales based on conceptual overlap at various levels of abstraction as defined in Dutch WordNet. The method is applied on a corpus of Dutch folktales and evaluated using a comparison to traditional folktale similarity analysis based on the Aarne–Thompson–Uther (ATU) classification system. Document similarity computed by the presented method is in agreement with traditional analysis for a certain amount of folktale pairs, but differ for other pairs. However, it can be argued that the current approach computes an alternative, data-driven type of similarity. Using WordNet instead of a domain-specific ontology or classification system ensures applicability of the method outside of the folktale domain.

1 Introduction

A folktale is a specific type of narrative that is particularly suitable for analysis of semantic structure. Although folktales may differ in various aspects, such as the characteristics of the main actors or the sequence of events, often similarities can be identified on a more general or more abstract level. In this paper similarity between folktales is investigated using an explicit abstraction of text according to the WordNet concept hierarchy. A comparison is provided to conventional folktale motif analysis. An example of folktale similarity on various levels of abstraction is provided by the folktales *Sleeping Beauty* and *Snow White*, which both feature a princess as specific character and a variable number of enchanted objects at a more general level. Similarities regarding events occur at various levels as well, for example the princess in *Snow White* is asked by the seven dwarfs to perform household tasks, whereas the girl protagonist from *Hansel & Gretel* is ordered by the witch to do housework. In this case both the actors and the actions are similar at various levels, depicted in Figure 1. This notion of abstraction-based semantic similarity can be computed automatically using a machine-readable concept hierarchy such as WordNet. The paper is structured as follows: Section 2 describes characteristics of folktales and provides an overview of the resources used in the analysis, Section 3 discusses related work, Section 4 provides details on similarity computation, Section 5 contains experimental results and a comparison to existing folktale analysis approaches, and Section 6 concludes.

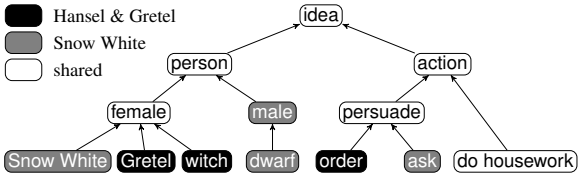


Figure 1: Example partial hierarchy showing concepts from *Snow White* and *Hansel & Gretel*.

2 Folktale similarity

The folktale texts used in the current research are extracted from the Dutch Folktale Database (Meder, 2010). This collection contains over 40,000 folktales (including jokes, urban legends, etc.) from written and oral sources, in Dutch, Frisian and several contemporary and historical Dutch dialects[1]. The database is partially annotated using Aarne-Thompson-Uther (ATU) codes (see the remainder of this section for details). The database, maintained by the Meertens Institute, is available for research purposes upon request. For the current research a pilot set of 16 traditional fairy tales is used from a single original source (van Dongen and Grooten, 2009), with a total of

[1] http://www.verhalenbank.nl, in Dutch

id	*description*	*category*	*ATU*	
F451.5.1.2	Dwarfs adopt girl as sister	Marvels → Spirits and demons → Underground Spirits	709	Snow White
Q2	Kind & Unkind	Rewards and punishments	440, 480, 513, 571, ...	Frog King, Kind & Unkind Girls, Wonderful Helpers, Golden Goose
F911.3	Animal swallows man (not fatally)	Marvels → Extraordinary occurrences → extraordinary swallowings	123, 333, 700	The Wolf & the Seven Kids, Little Red Riding Hood, Tom Thumb
F913	Victims rescued from swallower's belly	Marvels → Extraordinary occurrences → extraordinary swallowings	123, 333, ...	The Wolf & the Seven Kids, Little Red Riding Hood
J144	Well-trained kid does not open to wolf	The wise and the foolish → Wisdom aqcuisition → education	123	The Wolf & the Seven Kids
K1832	Disguise by changing voice	Deceptions → Deception by disguise	123	The Wolf & the Seven Kids

Table 1: TMI motif examples.

33,022 words. This set provides folktales in grammatically correct, modern Dutch, which increases applicability of natural language processing tools and methods. Several folktales in this set do not appear in the ATU catalog, illustrating the applicability of the current method on non-traditional folktale sources.

In the current research the Dutch Cornetto database is used (Vossen et al., 2013) to obtain term abstractions. Cornetto is modeled after the Princeton WordNet, which is a widely used ontology for English concepts (Fellbaum, 1998) containing a comprehensive set of terms and (hierarchical) relations for an extensive variety of domains. Concepts are organized in sets of (approximate) synonyms, called *synsets*, which are connected by relations such as hypernymy and meronymy. Cornetto contains over 92,000 lemmas and is available under academic license[2].

Traditionally, folktales are analyzed using the Thompson Motif Index (TMI). This index is a set of over 12,000 story elements (motifs), classified in semantic categories and subcategories (Thompson, 1960). Some examples are provided in Table 1. The motifs in this index are often specific to a single folktale (or *folktale type*, i.e., the set of variants of a story that are considered the same folktale), however more general motifs are used as well. The Aarne–Thompson–Uther (ATU) classification system (Uther, 2004) describes a folktale type as a list of motifs (typically two or three to about 20) from a subset of nearly 1,900 elements from the TMI, divided into thematic categories and subcategories. The ATU classification is centered around the type of protagonists and the general theme of the folktale, while the TMI is centered around events and relations. This may introduce semantic relatedness differences between the two systems, for example the classification of ATU 123 as *Animal tales–Wild animals and domestic animals* compared to ATU 333 which is classified as *Fairy tale–supernatural opponent*, while two out of the total of four motifs of ATU 123 are also found in ATU 333 (see Table 1).

[2]An open source database using the structure of Cornetto and translations of the content of English WordNet is available as an alternative (Postma and Vossen, 2014a).

3 Related work

Folktale similarity using WordNet-based term matching has been previously investigated (McIntyre and Lapata, 2010; Lestari and Manurung, 2015) using the hierarchical similarity measure of Wu and Palmer (1994). In this approach a folktale is considered sequential, with similarity computation based on alignment of the sequence of actions and actors. In contrast, the current approach considers the (non-sequential) presence of terms and term abstractions, similar to a bag-of-words approach, while preserving event or situation similarity by comparing folktales on a sentence level. Abstraction based on Dutch WordNet for folktale similarity has been used by Nguyen et al. (2013), using abstractions of verbs as one of several features involved in similarity computation. The abstraction feature did not improve the results significantly, which is attributed to limited coverage of the abstraction lexicon and inaccuracy of the grammatical analysis.

Characterizing semantic relations between folktales using TMI motifs is discussed by Karsdorp et al. (2012), presenting the conclusion that motif-based methods suffer from the limited amount of motif overlap between folktales. A search tool for TMI motifs using WordNet based semantic abstraction is presented in (Karsdorp et al., 2015). A mapping of nominal phrases to folktale actors using a domain-specific ontology for term abstraction is described by Declerck et al. (2012).

An unsupervised exploration and visualization method for concept clustering in folktales has been proposed by Honkela (1997), using self-organizing maps trained on word trigrams. Natural computing approaches using (phylo)genetic algorithms are used to study variation within folktale types and between closely related types, using TMI motifs and other story elements as features (Ross et al., 2013; Tehrani, 2013). A vector-based method for semantic folktale clustering using Latent Semantic Mapping is described in (Vaz Lobo and Martins de Matos, 2010).

Several semantic relatedness measures that use WordNet as knowledge base have been proposed, see, e.g., (Pedersen et al., 2004) for an overview. Considering hierarchy traversal, well-known approaches include the Wu-Palmer measure mentioned above, which defines similarity between two nodes as the path length from the first shared parent node to the root node of the hierarchy, and the Leacock-Chodorow measure, which finds the shortest path between two concepts (scaled for specificity of the hierarchy). Further graph-topological information is incorporated using PageRank (Agirre et al., 2009). Evaluation of graph-based semantic relatedness measures has been performed using comparison to human word-pair similarity ratings, e.g., (Postma and Vossen, 2014b). Recent approaches of similarity computation include path length weighting strategies (Gao et al., 2015) and domain-specific data (McInnes and Pedersen, 2015).

Using WordNet for similarity of documents has been investigated by, e.g., (Hotho et al., 2003; Sedding and Kazakov, 2004) for the task of document clustering. These approaches represent a document as a bag-of-words, consisting of terms in the document as well as term synonyms and hypernyms from WordNet. However, it is concluded that the investigated approach of adding WordNet relations does not improve clustering results significantly. Similar methods for document clustering do show improved results, e.g., (Wang and Taylor, 2007), suggesting a large impact of preprocessing and sense selection procedures. Further applications include information retrieval, matching a WordNet-expanded query to a set of (non-expanded) documents (Varelas et al., 2005).

Note that many approaches using WordNet for semantic similarity focus either on pairs of concepts (or synsets, words, lemmas, etc.), document clusters, or, in the folktale setting, variants of the same story. These tasks are generally motivated by the availability of evaluation resources, such as human concept similarity ratings, the Reuters categorized news corpus, or folktale corpora tagged by story type, respectively. In contrast, the current approach attempts to construct a network of documents based on semantic relatedness, by comparing document pairs on (non-sequential) sentence level. Evaluation of this approach is arguably less straightforward, however this task and the proposed WordNet-based method provide a shift in focus compared to traditional approaches.

4 Method

In the current approach a document collection is compared at sentence level. First, sentence boundaries, lemmas and part-of-speech tags are obtained using the Frog toolkit (van den Bosch et al., 2007). Lemmas tagged as noun (including proper names), adjective, or verb are selected (except for the common verbs *be, have, can* and *will*). The set of lemmas for a sentence is compared to the set of lemmas for all other sentences in all other folktales in the corpus. If a matching lemma is not found in the compared sentence then the WordNet hierarchy is consulted for a match at a higher level of abstraction, using the match level to adjust the similarity score. The similarity of two sentences is computed as the total of all match scores relative to the combined size of the lemma sets. Formally, the score $s \in [0, 1]$ equals $(\sum_i \frac{1}{level(a_i)} + \sum_j \frac{1}{level(b_j)})/(|A| + |B|)$ for sentences A and B as sequences of WordNet lemmas and $level(\ell)$ defined as the minimum level of the lemma ℓ that matches a lemma (at any level) in the compared sentence. After computing similarity scores for all sentence pairs, for each (ordered) folktale combination (f_A, f_B) the relative number of sentences in f_A is counted for which the most similar sentence in the corpus originated from f_B. The procedure is described formally in Algorithm 1, an example is provided in Figure 2. For the mapping of sentence lemmas to WordNet the synset with the lowest WordNet sense number is selected, corresponding to some extent to the 'default' sense. Incorrect senses are assumed to be related or to have a minimal effect given the document size (cf. (Hotho et al., 2003)). During hierarchy traversal a random hypernym is selected for a given synset to limit the amount of branching. In the example the

Algorithm 1 F×F document pair similarity.

function WORDNETLOOKUP (sentence S)
set of tuples $R \leftarrow \emptyset$
for all $(term, position, level = 0)$ in S **do**
  synset $syn \leftarrow$ WORDNET$(term)$
  while $syn \neq undefined$ **do**
    $R \leftarrow R \cup \{(syn, position, level)\}$
    $level \leftarrow level + 1$
    $syn \leftarrow$ WORDNETHYPERNYM(syn)
return R

function MAIN (document set F)
for all folktales $f_a \in F$ **do**
  for all sentences $A \in f_a$ **do**
    $score_{max} \leftarrow 0, f_m \leftarrow undefined$
    $SYN_A \leftarrow$WORDNETLOOKUP(A)
    for all folktales $f_b \in F - \{f_a\}$ **do**
      for all sentences $B \in f_b$ **do**
        $s \leftarrow 0, m_a \leftarrow [\infty], m_b \leftarrow [\infty]$
        $SYN_B \leftarrow$WORDNETLOOKUP(B)
        for all combinations $((t_a, p_a, \ell_a) \in SYN_A, (t_b, p_b, \ell_b) \in SYN_B)$ **do**
          if $t_a = t_b$ **and** $\ell_a < m_a[p_a]$ **then**
            $m_a[p_a] \leftarrow \ell_a$
          if $t_a = t_b$ **and** $\ell_b < m_b[p_b]$ **then**
            $m_b[p_b] \leftarrow \ell_b$
        for all matches m in m_a, m_b **do**
          $s \leftarrow s + \frac{1}{m}$
        if $\frac{s}{|A|+|B|} > s_{max}$ **then**
          $s_{max} \leftarrow \frac{s}{|A|+|B|}$
          $f_m \leftarrow f_b$
    $scores[f_a][f_m] \leftarrow scores[f_a][f_m] + \frac{1}{|A|}$
return $scores[\,]$

two occurences of the verb *do* are associated to the different synsets *d_v-2652* {*do,behave*} (sentence level) and *d_v-2045* {*do,work,execute*} (abstraction level). Synset matching succeeds at the shared hypernym *d_v-2859* {*act*}.

The distance measure applied in the current research uses elements from both Wu-Palmer and Leacock-Chodorow (see Section 3), by measuring the distance from a source synset to the first shared parent node. As a comparison, Table 2 provides the correlation of this measure to the Dutch gold standard human similarity ratings of Postma and Vossen (2014b). The asymmetrical definition of the similarity measure allows for several options for the score assigned to a concept pair, which has a marked influence on the correlation values. The

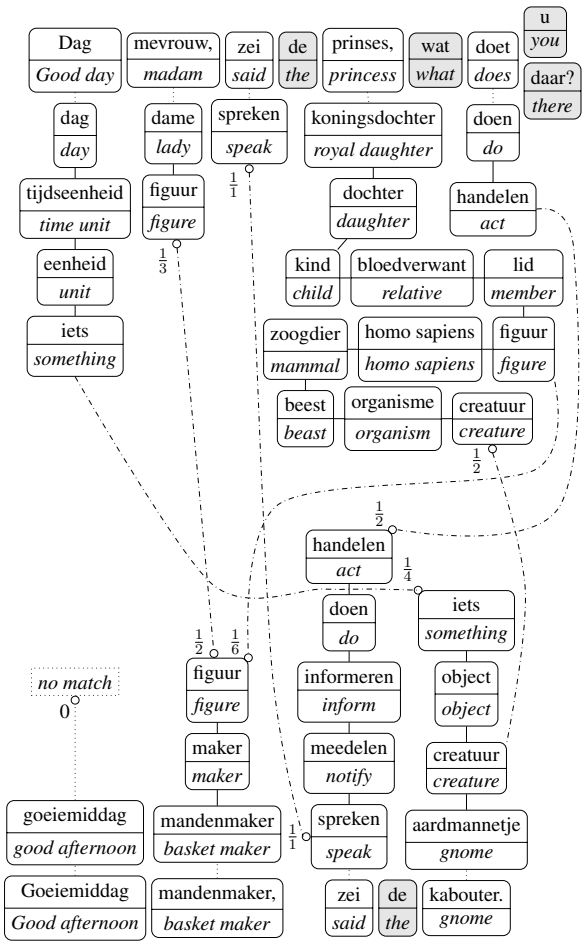


Figure 2: Sentence similarity example showing a score of $((\frac{1}{4} + \frac{1}{2} + \frac{1}{1} + \frac{1}{6} + \frac{1}{2}) + (0 + \frac{1}{3} + \frac{1}{1} + \frac{1}{2}))/(5 + 4) = 0.47$. English word translations in italics, grey nodes represent terms not listed in WordNet.

overall values are somewhat lower than the correlation for the hierarchy traversal measures reported by Postma and Vossen (2014b) on Dutch WordNet, which might be caused by the lack of hierarchy depth awareness of the current method. However, the current measure is not intended as a stand-alone word pair similarity computation, instead it is part of an asymmetrical sentence matching procedure intentionally designed for matching on any level of the hierarchy.

scored term	McNo	McRel	McSim	RgNo	RgRel	RgSim
source	**0.64**	**0.60**	0.64	0.54	0.48	0.55
target	0.44	0.39	0.49	0.53	0.53	0.54
lowest	0.59	0.54	0.63	0.53	0.52	0.55
average	0.62	0.56	**0.65**	**0.58**	**0.55**	**0.59**
highest	0.58	0.53	0.61	**0.58**	0.54	**0.59**

Table 2: Spearman's ρ correlation between the abstraction measure and human similarity ratings.

ATU	*title*	123	327	333	410	440	480	513	533	563	571	709	GTF	NG	RS	GF	WL
123	Wolf & Seven Kids	—	8.45	**17.61**	4.93	7.04	4.23	3.52	7.75	5.63	9.15	*11.27*	2.11	2.82	6.34	2.11	2.11
327	Hansel & Gretel	3.04	—	9.57	3.91	6.09	5.65	*11.30*	8.70	4.35	6.52	**14.78**	2.17	6.96	9.13	2.61	2.17
333	Red Riding Hood	**14.29**	**17.01**	—	3.40	3.40	6.12	7.48	9.52	2.72	4.76	3.40	2.72	2.04	8.16	4.08	2.04
410	Sleeping Beauty	1.42	5.67	2.84	—	7.09	4.26	**13.48**	9.93	5.67	4.26	*12.06*	7.80	8.51	9.93	4.96	0
440	Frog King	4.03	8.05	3.36	*11.41*	—	5.37	**14.09**	**14.09**	4.03	4.03	*10.74*	4.70	8.05	1.34	3.36	0
480	Kind & Unkind Girls	5.15	**22.06**	5.88	2.94	4.41	—	6.62	*10.29*	8.09	6.62	5.88	1.47	5.88	*10.29*	2.21	0
513	Wonderful Helpers	1.45	*11.64*	6.18	9.45	8.00	3.64	—	*11.27*	4.36	5.09	8.36	3.64	9.45	5.45	2.91	1.09
533	Speaking Horsehead	1.83	*10.98*	1.83	7.93	**13.41**	4.27	**16.46**	—	4.27	6.71	8.54	4.88	6.10	6.71	3.05	1.22
563	Table, Ass & Stick	3.29	*11.84*	3.95	1.32	9.21	4.61	9.87	4.61	—	9.21	*10.53*	5.92	3.29	7.24	7.89	1.32
571	Golden Goose	3.68	*11.76*	5.15	5.15	8.09	2.21	**14.71**	8.09	5.88	—	8.82	5.88	5.15	5.88	5.15	3.68
709	Snow White	4.21	*12.30*	6.47	8.09	7.12	5.83	9.06	5.83	4.21	4.85	—	6.80	6.47	6.15	4.85	2.27
GTF	Golden Tuning Fork	0.80	7.20	4.80	*11.20*	2.40	4.00	*12.80*	**16.00**	1.60	6.40	8.80	—	8.00	6.40	7.20	0
NG	Nightingale	4.10	**14.55**	4.10	7.46	5.60	3.36	*11.19*	7.84	2.24	6.34	8.58	6.72	—	4.48	8.96	0.75
RS	Red Shoes	2.68	8.72	**13.42**	8.72	3.36	6.04	7.38	6.04	4.03	6.04	*12.08*	4.70	8.72	—	2.01	3.36
GF	Gardener & Fakir	2.79	*11.73*	3.91	6.15	4.47	5.59	*11.73*	7.82	5.03	2.79	8.38	6.15	**14.53**	4.47	—	2.79
WL	Waterlilies	2.08	6.25	8.33	*10.42*	2.08	6.25	*10.42*	6.25	0	6.25	**14.58**	0	8.33	*12.50*	4.17	—

Table 3: Pair-wise WordNet-based similarity scores for the test corpus. Thresholds indicated in italics (10.0) and bold (13.0).

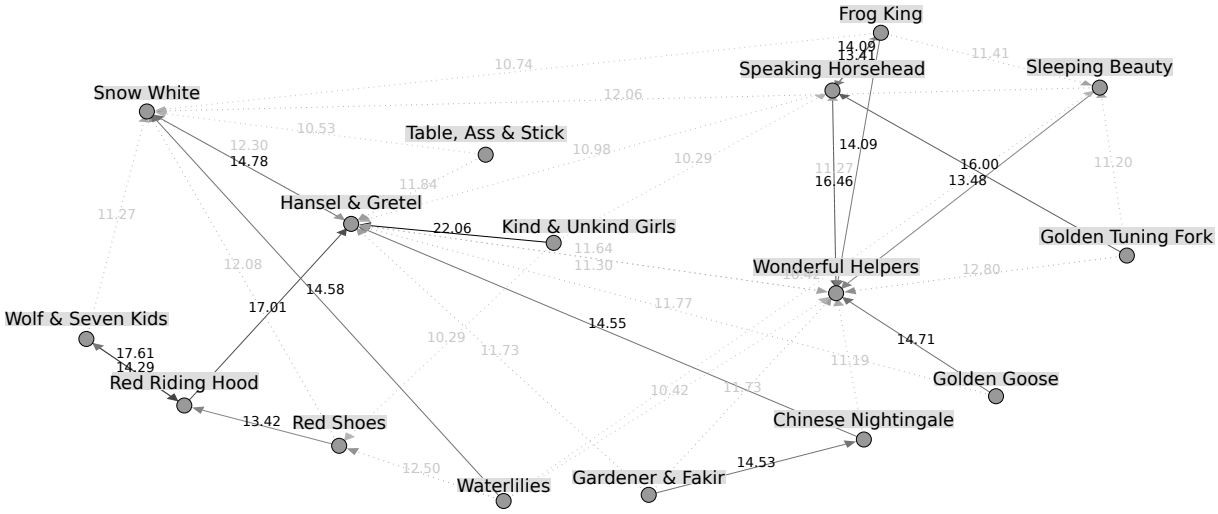


Figure 3: Graph of directed pairwise folktale similarity scores for threshold 10.0 (dotted edges) and 13.0 (solid edges).

5 Results and evaluation

Application of the current method on the folktale test corpus results in a matrix of pair-wise directed similarity scores, shown in Table 3. The graph of scores above a threshold of 10.0 (i.e., the most similar sentence for at least 10% of the sentences in folktale A was found in folktale B) is provided in Figure 3. The graph contains a number of central nodes, most notably *Snow White*, *Hansel and Gretel*, and *The Wonderful Helpers*. These nodes can be interpreted as representing a prototypical folktale, more specifically the fairy tale subgenre.

Increasing the similarity threshold to 13.0 (solid lines in Figure 3) reveals two clusters in the graph. The left cluster contains folktales featuring civilian protagonists, who find themselves in potentially harmful circumstances. The right cluster contains royal protagonists dealing with issues of moral values. The exception is *Snow White*, which has a royal protagonist, who is however banned from the royal court, living as a civilian house guest annex maid, and subject of murder attempts.

For comparison, the same method is applied without using WordNet abstractions, i.e., counting overlap in (lemmatized) terms as found in the text. This comparison (see Figure 4) shows that plain term overlap is less structured or partitioned in general and pair-wise relations display less topic overlap as compared to the abstraction method.

To provide an evaluation of the proposed similarity measure, a comparison is performed to the traditional ATU classification and associated TMI motif sets. To address the problem of limited motif overlap between folktales, the hierarchical TMI numbering system can be used for partial or abstract motif overlap using an approach similar to the term similarity computation described in Sec-

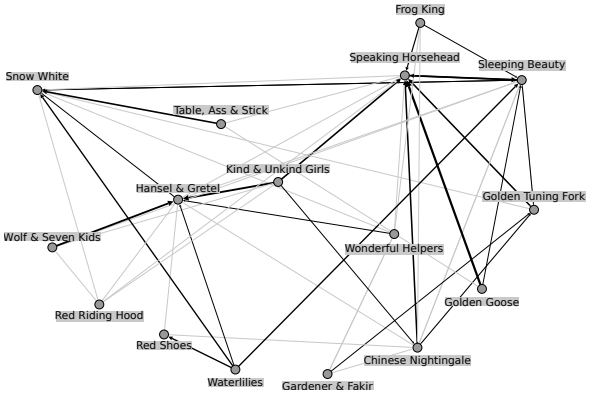


Figure 4: Plain term overlap scores for threshold 10.0 (grey edges) and 13.0 (black edges).

tion 4 (see Table 4 for examples of motif matching). Similarity scores based on the level of overlap of motif pairs are shown as a graph in Figure 5.

5.1 Graph comparison

In order to provide graph-theoretical support for the visual correspondence claim, the differences in degree distribution for corresponding nodes can be quantified. For this analysis the *assortativity coefficient* (Newman, 2002; Piraveenan et al., 2008) is used, which considers the number of edges of a node and the direct neighbors of this node and compares these numbers to the overall degree distribution of the network. In Figure 6 the assortativity values for both types of similarity computation are shown as a correlation graph. The figure shows correspondences for peripheral nodes and the central *Snow White* node, as well as differences for nodes which are central in one of the two graphs only. Note that assortativity is not a measure of centrality as such. The definition takes into account the difference in degree between neighboring nodes, i.e., a larger part of the network is measured as compared to single degree count.

6 Discussion and future work

The current WordNet-based similarity measure divides the example folktale set into two clusters corresponding to civilian protagonists in threatening circumstances and royal protagonists presented with moral choices, respectively. This result shows that the method is able to differentiate general topics in folktales based on overlap in terms and term abstractions. Using term ab-

ATU	*Title*	*Motif description*	*Motif code*	*match level*
123	The Wolf & the Seven Kids	Disguise by changing voice	K1832	
333	Little Red Riding Hood	Wolf puts flour on his paw to disguise himself	K1839.1	4
533	The Speaking Horsehead	Disguise as goose-girl (turkey-girl)	K1816.5	3
533	The Speaking Horsehead	Imposter forces oath of secrecy	K1933	2
709	Snow White	Compassionate executioner: substituted heart	K0512.2	1

Table 4: Example motif matches for *The Wolf & the Seven Kids*.

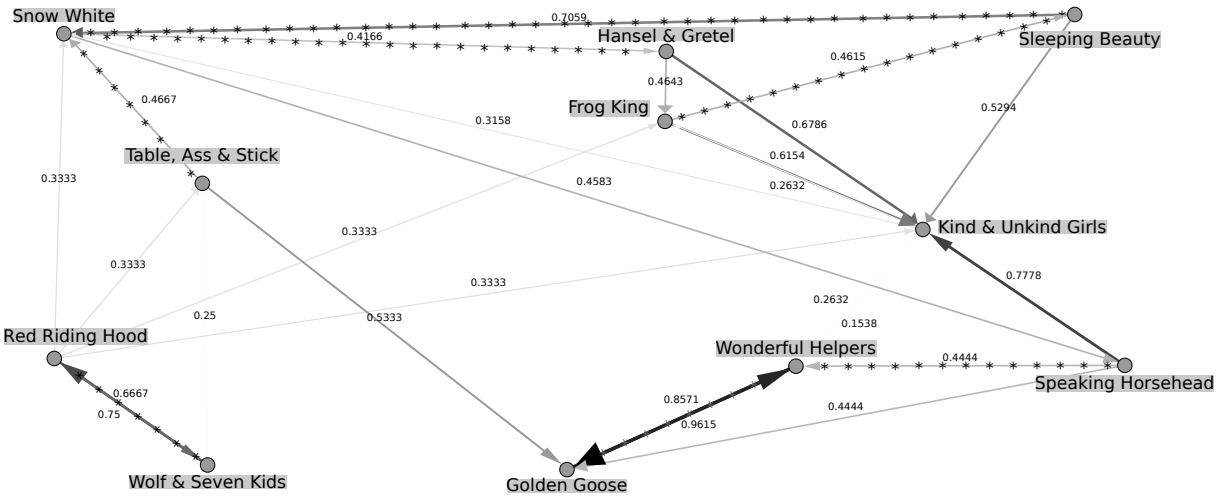


Figure 5: Graph of directed pairwise folktale similarity scores using the Thompson Motif Index. The number of motif pairs with the highest overlap is shown for pairs of documents (relative to the number of source document motifs), restricted to the top two highest ranked target nodes for each document.

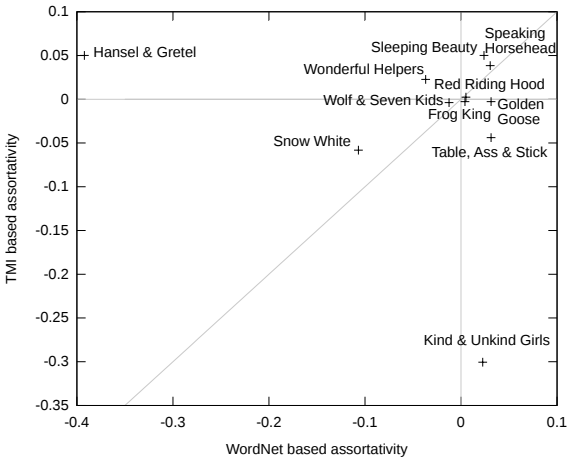


Figure 6: Correlation of assortativity of similarity graphs.

straction increases the level of clustering. The comparison with traditional folktale motif analysis shows corresponding similarity relations and centrality for a number of folktales, but deviating results for others. However, even though both analysis methods measure folktale similarity, the WordNet similarity measure considers the full text of a document, involving both syntax and semantics, while motif analysis is based on a small set of key events or themes, resulting in a highly specific semantic comparison on a considerably reduced and condensed representation of the document. The difference in approach leads to different results of similarity computation as well.

Rather than an alternative approach of computing TMI similarity, the WordNet method should be considered an alternative text-oriented measure of similarity of folktales. The current approach has the advantage that a domain-specific classification system is no longer required. Within the folktale domain this addresses the issue of selective motif attribution and differences in motif granularity for folktales featured in the existing catalogs, as well as the possibility to include folktales outside of the catalog, as demonstrated in the current test corpus. This advantage extends to potential use outside of the folktale domain, e.g., using general literary works or non-fictional narratives.

The current method is ranking-based, therefore a strong match between two documents (e.g., two variants of the same narrative) may cause less pronounced similarities to remain undetected. This behaviour can be exploited for incremental clustering, by leaving out the comparison of highly similar document pairs in subsequent iterations.

In future work, the granularity of the WordNet hierarchy and the relative position in the concept tree can be used to adjust term matching weights. Word sense disambiguation can be taken into account. The method can be applied on larger or more heterogeneous corpora, e.g., folktale documents lacking standardized spelling or grammatical sentences could be used to test the robustness of knowledge-based approaches. The approach could be extended towards discourse analysis to accomodate story element matching across sentence boundaries. Scalability issues resulting from the current method of comparing every pair of sentences for every pair of documents could be addressed using various precomputing, pruning or selection mechanisms. Finally, the development of an informative baseline (e.g., using existing clustering toolkits) and an automatic evaluation procedure tailored towards the current notion of narrative similarity (e.g., using story variants as in (Nguyen et al., 2013)) is desired to increase understanding and interpretation of current results.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the ACL (NAACL HLT)*, pages 19–27. ACL.

Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114. LOT Utrecht.

Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. 2012. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012)*, pages 30–34. ACL.

Gerrie van Dongen and Ad Grooten. 2009. *Sprookjesboek van De Efteling*. Ploegsma. *(in Dutch)*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jian-Bo Gao, Bao-Wen Zhang, and Xiao-Hua Chen. 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. *Engineering Applications of Artificial Intelligence*, 29:80–88.

Timo Honkela. 1997. Self-organizing maps of words for natural language processing applications. In *Proceedings of the International ICSC Symposium of Sof Computing*, pages 401–407. ICSC.

Andreas Hotho, Steffen Staab, and Gerd Stumme. 2003. Wordnet improves text document clustering. In *Proceedings of the Semantic Web Workshop at SIGIR-2003*. ACM.

Folgert Karsdorp, Peter van Kranenburg, Theo Meder, Dolf Trieschnigg, and Antal van den Bosch. 2012. In search of an appropriate abstraction level for motif annotations. In *Proceedings of the 2012 Computational Models of Narrative Workshop*, pages 22–26.

Folgert Karsdorp, Marten van der Meulen, Theo Meder, and Antal van den Bosch. 2015. MOMFER: A search engine of Thompson's motif-index of folk literature. *Folklore*, 126(1):37–52.

Victoria Lestari and Ruli Manurung. 2015. Measuring the structural and conceptual similarity of folktales using plot graphs. In *Proceedings of the 9th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2015)*. ACL.

Bridget McInnes and Ted Pedersen. 2015. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *Journal of Biomedical Informatics*, 54:329–336.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1562–1572. ACL.

Theo Meder. 2010. From a Dutch folktale database towards an international folktale database. *Fabula*, 51:6–22.

Mark Newman. 2002. Assortative mixing in networks. *Physical Review Letters*, 89(20).

Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2013. Folktale classification using learning to rank. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR 2013)*, pages 195–206. Springer.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::similarity – measuring the relatedness of concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04)*, pages 1024–1025. ACM.

Mahendra Piraveenan, Mikhail Prokopenko, and Albert Zomaya. 2008. Local assortativeness in scale-free networks. *Europhysics Letters*, 84.

Marten Postma and Piek Vossen. 2014a. Open source Dutch WordNet. Technical report, Free University of Amsterdam.

Marten Postma and Piek Vossen. 2014b. What implementation and translation teach us: the case of semantic similarity measures in wordnets. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*, pages 133–142.

Robert Ross, Simon Greenhill, and Quentin Atkinson. 2013. Population structure and cultural geography of a folktale in Europe. *Proceedings of the Royal Society B*, 280(20123065).

Julian Sedding and Dimitar Kazakov. 2004. WordNet-based text document clustering. In *Proceedings of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data*, pages 104–113. ACL.

Jamshid Tehrani. 2013. The phylogeny of Little Red Riding Hood. *PLoS One*, 8(11).

Stith Thompson. 1960. *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books and local legends*. Indiana University Press.

Hans-Jörg Uther. 2004. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson*. Finnish Academy of Science and Letters.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides Petrakis, and Evangelos Milios. 2005. Semantic similarity methods in WordNet and their application to information retrieval on the web. In *Proceedings of the 7th ACM International Workshop on Web Information and Data Management (WIDM 2005)*, pages 10–16. ACM.

Paula Vaz Lobo and David Martins de Matos. 2010. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *Proceedings of LREC 2010*, pages 1472–1475. ELRA.

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofman, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: A combinatorial lexical semantic database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the Stevin programme*, pages 165–184. Springer.

James Wang and William Taylor. 2007. Concept forest: A new ontology-assisted text document similarity measurement method. In *Proceedings of the IEEE/WIC/ACM International Confeence on Web Intelligence*, pages 395–401. IEEE.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics (ACL '94)*, pages 133–138. ACL.

The Predicate Matrix and the Event and Implied Situation Ontology: Making More of Events

Roxane Segers
The Network Institute
VU University Amsterdam
r.h.segers@vu.nl

Egoitz Laparra
IXA Group, UPV/EHU
egoitz.laparra@ehu.es

Marco Rospocher
Fondazione Bruno Kessler
rospocher@fbk.eu

Piek Vossen
The Network Institute
VU University Amsterdam
piek.vossen@vu.nl

German Rigau
IXA Group, UPV/EHU
german.rigau@ehu.es

Filip Ilievski
The Network Institute
VU University Amsterdam
f.ilievski@vu.nl

Abstract

This paper presents the Event and Implied Situation Ontology (ESO), a resource which formalizes the pre and post situations of events and the roles of the entities affected by an event. The ontology reuses and maps across existing resources such as WordNet, SUMO, VerbNet, PropBank and FrameNet. We describe how ESO is injected into a new version of the Predicate Matrix and illustrate how these resources are used to detect information in large document collections that otherwise would have remained implicit. The model targets interpretations of situations rather than the semantics of verbs per se. The event is interpreted as a situation using RDF taking all event components into account. Hence, the ontology and the linked resources need to be considered from the perspective of this interpretation model.

1 Introduction

In this paper, we present the new release of the Event and Implied Situation Ontology (ESO) that is matched with a new version of the Predicate Matrix (PM). Both resources rely on Semantic Role Labeling (SRL) descriptions and are used to detect and abstract over events, their participants and event implications in a large document collection about ten years of global automotive industries.

ESO (Segers et al., 2015) is a newly developed domain ontology to enhance the extraction and linking of dynamic and static events and their implications in text. Explicit modeling of event implications allows for extracting sequences of states and changes over time regardless of if this information was directly expressed in text, or inferred by a reasoner. Figure 1 shows such a chain of expressions for dynamic (*hire, starts at*, *fire, leave*) and static events (*works for, employs, is CEO*) and their implied situations. Lexicons that define implications of events, e.g. VerbNet (Kipper et al., 2000; Kipper et al., 2006), are rare and usually focus on the meaning of verbs in isolation. However, lexical structures do not make explicit how the meaning of a verb needs to be combined with other event components, such as the participants and the temporal properties for the purpose of semantic parsing. We therefore follow an ontological approach to interpret situations on the basis of text interpretation of all the event components to make the implications explicit. Though some research on deductive reasoning over Frame annotated text (e.g. (Scheffczyk et al., 2006)) and defining pre and post situations of predicates exist (Im and Pustejovsky, 2009; Im and Pustejovsky, 2010), to the best of our knowledge, ontologies that model both events, roles and implications do not. Most closest comes the extension to DOLCE-LITE (Hicks, 2009) that models property values as quality regions for reasoning. However, these quality regions are not connected to the events in the ontology as pre and post situations. Axioms in generic and top ontologies such as SUMO (Niles and Pease, 2001) and DOLCE (Masolo et al., 2002) provide a comprehensive semantic specification of the concepts, but these axioms do not always provide the information relevant and specific for our domain. Furthermore, such ontologies need to be integrated with semantic parsing systems that deal with expressions on natural language to be able to test these models. We therefore decided to develop a new ontology for modeling static and dynamic events and their implications that is tailored to a semantic parsing system for text.

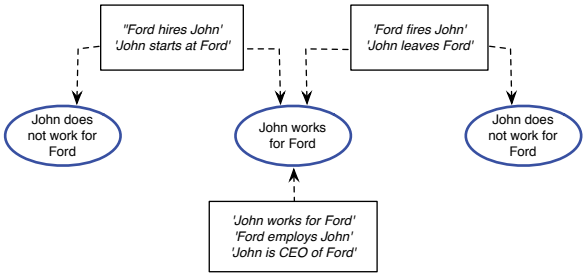


Figure 1: A chain of dynamic and static event expressions and their implied situation.

Version 2 of ESO was released in July 2015 and now includes modeling of scalar values and an extended expressivity of the assertions that define the situation that holds before, during and after the event. It also includes updated mappings to SUMO classes and to FrameNet frames and Frame Entities.

The Predicate Matrix[1] (de Lacalle et al., 2014a; de Lacalle et al., 2014b) is the second resource presented in this paper. It integrates predicate and role information from e.g. FrameNet, VerbNet, PropBank, NomBank and WordNet. This resource is used to assign role and predicate annotations at sentence level. All classes and roles in ESO are fed back into to the Predicate Matrix. As such the ontology provides an additional layer of annotations in text that allow for inferencing over events and implications. The current version of the PredicateMatrix contains 8,495 predicates from PropBank and NomBank connected to 4,704 synsets of WordNet, 554 frames of FrameNet and 55 ESO classes. On the other hand, this resource contains 23,386 roles of PropBank and NomBank mapped to 2,343 frame-elements of FrameNet and 53 ESO roles.

The remainder of this paper is organized as follows. Section 2 presents the ontological meta model and the content of ESO. Section 3 describes the Predicate Matrix and the integration with ESO. In section 4 we provide a preliminary overview of the Predicate Matrix and ESO in our document collection. In section 5 we report on an experiment carried out on a small corpus. We conclude in section 6 with a discussion and some outlines for future work.

[1] http://adimen.si.ehu.eus/web/PredicateMatrix

2 The Event and Implied Situation Ontology

In this paragraph, we briefly describe the meta model of the ontology with a focus on the instantiation of the rules that define what situation holds before, during and after some event. Next, we describe how the ontology was built and we provide an overview of the current content. The ESO ontology and a detailed documentation can be found online: https://github.com/newsreader/eso

ESO is an OWL 2 ontology.[2] It assumes that the semantic representation of text is converted to an RDF representation of event and entity instances, between which relations are expressed as triples. For instance, the statement

```
:obj-graph-eventX {
    :eventX
        a     eso:Translocation;
        eso:translocation-theme    :Enzo Ferrari;
        eso:translocation-goal     :Rome;
        eso:translocation-source   :Napels;
        sem:hasTime  :time_eventX.
}
```

specifies that the event (X) is of a certain type (eso:Translocation), that it involves an entity playing the role of a moving thing (:Enzo Ferrari), an entity playing the role of goal (:Rome), an entity playing the role of source (:Napels) and that it occurred at a certain time (:time_eventX). From these representations, we derive the statements that express the pre, post and during event situations.

For this purpose, we defined five core classes in ESO: 1) **Event**: this class is the root of the taxonomy of event types. Any event detected in a text will be an instance of some class of this taxonomy; 2) **DynamicEvent**: this is a subclass of Event for which dynamic changes are defined; 3) **StaticEvent**: this is another subclass of Event for "static" event types which capture more stable circumstances; 4) **Situation**: the individuals of this class are actual pre, post and during situations that will be instantiated starting from the event instances detected in the text; 5) **SituationRule**: the individuals of this class enable to encode the rules for instantiating pre/post/during situations when a certain type of event is detected.

Further, ESO includes mapping properties to match ESO roles to FrameNet roles, and properties to match ESO classes to FrameNet frames and SUMO classes. The mappings to FrameNet are necessary to translate the annotations provided by the SRL module using the Predicate Matrix to our

[2] http://www.w3.org/2001/sw/wiki/OWL

ontology. This is then exploited by the reasoning module to instantiate situations from events.

2.1 Formalization of the rules for instantiating situations from events

For all event classes in ESO an eso:SituationRule is defined; the individuals of this class trigger the pre, post and during situation related to a class or a set of event classes. For instance, the class eso:Translocation has two specific individuals: pre_Translocation and post_Translocation. Each eso:SituationRule individual defines exactly how the triples inside the Situation named graph have to be defined. This is done by defining an individual for each assertion to be created, which has three annotation properties: eso:hasSituationAssertionSubject (a role to be used as subject in the assertion), eso:hasSituationAssertionObject (a role to be used as object in the assertion) and eso:hasSituationAssertionProperty (a property relating the subject and object). In the case of e.g. eso:Translocation, the individual pre_Translocation has two eso:SituationRuleAssertions, where e.g. eso:pre_Translocation_assertion_1 states:

```
eso:pre_Translocation_assertion1
    eso:hasSituationAssertionSubject      eso:translocation-theme;
    eso:hasSituationAssertionProperty     eso:atPlace;
    eso:hasSituationAssertionObject       eso:translocation-source.
```

Based on all class assertions, the ESO reasoner[3] can now infer that some event belongs to the class eso:Translocation and that it has entity instances in certain roles where some entity is at some place before the event and not at this place after the event. The instantiation of the defined situations for the example event instance of eso:Translocation will then look as follows:

```
:eventX_pre {
     :Enzo Ferrari        eso:atPlace        :Napels
     :Enzo Ferrari           eso:notAtPlace     :Rome
:eventX_post
     :Enzo Ferrari        eso:atPlace        :Rome
     :Enzo ferrari           eso:notAtPlace     :Napels
}
```

Instantiation of events that express a change in a scalar value By default, situation assertions will only fire if some instance for an ESO role is found by the SRL module. However, in specific cases we also allow that assertions are instantiated even though no instance exists for the ESO role. We do this by adding an OWL existential restriction on the event class for the role considered. The reasoner will check if an instance of the role exists, if not it will create a blank node. This OWL existential restriction is applied in ESO for event classes that express a relative change in the value of an attribute (e.g. eso:Damaging, eso:Increasing, eso:Attacking) where the attribute itself such as 'price' or 'damagedness' often remains implicit. As such, it is possible to assert statements based on 'incomplete' information if needed. For eso:Increasing, the existential restriction is defined as follows:

```
eso:Increasing  rdfs:subClassOf [
    a owl:Restriction ;
    owl:onProperty        eso:triggersPreSituationRule   ;
    owl:hasValue          eso:pre_Increasing ] .
eso:Increasing  rdfs:subClassOf [
    a owl:Restriction ;
    owl:onProperty        eso:triggersPostSituationRule   ;
    owl:hasValue          eso:post_Increasing ] .
eso:Increasing  rdfs:subclassOf [
    a owl:Restriction ;
    owl:onProperty       eso:quantity-attribute   ;
    owl:someValuesFrom   owl:Thing] .

eso:pre_Increasing    a eso:SituationRule .
eso:post_Increasing   a eso:SituationRule .
```

These are the situation rule assertions defined for the pre an post situation of eso:Increasing:

```
eso:pre_Increasing_assertion1
    eso:hasSituationAssertionSubject      eso:quantity-item;
    eso:hasSituationAssertionProperty     eso:hasAttribute;
    eso:hasSituationAssertionObject       eso:quantity-attribute.

eso:pre_Increasing_assertion2
    eso:hasSituationAssertionSubject      eso:quantity-attribute;
    eso:hasSituationAssertionProperty     eso:hasRelativeValue;
    eso:hasSituationAssertionObjectValue  '-'

eso:post_Increasing_assertion1
    eso:hasSituationAssertionSubject      eso:quantity-item;
    eso:hasSituationAssertionProperty     eso:hasAttribute;
    eso:hasSituationAssertionObject       eso:quantity-attribute.

eso:post_Increasing_assertion2
    eso:hasSituationAssertionSubject      eso:quantity-attribute;
    eso:hasSituationAssertionProperty     eso:hasRelativeValue;
    eso:hasSituationAssertionObjectValue  '+'
```

The pre and post situation named graphs for the example sentence "Ford increased the production" can now be instantiated as follows:

```
:eventX_pre {
     :production        eso:hasAttribute        :xyz123
     :xyz123            eso:hasRelativeValue    ' - '
:eventX_post
     :production        eso:hasAttribute        :xyz123
     :xyz123            eso:hasRelativeValue    ' + '
}
```

These instantiations can be paraphrased as follows: the production has some unknown attribute and the value of this attribute has become more (+) after the event then it was before the event (-), meaning that the production goes from less (-) to more (+).

Alternatively, if the attribute is known, the assertions will instantiate the role that models the actual attribute. For a sentence like "Ford increased the price of the components", the event will look as follows:

```
:eventX_pre  a  eso:Increasing  ;
eso:quantity-item       :component  ;
eso:quantity-attribute   :price  ;
```

and the assertions will be instantiated as:

[3] Implemented as a processor of RDFpro (Corcoglioniti et al., 2015b). See also: `http://bit.ly/ESOreasoner`

```
:eventX_pre {
    :component        eso:hasAttribute        :price
    :price            eso:hasRelativeValue    ' - '
:eventX_post
    :component        eso:hasAttribute        :price
    :price            eso:hasRelativeValue    ' + '
}
```

Even though it may appear that these assertions for relative values are superfluous, we argue that finding *multiple* mentions of such an event and assertions over time, either with or without explicit values and attributes, allows for estimating the fluctuation of a certain value and the speed of the value change. We also need these values to determine that different event descriptions are coreferential even if one does not make the value explicit, while the other does. An existential representation of a value thus can match with an explicit value but two different explicit values cannot.

2.2 Mappings from external resources to ESO

A key ingredient of the ESO ontology is the mapping of FrameNet frames and Frame elements to the event types and roles that we defined. This mapping is necessary to translate the role annotations provided by the SRL module to our ontology vocabulary, which is then exploited by the reasoning module to instantiate situations from events. For each ESO event class and each ESO role we defined mapping properties representing the corresponding frames and frame elements. For instance, eso:Giving has three mappings to the frames fn:Giving, fn:Sending, and fn:Supply, meaning that if a frame of type fn:Supply or any of the others is identified in the text, it has to be considered as an event of type eso:Giving, and therefore pre and post situation rules defined for eso:Giving should be triggered. Similarly, the role eso:possession-owner_1 is mapped to a set of frame elements. These mappings make clear that our ontology is providing only a partial definition for concepts. We only define those elements necessary for capturing salient pre and post situations of events and not any other meaning aspect. As such the implications of a change in ownership of something are similar for all instances of eso:ChangeOfPossession, such as *stealing*, *giving* or *seizing*.

2.3 Development and content of ESO (Vers. 2)

Version 2 of ESO was released in July 2015. It contains: a hierarchy of event classes; a set of properties for the defining the pre, post and during situations of an event, and a set of roles for the entities affected by an event. In this section, we report how these structures were built and we conclude with an overview of its content.

The ESO ontology is a hand-built resource, based on high-frequent FrameNet frames that were extracted from a large domain-specific document collection. Frames that denote events pertaining to communication, feelings and perception were not taken into account. For deriving an initial conceptual structure for the frequent frames, we decided to map the frames manually to the SUMO ontology. The choice for using the SUMO ontology[4] as a background model was based on the fact that it is freely available, well-documented, has a good coverage and is mapped to English WordNet and also the Predicate Matrix. As such, we derived four main conceptual clusters that formed the backbone of ESO: 'changes in possession', 'translocations', 'internal changes' and 'intentional events'. Next, we modeled 103 FrameNet frames into 63 distinct ESO event classes. Frames that denote fine-grained semantic distinctions are often grouped into one class in ESO since these distinctions do not influence the modeling of a salient set of pre and post situations. As such we build an event class hierarchy that reuses and maps to groups of FrameNet frames, which deviates from the approach taken in e.g. (A. Nuzzolese, 2012) where FrameNet frames and frame relations are converted to RDF and partly to OWL. The second and third component of the ontology consists of properties and roles which are used for defining the assertions of the pre, post and during situations. All properties are hand-built, based on the shared semantics of the predicates related to a FrameNet frame and ESO class. The ESO roles define what entities are affected by a change and serve as the domain and range of properties. The majority of the ESO roles is mapped to a selection of FrameNet Frame Elements (FEs); these were selected manually from the FrameNet frames that correspond to an ESO class. This implies that not all FEs of a frame are mapped to ESO but only those that play a role in the assertions.

An important modeling decision is that assertions are defined at the highest possible level in the ontology. This way, all subclasses will inherit the same assertions and roles, which reduces re-

[4] http://www.ontologyportal.org

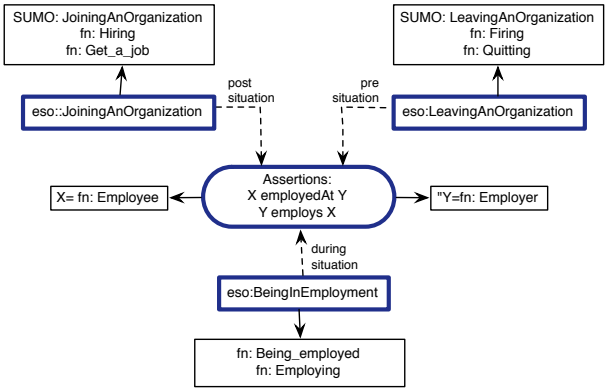


Figure 2: The shared assertion properties of a static and a dynamic event

dundancy. As such, many ESO roles have mappings to FEs that are aggregated from all mappings from ESO classes to FrameNet frame in a given sub-hierarchy. Another notable modeling choice is that the assertion properties for static event classes are partially shared with the assertion properties of the dynamic event classes. This is illustrated in Figure 2. Here, the same properties (eso:employedAt and eso:employs) are used in the pre situation assertion for the dynamic event class eso:LeavingAnOrganization, in the post situation assertion for the dynamic event class eso:JoiningAnOrganization and in the during situation assertion of the static event class eso:BeingInEmployment. As a result, the relation between the inferred situation of a dynamic event and the explicit mention of some state by a static event becomes explicit. Modeling the properties this way facilitates querying for chains of related changes and states (See also section 5).

To illustrate the expressivity of the assertions, in figure 3 we provide a non-formal transcription of a typical class in ESO, including the class mappings to SUMO and FrameNet, the aggregated role mappings to FEs, the inherited and class specific situation assertions and an example of the instantiation. From the ”knowledge” in the example sentence, we are able to infer that a) Marie has 600 dollar and not the car before the event, while John does have the car but not the 600 dollar, b) after the event, the money and the car have changed ownership while c) the car itself has a value of 600 dollar during the exchange.

In table 1 we provide an overview of the content of ESO, including the number of mappings to FrameNet frames (103), SUMO classes (46) and

-FinancialTransaction: subclassOf: ChangeOfPossession
"The subclass ofChangeOfPossession where some item changes of ownership in exchange for money."

Class mappings:
closeMatch: fn:CommercialTransaction
closeMatch: sumo:FinancialTransaction

Role mappings:
possession-financial-asset: fn:Money

Inherited role mappings:
possession-owner_1: fn:Supplier, fn:Exporter, fn:Donor, fn:Victim, fn:Source, fn:Lender, fn:Exporting_area, fn:Sender, fn:Seller
possession-owner_2: fn:Perpetrator, fn:Importing_area, fn:Importer, fn:Lessee, fn:Buyer, fn:Recipient, fn:Borrower, fn:Agent
possession-theme: fn:Theme, fn:Goods, fn:Possession
possession-financial-asset: fn:Money

Assertions:

pre situation	possession-owner_1	notHasInPossession	poss.-financial-asset
	possession-owner_2	hasInPossession	poss.-financial-asset
post situation	possession-owner_1	hasInPossession	poss.-financial-asset
	possession-owner_2	notHasInPossession	poss.-financial-asset
during situation	possession-theme	hasValue	possession-value

Inherited assertions from ChangeOfPossession:

pre situation	possession-owner_1	hasInPossession	possession-theme
	possession-owner_2	notHasInPossession	possession-theme
post situation	possession-owner_1	notHasInPossession	possession-theme
	possession-owner_2	hasInPossession	possession-theme

EXAMPLES:

"Marie bought the car from John for 600 dollars"

pre situation	Marie	hasInPossession	600 dollar
	Marie	notHasInPossession	the car
	John	hasInPossession	the car
	John	notHasInPossession	600 dollar
post situation	Marie	hasInPossession	the car
	Marie	notHasInPossession	600 dollar
	John	hasInPossession	600 dollar
	John	notHasInPossession	the car
during situation	the car	hasValue	600 dollar

Figure 3: Non-formal transcription of the mappings, assertions and instantiation for the ESO class FinancialTransaction

from ESO roles (65) to FrameNet Frame Elements (131). The properties in this table pertain to those properties that are used in the situation rule assertions.

3 Predicate Matrix

The PredicateMatrix (PM)(de Lacalle et al., 2014a; de Lacalle et al., 2014b) is an automatic extension of SemLink (Palmer, 2009) that merges several models of predicates such as VerbNet (Kipper et al., 2000), FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) and WordNet (Fellbaum, 1998). The PM also contains for each predicate features of the ontologies integrated in the Multilingual Central Repository (Gonzalez-Agirre et al., 2012) like SUMO (Niles and Pease, 2001), Top Ontology (Álvez et al., 2008) or WordNet domains (Bentivogli et al., 2004). The mappings between such knowledge bases allow to take advantage from their individual strengths. For example, the coverage of PropBank or the semantic relations among events and participants of FrameNet.

The semantic interoperability offered by the PM allows to translate the output of a SRL analysis to

Component	Number
Event classes	63
– DynamicEvent classes	50
– StaticEvent classes	13
SUMO class mappings	46
FrameNet Frame mappings	103
Situation rules	50
Situation rule assertions	123
– Pre situation rule assertions	41
– Post situation rule assertions	52
During situation rule assertions	30
Properties	58
– Unary properties	11
– Binary properties	47
ESO roles	65
Mappings to FrameNet FEs	131

Table 1: Overview of the content in ESO Vers. 2

a representation based on any resource connected to the PM like FrameNet, SUMO or the Domain Ontology. For this reason, we have connected the classes and roles of ESO to the predicates and roles of the PM. We have performed this alignment in two different steps. First, defining a set of new manual mappings between ESO and WordNet. Second, applying an automatic strategy that makes use of the existing mappings between ESO and FN and SUMO. Table 2 contains the number of predicates and roles mapped to ESO by each method.

	Manual	Automatic	Total
predicates	1,702	2,228	3,930
roles	4,831	6,026	10,857

Table 2: Number of predicates and roles mapped to ESO in the PM.

3.1 Manual mappings

For connecting ESO and the PM, manual mappings to Princeton Wordnet 3.0 have been created for all lexical units in a FrameNet frame associated to ESO. In total, 1,614 lexical units from FrameNet have been mapped to WordNet, covering 1,918 synsets. The mappings have been kept outside ESO in order not to overburden the ontology. Additionally, to increase the coverage of ESO in the Predicate Matrix, we manually mapped ESO classes to WordNet Base Level Concepts (BLC). BLCs are important WordNet concepts that cover all WordNet nominal and verbal concepts. In WordNet there are 616 verbal BLCs that cover all 13,151 verbal synsets. The PM can be mapped to 398 of these BLCs which covers 12,722 verbal synsets. The full set of BLCs have been manually checked for their correspondence to an ESO class and for 75 BLCs a mapping to an ESO class could be made which covers 4,306 synsets.[5]

3.2 Automatic mappings

Both FrameNet and SUMO labels integrated in ESO are used to connect ESO to the PM. For example, the predicate **sell.01** of PropBank belongs, according to its mappings in the PM, to the frame *Commerce_sell* of FrameNet. Consequently, this predicate and its arguments could also be mapped to ESO as shows table 3. Moreover, the frame can also be linked through the SUMO classes. For instance, the predicate **drain.01** of PropBank belongs to the frame *Emptying* that is not considered in ESO. However, it also belongs to the class *Removing* of SUMO and, as a consequence, the mappings in table 4 can be obtained.

PB-pred	PB-arg	FN-frame	FN-fe	ESO-class	ESO-role
sell.01	arg_0	Commerce_sell	Seller	Selling	possession-owner_1
sell.01	arg_1	Commerce_sell	Goods	Selling	possession-theme
sell.01	arg_2	Commerce_sell	Buyer	Selling	possession-owner_2

Table 3: Mapping between PropBank and ESO through FN.

PB-pred	PB-arg	SUMO-class	FN-fe	ESO-class	ESO-role
drain.01	arg_0	Removing	Theme	Removing	translocation-theme
drain.01	arg_1	Removing	Source	Removing	translocation-source

Table 4: Mapping between PropBank and ESO through SUMO.

4 Current Output

At the time of submission, about 2.1 million articles on the automotive industry were processed with the NewsReader English pipeline (Agerri et al., 2015) that incorporates the PM and ESO for semantic parsing. Table 5 provides an overview of the number of roles and predicates found, and the number of labels assigned to them per resource in the Predicate Matrix. Note that predicates and roles can receive multiple labels from one resource.

5 Experiment on the WikiNews Corpus

The WikiNews Corpus consists of 120 manually annotated news articles selected from WikiNews[6] and is used within NewsReader as an evaluation corpus.[7] The evaluation of the Mate tool that is

[5] All mappings can be downloaded from `https://github.com/newsreader/eso`.

[6] `https://en.wikinews.org`

[7] The corpus will be made available soon at `http://www.newsreader-project.eu/results/data/`

Resource	label frequency
Total predicates	138,695,190
WordNet	293,249,984
VerbNet	236,497,891
PropBank	197,331,322
FrameNet	232,685,360
ESO	85,831,344
Total roles	300,544,817
VerbNet	277,233,904
PropBank	202,134,061
FrameNet	336,248,141
ESO	55,787,300

Table 5: Overview of the number of predicates and roles in a subset of the automotive industry corpus labeled by the Predicate Matrix and ESO

Resource	Label frequency
Total predicates	7,060
WordNet	15,157
VerbNet	12,294
PropBank	10,018
FrameNet	12,330
ESO	4,337
Total roles	15,652
VerbNet	14,474
PropBank	10,312
FrameNet	17,680
ESO	3,230

Table 6: Overview of the number of predicate and role labels in the WikiNews corpus labeled by the Predicate Matrix enriched with ESO

Component	Number
Events	5443
ESO events	2508
ESO events with ESO roles	736
ESO events with pre and post situations	444
ESO events with at least one inferred situation	498
ESO events with a during situation	52

Table 7: ESO related statistics of the populated KnowledgeStore of the WikiNews corpus

used for the Semantic role labeling scores an F1 of 34.74 for this corpus.[8] WikiNews has not yet been annotated with ESO classes and roles, as such we used this corpus to test the expressivity and coverage of the Predicate Matrix and ESO first. In short, we followed the same procedure that is also used for the Automotive Corpus. First, all 120 WikiNews articles were processed by the NewsReader Pipeline (Agerri et al., 2015) using the Predicate Matrix and ESO; next, a module called NAF2SEM merged identical events across documents and translated all events into SEM-RDF and finally, all events were loaded into the KnowledgeStore (Corcoglioniti et al., 2015a) and further enriched by the ESO reasoner that infers all ESO assertions, based on the class and role labels. In the KnowledgeStore, the data can be queried via SPARQL queries or simple look-ups.

In table 6 we provide an overview of the results of the first step, the output of the pipeline with respect to the labels for roles and events found. In total, 7,060 predicates were found in the WikiNews corpus. These predicates are assigned one or multiple labels by the Predicate Matrix such as WordNet synset IDs (15,157), FrameNet Frames (12,330) and ESO classes (3,405). The relatively low number of predicates with an ESO class is due to the fact that ESO covers a limited set of concepts and ignores e.g. all speech acts. This table also shows the number of labels found for the roles. In total, 15,652 roles were found that each can again have one or multiple labels.

Next, we derived some basic statistics from the KnowledgeStore that contains all events derived from the corpus. In table 7 we provide an overview. As is shown, 5,443 distinct events were found of which 2,508 events with an ESO class. Of these events, 736 have at least also an ESO role which is necessary to trigger the situation rules defined in ESO. In total, 444 events were found with inferred pre and post situations and 52 events with inferred during situations. Note that the number of ESO classes that trigger a during situation is smaller (12) than the set of classes that can trigger pre and post situations (46).

Finally, we manually inspected 52 ESO events in the KnowledgeStore with both a pre and post situation (43) and ESO events with a during situation (9).[9] For this, we randomly selected one or two ESO events per class, depending on the number of occurrences. The result of this inspection are shown in table 8. We found 37 events (71.1%) with a correct class label and 18 events (41.8%) with correct pre and post situations, meaning that the assertions made sense with respect to the original sentences in the document and that the correct role instances were found, if applicable. The set of events with a during situation was correct in 66,6% of the cases. Overall, 21 out of 52 inspected ESO events were found to be correct.

Additionally, we performed an error analysis

[8] see (Agerri et al., 2015) for an overview and discussion of these results

[9] The data and analysis can be found at `https://github.com/newsreader/eso`

ESO events with pre/post or during situation	495
Number of events inspected	52 (10.5%)
Number events insp. with a pre/post situation	43
Number events insp. with a during situation	9
Correct class label	37 (71.1%)
Correct pre and post situation(s)	18 (41.8%)
Correct during situation(s)	6 (66.6%)
Correct ESO events	21 (50%)

Table 8: Results of the analysis of ESO events with during or pre/post situation assertions derived from the WikiNews corpus

Error in interpretation sentence (multiple causes)	3
Error in interpretation predicate	9
Multiple conflicting ESO classes assigned	8
Wrong role instance (non-entities)	5
Wrong role instance (entities)	10
Role instance duplication	6
Conflicting assertions	1

Table 9: Results of the error analysis of the inspected ESO events derived from the WikiNews corpus

to investigate where errors or omissions stemmed from. The results of the error analysis can be found in table 9. In general, each of the 16 modules in the pipeline introduces some errors, which is reflected in the outcome of the error analysis. For nine events we found that the sense of the predicate was misinterpreted, for eight events multiple and conflicting ESO classes were assigned due to some unavoidable level of ambiguity in the Predicate Matrix. In five cases, we found that the Semantic Role Labeler picked up the wrong role; for ten events DBpedia Spotlight assigned a wrong label for a named entity. These errors also resulted in 6 role duplications where subject and object of an assertion are identical while they should not. For one event, it caused conflicting assertions.

6 Discussion and Future Work

In this paper, we have presented the new release of the Event and Implied Situation Ontology (ESO) and the PredicateMatrix (PM). Both resources augment Semantic Role Labeling techniques and are applied to a very large document collection to capture implications of events for a selected set of concepts, roles and properties. Through the WordNet backbone of ESO and PM, we were able to derive a formal model for event implications with a large coverage in English. Since wordnets in many languages are connected to WordNet, this model has also been projected to other languages in the NewsReader project: Spanish, Dutch and Bulgarian. ESO thus has shown to be used as a interoperable framework on reasoning over changes and their implications across different languages. This allows us to compare the content of text across languages, regardless of the way this content is expressed.

From the experiment on the WikiNews corpus, we conclude that ESO performs reasonably well on this dataset with 50% of correct ESO events with a pre/post or during situation. The ontology is not built in order to define all events in text which is shown in the coverage of all events found (5,443), and the ESO events (2,508) of which 496 have either both a pre and post situation or a during situation. The errors in the ESO events with assertions are mainly caused by an unavoidable degree of errors in the processing pipeline as was reported in the error analysis. The observation that not all ESO events come with assertions is likely due to the fact that a sentence does not always contain all roles necessary for an assertion rule to fire. A more in-depth analysis of the annotated texts will provide an answer for this.

We are currently processing about 2.1 million news articles on the automotive industry, where the ESO mapping are inserted in the SRL layers. The output is converted to RDF, after which we apply reasoning to derive new statements as was shown in the experiment. The output will be evaluated through inspecting samples, against benchmark data that will be developed on the WikiNews corpus and through end-user tasks on the data sets. Also, we planned additional experiments on the usability of the ESO assertions for tracking actual chains of property changes through time. Finally, the WikiNews corpus has been translated to Spanish, Dutch and Italian. The processing of the translated text through NewsReader pipelines in these languages, where these pipelines exploit the same ESO model and a language-specific PM, will allow us to do a cross-language comparison of the inferred properties.

Acknowledgments

The research for this paper was supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

References

V. Presutti A. Nuzzolese, A.Gangemi. 2012. Gathering lexical linked data and knowledge patterns from FrameNet. In *Proceedings of K-CAP '11*.

R. Agerri, I. Aldabe, Z. Beloki, E. Laparra, M. Lopez de Lacalle, G. Rigau, A. Soroa, A. Fokkens, R. Izquierdo, M. van Erp, P. Vossen, C. Girardi, and A. Minard. 2015. Event detection, version 3. Deliverable 4.2.3. NewsReader-ICT316404.

J. Álvez, J. Atserias, J. Carrera, S. Climent, A. Oliver, and G. Rigau. 2008. Consistent annotation of EuroWordNet with the Top Concept Ontology. In *Proceedings of GWC'08*.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings COLING-ACL*, ACL '98, Montreal, Canada.

L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. 2004. Revising the Wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*. ACL.

F. Corcoglioniti, M. Rospocher, R. Cattoni, B. Magnini, and L. Serafini. 2015a. The KnowledgeStore: a storage framework for interlinking unstructured and structured knowledge. *International Journal on Semantic Web and Information Systems*, 11(2):1–35, April-June.

F. Corcoglioniti, M. Rospocher, M. Mostarda, and M. Amadori. 2015b. Processing billions of RDF triples on a single machine using streaming and sorting. In *ACM SAC 2015 Proceedings*.

M. López de Lacalle, E. Laparra, and G. Rigau. 2014a. First steps towards a Predicate Matrix. In *Proceedings of GWC 2014*, Tartu, Estonia.

M. López de Lacalle, E. Laparra, and G. Rigau. 2014b. Predicate matrix: extending SemLink through WordNet mappings. In *Proceedings of LREC'14*, Reykjavik, Iceland.

C. Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press.

A. Gonzalez-Agirre, E. Laparra, and G. Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of LREC '11*, pages 2525–2529.

A. Hicks. 2009. Domain extension of the central ontology - final. Deliverable 8.3, KYOTO-ICT 211423.

S. Im and J. Pustejovsky. 2009. Annotating event implicatures for textual inference tasks. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*.

S. Im and J. Pustejovsky. 2010. Annotating lexically entailed subevents for textual inference tasks. In *Proceedings of FLAIRS-23*, Daytona Beach, USA.

K. Kipper, H. Trang Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. In *Seventeenth National Conference on Artificial Intelligence*, AAAI-2000.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC 2006*.

C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. 2002. Wonderweb deliverable d17. Technical report, ISTC-CNR.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of FOIS-Volume 2001*. ACM.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Palmer. 2009. Semlink: Linking Propbank, Verbnet and Framenet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15.

J. Scheffczyk, A. Pease, and M. Ellsworth. 2006. Linking FrameNet to the Suggested Upper Merged Ontology. In *Proceedings of FOIS 2006*.

R. Segers, P. Vossen, M. Rospocher, L. Serafini, E. Laparra, and G. Rigau. 2015. ESO: a frame based ontology for events and implied situations. In *Proceedings of Maplex 2015*, Yamagata, Japan.

Semi-Automatic Mapping of WordNet to Basic Formal Ontology

Selja Seppälä
University at Buffalo
Buffalo, NY, USA
seljamar@buffalo.edu

Amanda Hicks
University of Florida
Gainesville, FL, USA
aehicks@ufl.edu

Alan Ruttenberg
University at Buffalo
Buffalo, NY, USA
alanruttenberg@gmail.com

Abstract

We present preliminary work on the mapping of WordNet 3.0 to the Basic Formal Ontology (BFO 2.0). WordNet is a large, widely used semantic network. BFO is a domain-neutral upper-level ontology that represents the types of things that exist in the world and relations between them. BFO serves as an integration hub for more specific ontologies, such as the Ontology for Biomedical Investigations (OBI) and Ontology for Biobanking (OBIB). This work aims at creating a lexico-semantic resource that can be used in NLP tools to perform ontology-related text manipulation tasks. Our semi-automatic mapping method consists in using existing mappings between WordNet and the KYOTO Ontology. The latter allows machines to reason over texts by providing interpretations of the words in ontological terms. Our working hypothesis is that a large portion of WordNet synsets can be semi-automatically mapped to BFO using simple mapping rules from KYOTO to BFO. We evaluate the method on a randomized subset of synsets, examine preliminary results, and discuss challenges related to the method. We conclude with suggestions for future work.

1 Introduction

Ontologies are often used in combination with natural language processing (NLP) tools to carry out ontology-related text manipulation tasks such as automatic annotation of biomedical texts with ontology terms. These tasks involve categorizing relevant terms from texts under the appropriate categories. This requires coupling ontologies with lexical resources. Several projects have realized these kinds of mappings with upper-level ontologies that are extended by domain-specific ontologies (Gangemi et al., 2010; Laparra et al., 2012; Niles and Pease, 2003; Pease and Fellbaum, 2010). However, no such resource is available for the Basic Formal Ontology (BFO), which is widely used in the biomedical domain.

We describe and evaluate a semi-automatic method for mapping the large lexical network WordNet 3.0 (WN) to BFO 2.0 exploiting an existing mapping between WN and the KYOTO Ontology (hereafter 'KYOTO'). Our hypothesis is that a large portion of WN, primarily nouns and verbs, can be semi-automatically mapped to BFO 2.0 types by means of simple mapping rules exploiting the KYOTO Ontology.

In section 2, we give a brief overview of the ontological and lexical resources involved in the task: BFO, WN, and KYOTO. In section 3, we motivate and describe our methodology. In section 4, we evaluate the method and present preliminary results. In section 5, we discuss the major challenges related to this task. We conclude with suggestions for future work.

2 Ontological and Lexical Resources

The mapping methodology described below in section 3 takes as input WordNet 3.0, which is mapped to the KYOTO 3 Middle Ontology. The latter is itself based on the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE-Lite-Plus, version 3.9.7). The KYOTO 3 Top Ontology was extended to a middle level ontology KYOTO 3 Middle by manually mapping Base Concepts automatically generated from WN (Herold et al., 2009). We use those existing mappings to create mapping rules from KYOTO to BFO 2.0. Hereafter, we briefly describe each of these ontological and lexical resources. We briefly present the aspects of BFO, WN, and the KYOTO Ontology that are relevant to this work. For more details, see the cited references.

The **Basic Formal Ontology (BFO)** is a domain-neutral upper-level ontology (Arp et al., 2015; Smith et al., 2012; Spear, 2006). It represents the types of things that exist in the world and relations between them. BFO serves as an integration hub for mid-level and domain-specific ontologies, such as the Ontology for Biomedical Investigations (OBI) and Ontology for Biobanking (OBIB). It is widely used in biomedical and other domain-specific ontologies,[1] which thus become interoperable (Smith and Ceusters, 2010). BFO is subdivided into CONTINUANTS (e.g., OBJECTS and FUNCTIONS) and OCCURRENTS (e.g., PROCESSES and EVENTS). Continuants can be either independent (e.g., physical OBJECTS like persons and hearts) or dependent (e.g., the ROLE of a person as a physician and the FUNCTION of a heart to pump blood). The most recent version, BFO 2.0, represents 35 types to which previous versions (BFO 1.0 and BFO 1.1) have been mapped in Seppälä et al. (2014).

WordNet 3.0 is a large lexical network linking over 117000 sets of synonymous English words (synsets) by means of semantic relations; it is widely used in NLP tasks (Fellbaum, 1998; Miller, 1995). Noun and verb synsets are linked via the hypernym relation.[2] WN 3.0 distinguishes between types and instances, meaning named entities. It also links a subset of synsets to topic domains (e.g., 'medicine') and semantic labels (e.g., 'noun.artifact').

The **KYOTO Ontology** is part of a project aimed at representing domain-specific terms in a computer-tractable axiomatized formalism to allow machines to reason over texts in natural language (Vossen et al., 2010). It links WordNets of different languages to ontology classes, on the basis of a mapping of the English WN to KYOTO. The approximately 2000 classes of KYOTO are subdivided into three layers: (1) The top-most layer is based on DOLCE-Lite-Plus. DOCLE shares a number of relevant characteristics with BFO: domain neutrality; bi-partition into 'endurants' (CONTINUANTS) and 'perdurants' (OCCURRENTS); strict hierarchical *is_a* taxonomy; distinction between independent and dependent entities. (2) The second layer is composed of noun and verb synsets constituting a set of Base Concepts (BCs) as well as some adjectives or qualities. (3) The third layer contains domain-specific classes (e.g. from the environmental domain) and some corresponding synsets.

3 Mapping Method

Our semi-automatic mapping method involves three main steps:

1. Manually creating mappings:
   - from KYOTO to BFO on the basis of existing mappings of DOLCE to BFO 1.0 and BFO 1.1 (Grenon, 2003; Khan and Keet, 2013; Seyed, 2009; Temal et al., 2010), ignoring the axiomatization incompatibilities;
   - from BFO 1.0 and BFO 1.1 to BFO 2.0 on the basis of work in Seppälä et al. (2014);
   - from WN semantic labels to BFO 2.0.
2. Manually creating mapping rules using the above mappings and extending them with more specific rules from other KYOTO types. The rules map to BFO 2.0 leaf types or, when BFO has no leaf-level type to represent the referent of a synset, to intermediary types (e.g., MATERIAL ENTITY, the direct parent of three leaf types).
3. Implementing the resulting mapping rules in a Python pipeline using the natural language toolkit for Python that integrates WN 3.0 (NLTK 3.0).[3]

The rules are of the form `'KYOTO/WN > BFO 2.0'`, for example:

```
'#non-agentive-social-object
     > disposition'
'accomplishment > process'
'noun.act > process'
```

The implementation first lists all KYOTO types that subsume or otherwise characterize a WN synset using the WN-KYOTO mapping data files.[4] We only retained types related to the synsets through equivalence and subclass relations, plus the following ones deemed

[1] See http://ifomis.uni-saarland.de/bfo/users.

[2] Adjectives and adverbs are linked by way of other semantic relations.

[3] Natural Language Toolkit for Python (NLTK), version 3.0, http://www.nltk.org.

[4] http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index9c60.html?option=\\com_content&view=article&id=429&Itemid=156.

useful for creating appropriate mapping rules:

```
'DOLCE-Lite.owl#generically-dependent-on'
'DOLCE-Lite.owl#specifically-constantly
    -dependent-on'
'ExtendedDnS.owl#realized-by'
'ExtendedDnS.owl#realizes'
```

For example, the synset `immunity.n.02` is linked to the following types:

```
'Kyoto#condition__status
    -eng-3.0-13920835-n',
'Kyoto#state-eng-3.0-00024720-n',
'ExtendedDnS.owl#situation',
'ExtendedDnS.owl#non-agentive
    -social-object',
'ExtendedDnS.owl#social-object',
'DOLCE-Lite.owl#non-physical-object',
'DOLCE-Lite.owl#non-physical-endurant',
'DOLCE-Lite.owl#endurant',
'DOLCE-Lite.owl#spatio-temporal
    -particular',
'DOLCE-Lite.owl#particular'
```

Second, the mapping rules are applied starting from the more specific type in the types list: the program tests whether a given string on the left-hand side of the rule (e.g., `'#non-agentive-social-object'`) matches a string in the types list; if the strings match, the program assigns to that synset the corresponding BFO 2.0 type (e.g., `'disposition'`). Thus, the synset `immunity.n.02` is categorized as referring to a subtype of the BFO type DISPOSITION.

4 Evaluation and Results

In a first step, we evaluated the method on the 106 synsets in KYOTO marked with a 'medicine' topic domain (Seppälä, 2015a). The aim in this first step was to get a rough idea of the feasibility of the method, the results that could be expected, and the possible challenges. The medicine gold standard was created collectively by a BFO developer and a BFO expert. The experts followed an intuitive categorization criterion: assign the most specific BFO type of which the referent of the synset is a subtype. Following this principle, for each synset we may obtain a statement of the form "the WN synset X refers to a subtype of the (leaf) BFO type Y". For example, "the WN synset `immunity.n.02` refers to a subtype of the BFO type DISPOSITION". This task revealed difficult interpretation issues related to adjectives. 71.7% of the assigned BFO types were correct (63.2% of the synsets were assigned the expected BFO type; 8.5% a superclass). As hypothesized, all the correctly categorized synsets were nominal and verbal. 27.4% of the assigned BFO types were incorrect (mostly adjectives). One synset was not matched by any rule.

In a second step, we focused on nouns and verbs, and left adjectives for future work. After examining the erroneous cases in the first evaluation, we created a new ruleset, which we tested on a new randomly extracted sample of 100 nouns and 100 verbs (hereafter the 'POS-sample').

To create the corresponding gold standard, two of the authors, experts of BFO, first pre-annotated the POS-sample independently. They followed the same intuitive annotation criteria as with the medicine gold standard. The annotations were compared and the synsets separated into 'easy cases' (where both annotators agreed) and 'difficult cases' (in case of disagreement), respectively 101/200 and 99/200 synsets. Second, two BFO developers (annotators A & B) independently (i) reviewed the easy cases for validation and (ii) annotated the difficult cases. They were asked to apply the same intuitive annotation criteria as in previous steps. The annotators agreed on 2/3 of the latter sample. Finally, annotator B examined the cases on which they disagreed and decided on the final BFO type to assign considering the comments left by annotator A. Some difficult cases were collectively discussed to reach consensus. We discuss the challenging cases in section 5.

The baseline was created by a BFO developer and discussed with a BFO expert to resolve a few problematic cases (see, for example, section 5.5). The resulting mapping rules map, whenever possible, WN's top-level nouns to lowest level BFO 2.0 types and all verbs to BFO PROCESS.

In the following, we review the main results of our mappings. We limit the evaluation of the medicine sample to nouns and verbs to allow for comparisons with the performance of the rules on the POS-sample.

Figure 1 shows the overall performance of the first and the new rulesets compared to the baseline when applied to the medicine nouns and verbs sample. A total of 85% of the sample's synsets were correctly mapped with the new ruleset, which is considerably better than with the baseline rules (76%) and the first ruleset (77%). In the baseline and the first experiment, respectively 9% and 6%

% of WN-BFO mappings	medicine n-v sample								
	baseline			first ruleset			new ruleset		
	n	v	total	n	v	total	n	v	total
correct	55	100	76	70	85	77	72	100	85
partial	17	0	9	0	12	6	0	0	0
incorrect	28	0	15	28	2	16	26	0	14
no mapping	0	0	0	2	0	1	2	0	1
total	100	100	100	100	100	100	100	100	100

Figure 1: Performance of the rulesets on the medicine nouns and verbs sample.

of the synsets were mapped to a parent BFO type (see 'partial' row). There were no occurrences of partial mapping with the new ruleset. We did indeed correct some mapping rules with lower-level BFO types. However, assigning lower-level BFO types cannot be done with all of WN's top level categories. This means that with the baseline mapping method there will always be synsets that are not assigned an adequate (lowest level possible) BFO type. The percentage of incorrect mappings is steady across the applied mapping rules, around 15%. The proportion of synsets for which no rule was able to output a mapping is 1% for the KYOTO-based rules. As the baseline mapping rules were propagated all the way down the WN hierarchy, there are no such cases.

These results show that the performance of the KYOTO-based rulesets applied to the medicine nouns and verbs sample is (i) comparable to that of the baseline with the first ruleset and (ii) better with the new ruleset. This suggests that developing a more sophisticated mapping method, the KYOTO-based method, has advantages over a simple mapping of WN's top levels to BFO types.

% of WN-BFO mappings	pos sample					
	baseline			new ruleset		
	n	v	total	n	v	total
correct	41	99	70	42	86	64
partial	25	0	12.5	0	0	0
incorrect	34	1	17.5	53	7	30
no mapping	0	0	0	5	7	6
total	100	100	100	100	100	100

Figure 2: Performance of the new ruleset on the POS-sample.

The results of our second evaluation on a randomly extracted sample of 100 noun and 100 verb synsets are less encouraging. As shown in figure 2, the overall performance of the new ruleset on the POS-sample is lower than that of the baseline rules, respectively 64% and 70%. However, a closer look at the mappings reveals that while the new ruleset introduced some errors and non-matches, it also has the advantage of avoiding partial matches (when a synset is tagged with a superordinate BFO type instead of the lowest possible type). In 16 cases, the partial matches in the baseline correspond to the BFO type ENTITY, which is the uppermost level of the ontology and not relevant for a resource mapping WN to BFO.[5] With the baseline method, these cases could only be manually resolved; with the KYOTO-based method, with which they were mostly incorrectly mapped (11/15 synsets) or not mapped at all (3/15 synsets), we can test new rules to capture these cases.

Morevover, the analysis of these new results reveals useful information for improving the WN-BFO mapping method. A notable example is the case of verb synsets: the baseline rules systematically mapped them to the BFO type PROCESS. This yielded only one error due to the fact that the erroneously mapped synset does actually not refer to any BFO type (see the discussion in section 5.3). This suggests that the KYOTO-based rules can be improved with this verb mapping rule. To test this hypothesis, we performed a supplementary test. Figure 3 shows the prospective performance of the new ruleset complemented with the verb mapping rule on the POS-sample. The overall performance rose from 64% to 70.5%, slightly higher than that of the baseline.

WN-BFO mappings	n	v	total	%
correct	42	99	141	70.5
partial	0	0	0	0
incorrect	53	1	54	27
no match	5	0	5	2.5
total	100	100	200	100

Figure 3: Prospective performance of the new ruleset combined with a systematic mapping of verbs to BFO PROCESS on the POS-sample.

The rest of the results are not surprising. Indeed, nouns refer to a large array of entity types (to 10 BFO types in the POS-sample). Although the new ruleset did include rules for all the expected BFO types in the gold standard (as checked after the experiment), it did not capture the correct types in the KYOTO-types lists associated to the synsets. We suspect that the issue might be re-

[5] These mappings could as well be considered incorrect.

lated to the ordering of the rules. A total of 12 synsets (6%) were not mapped to BFO at all. This means that none of the rules was matched to the KYOTO-types lists associated with these synsets. These cases covered mostly deverbal nouns (8/12 synsets) that should be mapped to PROCESS. Further work is needed on the unmapped cases.

To summarize, these evaluations show that the KYOTO-mappings allow for creating more specific rules that map to BFO leaf- or lower-intermediate types. This is not possible with a mapping of WN's top level nouns to BFO. The results also show that the KYOTO-based rules can be successfully complemented with a baseline rule that consists in systematically mapping verb synsets to BFO PROCESSES.

5 Discussion

5.1 Usefulness of a Lexico-semantic Resource Linked to BFO

While BFO may be seen as too small and high level for ontology-related text manipulation tasks, we believe that a lexico-semantic resource linked to BFO is still useful. Such a resource can be used, for example, to semi-automatically create BFO-compliant domain ontologies from existing domain-specific terminological resources.[6]

BFO could also be used as an alternative upper level ontology in the KYOTO project, in which case it needs to be mapped to WN. Substituting DOLCE with BFO would allow for comparisons of the performances of these upper level ontologies in applied tasks.

The fact that BFO types are very general is not a problem either. BFO reveals fundamental distinctions between the types, which involve different kinds of relations to different types of entities. When associated to WN synsets, such information may, for example, help definition authors write better definitions, as proposed in Seppälä (2015b).

5.2 Non-trivial Mappings Between Ontologies

Mapping DOLCE to BFO is not trivial. The former is meant to capture our use of language and conceptualization of the world; the latter is a realist ontology and excludes from its scope unicorns and other putative non-real entities. Consequently, their categories do not align in every case and are in some cases governed by different axioms.

In the following, we describe the problems that arise from these mapping issues and our solutions to them.

Axiomatic divergence: The number of DOLCE and BFO types that are axiomatically mappable is relatively reduced compared to the total number of types (Khan and Keet, 2013). Our solution to the axiomatization issue is to ignore axioms. Indeed, unlike work carried out, e.g., in Gangemi et al. (2003), this work is neither aimed at mapping DOLCE to BFO, nor at axiomatizing WN. Instead, we attempt to answer the question: to what types of entities do WN synsets refer? As mentioned in section 4, the resulting mappings are to be read as "a WN synset X refers to something that is a subtype of BFO type Y" — we exclude instances for now.

Incomplete rule coverage: Incomplete mappings between DOLCE and BFO lead to an incomplete rule coverage. Our solution to this issue was to extend the coverage of the rules by mapping other types and relations included in KYOTO as well as relevant semantic labels in WN to BFO types. As we saw in section 4, the KYOTO-based rules can also be complemented with rules from the baseline mapping ruleset. In addition to the verb mappings, we can test other WN top-level mappings, for example, when no mapping rule applies. Additional mapping rules could further be tested using mappings of WN to other upper level ontologies, such as the Suggested Upper Merged Ontology (SUMO).[7] The potential issue with SUMO is that it allows for multiple inheritance, which might result in categorization problems. Using other upper level ontologies also involves creating mappings between their categories and BFO categories.

One-to-many mapping types: In some cases, a single DOLCE type maps to several BFO types. For example, DOLCE 'feature' ambiguously refers to a BFO INDEPENDENT CONTINUANT, FIAT OBJECT PART, or SITE. Our solution was to map, whenever possible, to BFO leaf types, i.e. in this example to FIAT OBJECT PART and SITE. For the DOLCE type 'non-physical object' which is mapped to both FUNCTION and ROLE, we chose the rule to output the lowest common genus REALIZABLE ENTITY. A further disambiguation step is required to chose between the two. This might be done using additional mapping

[6] See work in this direction in Seppälä (2015b).

[7] See `http://www.adampease.org/OP/`.

rules based on KYOTO and new sources.

Non-mapping types: Several DOLCE types have no matching type in BFO, as is the case for the DOLCE type 'abstract'. Conversely, some BFO types have no corresponding type in DOLCE, such as OBJECT BOUNDARY. A number of these cases might be captured by adding new rules using KYOTO and other sources.

While some of the above issues might not have a straightforward solution or no solution at all, even a partial mapping should be sufficient to cover a large portion of WN, leaving a smaller subset of problematic cases.

5.3 Heterogeneous Semantic Networks

One of the difficulties mapping WordNet to any ontology is that this task involves aligning semantic networks that were constructed with different aims and criteria. WN represents linguistic usage; BFO, entities in the world. While there are enough similarities between wordnets and ontologies to make this task possible, there are enough discrepancies to pose specific challenges. In this section, we discuss some of the challenges we encountered in our work.

Systematic polysemy: Systematic polysemy was one source of disagreement for generating the gold standard. For example, the synset `red-green_dichromacy.n.01` has the definition "confusion of red and green". Should this be mapped to BFO's DISPOSITION or PROCESS? On the one hand, a person with red-green dichromacy would not distinguish red from green if they were looking at a red and green object. This suggests that it is a disposition that inheres in the person and is realized by confusing red and green. On the other hand, 'confusion' is polysemous between a process and a result of a process. Similarly, we had nine cases of synsets such as `carpet_beetle.n.01` that can be used to describe a single organism, as in "I saw a carpet beetle in my bedroom", or an entire population, as in "The carpet beetle is not endangered." When we speak of a species being endangered, we are not speaking of a threat to individual organisms, but instead a threat to the population as a whole. For the gold standard, we tagged these as BFO OBJECTS (as individual organisms) rather than OBJECT AGGREGATES (as populations). Further investigation is needed to determine more nuanced ways of handling systematic polysemy.

Hierarchical discrepancies: Some of the mapping errors are the result of the discrepancy between WordNet's hypernym relation and `rdfs:subClassOf` used by BFO. For example, in WN `symptom.n.01` and `sign.n.06` are both descended from the Base Concept (BC) `cognition.n.01`. However, this is not ontologically precise. Symptoms such as a fever are not literally cognitions. However, WordNet's hypernym relations, in contrast to `rdfs:subClassOf`, are not intended to express relations among types of things but pyscho-linguistic intuitions of native English speakers. In the future, strategies for dealing with these cases need to be developed. Such strategies could include semi-automatic methods for ontologically evaluating WordNet's hierarchies, as in RUDIFY (Hicks and Herold, 2009). Another approach we are considering is to iteratively refine the mapping rules as errors are found.

Ontological distinctions: Other errors relate to ontological distinctions that WN is not intended to capture, e.g., `carrier.n.09`'s BC is correctly 'person', but this does not capture the distinction between rigid and non-rigid properties (Guarino and Welty, 2002). In the current version of the rules, these synsets are annotated with the BFO type ROLE, a non-rigid property, using WN's semantic type 'noun.person'.

Non-existent entities: Finally, we found that mapping a linguistic resource to a realist ontology raises the question of what to do with synsets that describe entities that are not real. For example, how can we map WN's `mythical_creature.n.01` and `metempsychosis.n.01` to a realist ontology such as BFO? This issue is particularly challenging for automating synset annotation since the system and the rules it uses have no way of telling apart existent entities from non-existent ones.[8]

While we have not resolved all questions around WN synsets that don't map to BFO, they raise interesting issues. A stimulating challenge will be to provide BFO-compliant interpretations of unmatched WN synsets.

5.4 Nature of Some Entity Types

Achieving consensus on the gold standard can also be challenging when the BFO community has on-

[8]The question of non-existent entities is itself an issue in ontology in general. For an overview, see: `http://plato.stanford.edu/entries/nonexistent-objects/`.

going discussions about the nature of some entity types such as language, measurements, and quantities. For example, measurements of temporal intervals are not modeled as such in BFO. So `track_record.n.01` would be mapped to ONE-DIMENSIONAL TEMPORAL REGION instead of the measurement of the time interval.

However, in the Information Artifact Ontology (IAO) and the Ontology for Biomedical Investigations (OBI), ontologies that extend BFO, measurements are categorized under GENERICALLY DEPENDENT CONTINUANTS.[9] Thus, for the time being and unless BFO proposes another way to represent these types of entities, we mostly consider these cases to refer to subtypes of GENERICALLY DEPENDENT CONTINUANTS.

5.5 Other Challenges

Issues arising from WN definitions: In a few cases, deciding what BFO type to assign was difficult due to vague, ambiguous, or unclear definitions of synsets. For example, the definition of `attribute.n.02` ("an abstraction belonging to or characteristic of an entity") is rather vague; it is thus difficult to determine which BFO type to assign to the corresponding synset and, consequently, to its hyponyms. When that happend in the elaboration of the baseline mapping rules, we examined the direct hyponyms of the category and assigned, whenever possible, a BFO leaf- or lower-intermediate type that seemed to cover most of the hyponyms.

Annotation mistakes: In two cases, our annotators made obvious mistakes, such as tagging the verb `die.v.02` with MATERIAL ENTITY instead of PROCESS. We confirmed with the annotators that these were in fact errors on their part and corrected the annotations with the appropriate BFO type.

Unknown issues: In some cases, we did not find any obvious explanation for the disagreement. For example, annotator A assigned the type REALIZABLE ENTITY to `federal_tax_lien.n.01` and annotator B, GENERICALLY DEPENDENT CONTINUANT. These cases were resolved by annotator B during the final examination of the annotations, as described in section 4.

[9] See OBI 'value specification' (OBI:0001933) and IAO 'measurement datum' (IAO:0000109).

6 Conclusion and Future Work

We presented a method to semi-automatically map WordNet 3.0 synsets to BFO 2.0 types via the KYOTO Ontology. Our preliminary results are encouraging, but reveal a number of challenges as addressed in the discussion section. More work is thus needed to see if the method scales to the full WN.

Future work will include:

- adding the verb-mapping baseline rule to improve verb mappings;
- examining our results and future development-sample results in more detail to investigate which parts of the rules are most productive, which ones cause errors, etc., and refine and reorder the rules;
- testing if complementing the KYOTO-based rules with other baseline rules improves the mapping results;
- testing if mapping all or part of the Base Concepts to BFO and propagating the mappings downwards would perform better or could be used in combination with the current method;
- resolving issues related to systematic polysemy by determining specific principles on their processing with BFO developers;
- studying the case of adjectives and their processing in terms of BFO types.

Acknowledgements

Work on this paper was supported in part by the Swiss National Science Foundation (SNSF) and by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Thanks also to the anonymous reviewers, Christopher Crowner, and Barry Smith.

References

Robert Arp, Barry Smith, and Andrew D. Spear. 2015. *Building Ontologies with Basic Formal Ontology*. MIT Press, Cambridge, MA.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 820–838.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2010. Interfacing WordNet with DOLCE: towards OntoWordNet. In Chu-ren Huang, Nicoletta Calzolari, and Aldo Gangemi, editors, *Ontology and the Lexicon: A Natural Language Processing Perspective*, pages 36–52. Cambridge University Press.

Pierre Grenon. 2003. BFO in a Nutshell: A Bicategorial Axiomatization of BFO and Comparison with DOLCE. IFOMIS Report 06/2003. Technical report, Institute for Formal Ontology and Medical Information Science (IFOMIS), University of Leipzig, Leipzig, Germany.

Nicola Guarino and Christopher Welty. 2002. Evaluating ontological decisions with ontoclean. *Commun. ACM*, 45(2):61–65, February.

Axel Herold, Amanda Hicks, German Rigau, and Egoze Laparra. 2009. Central Ontology Version - 1 Deliverable 6.2. Technical report.

Amanda Hicks and Axel Herold. 2009. Evaluating ontologies with rudify. In Jan L. G. Dietz, editor, *Proceedings of the 2nd International Conference on Knowledge Engineering and Ontology Development (KEOD'09)*, pages 5–12. INSTICC Press.

Zubeida Casmod Khan and C. Maria Keet. 2013. Addressing issues in foundational ontology mediation. In *Proceedings of KEOD'13*, pages 5–16, Vilamoura, Portugal, September 19–22. SCITEPRESS.

Egoitz Laparra, German Rigau, and Piek Vossen. 2012. Mapping WordNet to the Kyoto ontology. In *LREC*, pages 2584–2589.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

I. Niles and A. Pease. 2003. Linking Lexicons and Ontologies: Mapping Wordnet to the Suggested Upper Merged Ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416. Citeseer.

Adam Pease and Christiane Fellbaum. 2010. Formal ontology as interlingua: The SUMO and WordNet linking project and global WordNet. In Chu-ren Huang, Nicoletta Calzolari, and Aldo Gangemi, editors, *Ontology and the Lexicon: A Natural Language Processing Perspective*. Cambridge University Press.

Selja Seppälä, Barry Smith, and Werner Ceusters. 2014. Applying the realism-based ontology-versioning method for tracking changes in the basic formal ontology. In *8th International Conference on Formal Ontology in Information Systems (FOIS 2014)*, Rio de Janeiro, Brazil.

Selja Seppälä. 2015a. Mapping WordNet to the Basic Formal Ontology using the KYOTO ontology. In *Proceedings of ICBO 2015*.

Selja Seppälä. 2015b. An ontological framework for modeling the contents of definitions. *Terminology*, 21(1):23–50.

A. Patrice Seyed. 2009. Bfo/dolce primitive relation comparison. In *Nature Precedings*.

Barry Smith and Werner Ceusters. 2010. Ontological Realism: A Methodology for Coordinated Evolution of Scientific Ontologies. *Applied Ontology*, 5:139–188.

Barry Smith, Mauricio Almeida, Jonathan Bona, Mathias Brochhausen, Werner Ceusters, Melanie Courtot, Randall Dipert, Albert Goldfain, Pierre Grenon, Janna Hastings, William Hogan, Leonard Jacuzzo, Ingvar Johansson, Chris Mungall, Darren Natale, Fabian Neuhaus, Anthony Petosa Robert Rovetto, Alan Ruttenberg, Mark Ressler, and Stefan Schulz, 2012. *Basic Formal Ontology 2.0: DRAFT SPECIFICATION AND USER'S GUIDE*, July.

Andrew D. Spear, 2006. *Ontology for the Twenty First Century: An Introduction with Recommendations*. Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany.

Lynda Temal, Arnaud Rosier, Olivier Dameron, and Anita Burgun. 2010. Mapping BFO and DOLCE. *Studies In Health Technology And Informatics*, 160(Pt 2):1065–1069.

Piek Vossen, German Rigau, Eneko Agirre, Aitor Soroa, Monica Monachini, and Roberto Bartolini. 2010. KYOTO: an open platform for mining facts. In *Proceedings of the 6th Workshop on Ontologies and Lexical Resources*, pages 1–10.

Augmenting FarsNet with New Relations and Structures for verbs

Mehrnoush Shamsfard
Faculty of Computer Science and Engineering
Shahid Beheshti University,
Tehran, Iran.
m-shams@sbu.ac.ir

Yasaman Ghazanfari
NLP Research lab.
Shahid Beheshti University,
Tehran, Iran.
yasaman_ghazanfari@yahoo.com

Abstract

This paper discusses the semantic augmentation of FarsNet -the Persian WordNet- with new relations and structures for verbs. FarsNet1.0, the first Persian WordNet obeys the Structure of Princeton WordNet 2.1. In this paper we discuss FarsNet 2.0 in which new inter-POS relations and verb frames are added. In fact FarsNet2.0 is a combination of WordNet and VerbNet for Persian. It includes more than 30,000 lexical entries arranged in about 20,000 synsets with about 18000 mappings to Princeton WordNet synsets. There ae about 43000 relations between synsets and senses in FarsNet 2.0. It includes verb frames in two levels (syntactic and thematic) for about 200 simple Persian verbs.

1 Introduction

The Persian language, also known as Farsi, is a member of the Iranian group of the Indo-Iranian sub-family of the Indo-European languages. It is the official language of Iran, Afghanistan and Tajikistan with more than 100 million speakers.
In Persian verbs are the main carriers of a sentence meaning like many other languages. They may appear in simple or complex forms. Simple verbs have simple morphological structure, the verbal constituent. Compound verbs, on the other hand, consist of a nonverbal constituent, such as a noun, adjective, past participle, prepositional phrase, or adverb, and a verbal constituent.
In this paper we focus on the new relations and structures added to Persian WordNet (FarsNet) for verbs.
In the rest of the paper we first have an overview on FarsNet, the Persian WordNet and its features in the last two versions. Section 3 talks about verb argument structures and frames. Section 4 discusses the developed corpus management system in which argument structures are extracted and tagged. Section 5 is dedicated to results and discussion and at last section 6 concludes the paper.

2 FarsNet: The Persian WordNet

FarsNet project was announced with the release of FarsNet 1.0 at 2008. FarsNet 1.0 included the lexical, syntactic and semantic knowledge about more than 17000 Persian words and phrases organized in about 10000 synsets of nouns, adjectives and verbs. It was a medium scaled WordNet like the Arabic one (at that time). Table 1 shows the statistics of FarsNet 1.0.

Table 1. FarsNet 1.0 Statistics

POS Category	Word	Sense	Synset
Noun	9488	14079	5180
Verb	4402	6028	2306
Adjective	3950	4363	2526
Total	17842	24480	10012

As it can be seen for each word in FarsNet 1.0 we have an average of 1.5 senses and each synset includes an average of .1.7 words.
FarsNet 1.0 was developed by a semiautomatic approach. The base concepts covered in FarsNet were chosen from the base concepts BCS1 and BCS2 of BalkaNet (Tufis, 2004) with an equivalent in Persian to achieve compatibility with other WordNets. And also from the most frequent words of two Persian corpora: Peykareh (Bijankhan, 2004) and PLDB (Assi, 1997) to preserve the Persian specific structures (Shamsfard, et. al, 2010).
FarsNet 1.0 had two main classes of relations defined: inner language and inter-language relations. Synonymy, hypernymy and hyponymy,

different types of meronymy, Antonymy and cause were among the inner-language relations. The second class included the relations equal-to and near-equal-to between FarsNet and WordNet 3.0 synsets as inter-language relations. All inner-language relations were inner-POS; which means that their domain and range were from the same POS category. In other words FarsNet 1.0 did not cover inter-POS relations.

At 2010 a major restructuring of FarsNet began which resulted in FarsNet 2.0. The main goals of the changes were enlarging the size (improving the quantity) along with enhancing the quality. The new version was supposed to include new PoS category, new types of relations and new structures.

FarsNet 2.0 extends FarsNet 1.0 in the following dimensions:

- Size: FarsNet 2.0 includes more than 30,000 lexical entries organized in about 20,000 synsets with about 43,000 relations and 18000 mappings to Princeton WordNet 3.0. The size is approximately doubled comparing to FarsNet 1.0. In FarsNet 2.0 Princeton base concepts are included in addition to the base concepts of BalkaNet.
- POS categories: FarsNet 2.0 adds the adverb category to FarsNet 1.0. It includes nouns, verbs, adjectives and adverbs now.
- Number and type of relations: FarsNet 2.0 includes inter-POS as well as inner-POS relations. 'Derivational form', antonymy, 'verbal part of' and 'non-verbal part of' are relations between word senses. 'Verbal (non-verbal)-part-of' is a new relation between a compound verb and its verbal (non-verbal) component.

  From the synset relations, in addition to hypernym (as between peach and fruit), hyponym (as between food and hamburger) , various types of meronym (as between apple and apple juice) and holonym (as between car door and car) entailment (as between snore and sleep) and cause (as between kill and die) which were all present at FarsNet 1.0 as well as at Princeton WordNet (Fellbaum, 1998), FarsNet 2.0 includes the following relations:

  - Has-attribute / is-attribute-of: the relation between a quantitative adjective and the attribute whose value is the adjective. For example the relation between heavy and weight or between warm and temperature
  - Domain / is-domain-of: the relation between a domain specific term and its corresponding domain. For example between Carbid and chemistry or between arthritis and medicine.
  - Agent/ Is-agent-of: the relation between a predicate (verb) and the potential agent of it. For example between author and writing or chef and cooking.
  - Patient/ Is-patient-of: the relation between a predicate (verb) and the potential patient or theme of it. For example between eat and edible thing or write and letter.
  - Instrument/ Is-instrument-of: the relation between a predicate (verb) and the potential instrument of it. For example between eat and spoon or write and pen.
  - Corresponding adjective: The relation between an adjective and the noun it often/ mainly describes. For example the relation between Stale and bread.
  - 'Related to'- the relation between any two synsets which has a semantic relation other than the previous named relations. For example the relation between author and book or between school and teaching.

  The above relations except the "domain/is domain of" and "has attribute/is attribute of" are new to both FarsNet and Princeton WordNet. Their creation is motivated by various NLP tasks. For example the relations between a predicate and its arguments such as agent, patient and instrument help semantic role labelers, word sense disambiguation (WSD) modules and information/ knowledge extraction systems to better find the corresponding relations and do their jobs.

  "related-to" relation is used to relate any two synsets which has a sort of relation not included in the above named relations. Although the relation between some of the related concepts could be extracted by traversing the links in Princeton WordNet or FarsNet 1.0, the new relation specifies the important ones explicitly. It is mostly used in information retrieval and also in finding similarity between text components for example in text summarization.
- New structure- FarsNet 2.0 is actually a combination of Persian WordNet and Persian VerbNet. It includes the verb frames (argument structure) of about 200 Persian simple verbs along with the selectional restrictions of their arguments. In the rest of the paper we discuss this new feature in more details.

3 Augmenting FarsNet with Verb Frames

3.1 Argument structures and Verb frames

FarsNet 2.0 includes the information about the argument structure of verbs and their selectional restrictions. In this part it is somehow similar to resources like VerbNet (Kipper, et al., 2006) developed for English language.

When talking about the semantic relationships among different entities within a sentence, the most relevant term is proposition. The core semantic content of every sentence is called a proposition which in turn consists of a predicate and one or more arguments (Brinton & Brinton, 2010). The arguments may appear in the form of a noun phrase, a propositional phrase, an adjective or adverb phrase or a sentence.

The argument structure (or frame) of a verb can be defined as the representation of that verb regarding the nature and number of participants it requires. In other words, it is considered as the kind of semantic relationship which holds among verb and other obligatory constituent within a sentence [Ghazanfari, 2014]. Other expressions in the sentence whose existence are optional are called adjuncts. The number of arguments of a verb makes its valency. Verb valence may be from zero to 4 (Dixon, 2000).

In many NLP applications, knowing the verb arguments can help parsers and analyzers to process and disambiguate the text. The arguments are the constituents of a sentence which complete the meaning of its verb.

Arguments can be defined in different levels: syntactic (such as NP, PP,…), grammatical (such as subject, object, …) and thematic or semantic (such as agent, patient, theme, …). In syntactic level, arguments are represented by their POS categories. For example the verb خندیدن (khandidan) 'to laugh' has one NP argument while دیدن (didan) 'to see' has two NP arguments regardless of their grammatical or semantic relations to the verb. Syntactic arguments can be used by syntax parsers to resolve the ambiguities.

On the other hand arguments may be defined at grammatical level showing grammatical roles such as subject and object of a verb. In the above example the verb 'to see' has two grammatical arguments, a subject and an object. These argument structures may be used by dependency parsers for disambiguation. We don't consider this level in our work.

The third level is semantics. Semantic arguments known as semantic roles, thematic roles or Θ-roles (theta roles) are used for semantic processing of texts. The verb 'to see' has two thematic roles; agent and theme as semantic arguments.

By these considerations, we define the argument structure or the frame of a verb in two levels: syntactic and semantic.

Syntactic tags include NP, VP, PP, Sentence, ….

For more than half a century, linguists have been trying to come up with a neat comprehensive set of universal semantic roles; however, there has not been a general agreement regarding the inventory of them yet. In this paper we use the role list proposed by Ghazanfari (2014). She has modified the list of Brinton & Brinton (2010) in order to fit the requirements to be used in different wordnets and especially to be applied in a convincing manner in FarsNet. Her list consists of the following roles [Ghazanfari, 2014] (in each case the role holder is shown in *italic*):

1. Agent: the human initiator, causer, doer or instigator of an action who acts by will or volition. The *logger* felled the tree. The tree was felled by the *logger*.
2. Actor: the animate entity who or which acts or causes an action. The *boy* broke the window accidentally. The *dog* barks.
3. Force: the inanimate cause of an action and its direct cause. The *wind* felled the tree. The window was broken by the *wind*.
4. Instrument: the means by which an event is caused or the tool generally inanimate used to carry out an action. The tree was felled with an *axe*. He used an axe to fell the tree.
5. Stimulus: The entity which causes a kind of psychological effect in another entity, the experiencer. The *noise* frightened the students.
6. Experiencer: the animate being affected inwardly by a state or action. *Mina* feels lonely. *I* like apple. The noise frightened the *students*. The news is pleasing to *me*.
7. Source: the place-from-which or person-from-whom an action emanates. I got the book from the *library/* my *friend*.
8. Goal: the place-to-which an action is directed, including indirect objects and directional adverbs. She reached the *coast*.
9. Recipient: an animate or some kind of quasi-animate entity, the person who gets or receives something. My *mother* was sent a gift. A new idea came to *me*. Daniel wrote a letter to the *bank*.

10. Path: the path taken in moving from one place to another in the course of an action. Hannibal travelled *over the mountains*. The package came *via Tehran*.
11. Location: the place-at/in-which an action occurs. The cat is in the *room*/ under the *table*. The *room* has many people in it.
12. Temporal: the time at which something happens or an action occurs. I will call on *Tuesday*/ at *noon*.
13. Possessor: the possessor of a thing, *He* has/ owns/possesses a house. The bag belongs to *minoo*.
14. Benefactive: the person or thing for which an action is performed or the person derives something from the actions of another. He ordered the book for *me*.
15. Patient: the person or thing affected by an action or the entity undergoing a change. I baked the *chicken*. He ate the *cake*.
16. Theme: The person or thing which undergoes an action or that which is transferred or moved by an event otherwise unchanged. I put the *book* on the table. The *paper* flew out of the window.
17. Neutral: The person or thing which is not changed or even acted upon but is simply present at an action. The *house* costs a lot. The *table* measures three feet by three feet.
18. Range: The specification or limitation of an action. The dress costs *a hundred dollars*. We drove *ten miles*.
19. Role: a person playing a role or part in an action or state. We made *Lise* treasurer of the club. *Hilda* is the principal of the school.
20. Associate: the entity having an equal status (role) with another argument in the sentence. They made Reza *the head of department*. She calls her doll *Juju*.
21. Reason: This refers to the reason or purpose for which an action takes place. Robin called the police for *help*. She returned to class to *take her book*.
22. Accompaniment: the entity which participates in close connection with the agent, actor, force, patient or theme but has a secondary role in the event. I went to the movies with *my friends*.
23. Manner: the qualification of an event, the way in which an action is performed or an event takes place. He lived out his life happily. Tom left in a hurry.

To extract the argument structures of verbs and the selectional restrictions of arguments, we used a corpus driven approach. For this reason we developed a corpus management system called Samp. First we tagged the arguments of various occurrences of the candidate verbs in the corpus by both syntactic and semantic roles. Then using the developed tool the argument structure and also the selectional restrictions are concluded semi-automatically and confirmed by linguist before adding to FarsNet.

Next sections discuss the corpus management system and the process of extracting the argument structures for FarsNet in more details.

4 The Corpus Management System

To extract the verb argument structures we developed a corpus management system (CMS) called Samp [Shahriyari, et al., 2014]. Samp like other corpus management systems (such as BNCweb) is able to receive a corpus as input, search in it and find and show all occurrences of a word along with its surrounding words in the corpus and prepare various types of reports about it.

Besides the above ordinary capabilities of a corpus management system, Samp has the following features:

- Samp accepts any Persian corpus, and changes its format to the desirable standard.
- Samp is a web based system capable of handling multiple synchronous users enabling cooperative corpus tagging. It creates a log of users' activities over the net.
- Samp is able to tokenize a raw corpus and tag it by POS categories either automatically or help to tag manually.
- Samp helps users to tag the corpus by senses provided by FarsNet or user. In fact, Samp provides a cooperative environment to let users tag the corpus semantically by FarsNet senses or by new user defined senses.
- Samp is able to search for a word and all of its inflections, derivations and also multi part words in which the search keyword is involved. For each search the word within its surrounding context is returned. The size of the surrounding window can be determined by user.
- Samp helps users to tag sentences by their verb's syntactic and thematic arguments.
- Samp helps the linguist to extract the verb frames and determine the selectional re-

strictions of arguments. Actually, it recommends the verb frames by summarization and generalization (mining) of tags users created for verbs and their arguments and let the linguist to confirm or correct it (more details in the next subsection).

4.1 Extraction of verbs' argument structures

Tagging the corpus

Tagging the corpus by senses and arguments of verbs has the following steps:

1- User enters the corpus to be tagged.
2- Samp reformats the corpus into its standard and makes it ready to be tagged.
3- User enters the word (verb) into the search pane.
4- Samp provides the list of sentences (evidences) in which the word (verb) or its inflections or its stem or its derivations are present by applying morphological analysis.
5- For the sentences in the list Samp asks the user to tag the verb by its meaning. It shows the list of senses provided by FarsNet. User can select the appropriate sense or add a new sense. User defined senses will be then evaluated to be added to FarsNet if necessary. This way we can complete the missing senses of FarsNet while tagging the corpus. Currently this task is performed manually. We are going to use WSD algorithms to tag word senses automatically in the future.
6- In the selected sentence, according to the determined sense, the arguments of the verb are found and tagged by syntactic (NP, PP, …) and semantic roles (Agent, Patient, …). More details are discussed in the next subsection.
7- Samp saves the tags and repeats steps 5 and 6 to complete the task for a verb.

After completing tagging the corpus, it's time to make a conclusion on the tags and extract the argument structure of a verb and the arguments' selectional restrictions. This task is discussed in following section.

Determining Syntactic and Semantic Arguments

For each evidence (sentence in which the desired verb is occurred) the verb arguments should be extracted. Then for each argument it is determined if it is obligatory or optional. Also the arguments are tagged by their selectional restrictions which show the properties of the filler of each argument slot.

For example suppose the verb بردن (bordan).
One of its meanings (senses) is 'to win' and the other one is 'to take'. For the first sense we may tag the following sentence in the corpus as follows:
Sentence: Iranian films won some prizes in the festival.
Force= Iranian Films and theme=prize
And for the second sense the following is an example.
Sentence: he took Reza from home to school at noon.
Agent = he, theme=Reza, source= home, goal= school, temporal= at noon.
As an instance the selectional restriction of the theme argument of this verb is 'to be portable".
Extracting syntactic and semantic arguments can be done in two modes; manual or semiautomatic.

In the manual mode (which is the main focus of this paper) Samp provides the environment for user to tag arguments and select their selectional restrictions in each sentence. The restrictions are recommended to the user by upward traversing the inclusion hierarchy of FarsNet from the argument node (finding its ancestors).

For semiautomatic mode we used a syntax parser to extract syntactic arguments and a semantic role labeler (SRL) (Jafarinejad & Shamsfard, 2012) to extract semantic arguments of the verb.

Concluding the Structure

In this part, the final argument structure and the most general selectional restrictions for its components are determined by Samp automatically. In the concluding subsystem, for each sense of a verb, Samp shows user a list of all of its assigned arguments in all sentences (evidences) with their selectional restrictions. This list shows the frequency of cooccurrency of each argument with the corresponding verb sense. It also shows the number of times each argument for a specific sense has been obligatory or optional.

According to this report Samp can suggest the final argument structure of a verb to be confirmed or corrected by user. This structure is built by getting union among all argument sets of the verb sense in all the evidences. In this task similar or identical sets are recognized and merged and different sets whose frequency of

occurrence is more than a threshold are added to the union set.

To determine the selectional restrictions, Samp finds the most general class among various classes introduced as the restriction of the arguments which are being merged.

For example suppose the verb خوردن (khordan). It is a polysemous verb for which 'to eat' and 'to hit' are two of possible meanings (senses). For the first sense we may tag the following sentences in the corpus as follows:

S1: To become healthy one should eat an apple a day.

S2: Babies eats milk as the main course before the age of 6 months.

S3: eating breakfast is important in having a successful day.

In S1 eat needs agent and patient as obligatory arguments and temporal (time) and reason as optional ones (adjuncts). In this sentence the selectional restriction of patient is being apple or its superclass: 'fruit'. Similarly in S2 the patient is milk and its selectional restriction can be 'drinks'. And in S3 the selectional restriction of the patient (breakfast) is meal.

In other words the patient of khordan in the meaning of 'to eat' may be a fruit, a dink[1] or a meal. Samp can infer from these evidences besides other sentences for this sense of khordan that the patient of 'khordan' may be an 'edible'.

It also concludes that 'khordan' ('to eat') has obligatory agent and patient and may have optional temporal, associate and reason.

In some cases more than one argument structure may be inferred for a unique sense of a verb. This may happen for one of the following reasons:

1- The argument sets may not be merged. For example for a unique sense, we may have agent and patient in some sentences and force and patient in some other sentences. In this case we may merge agent and force in a broader class as undergoer or keep the original structures and so have more than one legal argument structure.
2- The differences of two sets are in the obligatory arguments and have never co-occurred in the sentences. For example suppose a verb with agent, patient and source in some sentences and with agent and goal in some others but the patient and goal has never co-occurred for this verb in the corpus. In this case the two structures are kept separately to ask the user to see if they should be merged or not.
3- In case of having more than one argument set for a verb sense, the user may decide to split the sense into two more specific senses or add the argument sets 'as is' into FarsNet.

The final concluded argument structure is represented in a specific language and added to FarsNet

A sample of data added to FarsNet is following. (for verb bordan meaning to take). Anything within parenthesis is optional.

Syntactic arguments: NP&NP&(PP)&(PP)
(it means that the verb has 4 syntactic arguments , two obligatory noun phrases and two optional prepositional phrases)

Thematic Roles : Agent&Theme&(Source)&(Goal)
It means that the verb has 4 thematic arguments an obligatory agent, an obligatory theme and optional source and goal.

Relations :
NP&Agent
/NP&Theme/
(PP)&(Source)
(PP) & (Goal)

This shows the correspondence between the syntactic and the thematic arguments.

5 Results and Discussion

In this paper we talked about some new features developed in FarsNet 2.0. Table 2 shows the last statistics for FarsNet 2.0.

Table 2-some statistics on FarsNet 2.0

	Noun	Adj.	Adv.	Verb	Total
Word	16008	6560	2014	5679	30261
senses	19773	6904	2023	7438	36138
Synset	10954	4261	923	3266	19403
Sense relation	3096	345	22	3585	7048
Synset relation	31333	6733	1100	5492	36749
Mapped synsets	10108	4518	929	3023	18576

Besides extending the Persian WordNet we have had some studies (corpus based) on verbs.

In this study we selected 187 simple distinct Persian verbs. For these verbs, we extracted about 4118 distinct evidence sentences from the corpus and tagged them by the meaning (sense) of verb

[1] In Persian, It is usual to use the verb 'to eat' for drinks instead of 'to drink'

and its arguments. From these sentences we extracted 847 sets of verb-sense-argument structure which are all entered into FarsNet 2.0. In other words we completed the information of 187 verbs in FarsNet with their verb frames. considering that each verb has some senses and each sense may have more than one frame we entered 847 verb frames with their selectional restrictions into FarsNet.

To extract the arguments we considered the valency of verbs too. Valency refers to the capacity of a verb to take a specific number and type of arguments. Our study showed that there is no zero-valence verb in Persian. The statistics of the studied 190 simple verbs regarding their valence is shown in table 3.figure 1 is about the frequency of arguments in the test data.

Table 3-statistics on Persian simple verbs regarding their valence in the test set

type	Percentage
0-valence	%0
1-Valence	%17
2-Valence	%60
3-Valence	%23
4-valence	%0

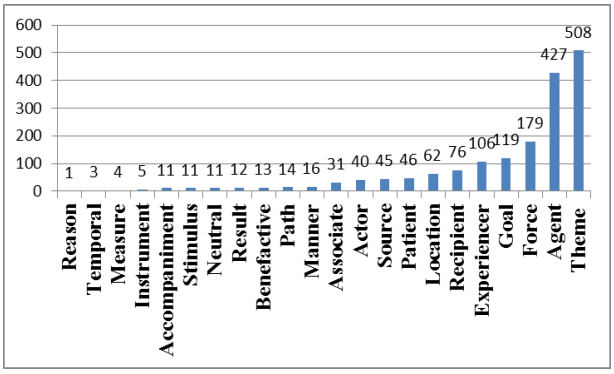


Figure 1- frequency of the arguments in the selected set

Enhancing the automatic part of our work especially in applying WSD algorithms to find the verb sense, SRL methods to extract semantic roles and the reasoning (concluding) part of extracting the argument structures besides using the extracted data in real world applications are among our further works.

References

Bijankhan, M. (2000). Bijankhan Corpus, http://ece.ut.ac.ir/dbrg/Bijankhan.

Brinton, L. J. and D. M. Brinton. (2010). The Linguistic Structure of Modern English. Amsterdam and Philadelphia: John Benjamins publishing Company.

Dixon, R.M.W. (2000). A Typology of Causatives: Form, Syntax, and Meaning. In Dixon, R.M.W. & Aikhenvald, Alexendra Y. Changing Valency: Case Studies in Transitivity. Cambridge University Press.

Fellbaum, C. (ed.) 1998. WordNet: An Electronic Lexical Database. MIT Press.

Ghazanfari, Y. (2014) A survey on semantic roles for inclusion in Persian WordNet, International Journal of Language Learning and Applied Linguistics World (IJLLALW) Volume 6 (2), June 2014; 150-158

Jafarinejad, F.,Shamsfard M.,(2012) Extracting Generalized Semantic Roles from Corpus, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006) Extending VerbNet with Novel Verb Classes. Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. June, 2006.

Scott, M. (2012). WordSmith Tools version 6, Liverpool: Lexical Analysis Software.

Shahriyarifard, A., Sharifzadeh A., Shamsfard M.,(2014), Introducing Samp, the corpus Management system, 3rd Iranian computational linguistics, Iran.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., et al. (2010). Semi Automatic Development of Farsnet; the Persian Wordnet. Proceedings of 5th Global WordNet Conference (GWA2010). Mumbai, India.

Dan Tufis. 2004. Balkanet: Aims, Methods, Results and perspectives. Romanian journal of Information Science and Technology. V7, pp.9-43.

High, Medium or Low? Detecting Intensity Variation Among Polar Synonyms in WordNet

Raksha Sharma and Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
IIT Bombay, Mumbai, India
{raksha,pb}@cse.iitb.ac.in

Abstract

For fine-grained sentiment analysis, we need to go beyond zero-one polarity and find a way to compare adjectives (synonyms) that share the same sense. Choice of a word from a set of synonyms, provides a way to select the exact polarity-intensity. For example, choosing to describe a person as *benevolent* rather than *kind*[1] changes the intensity of the expression.

In this paper, we present a sense based lexical resource, where synonyms are assigned intensity levels, *viz., high, medium* and *low*. We show that the measure $P(s|w)$ (probability of a sense s given the word w) can derive the intensity of a word within the sense. We observe a statistically significant *positive correlation* between $P(s|w)$ and intensity of synonyms for three languages, *viz., English, Marathi* and *Hindi*. The average correlation scores are 0.47 for English, 0.56 for Marathi and 0.58 for Hindi.

1 Introduction

Sentiment analysis is a crucial task for various Web and media outlets, such as, e-commerce websites, blogs and newspapers. The general approach of Sentiment Analysis is to summarize the semantic polarity (i.e., positive or negative) of sentences/documents (Riloff and Wiebe, 2003; Pang and Lee, 2004; Danescu-Niculescu-Mizil et al., 2009; Takamura et al., 2005; Baccianella et al., 2010a; Guerini et al., 2013). However, sentence intensity becomes crucial when we need to compare sentences having the same polarity orientation. In such scenarios, we can use intensity of words to judge the intensity of a sentence. Words that bear the same sense can be used interchangeably to upgrade or downgrade the intensity of the expression. The following example helps illustrate the problem we attempt to address.

- the synset (set of synonyms), {sound, levelheaded, intelligent, healthy} (Gloss: exercising or showing good judgment), are assigned a fixed positive polarity of *0.75* in SentiWordNet[2], while most people would agree that all the synonymous words are not equally positively intense. The use of *levelheaded* or *sound* makes a sentence more intensely positive in comparison to *healthy*, given that the sentence expresses the sense *exercising or showing good judgment*.

In addition to English, there exists polarity-intensity variation across synonyms in other languages also. Consider the following example from Hindi:

- The word सद्गुणी (Transliteration: Sadgunee, Translation: Virtuous) and लायक (Transliteration: Layak, Translation: Worthy) are synonymous words according to HindiWordNet for the sense *morally excellent*. Hindi native speakers confirm that the word सद्गुणी , is more intense than the word लायक in terms of polarity.

There are several manually or automatically created sense based lexical resources (Agerri and García-Serrano, 2010; Baccianella et al., 2010b) that assign the same positive or negative polarity to all synonymous words, making no distinction among them in terms of their intensity.

In this paper, we address the concept of polarity-intensity variation among synonyms and come up with a measure to predict the polarity intensity of a word for the given sense. We show that

[1]The words, *Benevolent* and *kind* are synonyms for the sense *well meaning and kindly* as per Oxford English dictionary.

[2]Available at: http://sentiwordnet.isti.cnr.it/

there is a statistically significant *positive correlation* between $P(s|w)$ (probability of sense s given word w) and intensity of a word w within the sense s (Section 3). Hence, the measure $P(s|w)$ can be used to predict the intensity of a word within the sense. We extensively validate this *positive correlation* in three languages[3], *viz.*, English, Marathi and Hindi (Section 5). We observe a statistically significant *positive correlation* of 0.47 for English, 0.56 for Marathi and 0.58 for Hindi (Section 7).

Our Contribution: Our work contributes an automatically generated sense based lexical resource where words which belong to the same sense are assigned three intensity levels, *viz., high, medium* and *low*. This resource can be used to derive intensity information of a subjective sentence or document, which essentially empowers existing sentiment analysis systems. In addition to this, intensity information of words can be used to reduce or enhance an over-expressed or under- expressed text respectively.

2 Related Work and Discussion

Several researchers have made successful attempts for finding opinion words (Wiebe, 2000; Taboada and Grieve, 2004; Takamura et al., 2005; Wilson et al., 2005; Kanayama and Nasukawa, 2006; Liu, 2010; Dragut et al., 2010; Ohana and Tierney, 2009; Agerri and García-Serrano, 2010; Sharma and Bhattacharyya, 2013); however, finding intensity of words still considered as a challenging task.

There have been some works on scaling adjectives by their strength, independent of the sense they express. The first work in the direction of adjectival scale was done by Hatzivassiloglou and McKeown (1993). They exploited linguistic knowledge available in the corpora to compute similarity between adjectives. However, their approach did not consider polarity orientation of adjectives, they provided ordering among non-polar adjectives like, *cold, lukewarm, warm, hot*. Kim et al. (2013) demonstrated that vector off-set can be used to derive scalar relationship amongst adjectives. De Melo and Bansal (2013) used a pattern based approach to identify intensity relation among adjectives, but their approach had a severe coverage problem. Ruppenhofer et al. (2014) provided ordering among polar adjectives that bear the same semantic property.

None of the existing works address intensity variation among synonyms. However, choice of a word from a set of synonyms provides a way to intensify the expression. Our approach pin-points the polarity-intensity variation across synonyms.

[3]A person who is a linguist as well as a native speaker of the language can annotate words with more accuracy. The availability of the linguists, who are also native speakers of Hindi and Marathi made us to choose these two languages other than English. Hindi and Marathi are two of the 23 official languages of India, which have approximately 258 and 73 million speakers respectively. English is chosen, because most of the lexical resources which we had pointed out in our work are in English only.

Language	Variables	Cor-value
English	Annotator-1∼TFC	-0.09
English	Annotator-2∼TFC	-0.09
Hindi	Annotator-1∼TFC	-0.04
Hindi	Annotator-2∼TFC	-0.09
Marathi	Annotator-1∼TFC	-0.11
Marathi	Annotator-2∼TFC	-0.10

Table 1: Correlation between Total Frequency Count (TFC) and intensity score assigned by two annotators in each language.

3 Polarity-intensity Variation and Synonymous Words

The classical *semantic bleaching theory*[4] states that a word which has *high* frequency of use tends to have *low* intensity in comparison to a word having less frequency of use. For example, the frequent use of the word *good* makes it less intense, while rare use of the word *great* makes it more intense (Kim and de Marneffe, 2013). However, *good* and *great* are not synonyms according to SentiWordNet. The *semantic bleaching phenomenon* throws light on the *positive* association between frequency and intensity regardless of any semantic relation (for example, *synonymy*). But, when we computed correlation between total-frequency (Section 5) and polarity-intensity within a sense (Section 6), we observed a negative correlation. Table 1 shows the correlation values obtained for three languages, *viz.,* English, Hindi and Marathi. The negative correlation shown in table 1, substantiates that total-frequency of a word cannot predict the polarity-intensity of a word within a particular sense. The *semantic bleaching* phenomenon compares total-frequency of words (sum

[4]The *semantic bleaching phenomenon* in words was reported in US edition of *New York Times*:`http://www.nytimes.com/2010/07/18/magazine/18onlanguage-anniversary.html?\_r=0`

of the word's count in all its senses), while count of a word in the sense is the potential clue for intensity of the word in the sense.

In this paper, we define polarity-intensity variation among synonyms (words having the same sense), specifically for the polar words. Words having the same sense cannot be directly compared on the basis of their frequency count in the sense, because their total frequency of usage (total count of the words in all its senses) are different. We need a relative count of the synonymous words, that is, (*Count of the word with the sense* / *Total count of the word*).

The cause of overuse of a word is its use in multiple senses (Durkin, 2009). Therefore, use of a word in multiple senses increases the total frequency of use, but the word loses its frequency count with a particular sense relative to the total frequency count of the word. Considering this frequency distribution as a base, we hypothesize polarity-intensity variation among words belonging to a particular sense.

{*A word which has high relative frequency for a sense is high intense in comparison to a word which has low relative frequency for the sense.*}

Consider the following derivation that validates our proposed hypothesis for polarity-intensity variation across synonyms. According to the *semantic bleaching* phenomenon:

$$TC(w_1) < TC(w_2) \Rightarrow I(w_1) > I(w_2) \quad (1)$$

Where, TC is a function that gives total-count of a word and I is a function that gives intensity of a word.

Since w_2 has higher total-frequency (overused) than w_1, we can deduce that w_2 has more senses in comparison with w_1. Let us assume that w_1 and w_2 are synonyms for i^{th} sense and w_1 has only one sense, that is, i^{th} sense and w_2 has n ($n > 1$) senses. Now, we rewrite equation 1 in terms of count of words in their senses in equation 2. Here, $SC^k_{W_j}$ represents sense-count, that is, count of the word 'w_j' with the sense 'k'.

$$SC^1_{w_1} < \sum_{i=1}^{n} SC^i_{w_2} \Rightarrow I(w_1) > I(w_2) \quad (2)$$

Now, to compare the synonymous words w_1 and w_2 in the i^{th} sense, we need their relative counts in the sense (Equation 3). Relative count is the count of the word with the sense divided by total-count of the word. Since w_1 has only one sense, so its sense-count and total-count will remain the same. Hence, the fraction $\frac{SC^i_{w_1}}{SC^1_{w_1}}$ will be 1, which is the maximum possible value for the fraction. While, w_2 has more than one senses, so its sense-count will always be less than total-count. Hence, the fraction $\frac{SC^i_{w_2}}{\sum_{i=1}^{n} SC^i_{w_2}}$ will always be less than 1. On the other hand, w_1 has only one sense, so the intensity relation between w_1 and w_2, given by the *semantic bleaching phenomenon* will remain the same.

$$\begin{aligned} \frac{SC^i_{w_1}}{SC^1_{w_1}} &> \frac{SC^i_{w_2}}{\sum_{i=1}^{n} SC^i_{w_2}} \\ \Rightarrow I^1(w_1) &> I^i(w_2) \end{aligned} \quad (3)$$

We observe a reversal of the sign $<$ to $>$ in case of relative frequency comparison of w_1 and w_2, but the intensity relation remains intact. Essentially, a word that shows its majority occurrence with the sense or has a higher relative frequency count, is more intense for the sense than the other synonymous words.

A few such instances of polarity-intensity variation in a sense are shown in table 2. We asked two linguists in each language to compare the polarity-intensity of the exemplified synonymous words for the given sense. They mutually agreed on the fact that the first word is more intense than the second word for the considered sense. The same intensity relation between the synonyms can be inferred from the relative frequency counts (sense-count/total-count) of words. The relative frequency count of the first word is higher than the second word for all the senses given in table 2. The total-count and sense-count values are obtained from English and Hindi sense annotated corpus (section 5).

4 Probability of Sense Given Word

Statistically, a relative frequency count of a word is nothing but the *probability of sense given word* ($P(s|w)$). The function $C(w_i, s_j)$ gives count of w_i with the sense s_j, while the function $C(w_i)$ gives total-count (aggregation of count in all senses). The measure $P(s_j|w_i)$ is defined as follows:

$$\begin{aligned} P(s_j|w_i) &= P(s_j, w_i)/P(w_i) \\ &= C(w_i, s_j)/C(w_i), \\ \text{Where, } C(w_i) &= \sum_{K} C(w_i, S_k) \end{aligned} \quad (4)$$

Synonymous-Words(w_1,w_2)	Sense-Definition	Total-Count(w_1)	Total-Count(w_2)	Sense-Count(w_1)	Sense-Count(w_2)
Awful, Painful	exceptionally bad or displeasing	14	10	12	1
Proficient, Good	Having or showing knowledge and skill and attitude	4	263	3	2
लाभदायक, उपयोगी (Translation: Beneficial, Useful)	Giving an advantage	22	35	22	15
उत्तम, अच्छा (Translation: Exquisite, Substantial)	Having or marked by unusual and impressive intelligence	181	270	181	1

Table 2: Examples of polarity-intensity variation from English and Hindi. In all cases, first word is more intense than the second word for the given sense.

Hence, we deduce that if a word possess higher value for the measure $P(s|w)$, then it is more intense than other synonymous words. Equation 5 generalizes the proposed hypothesis.

$$\boxed{P(s|w_1) > P(s|w_2) \Rightarrow I^s(w_1) > I^s(w_2)}$$

Where, w_1 and w_2 belong to the same sense s. (5)

In summary, when we compare words within a sense, we need to account for the participation of these words in other senses also. The proposed probabilistic measure, *probability of sense given word* considers the participation of a word in other senses also in the form of its total-count. We observe a statistically significant *positive correlation* between polarity-intensity levels assigned by linguists and the value of $P(s|w)$ (relative frequency of a word w in a sense s) (Section 7).

A high value of $P(s|w)$ is possible in the following scenarios.

- If w is rarely found with the sense s, then it should be rare in all.
- If w is very frequent, then the majority part of its total occurrences should be with the sense s only.

5 Dataset

We validate our hypothesis using three languages, *viz., English, Hindi*, and *Marathi*.

English: For English, we extracted all the adjective synsets whose polarity (positive or negative) value is greater than 0.5 as per SentiWordNet, except the synsets that have only one word. We ignored the synsets having polarity values less than or equal to 0.5, considering them a weak candidate for polarity-intensity variation phenomenon. With the threshold value of 0.5, we extracted a total of 1116 synsets. However, SentiWordNet is an automatically compiled lexical resource, which assigns polarity values based on corpus dependent probabilistic measures. To make our English dataset potentially conclusive, we asked two linguists in English to manually inspect the polarity orientation of synsets (senses). Table 3 is a confusion matrix, that summarizes the results of manual inspection of English dataset extracted from SentiWordNet (SWN).

		Polarity Orientation in SWN		
		Negative	Positive	Objective
Actual	Negative	599	37	0
	Positive	77	311	0
	Objective	84	8	–

Table 3: Confusion matrix

A few examples of wrong polarity orientation by SentiWordNet are given in table 4. We considered the correct synsets for our experiment. Consequently, intensity ordering is demonstrated for 1024 $(1116 - 92)$ English synsets.

Hindi and Marathi: For Hindi and Marathi, we asked two linguists in each language to extract polar synsets (senses) from HindiWordNet and MarathiWordNet.[5] Manual extraction of

[5]Available at: http://www.cfilt.iitb.ac.in/wordnet/

Synonymous-Words	Sense-Definition	Polarity by SWN	Actual Polarity
Murderous, homicidal	Having a tendency towards killing another human beings	Positive(0.625)	Negative
Enthralled, entranced	Filled with wonder and delight	Negative(0.75)	Positive
Unmarried, Single	Not married	Negative(0.75)	Objective

Table 4: Examples of synsets (Senses), which are assigned wrong polarity by SentiWordNet.

senses were required, because HindiWordNet and MarathiWordNet do not have polarity information for synsets. The total number of observed senses and words in each language are specified in table 5.

Language	Senses	Words
English	1024	3397
Hindi	172	2614
Marathi	325	1346

Table 5: Observed synset statistics

C(w_i,s_j): For English words, the value of the function C is obtained from the English WordNet database file, that is, 'cntlist'.[6] For Hindi and Marathi, we used a sense marked corpus in tourism and health domain.[7] (Khapra et al., 2010) The total number of sense marked words in each domain are depicted in table 6. If a word shows zero frequency of use for any particular sense, we replace it with 0.1 according to a standard smoothing technique (Han et al., 2006).

POS category	Tourism	Health
Noun	72932	52230
Verb	26086	24291
Adjective	32499	22699
Adverb	9820	855

Table 6: Hindi/Marathi sense marked corpus statistics

6 Gold Standard Data Preparation

We asked two linguists in each language to assign words to different intensity levels, *viz., high* (3), *medium* (2), and *low* (1) within a synset. A discrete scale with only three intensity levels is chosen to reduce the subjectivity issue in annotation, consequently complexity of annotation. Consider the following example of synonymous words, where intensity levels are assigned by English linguists.

- Grievous (Intensity: 3) $>$ dangerous (Intensity: 2) $>$ serious (Intensity: 1) for the sense, *causing fear or anxiety by threatening great harm*.

Table 7 shows the inter annotator agreement for each language, computed using *weighted Cohen's kappa* measure.

Language	Inter-annotator Agreement
English	53%
Hindi	69%
Marathi	64%

Table 7: Weighted Cohen's Kappa in percent

7 Empirical Validation

We validate the hypothesized relation between polarity-intensity and probabilistic measure: $P(s|w)$ by finding Pearson product-moment correlation. To test the significance of correlation value, we perform a directional test, that is, *t-test* using *cor.test* function of R.[8] We obtain a statistically significant *positive correlation* between gold standard intensity levels and $P(s|w)$ for all the three languages. Table 8 shows the correlation values, t-values, p-values and confidence intervals. The statistically significant positive correlation parameter allows us to conclude that the polarity-intensity of a word in a sense can be inferred by the relative frequency ($P(s|w)$) of the word in the sense.

8 Error Analysis

The observed scenarios that affect the proposed hypothesis negatively are as follows.

[6] Detail available at: http://wordnet.princeton.edu/wordnet/man/cntlist.5WN.html

[7] Available at: www.cfilt.iitb.ac.in.

[8] R is a language and environment for statistical computing and graphics. Detail available at: http://www.r-project.org/

Lang.	Variable	Cor-value	t-Value	p-value	95% Confidence-interval
English	$P(s\|w) \sim$ Linguist1	0.48	32.36	$< .0001$	0.46 to 0.51
	$P(s\|w) \sim$ Linguist2	0.44	28.96	$< .0001$	0.42 to 0.47
Marathi	$P(s\|w) \sim$ Linguist1	0.58	26.10	$< .0001$	0.54 to 0.62
	$P(s\|w) \sim$ Linguist2	0.53	22.91	$< .0001$	0.49 to 0.58
Hindi	$P(s\|w) \sim$ Linguist1	0.60	38.33	$< .0001$	0.56 to 0.63
	$P(s\|w) \sim$ Linguist2	0.55	33.66	$< .0001$	0.53 to 0.58

Table 8: Statistically significant correlation values with the results of t-test

1 There are words which do not have their all senses in WordNets. For instance, the word *bastard* as an adjective has only one sense, that is, *fraudulent; having a misleading appearance* as per WordNet, but according to Oxford dictionary, it has one more sense, that is, *born of parents not married to each other; illegitimate*. The exclusion of such senses leads to wrong total-count of the word in the English WordNet database file.

2 There are words which are not found in the corpus with any sense of them. In this case, besides the frequency count of the word with the sense, we fail to get evidence for total-count of the word.

In such cases, the probabilistic measure, that is, $P(s|w)$ fails to result in a strong value that can insinuate the correct polarity-intensity of a word, which leads to fall in the correlation estimate.

9 Conclusion

In this paper, we addressed the concept of polarity-intensity variation among synonyms. We show that the relative frequency of a word w in a sense s, that is, $P(s|w)$ is a predictor of polarity-intensity of the word in the sense. We present a sense based lexical resource in three languages, where polar synonyms are annotated with the intensity levels, *viz., high, medium* and *low*.

Manual checking of sentiment WordNets for intensity variation is a difficult endeavor. Therefore, a by-product of our polarity-intensity analysis is that sentiment WordNets can become more informative resource for sentiment analysis. In addition to this, intensity information of words can be used to reduce or enhance an over-expressed or under-expressed text respectively.

10 Acknowledgment

We heartily thank linguists Rajita Shukla, Jaya Saraswati, Gajanan Rane, and Lata Popale from CFILT Lab, IIT Bombay for giving their valuable contribution in gold standard data creation.

References

Rodrigo Agerri and Ana García-Serrano. 2010. Q-wordnet: Extracting polarity from wordnet senses. In *LREC*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010a. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010b. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 141–150, New York, NY, USA. ACM.

Gerard De Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1.

Eduard C Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. 2010. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1761–1764. ACM.

Philip Durkin. 2009. *The Oxford guide to etymology*. Oxford University Press.

Marco Guerini, Lorenzo Gatti, and Marco Turchi. 2013. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*.

Jiawei Han, Micheline Kamber, and Jian Pei. 2006. *Data mining: concepts and techniques*. Morgan kaufmann.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics.

Mitesh M. Khapra, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. All words domain adapted wsd: Finding a middle ground between supervision and unsupervision. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1532–1541, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630, Seattle, Washington, USA, October. Association for Computational Linguistics.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:568.

Bruno Ohana and Brendan Tierney. 2009. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. *EACL 2014*, 117.

Raksha Sharma and Pushpak Bhattacharyya. 2013. Detecting domain dedicated polar words. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 661–666, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press*.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 133–140, Stroudsburg, PA, USA. Association for Computational Linguistics.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.

The Role of the WordNet Relations in the Knowledge-based Word Sense Disambiguation Task

Kiril Simov **Alexander Popov** **Petya Osenova**
Linguistic Modeling Department
IICT, Bulgarian Academy of Scinces
{kivs|alex.popov|petya}@bultreebank.org

Abstract

In this paper we present an analysis of different semantic relations extracted from WordNet, Extended WordNet and SemCor, with respect to their role in the task of knowledge-based word sense disambiguation. The experiments use the same algorithm and the same test sets, but different variants of the knowledge graph. The results show that different sets of relations have different impact on the results: positive or negative. The beneficial ones are discussed with respect to the combination of relations and with respect to the test set. The inclusion of inference has only a modest impact on accuracy, while the addition of syntactic relations produces stable improvement over the baselines.

1 Introduction

Knowledge-based methods for Word Sense Disambiguation (WSD) are attractive to the NLP community because they do not require manually annotated corpora. On the other hand, these methods are not considered completely unsupervised, because they do need information about senses of words in texts, and about the relations that hold between them, represented in the form of a directed or undirected graph, called **knowledge graph** (KG). The most frequently used knowledge graph is based on WordNet (WN) (Fellbaum, 1998) or Extended WordNet (XWN) (Mihalcea and Moldovan, 2001), where synsets constitute the vertices of the graph and relations between synsets are represented as edges within it. Simov et al. (2015) provided evidence that the addition of linguistically motivated semantic relations to the KG improves the performance of Knowledge-based WSD (KWSD). In the current work we perform an analysis of the various semantic relations in WN and XWN knowledge graphs. The analysis is performed via experiments with different subgraphs that include only some of the semantic relations in WN and XWN. Some of the relation types allow for inference to be applied over them. Thus, inferred semantic relations have been included in some of KGs as well. The experiments were performed on the manually annotated SemCor corpus (Miller et al., 1993). In order to test the semantic relations extracted from the syntactically annotated corpus, the same was divided into four parts. We used three of the divisions for the extraction of new relations and one part for testing.

The structure of the papers is as follows: the next section discusses related work on the topic. Section 3 describes the experimental setup. Section 4 focuses on the experiments with the semantic relations in WordNet. Section 5 presents the experiments with the semantic relations in Extended WordNet. Section 6 gives an overview of the experiments with syntactic relations. Section 7 concludes the paper.

2 Related Work

Knowledge-based systems for WSD have proven to be a good alternative to supervised systems, which require large amounts of manually annotated training data. In contrast, knowledge-based systems require only a knowledge base and no additional corpus-dependent information. An especially popular knowledge-based disambiguation approach has been the use of successful graph-based algorithms known under the name of "Random Walk on Graph" (Agirre et al., 2014). Most methods exploit variants of the PageRank algorithm (Brin and Page, 2012). Agirre and Soroa (2009) apply a variant of the algorithm to Word Sense Disambiguation by translating WordNet into a knowledge graph in which the synsets are represented as vertices and the relations between them are represented as edges between the ver-

tices. Calculating the PageRank vector **Pr** is accomplished through solving the equation:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v} \quad (1)$$

where *M* is an *N x N* transition probability matrix (*N* being the number of vertices in the graph), *c* is the damping factor and **v** is an *N x 1* vector. In the traditional, static version of PageRank the values of **v** are all equal (*1/N*), which means that in the case of a random jump each vertex is equally likely to be selected. Modifying the values of **v** effectively changes these probabilities and thus makes certain vertices more important. The version of PageRank for which the values in **v** are not uniform is called *Personalized PageRank*.

The words in the text that are to be disambiguated are inserted as nodes in the knowledge graph and are connected to their potential senses via directed edges (by default, a context window of at least 20 words is used). These newly introduced nodes serve to inject initial probability mass (via the **v** vector) and thus to make their associated sense nodes especially relevant in the knowledge graph. Applying the Personalized PageRank algorithm iteratively over the graph determines the most appropriate sense for each ambiguous word. The method has been boosted by the addition of new relations and by developing variations and optimizations of the algorithm. It has also been applied to the task of NED (Agirre et al., 2015).

The success of KWSD approaches apparently depends on the quality of the knowledge graph – whether the knowledge represented in terms of nodes and relations (edges/links) between them is sufficient for the algorithm to pick the correct senses of ambiguous words. Several extensions of the knowledge graph, constructed on the basis of WordNet, have been proposed and implemented. In (Simov et al., 2015), semantic and syntactic relations from the sense annotated BulTreeBank have been extracted and the algorithm has been applied to Bulgarian data. In order to do that, the treebank was first annotated with synsets from the BulTreeBank WordNet[1], aligned to the Princeton WordNet. The word forms annotated with senses at this point are 69,333, consisting of nouns and verbs. Out of these, 12,792 sense-tagged word forms have been used for testing, and the rest have been used for relation extraction.

The WordNet ontological relations that have been used are 252,392, and the relations derived from the synset glosses are 419,387. Additionally, the following relations have been extracted: inferred hypernymy relations; syntactic relations from the gold corpus; extended syntactic relations; domain relations from WordNet. Thus, 590,272 new relations have been added. The newly added relations introduce syntagmatic information into the graph, which was originally constructed out of paradigmatic relations. The results from the experiments with paradigmatic relations alone (done on the the whole corpus) show highest accuracy (0.551) for the combination of: WordNet relations + relations from the glosses + inferred hypernymy relations + domain relations of the kind synset-to-synset + domain hierarchy relations. The results from the experiments with mixed – paradigmatic and syntagmatic – relations (done on a test portion of one fourth of the corpus) show highest accuracy (0.656) for the combination of: WordNet relations + relations from XWN + inferred hypernymy relations + dependency relations from the golden corpus + extended dependency relations starting from one level up + domain relations of the kind synset-to-synset + domain hierarchy relations.

Kdzia et al. (2014) present work on WSD for Polish using the Polish WordNet, extended with relations between semantically similar words. The authors use the Measure of Semantic Relatedness which assigns a numerical value to pairs of words. This numerical value reflects the degree of closeness between two words. For each word w_i, a list of most closely related words w_j is constructed (length of the list is 20). Then the synsets that contain w_i and synsets containing some of w_j are connected with new links. The evaluation, based on the extended knowledge graph, shows improvement on the sentence level.

3 Experimental Set-Up

The experiments presented here were carried out with the UKB[2] tool, which provides graph-based methods for Word Sense Disambiguation and measuring lexical similarity. The tool uses the Personalized PageRank algorithm, described in Agirre and Soroa (2009). It builds a knowledge graph over a set of relations that can be induced

[1]The Core WordNet is freely available at: http://compling.hss.ntu.edu.sg/omw/. The extended one will be released soon. For more details about the sense annotated BulTreeBank, see (Popov et al., 2014).

[2]http://ixa2.si.ehu.es/ukb/

from different types of resources, such as WordNet or DBPedia; then it selects a context window of open class words and runs the algorithm over the graph. There is an additional module called NAF UKB[3] that can be used to run UKB with input in the NAF format[4] and to obtain output structured in the same way, only with added word sense information. For compatibility reasons, NAF UKB was used to perform the experiments reported here; the input NAF document contains in its "term" nodes lemma and POS information, which is necessary for the running of UKB. We have used the UKB default settings, i.e. a context window of 20 words that are to be disambiguated together, 30 iterations of the Personalized PageRank algorithm.

The UKB tool requires two resource files to process the input file. One of the resources is a dictionary file with all lemmas that can be possibly linked to a sense identifier. In our case, WordNet-derived relations were used as our knowledge base; consequently, the sense identifiers are WordNet IDs. For instance, a dictionary line compiled from WordNet synsets looks like this:

predicate 06316813-n:0 06316626-n:0 01017222-v:0 01017001-v:0 00931232-v:0

It comprises of a lemma followed by the sense identifiers it can be associated with. Each ID consists of eight digits followed by a hyphen and a label referring to the POS category of the word. Finally, a number following a colon indicates the frequency of the word sense, calculated on the basis of a tagged corpus. When a lemma from the dictionary has occurred in the analysis of the input text, the tool assigns all associated word senses to the word form in the context and attempts to disambiguate its meaning among them. The Bulgarian dictionary comprises of all the lemmas of words annotated with WordNet senses in the BTB. It has 8,491 lemmas mapped to 6,965 unique word senses.

The second resource file required for running the tool is the set of relations that is used to construct the knowledge graph over which Personalized PageRank is run. The distribution of the tool provides data (dictionary and relation files) for WordNet 1.7 and 3.0. Since the BTB has been annotated with word senses from WordNet 3.0, the resource files for version 3.0 have been used in our experiments. The distribution of UKB comes with a file containing the standard lexical relations defined in WordNet, such as hypernymy, meronymy, etc., as well as with a file containing relations derived on the basis of common words found in the synset glosses, which have been manually disambiguated. The format of the relations in the knowledge graph is as follows:

```
u:SynSetId01 v:SynSetId02 s:Source d:w
```

where `SynSetId01` is the identifier of the first synset in the relation, `SynSetId02` is the identifier of the second synset, `Source` is the source of the relation, and `w` is the weight of the relation in the graph. In the experiments reported in the paper, the weight of all relations is set to 0. Here is one concrete example:

```
u:01916925-n v:02673969-a s:30glc d:0
```

All the experiments use the same algorithm and the same test data. Only the knowledge graph differs in the different cases, as it is generated out of various sets of relations.

The experiments, reported in Table 1, are considered baselines for the two semantically annotated corpora: the first 49 documents of SemCor (about 1/4 of the data) and the three selected documents from BulTreeBank (about 1/4 of the data). The baseline results include WordNet relations (WN), gloss-derived relations (GL) and the combination of WN and GL — WNG:

KG	$SemCor$	BTB
WN	49.24	51.72
GL	51.48	47.02
WNG	58.83	53.82

Table 1: Experimental results when using the original knowledge graphs (WN, GL, WNG) on the two test corpora.

Some considerations are in order. It is apparent that the results for the English corpus increase monotonically, while for the Bulgarian one they are non-monotonic. Also, the combined WordNet and gloss-derived relations increase the SemCor results a lot more than the BTB ones. This probably reflects the fact that these are, after all, glosses in English and they capture better meanings encoded in the English corpus.

[3] https://github.com/asoroa/naf_ukb

[4] http://www.newsreader-project.eu/files/2013/01/techreport.pdf

4 Experiments with Semantic Relations in WordNet

The WordNet-based KG (WN) has been constructed out of the relations in the Princeton WordNet (PWN3.0). PWN3.0 groups together words in synsets, which we consider as concepts, and thus as units. The relation types possible between the different synsets are 16. In our experiments we separated the relations in WN into 16 sets of relations corresponding to the relations in PWN3.0:

1. **WN-Hyp** *(hypernymy)* **89089**. (N-N), (V-V)[5].
2. **WN-Ant** *(antonymy)* **8689**. (A-A), (N-N), (R-R), (V-V).
3. **WN-At** *(attribute relation between noun and adjective)* **886**. (N-A), (A-N).
4. **WN-Cls** *(a member of a class)* **9420**. (A-N), (N-N), (R-N), (V-N).
5. **WN-Cs** *(cause)* **192**. (V-V).
6. **WN-Der** *(derivational morphology)* **74644**. (A-N), (N-A), (N-N), (N-V).
7. **WN-Ent** *(entailment)* **408**. (V-V).
8. **WN-Ins** *(instance)* **8576**. (N-N).
9. **WN-Mm** *(member meronym)* **12293**. (N-N).
10. **WN-Mp** *(part meronym)* **9097**. (N-N).
11. **WN-Ms** *(substance meronym)* **797**. (N-N).
12. **WN-Per** *(pertains/derived from)* **8505**. (A-N), (R-A).
13. **WN-Ppl** *(participle of the verb)* **79**. (A-V).
14. **WN-Sa** *(additional information about the first word)* **3269**. (A-A), (V-V).
15. **WN-Sim** *(similar in meaning)* **21386**. (A-A).
16. **WN-Vgp** *(similar in meaning verb synsets)* **1725**. (V-V).

These classes differ in the type of semantic relations they represent, the number of relations in each class, the parts-of-speech of the words in the synsets that are connected by the relation. Obviously, isolated vertices do not play a role in the disambiguation process. Thus, if we exploit only relations between nouns, we cannot expect that the system could select appropriate senses for other parts-of-speech. Nevertheless, we performed some experiments with only some of the relations in order to have a basis for comparison with larger combinations. As a basic relation we consider the superordinate-subordinate relation (hypernymy), because it provides relations between the biggest groups of synsets: nouns and verbs. Thus, we assume that this set of relations always has to be used in the knowledge graph.

KG	$SemCor$	BTB
WN	49.24	51.72
GL	51.48	47.02
WNG	58.83	53.82
WN-Hyp	33.38	44.89
WN-Hyp+WN-Ant	**39.79**	47.55
WN-Hyp+WN-At	**35.77**	46.18
WN-Hyp+WN-Cls	**34.12**	46.11
WN-Hyp+WN-Cs	33.30	40.94
WN-Hyp+WN-Der	**38.93**	49.26
WN-Hyp+WN-Ent	33.09	44.29
WN-Hyp+WN-Ins	33.89	45.00
WN-Hyp+WN-Mm	33.42	44.61
WN-Hyp+WN-Mp	**35.60**	45.03
WN-Hyp+WN-Ms	33.32	45.00
WN-Hyp+WN-Per	**39.62**	47.29
WN-Hyp+WN-Ppl	33.29	40.57
WN-Hyp+WN-Sa	**38.07**	44.48
WN-Hyp+WN-Sim	**42.71**	44.49
WN-Hyp+WN-Vgp	33.96	41.11

Table 2: Experimental results when using the sets of relations from the WordNet knowledge graph on the two test corpora.

In Table 2 we present the results for combinations between the hypernymy relation and all other relations. The biggest improvement is observed for the combination **WN-Hyp+WN-Sim**. It shows 9 % of improvement over the **WN-Hyp** relation alone. In our view, the great difference is due to the different coverage of the relations over the synsets in WordNet. Hypernymy relation covers only noun and verb synsets, but not adjective and adverb synsets. Thus, a KG based only on hypernymy relation does not provide any knowledge about adjectives and adverbs. Additionally, it does not contain any knowledge about the interactions between verbs and nouns. The relations that improve over hypernymy ones in fact introduce knowledge about adjectives or interaction

[5]Here we present the combination of synsets in each relation as parts-of-speech. The parts-of-speech are: A — adjective, N — noun, R — adverb, and V — verb. Also we present the number of links for the relation in WordNet.

across parts-of-speech. We have performed some more experiments in order to check whether we could exclude some relations without considerable loss. For instance, the combination of the following eight sets: **WN-Hyp + WN-Ant + WN-Der + WN-Per + WN-Sa + WN-Sim + WN-Mp + WN-Cls**, gives accuracy of **49.10** % on the $SemCor$ test corpus, which is **0.14** % less than the accuracy obtained with the whole KG of WordNet. The results also show the differences between the corpora. BTB seems more compact with respect to sub-domains, while SemCor introduces a big variety of sub-domains. Also, it is mainly annotated with noun and verb synsets. Thus, the impact of the relations is different from the impact they have over the SemCor corpus.

The general conclusion from these experiments is that the addition of relations to the knowledge graph does not contribute monotonically to the accuracy of the KWSD. It shows that some of the relations in the original graph lower the accuracy.

In the next sections we report only experiments performed over SemCor corpus for brevity.

4.1 Inference over WordNet Relations

Under inference in our experiments we consider the application of rules, given relations in the knowledge graph, which produce new relations to be added to the knowledge graph. In this section we consider some rules applicable to the relations from WordNet. Having in mind that WordNet is not a fully formalized lexical database, we cannot expect that the inferences proposed below are always correct. The main inference rule is the hypernymy hierarchy inheritance: if some relation includes a noun as an argument, then the hyponyms of the noun also could be arguments in the relation. The situation is similar for verbs. Sometimes the appropriate inference includes their hypernyms.

1. **WN-Hyp**. The hypernymy relation is transitive. Thus, we could construct its transitive closure: if **doctor** is a hypernym of **surgeon** and **professional** is a hypernym of **doctor** then **professional** is a hypernym of **surgeon**. Similarly, for the verb hierarchy.

2. **WN-Ant**. Antonymy relations between adjectives and adverbs cannot participate in the inference, because there is no support in WordNet. For nouns and verbs it is possible, if we assume that the antonymy relation means that corresponding synsets are disjoint. The disjointedness is preserved by the hyponymy relation: if we have two disjoint concepts, then their subconcepts are also disjoint. For example, **man** and **woman** do not have common instances. Then we could infer that **man** and **girl** are disjoint.

3. **WN-At**. The attributes of a noun usually can be inherited by its hyponyms. For example, **measure** as a quantity of something has attributes — **standard** and **nonstandard**. These attributes can be inherited by all kinds of measures like **time interval** and others.

4. **WN-Cls**. The general understanding of the relation *a member of a class* is that each hyponym of the **member** could be a member of each of the hypernyms of the **class**. For instance, **desktop publishing** is a member of **computer science** as a branch, but also it is a branch of **engineering**, which is a hypernym of computer science.

5. **WN-Cs**. The *cause* relation between verbs naturally allows for inference on both arguments — each hyponym of the first argument could be a cause for each hypernym of the second argument. The sets resulting from the inference on the first and second arguments are denoted with WN-Cs1stVerbInfer and WN-Cs2ndVerbInfer.

6. **WN-Der**. The derivational relation is quite diverse, connecting adjectives and nouns, nouns and nouns, and nouns and verbs. We consider this relation as denoting an event or a state in which the noun determines a participant of the event or a state. Thus, a noun can be substituted with its hyponyms, and a verb can be substituted with its hypernyms.

7. **WN-Ent**. If a verb entails another verb, then we assume that each hyponym of the first verb entails each hypernym of the second verb. The sets resulting from the inference on the first and second arguments are denoted with WN-Ent1stVerbInfer and WN-Ent2ndVerbInfer.

8. **WN-Ins**. An instance of a class is an instance of its super classes. Thus, we perform substitution of the second noun with its hypernyms.

9. **WN-Mm**. Each hyponym of a member of a set is a member of each hypernym of the set.

10. **WN-Mp**. The transitive closure over the part meronym relation is a feasible inference rule. In these experiments we do not perform it.
11. **WN-Ms**. Substitution with hyponyms of the substance noun is a feasible inference rule. Similarly to the previous relation, in these experiments we do not perform it.
12. **WN-Per**. Similarly to the derivational relation, we perform substitution with hyponyms on the noun synset.
13. **WN-Ppl**. We do not perform any inference for this relation.
14. **WN-Sa**. The additional information about the first word can be inherited by its hyponyms.
15. **WN-Sim**. We do not perform any inference for this relation.
16. **WN-Vgp**. Because the definition "verb synsets that are similar in meaning" allows for very wide interpretation, we do not perform any inference on this relation.

Some of the above inferences produce a huge amount of new relations, which prevents us from effectively experimenting with them. We have used the inference rules only partially. These experiments have been performed only on the SemCor test corpus. We consider only combinations in which the knowledge graphs of the original WordNet and the Extended WordNet are included as a basis. Table 3 presents some of the results. There are few cases in which the inferred new relations add accuracy above the baselines (more substantial for the combination WN+WN-HypInfer). In most of the cases, however, the additional relations decrease the accuracy. For the WordNet relations, these improvement-inducing combinations include inference over the hypernymy relation (54.15) and inference over the second verb of the cause relation (49.25). For the Extended WordNet relations, one of the sets that outperforms the baseline includes inference over hypernymy, but the other one includes inference over antonymy.

KG	*SemCor*
WN+WN-HypInfer	**54.15**
WN+WN-AntInfer	48.49
WN+WN-ClsInfer	48.48
WN+WN-Cs1stVerbInfer	49.21
WN+WN-Cs2ndVerbInfer	**49.25**
WN+WN-DerNAInfer	48.49
WN+WN-DerNNInfer	47.82
WN+WN-DerNVInfer	47.79
WN+WN-DerVNInfer	48.69
WN+WN-Ent1stVerbInfer	49.21
WN+WN-Ent2ndVerbInfer	49.21
WN+WN-InsInfer	48.89
WNG+WN-HypInfer	**58.93**
WNG+WN-AntInfer	**59.08**
WNG+WN-ClsInfer	57.66
WNG+WN-Cs1stVerbInfer	58.85
WNG+WN-Cs2ndVerbInfer	58.80
WNG+WN-DerNAInfer	58.41
WNG+WN-DerNNInfer	58.62
WNG+WN-DerNVInfer	55.68
WNG+WN-DerVNInfer	58.89
WNG+WN-Ent1stVerbInfer	58.84
WNG+WN-Ent2ndVerbInfer	58.79
WNG+WN-InsInfer	58.23

Table 3: Experimental results when using some of the inferred sets of relations. The results that are above the baselines from Table 1 are bolded.

5 Experiments with Semantic Relations in Extended WordNet

The Extended WordNet (Mihalcea and Moldovan, 2001) is constructed on the basis of analyses of the glosses of the synsets. During this analysis, the open class words were annotated with word synsets from PWN3.0. For example, the synset {stony coral, madrepore, madriporian coral} — 01916925-n, is defined by "corals having calcareous skeletons aggregations of which form reefs and islands." After the analysis, the following synsets are selected: 02673969-a — *calcareous*, 01917882-n — *mushroom coral*, 05585383-n — *skeleton*, 07951464-n — *aggregation*, 09316454-n — *island*, 09406793-n — *reef*, and 02621395-v — *form*. Each of these synsets is related to the synset to which the gloss belongs to:

```
u:01916925-n v:02673969-a s:30glc d:0
u:01916925-n v:01917882-n s:30glc d:0
u:01916925-n v:05585383-n s:30glc d:0
u:01916925-n v:07951464-n s:30glc d:0
u:01916925-n v:09316454-n s:30glc d:0
u:01916925-n v:09406793-n s:30glc d:0
u:01916925-n v:02621395-v s:30glc d:0
```

The first division of the relations in WNG into groups is on the basis of the parts of speech of the main synset. The four sets are: WNG-A (first synset is for adjectives), WNG-N (first synset is

for nouns), WNG-R (first synset is for adverbs), and WNG-V (first synset is for verbs).

KG	$SemCor$
WN+WNG-A	52.80
WN+WNG-N	**56.85**
WN+WNG-R	51.56
WN+WNG-V	52.61

Table 4: Experimental results when using the sets of relations from the XWN knowledge graph.

In Table 4 we present the impact of each of these sets of relations on the knowledge graph of WN. As can be seen, each set adds accuracy above the baseline of WN. When comparing the inferred relations (Table 3) and the WNG sets, it can be observed that the set WNG-N improves accuracy even over the WN-HypInfer set.

Additionally, each of the groups — WNG-A, WNG-N, WNG-R, and WNG-V — was divided into four subgroups on the basis of the part of speech of the second synset in the relation. Thus, we created 16 new sets: WNG-AA, WNG-AN, ..., WNG-VV. After experimenting with each of them, we arrived at the following combination: WN, WNG-AN, WNG-NN, WNG-RN, and WNG-VN. The accuracy for this combination is **56.99**, which is higher than the results for each individual set.

6 Syntax-based Relations

As was mentioned above, in our experiments we have also used semantic relations from a syntactically annotated corpus. To achieve this, we parsed SemCor with a dependency parser included in IXA pipeline. Then we divided the corpus in a proportion one-to-three: first part comprises of 49 documents (from br-a01 to br-f44) and it was used as a test set in the experiments reported here. The rest of the documents formed the training set from which the new relations were extracted. First, we defined patterns of dependency relations. For example, we used patterns like the following: $s_1 subj s_2$, which defines a relation between a noun synset s_1 and a verb synset s_2; $s_1 mods_2$, which defines a relation between an adjective synset s_1 and a noun synset s_2; $s_1 modxpobj s_2$, which defines a relation between a noun synset s_1 and a noun synset s_2; etc. We extracted the following sets of relations: SC-AA, SC-AN, SC-AV, SC-NN, SC-NV, SC-RA, SC-RN, SC-RR, SC-RV, SC-VN, SC-VV, where the suffixes — AA, AN, AV, etc. — denote the parts of speech of the related synsets. The results from the experiments performed are presented in Table 5. As can be seen, many of the extracted new sets increase the accuracy above the baseline for the original knowledge graph — WNG.

KG	$SemCor$
WNG+SC-AA	**59.08**
WNG+SC-AN	**59.13**
WNG+SC-AV	**59.28**
WNG+SC-NN	58.69
WNG+SC-NV	**59.20**
WNG+SC-RA	**59.35**
WNG+SC-RN	58.77
WNG+SC-RR	58.92
WNG+SC-RV	**59.24**
WNG+SC-VN	**58.92**
WNG+SC-VV	**59.09**

Table 5: Results from experiments using the sets of relations from syntax.

We have combined most of these sets in joint combinations. The combination of all the sets with the original knowledge graph: WNG, SC-AA, SC-AN, SC-AV, SC-NN, SC-NV, SC-RA, SC-RN, SC-RR, SC-RV, SC-VN, SC-VV gives accuracy of **60.13**. The best combination is WNG, SC-AA, SC-AN, SC-AV, SC-NV, SC-RA, SC-RN, SC-RR, SC-RV, SC-VN, SC-VV. The accuracy for this combination is **60.14**.

We also preformed inference over these sets of relations using hypernymy and hyponymy hierarchies for nouns and verbs. The best result was achieved for the combination WNG, SC-AA, SC-AN, SC-AV, SC-NV, SC-RA, SC-RR, SC-RV, SC-VN, SC-VV, WN-HypInfer, WN-AntInfer. Its accuracy is **60.42**. This result is 1.5 % higher than the baseline for the original knowledge graph. This improvement is statistically significant.

7 Conclusion

In this paper we have evaluated the performance of different relations encoded in the knowledge graph, for the purposes of the knowledge-based Word Sense Disambiguation. Each of the sets of relations reflects an important linguistic piece of knowledge. Thus, each of them is important for the description of languages. However, from the

point of view of knowledge-based WSD each of these relations, as well as their various combinations, seem to have a different impact on the performance of the task.

The results from the experiments show that the addition of whole sets of relations might have a positive or a negative effect. In our view, at least two factors are of importance: (i) the number of relations assigned to each synset. Following Zipf's law, we can conclude that the distribution of relations per synset is very uneven. For many synsets there is not sufficient information present in the context, in order for a good decision to be taken. For many ambiguous words the context provides no information for disambiguation, and the decision is taken arbitrarily. (ii) The second factor is that the inference rules applied to the explicit relations do not produce the expected improvement. This might be due to the fact that WordNet is not the right place to store the inference information. Our expectations about the positive influence of inference are not always realized in practice. For intance, we expected to get relations between events and their participants from the derivational relations, but this was often not the case. If we take the verb "to kiss" and the derived noun "kisser", we would expect that "kisser" is a more general synset than the synsets for any specific kisser. But the synset for "kisser" had no single hyponym in WordNet. The gloss is *someone who kisses* and it determines the connection from "kisser" to "someone" who is the most general agent of the verb "to kiss". The connection is stated in XWN via the gloss of the noun "kisser". But for this configuration of relations in the original graph there is no inference rule defined. It seems that the systemic and monotonic knowledge that is needed for WSD and other NLP tasks is not always considered interesting enough to encode in various lexical resources.

Our future goals are the following: (i) the application of more complicated inference rules; (ii) the modification of relations per synset in order to ensure enough disambiguation relations. We expect these modifications of the relations to be performed via machine learning techniques over the contexts of the words in large corpora.

The number of synsets in the knowledge graph is 136,334, thus the possible links between them are in total 18,586,823,222. At the same time, the number of the actual links in the biggest graphs used in practice, is less than 5 million, which is only 0.027 % of all possible combinations. Probably we need much more than 0.027 % of the links in order to capture all the available, and also all the necessary, knowledge for WSD. However, in such cases a faster algorithm must be employed.

Acknowledgements

This research has received support by the EC's FP7 (FP7/2007-2013) project under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches."

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Comp. Ling.*, 40(1):57–84.

Eneko Agirre, Ander Barrena, and Aitor Soroa. 2015. Studying the wikipedia hyperlink graph for relatedness and disambiguation. *arXiv preprint arXiv:1503.01655*.

Sergey Brin and Lawrence Page. 2012. The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Pawe Kdzia, Maciej Piasecki, Jan Koco, and Agnieszka Indyka-Piasecka. 2014. Distributionally extended network-based word sense disambiguation in semantic clustering of polish texts. {*IERI*} *Procedia*, 10:38 – 44. International Conference on Future Information Engineering (FIE 2014).

Rada Mihalcea and Dan I. Moldovan. 2001. extended wordnet: progress report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT '93*, pages 303–308.

Alexander Popov, Stanislava Kancheva, Svetlomira Manova, Ivaylo Radev, Kiril Simov, and Petya Osenova. 2014. The sense annotation of bultreebank. In *Proceedings of TLT13*, pages 127–136.

Kiril Simov, Alexander Popov, and Petya Osenova. 2015. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. In *Proc. of RANLP 2015.*, pages 596–603.

Detection of Compound Nouns and Light Verb Constructions using IndoWordNet

Dhirendra Singh Sudha Bhingardive Pushpak Bhattacharyya

Department of Computer Science and Engineering,
Indian Institute of Technology Bombay.
{dhirendra,sudha,pb}@cse.iitb.ac.in

Abstract

Detection of MultiWord Expressions (MWEs) is one of the fundamental problems in Natural Language Processing. In this paper, we focus on two categories of MWEs - *Compound Nouns* and *Light Verb Constructions*. These two categories can be tackled using knowledge bases, rather than pure statistics. We investigate usability of IndoWordNet for the detection of MWEs. Our *IndoWordNet based approach* uses semantic and ontological features of words that can be extracted from IndoWordNet. This approach has been tested on Indian languages *viz.,* Assamese, Bengali, Hindi, Konkani, Marathi, Odia and Punjabi. Results show that ontological features are found to be very useful for the detection of *light verb constructions*, while use of semantic properties for the detection of *compound nouns* is found to be satisfactory. This approach can be easily adapted by other Indian languages. Detected MWEs can be interpolated into WordNets as they help in representing semantic knowledge.

1 Introduction

MultiWord Expressions or MWEs can be described as idiosyncratic interpretations that crosses word boundaries or spaces (Sag et al., 2002). MWE is formed by atleast two words which are syntactically and/or semantically idiosyncratic in nature. For example, *swimming pool*, *telephone booth*, *strong coffee*, *pay attention*, *fast food*, etc. are some of the MWEs in English, while धन दौलत (*Dhana daulata*, wealth), वादा करना (*vaadaa karanaa*, to promise), मार डालना (*maara Daalanaa*, to kill), धीरे धीरे (*dhiire dhiire*, slowly), *etc.* are some of the MWEs in Hindi. In past, ample number of approaches have been proposed in literature for the detection of MWEs (Calzolari et al., 2002),(Baldwin et al., 2003), (Guevara, 2010), (Al-Haj and Wintner, 2010), (Tsvetkov and Wintner, 2012). However, for Indian languages, many researchers have proposed statistical and rule based approaches (Sinha, 2009), (Kunchukuttan and Damani, 2008), (Chakrabarti et al., 2008), (Mukerjee et al., 2006), (Sinha, 2011), (Singh et al., 2012), (Sriram et al., 2007).

This paper focuses on Indian languages *viz.,* Assamese, Bengali, Hindi, Konkani, Marathi, Odia and Punjabi for the detection of MWEs. These languages are part of the IndoWordNet[1]. To the best of our knowledge, the IndoWordNet based approach is being used for the first time for detecting MWEs. This approach is restricted for two categories of MWEs: *compound nouns* (Noun+Noun) and *light verb constructions* (Noun+Verb, Adjective+Verb, Verb+Verb). Semantic features of words are used for detecting *compound nouns*, while ontological features are used for detecting *light verb constructions*. The motivation behind this work is that,

- If we add suitable amount of MWEs in WordNet, its coverage will be increased in terms of vocabulary and linguistic phenomenon.
- Improper handling of MWEs is one of the

[1]IndoWordNet is available in following Indian languages: Assamese, Bodo, Bengali, English, Gujarati, Hindi, Kashmiri, Konkani, Kannada, Malayalam, Manipuri, Marathi, Nepali, Punjabi, Sanskrit, Tamil, Telugu and Urdu. These languages cover three different language families, Indo-Aryan, Sino-Tibetan and Dravidian. http://www.cfilt.iitb.ac.in/indowordnet/

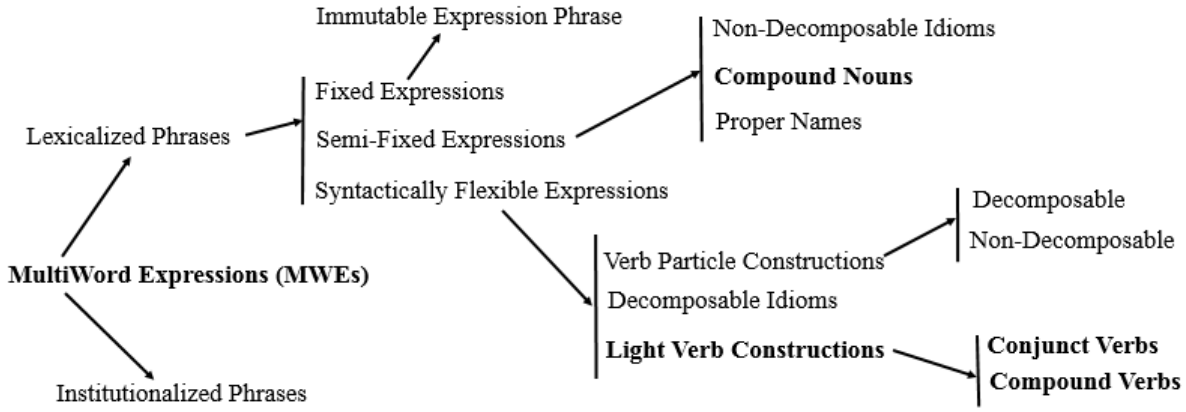


Figure 1: Classification of MWEs

major sources of error in various NLP applications. Hence, correct detection of MWEs will show improvement in performance of these applications, as reported by Finlayson et al. (2011) for word sense disambiguation, Ren et al. (2009) and Bouamor et al. (2011) for machine translation, etc.

The roadmap of the paper is as follows. Section 2 covers the classification of MWEs. The IndoWordNet based approach is explained in Section 3. Section 4 details the experimental setup. Results are presented in section 5 and discussed in section 6. Related work is given in section 7, followed by conclusion and future work.

2 MWEs Classification

MWEs are classified based on their lexical and semantic characteristics (Sag et al., 2002). This has been further studied from Indian language perspective and expanded as shown in Figure 1. As we can see in figure 1, we modified the Sag et al., (2002) classification by adding Light Verb Constructions and its further classification which is needed for Indian languages. MWEs are classified into two broad categories. They are Lexicalized Phrases and Institutional Phrases. The meaning of lexicalized phrases cannot be construed from its individual units that make up the phrase, as they exhibit syntactic and/or semantic idiosyncrasy. On the other hand, the meaning of institutional phrases can be construed from its individual units that make up the phrase. However, they exhibit statistical idiosyncrasy. Institutional phrases are not in the scope of this paper. Lexicalized phrases are further classified into three sub-classes *viz.,* Fixed, Semi-fixed and Syntactically flexible expressions.

In this paper, we focus on *compound nouns* and *light verb constructions* which fall under the semi-fixed and syntactically flexible categories respectively.

2.1 Compound Nouns

Compound Nouns (CNs) are syntactically-unalterable units that inflect for number. A word-pair forms CN if its meaning cannot be composed from the meanings of its constituent words. CNs are formed by either Noun+Noun or Adj+Noun word combinations. For example, पेड़ पौधे (*peda paudhe*, flora), बाग बगीचा (*baaga bagiichaa* , garden), काला धन (*kaalaa dhana* , black money), *etc.*

2.2 Light Verb Constructions

Light Verb Constructions (LVCs) show high idiosyncratic constructions with nouns. It is difficult to predict which light verb chooses which noun and why the light verb cannot be substituted with another. LVCs are further classified into Conjunct Verbs (CjVs) and Compound Verbs (CpVs). *CjVs* are formed by Noun+Verb and Adj+Verb word combinations, while *CpVs* are formed by Verb+Verb

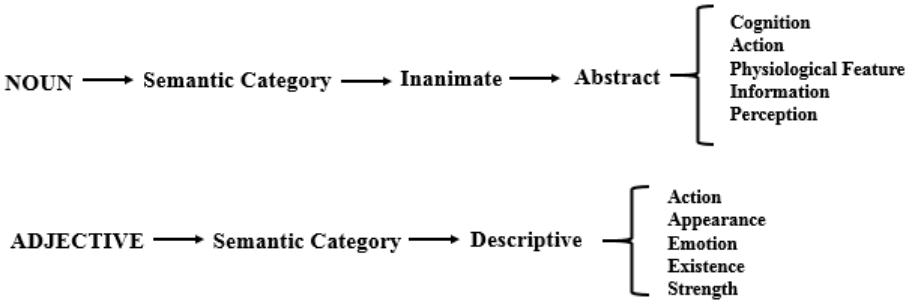


Figure 2: Noun and Adjective ontological features needed to form Conjunct Verbs

Figure 3: Verb ontological features needed to form Compound Verbs

word combinations. Examples of *CjVs* are गुजर जाना (*gujara jaanaa*, passed away), काम करना (*kaama karanaa*, to work), प्यार करना (*pyaara karanaa*, to love), etc. and examples of *CpVs* are भाग जाना (*bhaaga jaanaa*, run away), उठ जाना (*uTha jaanaa*, to wake up), खा लेना (*khaa lenaa*, to eat), etc.

3 IndoWordNet Based Approach

Our IndoWordNet based approach uses various semantic and ontological features from the IndoWordNet. The semantic features are used for *CN* detection while ontological features are used for *LVC* detection. Now, we explain the IndoWordNet based approach for each of these categories.

3.1 Detection of Compound Nouns

The semantic features of words such as *synonyms*, *definition/gloss*, *example sentence*, *hypernyms*, *antonyms, etc.* are used for detection of *CNs*.

The *bag of words* (*BOW*) for a word w_i is created using the semantic features of IndoWordNet, as follows.

$$BOW(w_i) = \{x | x \in WordNetFeatures(w_i)\}$$

where, $WordNetFeatures(w_i)$ contains all content words from *synonyms*, *gloss*, *example(s)*, *hypernyms*, *hyponyms*, *meronyms*, *antonyms* with respect to the word w_i. We considered only one level of hierarchy for extracting these semantic features.

Consider a word-pair w_1w_2 to be detected as a MWE. As per the IndoWordNet based approach, the given pair can be treated as *compound noun* MWEs when any one of the following condition holds -

- if $w_1 \in BOW(w_2)$, then w_1w_2 is a *CN*
- if $w_2 \in BOW(w_1)$, then w_1w_2 is a *CN*

For instance, consider a word-pair in Hindi, धन दौलत (*dhanaa daulata, wealth*). The *BOWs* for *dhana* and *daulata* are as follows,

$$BOW(dhana) = \{paisaa, daulata, vaibhava, ..\}$$

$$BOW(daulata) = \{sampatti, laxmi, dhana, ...\}$$

Since, $dhana \in BOW(daulata)$, the word-pair *dhana daulat* is considered as a *CN*.

3.2 Detection of Light Verb Constructions

The ontological features of words such as *abstract, inanimate, action, information, etc.* (refer figure 4) are used for detection of LVCs. There are two types of LVCs, Conjunct Verbs (CjVs) and Compound Verbs (CpVs).

3.2.1 Conjunct Verbs

As mentioned earlier, conjunct verbs are formed by Noun+Verb and Adj+Verb word combinations. However, it is very difficult to predict which type of nouns or adjectives form *CjVs*. Previous approaches tried to detect

such nouns or adjectives based on their statistical collocation with restricted sets of verbs (most frequently used, manually selected, etc.) (Sidhu et al., 2010). This limitation results in less coverage at *CjV* detection.

We claim that whether a noun or an adjective forms *CjVs* depends on its ontological properties. Figure 2 shows some ontological properties of nouns and adjectives that are available in IndoWordNet and needed to form *CjVs*. This removes the dependence on the restricted set of verbs, thereby increasing the upper bound of coverage that we can achieve. Algorithm 1 details the detection of *CjVs*.

Algorithm 1 Conjunct Verb Detection

```
1: procedure CJV–DETECTION (w1,w2)
2:     if w1 is Noun and w2 is Verb then
3:         if w1 is abstract Noun then
4:             print "CjV detected"
5:         end if
6:     end if
7:
8:     if w1 is Adj and w2 is Verb then
9:         if w1 is descriptive Adj then
10:            print "CjV detected"
11:        end if
12:    end if
13: end procedure
```

3.2.2 Compound Verbs

As mentioned earlier, compound verbs are formed by Verb+Verb word combinations. The first verb gives lexical information whereas the second verb provides grammatical information about the expression. Just as in the case of *CjVs*, formation of *CpVs* also depends on the ontological properties of the constituent verbs. Figure 3 shows some ontological properties of verbs that are available in IndoWordNet and needed to form *CpVs*. Algorithm 2 details the detection of *CpVs*.

4 Experiments

We performed experiments on some Indian languages *viz.,* Hindi, Marathi, Bengali, Punjabi, Konkani, Odia, Assamese for the detection of *compound nouns* and *conjunct verbs*. However, for *compound verb* detection, we performed experiments only on Hindi and Marathi due to unavailability of gold data for other languages.

The gold data for these experiments is created by automatically extracting Noun+Noun, Noun+Verb, Adj+Verb and Verb+Verb word-pair combinations. These word-pairs are extracted from the generic domain in-house corpus. Out of these word-pairs, 1000 Noun+Noun word-pairs are detected as *CNs* for each of the seven languages mentioned above, while 399 and 504 Verb+Verb word-pairs are detected as *CpVs* for Marathi and Hindi respectively. Also, 457, 404, 797, 1017, 879, 832, 703 Noun+Verb and 577, 502, 303, 307, 269, 368, 259 Adj+Verb word-pairs are detected as *CjVs* for Hindi, Marathi, Bengali, Punjabi, Konkani, Odia, Assamese languages respectively. Three lexicographers were engaged in this activity and the inter-annotator agreement is found to be 0.8.

Algorithm 2 Compound Verb Detection

```
1: procedure CPV–DETECTION (w1,w2)
2:     if w1 is action verb then
3:         if w2 is action verb or
4:             w2 is occurrence verb then
5:             print "CpV detected"
6:         end if
7:     end if
8:
9:     if w1 is occurrence verb then
10:        if w2 is action verb then
11:            print "CpV detected"
12:        end if
13:    end if
14: end procedure
```

5 Results

In this section, results of the experiments are presented and discussed in detail. Table 1 shows the results obtained for the detection of *CNs*, while Table 2 and Table 3 show the results obtained for the detection of *CjVs* and *CpVs* respectively. It has been observed that results of *CN* detection are found to be considerably good only for Marathi as compared to other languages. However, the results of *CjV* and *CpV* detection are found to be promising for languages under consideration. Hence, we can say that, IndoWordNet based approach using ontological properties are found to be very effective for the detection of *light verb constructions* such as *CpVs* and *CjVs*.

6 Discussions

As we have observed that the results of *CN* detection are found to be unsatisfactory for languages other than Marathi. This may be be-

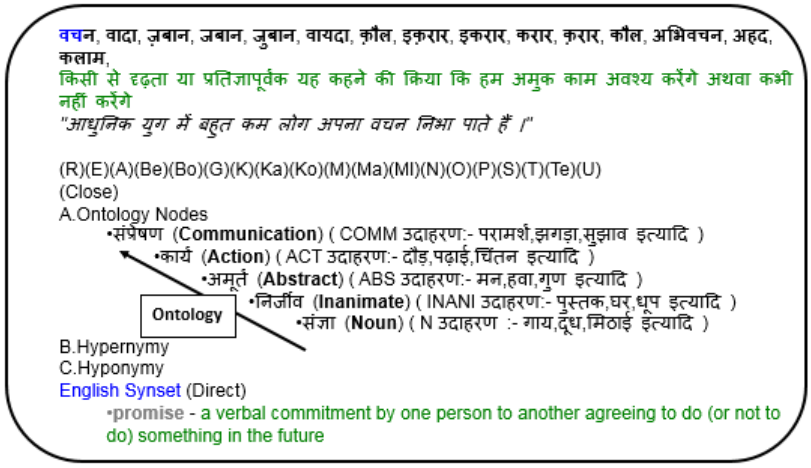


Figure 4: IndoWordNet ontological properties for a Hindi word 'vachanaa' (promise)

	Compound Nouns (CNs)	
	Total pairs(N+N)	**F-score**
Hindi	1000	0.58
Marathi	1000	0.72
Bengali	1000	0.53
Punjabi	1000	0.43
Konkani	1000	0.52
Odia	1000	0.38
Assamese	1000	0.40

Table 1: Results of Compound Noun Detection

	Compound Verbs (CpVs)	
	Total pairs(V+V)	**F-score**
Hindi	399	0.99
Marathi	504	0.88

Table 2: Results of Compound Verb Detection

	Conjunct Verbs (CjVs)			
	Total pairs(N+V)	**F-score**	**Total pairs(Adj+V)**	**F-score**
Hindi	457	0.87	577	0.89
Marathi	404	0.86	502	0.88
Bengali	797	0.87	303	0.92
Punjabi	1017	0.8	307	0.9
Konkani	879	0.84	269	0.95
Odia	832	0.85	368	0.91
Assamese	703	0.84	259	0.94

Table 3: Results of Conjunct Verb Detection

cause, our IndoWordNet based approach completely depends on the semantic properties of words and do not rely on the statistical co-occurrence. Also, in IndoWordNet, there are some word-pairs which are not semantically related but can form *compound nouns* due to their high statistical co-occurrence in the corpus. For example, काला धन (*kaalaa dhana*, black money) is a *CN* even though काला (*kaalaa*,black) and धन (*dhana*, money) do not exhibit any semantic relation in the IndoWordNet.

Results of *CjV* detection for Noun+Verb and Adj+Verb combinations are found to be promising. This may be because, our IndoWordNet based approach uses ontological properties of words wherein coverage of nouns and adjectives is high in IndoWordNet. While, the results of the detection of *CpVs* are found to be almost 100% for Hindi and 88% for Marathi. This also used ontological properties of words. Hence, we can say that IndoWordNet based approach is very useful for the detection of *CjVs* and *CpVs*.

7 Related Work

Most of the proposed approaches for the detection of multiword expressions are statistical in nature. They are based on association methods (Church and Hanks, 1990), deep linguistics based methods (Bansal et al., 2014), word embeddings based methods (Salehi et al., 2015), *etc.* The detection of MWEs for Indian languages is not explored much by researchers due to the reasons such as unavailability of gold data (Reddy, 2011), unstructured classification of MWEs, improper universal theory, *etc.* In literature, Gayen and Sarkar et al. (2013) used Random Forest approach for Compound Noun detection for Bengali language. Sriram et al. (2007) used a classification based approach for extracting Noun-Verb collocations for Hindi language. Mukerjee et al. (2006) used parallel corpus alignment and Part-Of-Speech tag projection to extract complex predicates. However, our IndoWordNet based approach uses ontological and semantic features of words to detect MWEs. The focus is restricted for the detection of *compound nouns* and *light verb constructions.*

8 Conclusion and Future Work

Detection of MultiWord expressions is the fundamental problem and a challenging task in the area of NLP. To address this problem, an IndoWordNet based approach is proposed in this paper. The focus is restricted to the detection of *compound nouns* and *light verb constructions.* Semantic features of words from IndoWordNet are used for the detection of *compound nouns*, while ontological features of words are used for the detection of *light verb constructions.* The IndoWordnet based approach is tested on some Indian languages *viz.,* Assamese, Bengali, Hindi, Konkani, Marathi, Odia, punjabi. It has been observed that our approach gives encouraging results for the detection of light verb constructions as compared to compound nouns. In future, the detected MWEs can be incorporated in IndoWordNet as they can help to represent the lexical knowledge. This approach can be used in NLP applications *viz.,* word sense disambiguation, machine translation, information retrieval, question answering, sentiment analysis, *etc.* It can be implemented and tested for other Indian languages.

References

Hassan Al-Haj and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 10–18. Association for Computational Linguistics.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL.*

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2011. Improved statistical machine translation using multiword expressions. International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011).

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine

MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L02-1259.

Debasri Chakrabarti, Hemang Mandalia, Ritwik Priya, Vaijayanthi M Sarma, and Pushpak Bhattacharyya. 2008. Hindi compound verbs and their automatic extraction. In *COLING (Posters)*, pages 27–30.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Mark Alan Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24. Association for Computational Linguistics.

Vivekananda Gayen and Kamal Sarkar. 2013. Automatic identification of Bengali noun-noun compounds using random forest. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 64–72, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Anoop Kunchukuttan and Om Prakash Damani. 2008. A system for compound noun multiword expression extraction for hindi. In *6th International. Conference on Natural Language Processing*, pages 20–29.

Amitabha Mukerjee, Ankit Soni, and Achla M Raina. 2006. Detecting complex predicates in hindi using pos projection across parallel corpora. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 28–35. Association for Computational Linguistics.

Siva Reddy. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 977–983.

Brahmaleen K Sidhu, Arjan Singh, and Vishal Goyal. 2010. Identification of proverbs in hindi text corpus and their translation into punjabi. *Journal of Computer Science and Engineering*, 2(1):32–37.

Smriti Singh, Om P Damani, and Vaijayanthi M Sarma. 2012. Noun group and verb group identification for hindi. In *COLING*, pages 2491–2506. Citeseer.

R Mahesh K Sinha. 2009. Mining complex predicates in hindi using a parallel hindi-english corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 40–46. Association for Computational Linguistics.

R Mahesh K Sinha. 2011. Stepwise mining of multi-word expressions in hindi. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 110–115. Association for Computational Linguistics.

V Sriram, Preeti Agrawal, and Aravind K Joshi. 2007. Relative compositionality of noun verb multi-word expressions in hindi. In *published in Proceedings of International Conference on Natural Language Processing (ICON)-2005, Kanpur*.

Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Mapping it differently: A solution to the linking challenges

Meghna Singh, Rajita Shukla*, Jaya Saraswati, Laxmi Kashyap,
Diptesh Kanojia and Pushpak Bhattacharyya
Center for Indian Language Technology,
Department of Computer Science and Engineering,
Indian Institute of Technology, Bombay,
Mumbai.
{meghnak, jayas, yupu, diptesh, pb}@cse.iitb.ac.in
*rajita.shukla38@gmail.com

Abstract

This paper reports the work of creating bilingual mappings in English for certain synsets of Hindi wordnet, the need for doing this, the methods adopted and the tools created for the task. Hindi wordnet, which forms the foundation for other Indian language wordnets, has been linked to the English WordNet. To maximize linkages, an important strategy of using direct and hypernymy linkages has been followed. However, the hypernymy linkages were found to be inadequate in certain cases and posed a challenge due to sense granularity of language. Thus, the idea of creating bilingual mappings was adopted as a solution. A bilingual mapping means a linkage between a concept in two different languages, with the help of translation and/or transliteration. Such mappings retain meaningful representations, while capturing semantic similarity at the same time. This has also proven to be a great enhancement of Hindi wordnet and can be a crucial resource for multilingual applications in natural language processing, including machine translation and cross language information retrieval.

1 Introduction

Wordnets are online lexical resources which are easily accessible, free to use, and fairly accurate. They play a dominant role in the field of text processing applications, such as machine translation, information extraction, information retrieval and natural language understanding systems. Among the Indian language wordnets, Hindi wordnet[1] was the first one to come into existence from the year 2000 onwards. It was inspired by the English WordNet[2] which contains nouns, verbs, adjectives and adverbs organized into synonym sets called synsets, each representing one underlying lexical concept (Fellbaum, 1998). Different relations like hypernymy, hyponymy, etc. link the synonym sets to each other. Soon, other Indian language wordnets started getting created, with Hindi wordnet as the pivot, inheriting all the relations. Hindi wordnet is linked to the English WordNet and the other Indian language wordnets are linked to Hindi wordnet, in turn. This has led to the creation of a wide grid of shared concepts, thus creating an important knowledge base for the NLP community. To achieve maximum linkage between the English and Hindi wordnets, the policy of having direct and hypernymy linkage (Saraswati et al, 2010) has been adopted. However, it was observed that the hypernymy linkage does not lead to an accurate word and concept in all cases. Thus, to overcome this challenge the idea of creating complimentary bilingual mappings in English came up.

The roadmap of the paper is as follows: Section 2 presents a comprehensive view of related work done earlier, while in Section 3 the need for this approach is discussed. Section 4 deals with the methodology. Section 5 presents the qualitative analysis of the challenges encountered and the solutions put forth. In Section 6 the interface used for this task is discussed. The overall statistics is given in Section 7 and Section 8 mentions some of the words sent to the English WordNet. Section 9 winds up the paper with the conclusion and future work.

[1]http://www.cfilt.iitb.ac.in/wordnet/webhwn/
[2] http://wordnet.princeton.edu/

2 Related Work

Wordnets have been built for around 100 different languages. Efforts towards mapping synsets across wordnets have been going on for a while in various parts of the world. Many languages have been trying to link their wordnets to the English WordNet for a universal set of linked concepts, enabling translation on the lexical level as well as cross-lingual WSD and other applications. Usually, a concept in one wordnet is directly linked to a similar one in the English WordNet, but in many cases, some kind of mapping is required. CoreNet (Kang et al, 2010) has made one such effort. It is a multilingual lexico-semantic network constructed in KAIST for the Korean, Chinese, and Japanese languages and many of its words/concepts have been linked to the English WordNet. To ameliorate translation problems between CoreNet (mostly written in Korean) and the English WordNet and to enhance recall of WordNet equivalents, the two are partially indirectly linked via KorLex (Yoon et al., 2009) to the English WordNet. When this kind of indirect mapping is also not available, the concepts in CoreNet are manually mapped to the concepts in the English WordNet. In EuroWordNet (Vossen et. al., 1999)[3], multilinguality is achieved by storing the language-specific wordnets in a central lexical database in which equivalent word meanings across the languages are linked to a so-called Inter-Lingual-Index (ILI) to get a comprehensive conceptual match of concepts across languages. Another effort towards wordnet linking can be found in the MultiWordNet (Pianta et. al., 2002)[4] which aligns the Italian and the English language wordnets.

Another such effort to create a multilingual wordnet is WWDS (Redkar et. al., 2015). A similar task was performed for Basque Wordnet (Pociello et. al., 2010). Bilingual mappings are a special case of wordnet linkage by which Hindi wordnet deals with this problem. Here, the concepts are translated, and, at times, the synset members are translated / transliterated, in English and this task is carried out manually. To the best of our knowledge this is a novel method and has not been implemented elsewhere.

[3] http://www.illc.uva.nl/EuroWordNet/

[4] http://multiwordnet.fbk.eu/english/home.php

3 Motivation - Need for Bilingual Mapping

The task of linking synsets of Hindi wordnet to those of the English WordNet has been undertaken to create valuable parallel data for various NLP applications. However, languages are mirrors of the society in which they develop and are used. They are, therefore, unique and specific to particular geographical regions and cultural milieus. When two languages, which are as far set apart as Hindi and English, have to be linked at the conceptual level, along with word transfers, it is bound to throw up the challenge of lexical and conceptual gaps. To overcome these challenges, the idea of having two types of linkages – direct and hypernymy – has been followed. Direct linkage provides exact matching concept and lexical item/items in English. For example, for the Hindi word *गंधयुक्त* (*gandhayukta*), which means

- जो गंध से युक्त हो
- which fragrance with is
- *jo gandha se yukta ho*
- which has fragrance,

there is a direct linkage to the English synset of *odorous* which means *having a natural fragrance*. Those concepts in Hindi for which there are no direct linkages in the English WordNet, we adopted the EuroWordnet methodology to link them to a hypernymy synset in English. The idea was that instead of having no linkage at all there would be at least a super-ordinate concept and lexical item/items with which the Hindi concept could be linked. This would provide translation candidates which could be exploited for various NLP tasks. An example of this is the concept of *सदावर्त* (*sadaavarta*), which means

- लिए गए व्रत के अनुसार गरीबों में एक निश्चित समय सीमा तक प्रतिदिन भोजन और अन्य जरूरी वस्तुएँ बाँटने का कार्य
- taken vow according to poor people a definite time period till daily food and other essential items distribution work
- *liye gaye vrat ke anusaar gareeboM meM ek nishchit samaya siimaa tak pratidin bhojan aur anya zaroorii vastueM baaMtane kaa kaarya*
- the act of distributing food and other essential items to poor people for a specific time period according to a vow undertaken.

Although the Hindi concept is very specific, yet it has been linked to the synset of *charity* which means *'an activity or gift that benefits the public at large'* and is marked as a hypernymy linkage.

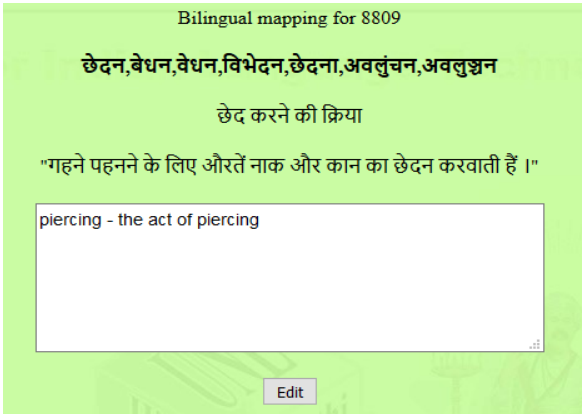
However, consider the example of the synset *छेदन* (*chhedan*), which means

- छेद करने की क्रिया
- make hole act
- *chhed karane kii kriyaa*
- act of piercing.

In the absence of a matching synset in the same POS category in the English WordNet, it had been given a hypernymy linkage to the English synset of *deed*, which means *something that people do or cause to happen*. What this would have implied was that each time the Hindi word *छेदन* would occur in the corpus, the parallel English word *deed* would be its lexical counterpart in English. This is not only too far-fetched but may prove to be insufficient in the translation process. It is in such cases that the hypernymy linkage was substituted for a bilingual mapping. Thus, for *छेदन* the lexical item was kept as *piercing* and the gloss as *the act of piercing*. Moreover, the proper nouns in Hindi wordnet, also pose a problem with hypernymy linkage. It is because of these issues that the idea of creating bilingual mappings of the Hindi synset into English was adopted. Here, mapping indicates the linking of two data-sets, in this case, between Hindi Wordnet and the data set containing English translations of glosses and words.

4 Method

The method adopted for creating bilingual mappings is translation / transliteration of the synset member and the translation of the gloss in English. For this we search various lexical resources and look for valid usages on the internet. After verifying, we create the bilingual mapping. As far as possible, we do not coin words for this purpose, but there are some exceptions. The mapping is created in a dialogue box where the lexicographer manually types the required text, which is then stored in the database of Hindi wordnet.



Screenshot 1: Bilingual mapping creation interface for lexicographers

Users can see the mappings on the online Hindi wordnet interface by querying for the English linkage. In cases where the hypernymy linkage is too distant then it is removed but where the hypernymy is close-enough, it is retained along with the mapping (see Screenshot 2). The retention of hypernymy linkage is also motivated by the fact that it may prove useful for the general users, who may not be familiar with the language and the culture it represents.

5 Challenges and Solutions

For the creation of mappings we have divided words into four major categories based upon the problems faced. These categories and their treatment are as follows:

5.1 Words / Concept not available in English WordNet

There are two types of methods used to deal with such cases. These are the following:

a. Transliteration - When no suitable word in the English WordNet is found to represent the Hindi concept, we transliterate the word and translate the gloss accordingly. For example, *पदयात्रा* (*padayatra*), which means

- किसी विशेष उद्देश्य (विशेषकर राजनैतिक या धार्मिक) से पैदल की जानेवाली यात्रा
- Some special purpose (especially political or religious) for being done foot journey
- *Kisii vishesh uddeshya (visheshkar raajnaitik yaa dhaarmik) se paidal kii jaane walii yaatraa*
- a foot journey undertaken for some special purpose (especially political or religious).

This word had initially been given a hypernymy linkage to *hike* which had the gloss as

a long walk usually for exercise or pleasure. This was found to be inadequate to convey the sense of the Hindi synset. Thus, the hypernymy linkage was removed and a bilingual mapping was created. The synset member was transliterated as *padayatra,* as found in Wikipedia[5], and the gloss was translated as *a journey undertaken for some special purpose (especially political, religious*).

b. Translation - the synset members are translated along with the gloss in English. An example is *अप्सरा* (apsaraa), which means

- स्वर्ग में इंद्र की सभा में नाचने-गाने वाली सुंदरियाँ
- heaven in Indra's court dancing singing beautiful ladies
- *Swarga meM Indra kii sabhaa meM naachane-gaane walii sundariyaan*
- beautiful ladies who dance and sing in Indra's court in the heaven.

Initially, it had been given the hypernymy linkage to the synset of *nymph* which means *a minor nature goddess usually depicted as a beautiful maiden*. This has now been given a bilingual mapping, where the word has been translated as *celestial dancer* and the gloss has been translated as *beautiful ladies who dance and sing in heaven in the court of Indra,* which is much more precise.

5.2 Required sense missing in the English WordNet

Here, the synset is present in the English WordNet, but the given sense/s does not match the one required for the Hindi synset. For example, *फूँकना* (*phuunkana*), which means

- फूँक मार कर दहकाना या प्रज्ज्वलित करना
- by blowing to ignite or to aflame
- *phuunk maar kar dahakaanaa yaa prajjawalit karanaa*
- to light or inflame by blowing.

Although, the English WordNet has four senses of the word *ignite* but this particular sense is not there. So we assigned the word *ignite* as the bilingual mapping and translated the gloss as *cause to start burning by exhaling hard through mouth*. Thus, an accurate meaning transfer is obtained.

5.3 Culture Specific Words

These are the words specific to Indian culture and hence not found in the English WordNet. For example, *बिछिया* (*bichhiyaa*), which means

- पैर की उँगलियों में पहनने का छल्ला
- toes in wearing ring
- *pair kii ungaliyoM meM pahanane kaa chhallaa*
- ring worn on toes.

This has a hypernymy linkage to *jewelry*, which means *an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems*). This gloss does not convey the meaning accurately. Since the word *toe ring* is commonly used for this object and its sense is found in other lexical resources[6], we assigned it as the bilingual mapping with the gloss translated as *a ring worn on any of the toes.*

5.4 Language Specific Words

There are many words in Hindi wordnet which capture the peculiar grammar of the language. It is but natural that their counterparts will not be available in English. Hence, these words require bilingual mappings. There are three categories of such words. These are as follows:

5.4.1 Causative Verbs

As the name implies, causative verbs indicate an action that the subject does not directly perform, but rather causes to happen, perhaps by causing some other agent to perform the action. Such verbs are a well-known feature of Hindi and are represented in English as a phrase. For example, *बरसाना* (*barsaanaa*), which means

- बादल से जल नीचे गिराना
- cloud from water below make fall
- *baadal se jal neeche giraanaa*
- to cause to rain.

For such a sense, finding even a hypernymy linkage was difficult. So we assigned it a bilingual mapping as *to make it rain* with the gloss translated as *to make water fall from clouds.*

[5] https://en.wikipedia.org

[6] http://www.thefreedictionary.com/

5.4.2 'Be' Form of Conjunct Verbs

A large number of Hindi verbs are formed by conjoining a noun or an adjective with a verb. Such verbs are called conjunct verbs. The most common verb used to form conjunct verbs is *करना* (to do/to make). Many conjunct verbs have corresponding intransitive forms which employ *होना* (to be). Hindi wordnet stores these intransitive forms which do not have corresponding English verbs. This is because English makes use of a phrase to convey the same meaning. In such cases, bilingual mapping is the only option. Take the example of *अर्पित होना* **(***arpit honaa***),** which means

- किसी के द्वारा श्रद्धापूर्वक देवता, समाधि आदि पर कुछ रखा जाना
- By someone respectfully deity, tomb, etc. on something to be kept
- *Kisii ke dvaaraa shraddhaapuurvak devataa, samaadhi aadi par kuchhh rakhaa jaanaa*
- be offered (something) by someone respectfully to a deity or on a tomb, etc.

Since such a sense does not exist in the English WordNet, the word is translated as *to be offered* and the gloss has been translated as *be offered (something) by someone respectfully to a deity or on a tomb, etc.*

5.4.3 Idiomatic Expressions

Idioms are words, phrases, or expressions that are either grammatically unusual or their meaning cannot be taken literally. They are highly culture specific and so they require special treatment, becoming perfect candidates for bilingual mapping, specifically those not available in English. For example, *हाथ खुला होना* (*haath khula honaa*), which would literally mean "to have an open hand", but the idiomatic sense is

- दान, व्यय आदि के संबंध में उदार प्रवृत्ति होना
- donation, expenditure, etc. with respect to generous tendency be
- *daan, vyaya, aadi ke saMbandha meM udaar pravritti honaa*
- to be of generous tendency towards donation, expenditure, etc.

It had a hypernymy linkage to *be* which means *have the quality of being*. This was too distant in meaning, did not convey the metaphorical sense, and would not have been accurate in translation. So it has been given the bilingual mapping as *to be big spender* with the gloss translated as *to be generous with respect to donation, expenditure, etc.*

5.5 Words for which Hypernymy Relation Unavailable

Wordnet does not have hypernymy relation for adjectives and adverbs. Thus these words in the Hindi wordnet when not linked to direct English words, do not have an option of hypernymy linkage. In such cases, they have to be invariably given bilingual mappings.

5.5.1 Adjectives

Those Hindi adjectives, which are participial adjectives in English, especially those which are formed by the present participle '-ing' and the past participle '-ed', rarely find exact matching synsets in English. All such adjectives are being assigned bilingual mappings in Hindi wordnet. An example of this is *आत्मवंचक* (*aatmavaMchaka*), which means

- स्वयं या अपने आप को धोखा देनेवाला
- self or to one self deceiving
- *swayam yaa apane aap ko dhokhaa dene waalaa*
- self deceiving.

Here the words translates to *self-deceiving* and the gloss is *who deceives oneself.*

5.5.2 Adverbs

Adverbs also do not have hypernymy relation. Hence, those Hindi adverbs which do not have a direct linkage to English, have to be given bilingual mappings. For example *सालों-साल* (*saaloM-saal*), which means

Part of Speech	Total Synsets in HWN	Direct Linkages	Hypernymy Linkages	Biligual Mappings	Total Linkages
Noun	29070	11582	8184	2110	21876
Adjective	6171	3541	0	331	3872
Verb	3303	1992	207	129	2328
Adverb	475	343	0	27	370
Total	39019	17458	8391	2597	28446

Table 1: POS wise statistics for HWN

- कई साल तक
- many years till
- *kaii saaloM tak*
- of many years

has the bilingual mapping as *of many years*. Here the gloss is omitted because the synset member, which is a phrase in itself, is also the gloss.

5.6 Proper Nouns

Hindi wordnet has more than 16,000 proper nouns, most of which are names of persons, places and organizations specific to India. All such words could not have been given a place in the English WordNet, making linkage difficult. Initially they were given hypernymy linkages to very distant synsets. For example, names of political leaders were linked to the synset of *leader* and characters from Indian epics and mythology were given hypernymy linkages to the synset of *mythical being*. It was felt that such names are better *transliterated* as they would occur in the corpus in the same manner. Thus, such entries are being transliterated with their glosses translated as per Hindi. We have currently mapped over 1,800 such proper nouns, and the work is going on. Some such examples are:

1. *सरदार वल्लभ भाई पटेल* (*sardaar vallabh bhai patel*), which has a gloss as
- भारत के स्वतन्त्रता संग्राम सेनानी जो स्वतन्त्र भारत के प्रथम गृहमंत्री तथा उपप्रधानमंत्री बने
- India of freedom fight soldier who free India of first home minister and deputy prime minister became
- *Bhaarat ke svatantrataa saMgraam senaanii jo savatantra bhaarat ke pratham grihamantrii tathaa upapradhaanmantrii bane*
- a freedom fighter of India who became the first Home Minister and Deputy Prime Minister of independent India.

This is assigned the bilingual mapping as *Sardar Vallabh Bhai Patel, Sardar Patel - a freedom fighter of India who became the first Home Minister and Deputy Prime Minister of independent India.*

2. *भारतीय वाणिज्य एवं उद्योग महासंघ* (*bhaaratiiya vaaNijya evam uddyog mahaasaMgha*) has a gloss as
- भारत की व्यापारिक संस्थाओं की एक मंडली
- India of business organizations of an association
- *Bhaarat kii vyaapaarik saMsthaaoM kii ek maMdalii*
- an association of business organizations of India.

It has been given a hypernymy linkage to the synset of *organization* which means *an organization formed by merging several groups or parties.* This is also assigned a bilingual mapping as *Federation of Indian Chambers of Commerce - an association of business organizations of India.*

6 Bilingual Mapping Interface

Hindi and English wordnets are in both MySQL and file format (text). Hindi wordnet is accessible via an online interface, which provides a login facility to administrators, thus enabling features like adding / editing of a bilingual mapping between Hindi and English over the web interface itself. We currently store the mappings

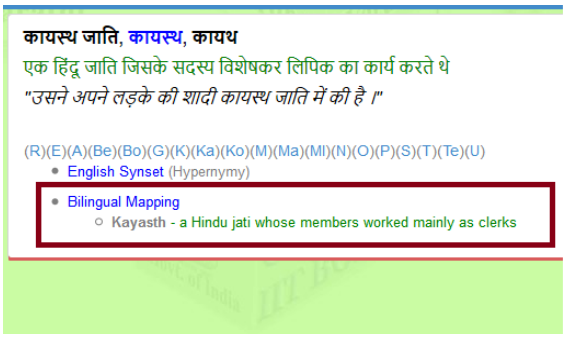
based on Hindi wordnet IDs as pivot. Mapping IDs are provided serially. While querying the database for a linkage, the interface also looks for a mapping, which, if present, is shown on the interface.

The bilingual mapping is stored in the database in the following format:

<word1, word2, ... , wordN> - <gloss>;<example> (for N number of words)

The lexicographers are familiar with this format, and update the database accordingly.

Following are some screenshots of the bilingual mapping interface for mapping / addition and querying of bilingual mapping.



Screenshot 2: Bilingual mapping querying

7 Statistics

The above statistics show that maximum numbers of synsets have direct linkages with the English WordNet. Although, around 8,000 hypernymy linkages have been done, yet these are under review. Some of these have been converted to bilingual mappings. Since assigning mappings is a very recent activity, the numbers are likely to go up as the task proceeds. A bulk of this would comprise of proper nouns which consists mainly of names of persons, places and organizations.

8 Words sent to the English WordNet for Inclusion

As a corollary to the linkage task, it was observed that there are many English language concepts that are missing in the English WordNet and can be easily assimilated therein. These concepts are available in other English dictionaries. We have sent lists of such words to the English WordNet team and have received assurance that these would be looked into. As and when such senses would be made available in the English WordNet, they will be utilized for the Hindi-English linkage task. Some examples of such words are given below:

1. page (Computer Science) - A quantity of memory storage equal to between 512 and 4,096 bytes.
2. flying - The piloting or navigation of an aircraft or spacecraft
3. occupier - one who seizes possession of and maintains control over forcibly or as if by conquest.
4. crisp - conspicuously clean or new

Screenshot 3: Bilingual mapping addition

5. shakingly - in a shaking manner
6. in hand - owned by or in possession

9 Conclusion and Future Work

In this paper we have discussed the process of creating bilingual mappings of the synsets of Hindi wordnet into English, the methods adopted and the tool used in creating them. It was observed that the problems occurred due to conceptual and lexical gaps between Hindi and English languages. The main problem areas are the following:

- Words/ Concept not available in English WordNet
- Required sense missing in the English WordNet
- Culture specific words
- Language specific words
  - Causative Verbs
  - Be' Form of Conjunct Verbs
  - Idioms
- Words for which Hypernymy Relation Unavailable
  - Adjectives
  - Adverbs
- Proper nouns

An online linking facility has been provided to incorporate the bilingual mappings in Hindi wordnet, which can be easily accessed by a user.

By using this method, it is hoped that the task of linking two language concepts can be accomplished with a high degree of accuracy. The

bilingual mappings in English can help clarify the Hindi concept for the lexicographers of the wordnets of the other Indian languages, who may not be very proficient in Hindi. Furthermore, in future, such a strategy may be adopted by wordnets of other Indian languages while linking their wordnets to Hindi wordnet. We can also provide the semantic and lexical relations that such mappings would carry. These mappings can also be tested on a small corpus to verify whether they provide better translation outputs than hypernymy linkages. As the task progresses we may come across other categories of concepts where such mappings may prove to be useful. Above all, it presents an interesting scenario in which two different languages are brought together in conceptual unity. This may in itself offer future research possibilities.

10 Acknowledgement

We gratefully acknowledge the support of the Department of Technology, Ministry of Communication and Information Technology, Government of India. We also acknowledge the work done in this task by Madhavi Khanal and Nootan Verma. Also, not to be missed, is the entire linguistic and computational Wordnet team at CFILT, IIT Bombay, which has provided its valuable input and critique, helping us refine our task.

References

Akshat Bakliwal, Piyush Arora, Vasudeva Varma *Hindi Subjective Lexicon : A Lexical Resource for Hindi Polarity Classification* .LREC 2012, Istanbul, Turkey.

Arun Karthikeyan Karra. 2010. *WordNet Linking.* Master of Technology Dissertation, CSE Department, IIT Bombay.

Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande and P. Bhattacharyya. 2002. *An Experience in Building the Indo WordNet- a WordNet for Hindi*. International Conference on Global WordNet (GWC 02), Mysore, India.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In In Proceedings of the First International Conference on Global WordNet, pages 293–302. Mysore, India.

Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Database*. The MIT Press.

In-Su Kang, Sin-Jae Kang, Se-Jin Nam, Key-Sun Choi. 2010. *Linking CoreNet to WordNet - Some Aspects and Interim Consideration* International Conference on Global WordNet (GWC 10), Mumbai , India.

Jaya Saraswati, Rajita Shukla, Ripple P. Goyal, Pushpak Bhattacharyya.2010. *Hindi to English WordNet Linkage: Challenges and Solutions*. Indowordnet Workshop, collated with ICON 2010, Kharagpur, India.

J. Ramanand, Akshay Ukey, Brahm Kiran Singh, Pushpak Bhattacharyya. 2007. *Mapping and Structural Analysis of Multi-lingual Wordnets.* IEEE Data Engineering Bulletin, 30(1).

Kamil Bulke. 1997. *An English-Hindi Dictionary* (ed.). S. Chand & Co, New Delhi, India.

Lewis Henry Morgan. 1871. *Systems of consanguinity and affinity of the human family*. Smithsonian Contributions to Knowledge; v. 218, Washington DC.

Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya. 2009. *Projecting Parameters for Multilingual Word Sense Disambiguation*. Empirical Methods in Natural Language Processing (EMNLP09), Singapore.

Piek Vossen. 1999. *EuroWordNet as a Multilingual Database*. In: Wolfgang Teubert (ed) TWC.

Hanumant Redkar, Sudha Bhingardive, Diptesh Kanojia, and Pushpak Bhattarcharyya. 2015. *World WordNet Database Structure: An Efficient Schema for Storing Information of Wordnets of the World*, Twenty-Ninth Association for the Advancement of Artificial Intelligence Conference (AAAI 2015), Texas, USA

Basque WordNet: Pociello E., Agirre E., Aldezabal I. 2010. *Methodology and construction of the Basque WordNet.* Language Resources and Evaluation (LREC). Springer, Netherlands.

Salil Joshi Arindam Chatterjee Arun Karthikeyan Karra Pushpak Bhattacharyya *Eating your own cooking: An on-line heuristic-based wordnet linking system using previously linked synsets*, COLING 2012, Mumbai, India, 10-14 Dec, 2012 (demo paper)

http://www.shabdkosh.com

http://pustak.org/bs/home.html

WordNet-based similarity metrics for adjectives

Emiel van Miltenburg
Vrije Universiteit Amsterdam
emiel.van.miltenburg@vu.nl

Abstract

Le and Fokkens (2015) recently showed that taxonomy-based approaches are more reliable than corpus-based approaches in estimating human similarity ratings. On the other hand, distributional models provide much better coverage. The lack of an established similarity metric for adjectives in WordNet is a case in point. I present initial work to establish such a metric, and propose ways to move forward by looking at extensions to WordNet. I show that the shortest path distance between derivationally related forms provides a reliable estimate of adjective similarity. Furthermore, I find that a hybrid method combining this measure with vector-based similarity estimations gives us the best of both worlds: more reliable similarity estimations than vectors alone, but with the same coverage as corpus-based methods.

1 Introduction

In this paper I present new WordNet-based (Fellbaum, 1998) measures to provide reliable estimates of human word similarity ratings. Ever since Hill et al. (2014) published their SimLex-999 data set, many people have tried to find a way to determine the similarity of all the word pairs without being affected by the relatedness of the words. Recently, Le and Fokkens (2015) showed that taxonomy-based approaches beat vector-based approaches (Turney et al., 2010) in the estimation of the SimLex data. This is because corpus-based approaches are more affected by association, while taxonomy-based approaches mainly use vertical relations that are well-suited for determining similarity. However, corpus-based approaches do have a big advantage in their coverage. Moreover, Le and Fokkens left adjectives out of consideration, for lack of a good WordNet-similarity measure. My aim was to fill this lacuna, and also to find a way to mitigate the coverage issue. In section 3, I propose three WordNet-based adjective similarity measures, and evaluate them on the SimLex-999 data.[1] Section 4 provides a more thorough discussion of our results. At the same time, we should acknowledge that the representation of the adjectives in WordNet could use some attention. Section 5 proposes future work, looking at some extensions to WordNet that might improve our proposed measures. Section 6 concludes.

2 Evaluation

It is important to note that similarity is a *relative* measure; we do not learn anything from the fact that the similarity between adjectives X and Y is 2.4 unless we also know the similarity between other pairs of adjectives. Only then do we learn whether X and Y are very similar or not similar at all. In other words, being able to *rank* adjective pairs in terms of their similarity is more important than having a specific number for each pair. This is why the Spearman rank correlation is typically used for evaluation. I follow this standard procedure in our general evaluation.

Le and Fokkens (2015) argue for the use of multiple different evaluation methods, since they may lead to different conclusions about the results. They propose to use *ordering accuracy* (an evaluation of the relative ordering between all combinations of pairs, following Agirre et al. (2009)), supplemented with tie correction, i.e. giving a partial score to word pairs having the same similarity score. This levels the playing field, as taxonomy-based similarity values are more prone to yield ties than corpus-based measures (discrete versus real scores). The intuition behind this proposal is that

[1] All the code and data is available for replication at https://github.com/evanmiltenburg/gwc2016-adjective-similarity

overall ranking is more important than arbitrary local differences. Therefore, we should not punish algorithms as much for getting specific pair orderings 'wrong' when they are too close to call. In the discussion (section 4), I will use Le and Fokkens' comparison by group, where *pairs of pairs* of adjectives are grouped by the difference in their similarity scores in the gold standard. This is useful to see how well different models perform at varying levels of granularity.

3 Current possibilities

In this section, I examine distance metrics for adjectives in WordNet. I will first look at two classical measures, *Hso* (Hirst and St-Onge, 1998) and *Lesk* (Lesk, 1986), and show that they perform reasonably well (although not state-of-the-art). Next, I propose a method based on derivationally related forms, that are associated with the adjective lemmas. Though this approach achieves good results, it does suffer from poor coverage. I will then look at an alternative approach using attributes, but conclude that it is not feasible to incorporate them in our distance metric. Finally, to remedy the coverage issue, I propose a hybrid approach using both WordNet and distributional vectors.

3.1 Classical measures

Two classical similarity measures are given by the *Lesk* and the *Hso* methods. The former uses word overlap between glosses as a similarity measure, while the latter uses path distance (with some restrictions on the path). Both are implemented in Perl by Pedersen et al. (2004). Banjade et al. (2015) evaluate these measures on the adjectives in SimLex-999 taking only the first sense in WordNet into account, achieving a Spearman correlation (ρ) of 0.42 for the Lesk measure, and $\rho = 0.236$ for Hso.

Following Resnik (1995), I evaluated these measures using *all* senses for each word form, and taking the highest similarity. Intuitively, this comes closer to what Hill et al.'s participants did during the judgment task: they were already primed to look for similarities, so they were likely to be biased towards selecting the most similar senses. This idea is reinforced by the Lesk results: now this method (taking the maximal Lesk similarity between all synsets) yields a stronger correlation of $\rho = 0.51$. The correlation of the Hso scores with SimLex almost doubled: $\rho = 0.45$.

3.2 Using derivationally related forms

For all adjectives that have derivationally related forms in WordNet, one can use the distance between those related forms as a measure of adjective similarity. This roughly equates to saying that similarity between adjectives is a function of the properties they describe. I again used the 111 adjective pairs in SimLex-999 to evaluate the performance of this measure. To perform the evaluation, I selected all pairs of adjectives for which WordNet 3.0 specifies derivationally related nouns (for at least the first sense of the adjective). This resulted in 88 (out of 111) pairs, consisting of 89 (out of 107) different adjectives. Our distance measure is defined as follows:

1. For both adjectives A and B, get a list of all synsets corresponding to A and B.
2. Then, generate two new lists of derivationally related nouns: DRN_A, DRN_B.
3. The distance between A and B is given by $min(\{distance(x, y) : \langle x, y\rangle \in DRN_A \times DRN_B\})$, where *distance* is the shortest-path distance.[2]

I predicted that there would be a (negative) correlation between the distance between A and B and the similarity between A and B (i.e. items that are further apart in WordNet should be less similar). This expectation is corroborated by the results: our similarity measure has a Spearman correlation (ρ) of -0.64 with the SimLex data, which is near human performance (overall human agreement $\rho = 0.67$). To compare this result, I used the best performing predict-vector from (Baroni et al., 2014)[3] to generate cosine similarities for the same pairs of adjectives, achieving $\rho = -0.59$.

3.3 Using attributes: negative results

A problem with using derivationally related forms is that only 41% of all adjective synsets *have* derivationally related nouns. For better coverage, can we apply a similar technique to measure similarity through each adjective's attributes? The answer seems to be negative. I took two types of

[2] I did not experiment with alternative measures, as performance is not the main goal of this paper.

[3] This model was trained using `word2vec` (Mikolov et al., 2013) on the UkWac corpus, the British National Corpus, and the English Wikipedia. It is available here: `http://clic.cimec.unitn.it/composes/semantic-vectors.html`.

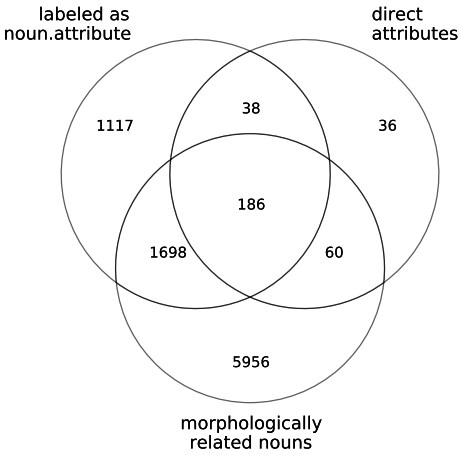


Figure 1: Nouns in WordNet that are, or could potentially be linked to adjectives in WordNet 3.0.

approaches, but neither produced any significant correlation with the SimLex data:

1. Take the shortest path distance between all attributes of the first/all senses of A and B.
2. Use the (relative) size of the overlap between the sets of attributes of A and B.

It is unclear why we get such a different result using attributes instead of derivationally related forms, but it probably has to do with the current status of WordNet attributes. A closer look at the adjectives in WordNet 3.0 teaches us that there are only 620 adjectives that even have attributes, and on average each adjective has 1.03 attributes. Furthermore, only a fraction of nouns that are labeled as `noun.attribute` is actually used as an attribute. Figure 1 provides an illustration of the current situation. In sum: it might be too soon to write off an attribute-based similarity measure, but getting such a measure to work requires a serious effort to link adjectives to all their possible attributes. Fortunately, there is already some work in this direction: Bakhshandeh and Allen (2015) describe a method to automatically learn from WordNet glosses which attributes an adjective can describe.

3.4 Going hybrid: WordNet plus vectors

What we *can* do, is make use of WordNet as much as possible, and only rely on vectors or other techniques if WordNet fails to provide a measure.[4] I used the following general algorithm, substituting Baroni et al's vectors for X:

1. Generate similarity values for all the pairs using WordNet, and other approach X, so that we have two lists of similarity values: L_W and L_X.
2. Sort both lists, so that we get a ranking for all pairs. In L_W, there will typically be many pairs with the same rank (i.e. ties).
3. Create a new output list L_O; initially a copy of L_W. Use the values from L_X as a tie-breaker, so that all pairs in L_O have a unique rank.
4. Iterate over all the pairs p in L_X that do not occur in L_W. The first pair is a special case: if p is the first item of L_X, put it at the start of L_O. Otherwise, treat it like the other pairs: get the pair immediately preceding p in L_X and look up its position in L_O. Insert p immediately after that position in L_O.

The result (L_O) is a sorted list that maintains the structure of L_W, but that also contains all the pairs under consideration. For the SimLex data set, the hybrid approach achieves a correlation of $\rho = -0.62$, compared to $\rho = -0.58$ for Baroni et al.'s vectors alone.

4 Discussion

From the Spearman correlations alone, it seems that we gain precision by involving derivationally related forms (DRF) in the estimation of similarity values. This picture changes when we look at ordering accuracy. I found that the DRF-based and vector-based approaches achieve comparable results. For the subset of 88 pairs where both adjectives have DRFs, I found a slight advantage for the vector-based method compared to the DRF-based method: 70% versus 71%. For the full dataset, this is exactly reversed, with a precision of 71% for the hybrid method and 70% for the vector-based method. That is not to say that both measures encode the same information; indeed we find interesting differences when we compare the pairs on a group-by-group basis.

Table 1 shows the ordering accuracy by group. When differences (in similarity scores) between two word pairs are small, the vector-based approach seems to have the upper hand in determining which is more similar. On the other hand, when differences between pairs are larger it seems that the hybrid approach is better at determining which pair is more similar. As the table shows,

[4] Banjade et al. (2015) also use a hybrid system to estimate similarity scores, but they use many different measures and combine them using a regression model.

Δ	WordNet	Vectors	Hybrid	Vectors
0	52	54	53	54
1	57	68	63	64
2	65	73	66	73
3	89	69	82	74
4	92	91	91	89
	Subset		Full dataset	

Table 1: Ordering accuracy scores by group, for the 88-pair subset from section 3.2 and the full dataset from section 3.4. The Δ-column indicates levels of granularity in the differences between pairs being compared. It runs from 0 (pairs with comparable similarity scores) to 5 (pairs with large differences in their similarity scores).

both effects are more pronounced in the 88-pair subset. Note especially the marked 20 percentage point difference with $\Delta = 3$.

Issues with tie-correction

The fact that with $\Delta \in \{0, 1, 2\}$ we find that vector-based approaches have a better ordering accuracy is interesting, but may also be an artifact of the tie-correction. Consider the way tie correction works: whenever a model predicts a tie, a score of 0.5 is awarded. In groups where the differences are small, the likelihood of a tie using the DRF-based method increases, and so the average score is drawn towards 50%. This is not what we want, as it actively biases the evaluation against coarse-grained measures in first group(s).

When we make the score linearly dependent on the difference between the pairs in SimLex-999 (punish the model for predicting a tie when there is actually a big difference, and reward the model for predicting a tie when there is little-to-no difference at all), the DRF-based method with the 88-pair subset gets an increased overall score of 74% whereas the vector-based method achieves the same score as before (71%).[5] More work is needed to determine whether this is a good way to do tie-correction, and whether it is at all possible to reliably compare fine-grained similarity measures with course-grained ones. But if we just ignore any ties between pairs in either the gold standard or in both of the similarity measures, then we are left with 3299 pairs where the DRF-based method has an accuracy of 74%, versus 73% for the vector-based approach.

[5]The updated scoring function returns the result of the following function if a tie is predicted (with P as the set of all pairs in the gold standard):

$$\text{score}_{\text{tie}}(p_1, p_2) = 1 - \frac{abs(p_1 - p_2)}{max(\{abs(p_i - p_j) : \langle p_i, p_j \rangle \in P \times P\})}$$

5 Future work: extensions to WordNet

There are several projects that add new information to the adjective synsets, which can be used to increase coverage. Below I discuss potential uses and the current limitations of this information.

Adjective hierarchy GermaNet (Hamp and Feldweg, 1997) contains a hierarchy for adjectives, structured using hyponymy relations. This means that it is possible to use any of the available WordNet distance metrics directly on the adjective synsets. Unfortunately, the mapping between GermaNet and Princeton WordNet is still incomplete, and there is no dataset similar to SimLex for German to test this idea.

Add new cross-POS relations In this paper we have used the two types of cross-POS links that are available in WordNet: attributes and derivationally related forms. Other projects have a more diverse set of relations between adjectives and nouns. EuroWordNet (Vossen, 1998) has the *xpos_near_synonym, xpos_has_hyperonym* and *xpos_has_hyponym*-relations that can be used as access points to the noun hierarchy. WordNet.PT (Mendes, 2006) has similar relations. These seem like a good addition to the *'derivationally related to'*-link that we have been using, as they encode very similar information without the requirement of the two words morphologically resembling each other. Adding these relations would give us a much better coverage, while hopefully still providing a good score, but this remains to be tested.

Add domain information a more general approach is WordNet-domains (Magnini and Cavaglia, 2000), where each synset is associated with a particular domain. Examples of domains are: ECONOMY, SPORT, MEDICINE, and so on. Like the *property-of* relation, domain information does not seem to be helpful in the actual ranking procedure, but the knowledge whether two adjectives are associated with the same domain may serve as a useful bias.

6 Conclusion

We have seen several different WordNet-based measures of adjective similarity: the classical

Lesk and Hso measures, and two new measures based on specific cross-POS links and the shortest-path distance between the nouns they are related to. It turns out that the *derivationally related forms*-link can be used to get state-of-the-art results on the SimLex-999 dataset. If coverage is an issue, then the hybrid method from section 3.4 is a better option than using vectors alone (though not by a large margin). We also noted that, on closer inspection, these measures do not seem to capture the same information. Therefore, future research should look at new ways to combine distributional and taxonomy-based measures.

Another way to improve similarity estimations would be to extend WordNet with new information. For example, the *attributes*-relation currently seems unusable for any similarity-related work, but may still be useful if more attribute links are added to WordNet. And looking at the literature, there is a lot of promising work being done with other WordNets, leaving us with many interesting avenues to explore the relation between WordNet and lexical similarity.

Acknowledgments

Thanks to Tommaso Caselli, Antske Fokkens, Minh Le, Hennie van der Vliet, and Piek Vossen for valuable comments on earlier versions of this paper. This research was supported by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of HLT*, pages 19–27. Association for Computational Linguistics.

Omid Bakhshandeh and James F Allen. 2015. From adjective glosses to attribute concepts: Learning different aspects that an adjective can describe. *IWCS 2015*, page 23.

Rajendra Banjade, Nabin Maharjan, Nobal B Niraula, Vasile Rus, and Dipesh Gautam. 2015. Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *Computational Linguistics and Intelligent Text Processing*, pages 335–346. Springer.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, volume 1, pages 238–247.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. Cambridge, MA: The MIT Press.

Minh Ngoc Le and Antske Fokkens. 2015. Taxonomy beats corpus in similarity identification, but does it matter? In *Proceedings of Recent Advances in NLP*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating subject field codes into wordnet. In *LREC*.

Sara Mendes. 2006. Adjectives in WordNet.PT. In *Proceedings of the GWA*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at hlt-naacl 2004*, pages 38–41. Association for Computational Linguistics.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Springer.

Toward a truly multilingual Global Wordnet Grid

Piek Vossen
VU University Amsterdam
Amsterdam, The Netherlands
piek.vossen@vu.nl

Francis Bond
Nanyang Technological University,
Singapore
bond@ieee.org

John P. McCrae
Insight Centre for Data Analytics
NUI Galway
Galway, Ireland
john@mccr.ae

Abstract

In this paper, we describe a new and improved Global Wordnet Grid that takes advantage of the Collaborative InterLingual Index (CILI). Currently, the Open Multilingal Wordnet has made many wordnets accessible as a single linked wordnet, but as it used the Princeton Wordnet of English (PWN) as a pivot, it loses concepts that are not part of PWN. The technical solution to this, a central registry of concepts, as proposed in the EuroWordnet project through the InterLingual Index, has been known for many years. However, the practical issues of how to host this index and who decides what goes in remained unsolved. Inspired by current practice in the Semantic Web and the Linked Open Data community, we propose a way to solve this issue. In this paper we define the principles and protocols for contributing to the Grid. We tested them on two use cases, adding version 3.1 of the Princeton WordNet to a CILI based on 3.0 and adding the Open Dutch Wordnet, to validate the current set up. This paper aims to be a call for action that we hope will be further discussed and ultimately taken up by the whole wordnet community.

1 Introduction

Princeton WordNet (PWN: Fellbaum, 1998) has existed for 25 years. It is a manually created resource that has proven its worth in many different aspects of linguistics, computational linguistics, industrial applications and last but not least lexicology and knowledge engineering, cited over 11,000 times in Google Scholar[1]. It models language based on a division between words and concepts (represented as synsets) and semantic relations between these synsets. WordNet provided a different perspective on lexical resources from the traditional view in which the lemmas are the basis for defining concepts. Since EuroWordNet Vossen (1998), the Princeton WordNet model has spread to many other languages all over the world and has been extended with inter-lingual relations through the InterLingual Index (ILI). By linking concepts across languages it became possible to compare wordnets across languages, raising fundamental issues with respect to the definition of a word and a concept.

Because synsets are based on sets of synonyms, they mainly represent concepts lexicalized in a particular language (although you can have a synset with a phrase rather than a single word). This implies that different language wordnets may define different concepts related to the network and in fact define semantic spaces that partially match and partially do not. Within EuroWordNet, two approaches were defined to build wordnets: **expand** and **merge**. The expand method takes the concepts from the Princeton WordNet (PWN) as a starting point and translates the synonyms in the synsets to equivalences in the target language. If the same word is a translation of synonyms in different synsets, this creates different senses for the translation. By default, the fund of concepts for this wordnet and the semantic space is identical to the PWN structure. Concepts that are not lexicalized in English cannot be represented, or will be added to the nearest possible synset, even if the denotation is slightly different. The merge method takes the words of a language as a starting point and independently creates the synsets and relations between them. This leads to an independently created semantic space, which can then be aligned with the PWN structure by providing equivalence relations. In the case of the merge approach, the spaces are usually partially aligned and

[1] 11,266 citations on 2015-09-12.

there may be concepts that are in the new wordnet but not in PWN.

Currently, there is no central registry for these new concepts. Wordnet builders for different languages have no control over the concepts included in PWN and cannot easily share their concepts to other wordnet builders. Some projects have created their own internal InterLingual indexes (for example MCR (Gonzalez-Agirre et al., 2012) and the Multilingual Wordnet (Pianta et al., 2002) but these have not been widely adopted, are also based on PWN and most importantly cannot be modified by the community.

The idea of a GlobalWordNet Grid (GWG): a platform for making all wordnets and their linkage available was proposed at the bi-annual business meeting of the Global Wordnet Conference in Jeju, Korea 2006. Such a platform would enable the discussion about what defines a word and a concept across the different wordnets and also enable concept-sharing in a more fundamental way.

The Open Multilingual Wordnet (OMW: Bond and Paik, 2012; da Costa and Bond, 2015) went a long way to making linked wordnets available. The key insight was that wordnets could only be legally linked if the data was freely available and allowed manipulation and redistribution. They showed that wordnets were cited more if released under an open license, and managed to persuade many projects to release under open licenses. As a result there are now open wordnets available for 33 languages, all linked to each other, as well as automatically constructed data for 150 languages. They postponed the question of how to link concepts across all languages by using PWN 3.0 as a *de facto* ILI, and dropping concepts that could not be linked.

Concluding: the essential problem of how to coordinate adding new concepts for multiple languages has not been realized until today. We want to start a new era for wordnets by establishing a framework so that the building and comparison of the different language wordnets may achieve another level: both theoretically and from an engineering point of view. This paper describes the details of this platform and opens up the discussion with the community how to proceed. The paper is further structured as follows. In Section 2, we give the background and motivation for the Grid, while section 3 describes the main principles for the GWG and Section 4 for the Collaborative ILI (CILI). Section 5 describes the procedures and the current status. In Section 6, we explain how the ILI is used to map to WordNet3.0 and WordNet3.1. We also discuss methods for gloss-comparison across synsets in wordnets to find matches and candidates for new concepts. Section 7 reports on an experiment to map the Open Dutch Wordnet to the Grid and the attempt to find new ILI concepts. Finally in Section 8, we discuss the future options to proceed and come to our conclusions.

2 Background and motivation

The Global Wordnet Association website currently lists 76 wordnet groups and projects for 47 languages and other initiatives such as IndoWordnet and Asian Wordnet with many more languages. Not all of these projects are at the stage where they have produced a working wordnet. These wordnets almost all have some relation with PWN, either through the expand method or through equivalence relations (merge). All wordnets implement the notion of a synset as the core structure with at least lexical semantic relations between these synsets. Although PWN has a well-defined structure, the development of wordnets for other languages shows a large variety of decisions and choices. Some of these choices relate to the content of the databases, whereas others apply to the way the resources are distributed. This variation seriously hampers the use and principled study of the wordnets, especially since it is not possible to obtain all wordnets and access them through a unified format and API, which is our main motivation for establishing the GWG platform. Further, different wordnet projects have extended the wordnet structure in different ways, adding different relations and using conventions. Because of this, it is hard to compare wordnets across languages.

In addition to these more fundamental problems, there are also various practical problems for usage of the collection of wordnets. Different wordnets are linked to different versions of the Princeton WordNet, released in different formats (e.g. Princeton offsets and sense-keys, EuroWordNet XML, Multiwordnet, WordnetLMF, RDF) and according to different licenses (from completely open source to commercially restricted). Further, many wordnets have added new concepts, but as there is no central ILI how many of these, if any,

are duplicates?

When Princeton releases a new version of WordNet, it immediately leads to a further decrease in compatibility of wordnets and all related tools and systems, in particular as synset identifiers cannot be preserved across versions, although sense keys are intended to be preserved. The fact that all wordnets and systems adhere to some version of the Princeton WordNet also means that the fund of concepts is biased towards an Anglo-Saxon worldview and is not open to concepts from other languages and cultures.

It could be argued that these problems would go away if a single multilingual database was developed instead. This would, in theory, solve problems of incompatible formats and coordination. In practice, however there is no single group that has expertise in all the world's languages. Further, much experimentation is done in the different projects; adding new relations (Vossen, 1998), adding richer domains (Bentivogli et al., 2004), adding new parts-of-speech (Seah and Bond, 2014) and so forth. This would be harder to do in one monolithic project.

As time passes, and PWN now celebrates its 25th anniversary, the need for implementing the GWG becomes more urgent. The GWG should be a platform for achieving linguistic and conceptual interoperability across wordnets and all related machinery. It should allow researchers to study the universals and idiosyncracies in lexicalisation across languages, to address fundamental questions about what is a word and what is a concept (Fellbaum and Vossen, 2010; Vossen and Fellbaum, 2011). Tools built on wordnets should enable the development of software that can process text in any language according to a common semantic backbone as was demonstrated by the KYOTO[2] (Vossen et al., 2013a) and NewsReader[3] (Vossen et al., 2014) projects.

3 The new Global Wordnet Grid

The global wordnet grid consists of:

- The individual wordnet **project**s
- The collaborative interlingual index (**CILI)**
- The platform that ties them together and allows for adaptation and collaboration

[2] `www.kyoto-project.eu`

[3] `www.newsreader-project.eu`

The projects contribute data for wordnets that they produce in an agreed upon format: WordnetLMF or a *lemon*-based WordnetRDF. These should be validated and checked by the projects, who will have the responsibility of clearly marking which synsets are ready to be included in the ILI (that is, hand checked to a good quality). Although most projects specialize in a single language, there are some that produce multiple languages: both LMF and *lemon* can handle this.

As the GWG will manipulate and redistribute the projects' data, it must be released under a suitable open license. The CILI is released under a the Creative Commons Attribution 4.0 (CC BY) license. However, some projects use the ShareALike license (CC BY SA). In order to keep compatibility across the grid, any projects in the Global Wordnet Grid must have a license compatible with CC BY SA (such as the original wordnet license, CC BY, MIT and many others), and the entire grid will be released under this license.

The individual projects, starting with PWN, are the foundations upon which the GWG is built, the CILI links them and the platform ties them together, allows for versioning and adaptation through the community.

4 The Collaborative ILI (CILI)

The Collaborative ILI is an extension of the ILI defined in EuroWordNet (see Bond et al., 2016, for more details). As a base for the CILI we take the synsets currently in Princeton Wordnet 3.0, the *de facto* ILI for the Open Multilingual Wordnet. This shows its central position in the current wordnet community. Each synset in PWN 3.0 gives rise to a concept in the CILI.

The CILI is just a collection of concepts to which all wordnets are linked. It does not duplicate the relations between these concepts as represented in any wordnet and it does not have any lexicalizations. Concepts and concept identifiers in the CILI are permanent. They will never be removed or changed. However, new concepts can be added to the CILI but only if:

- there is a synset in a wordnet in the GWG that represents this concept (that is, linked by a `owl:sameAs` relation)
- this synset is related to another concept in this wordnet that is already represented in the CILI with one of a set of known relations (`hypernymy`, `meronomy`, `antonymy`)

- It must have a unique English definition that complies with the definition guidelines

The CILI is expanded when a project commits a wordnet, or a new version of a wordnet, to the repository. A committed wordnet is analysed by the moderators of the site (the authors of this submission). If syntactically correct, we will update the ILI records for all synsets that have such a record as a value of the ILI-attribute so that the records get `owl:sameAs` mappings to the contributed wordnet.

All synsets without an CILI-attribute that fulfill the conditions given above (linked, uniquely defined) will generate a proposed new concept which is distributed to the wordnet community for feedback and voting.

Gloss similarity can be used to find CILI concepts that are similar, where we can limit the search space on the basis of the semantic relations (of any linked wordnet). This prevents orphan concepts to be added that cannot be positioned in the semantic space of any available wordnet. We will demonstrate this in the next sections for Princeton WordNet 3.1 and the Open Dutch Wordnet.

5 The community platform

The GWG platform consists of:

- the website providing the most important information and the status of the Grid: `http://globalwordnet.org/global-wordnet-grid/`
- the ILI hosted as an LOD repository with persistent identifiers for concepts: `globalwordnet.org/ili`
- the collection of wordnets in WordnetLMF, *lemon*-based WordnetRDF format in a version control platform (such as `https://github.com/globalwordnet`)

The versioning control system is used to keep track of changes and contributions.

Adapting the ILI within GWG is important to get a better mapping across wordnets, especially when following a merge approach. It enables us to bypass conceptual gaps in PWN and the English language and share related resources across languages such as parallel corpora, ontologies, terminologies and sense-tagged corpora. It should also tighten definitions of synonyms and relations through translation relations across texts in different languages or word embeddings derived for any language (such as Mikolov et al., 2013). Ultimately, it allows us to define what is a word and what is a concept across languages.

There should be no limit to the number of concepts. Phrasenets are equally legitimate as synsets to define a concept. We can allow for example for frequent adjective-noun, noun-prep-noun, verb-object combinations as well as for proverbs, idioms and compounds in languages. Whether and how these concepts are lexicalized is up to the wordnet builders in each language. Ultimately, we will be able to infer which and how many languages provide some type of lexicalization for these concepts. Concepts that are linked to many independently-built wordnets do matter, concepts linked to a single wordnet play a minor role within the Grid. The more `owl:sameAs` relations a concept gets, the more it is valued by the wordnet community. It is also possible to axiomatize concepts through any ontology, exploiting `owl:sameAs` relations between URIs. An ontology defines a semantic space just as any other wordnet, albeit more formally.

Given the fact that concepts with sufficiently different glosses can be added and can be adopted by others, we can imagine that the GWG forms different layers of concepts, starting from a core of concepts shared by many wordnets, possibly ontologized and applied to many different texts in different languages, up to concepts recently added and mapped to only a single wordnet. The linkage of the data can be seen as an *onion model*[4] of concepts based on:

- a kernel of fund consists of concepts that are:
  - shared by all associated wordnets
  - sufficiently voted for by different wordnets, built independently and with sufficient spread in language-families and cultures
  - axiomized through ontologies
  - passed various consistency checks
- an outer layer that contains:
  - most recently proposed new concepts with an `owl:sameAs` relation to a synset in a single wordnet that meets the minimal criteria described above

[4] as presented at the LREC-2014 workshop on Linked Data in Linguistics

- In between layers:
  - linked to more wordnets across languages and language families
  - while these wordnets express semantic relations for these concepts that are not in conflict
  - may have been moderated by the community for example through voting
- an external layer that contains:
  - synsets defined in project wordnets that do not fit the criteria for inclusion into the ILI (e.g. no English definition or unlinked). These concepts need more work to either link them or to be added as new concepts.

In addition to the CILI itself, we will host all public wordnets that are linked to the ILI and may have provided new concepts. We extended WordnetLMF (Vossen et al., 2013b) with some additional attributes to support the mappings of wordnets to the CILI. First of all, each synset element has an optional attribute *ili* for the CILI-record to which the synset is connected. Furthermore, the definition element has an obligatory language attribute and an optional provenance attribute to enable matching concepts. Below we show a WordnetLMF example for an Open Dutch Wordnet synset with a mapping to the CILI and different definitions:

```
<Synset id="eng-30-13956488-n" ili="i110277">
  <Definitions>
    <Definition gloss="overeenstemming met de werkelijkheid"
    language="nl" provenance="odwn"/>
    <Definition gloss="conformity to reality or actuality"
    language="en" provenance="pwn"/>
    <Definition gloss="agreement with reality" language="en"
    provenance="google-translate"/>
   </Definitions>
   <SynsetRelations>
     <SynsetRelation provenance="pwn" relType="has_hyperonym"
     target="eng-30-13954818-n"/>
   </SynsetRelations>
</Synset>
```

6 Mapping updates in PWN 3.1

One of our first checks was to ensure that the graph of the 3.1 version of PWN can be mapped to the CILI (which is based on version 3.0 of PWN). This should have been a trivial case as while the synset identifiers, which are based on the offset in a the release files, are not stable between versions, the sense keys used to identify the senses in Princeton WordNet should be. Using this as the basis of the mapping we found that 1,796 (1.5%) of all synsets were modified between version 3.0 and 3.1 and we manually mapped these synsets. The results were as follows:

- No equivalent in 3.1 (986 synsets):
  - Proper names, drug names, brand names and other proper nouns were systematically removed from 3.0
  - Many sexist, racist, homophobic etc. terms were removed (e.g., 'shirtlifter' and many much worse)
  - Some terms such as 'that much' or senses of terms were considered not be lexicalized concepts and thus erroneously introduced into PWN.

  In these cases, the concept in the CILI is marked as **deprecated**
- Multiple 3.0 synsets mapped to one in 3.1 (51 synsets)
  - Mostly duplicates, e.g., 'finish coat' (03342657-n and 03342863-n)

  In these cases, one of them (typically the one with a different definition from the one that was kept), should be marked as being **superceded** by the other and **deprecated**.
- Single 3.0 synset mapped to multiple in 3.1 (22 synsets)
  - In some cases a word is removed from a synset and put into a new synset, which may be either a hypernym, hyponym or co-hyponym of the previous synset. The definition of the original synset is preserved, e.g., the adjective 'documentary' was removed from the synset of 'objective' or 'documentary' ("emphasizing or expressing things as perceived without distortion of personal feelings, insertion of fictional matter, or interpretation") and given a new synset specifically stating that it must be a film or TV show.
  - In some cases an existing synset is split and both new meanings appear to be more specific, e.g., the heraldic terms 'annulet' and 'roundel' were given new synsets and the previous definition was removed.

  In the first case, a new concept is created and no other change is necessary. In the second

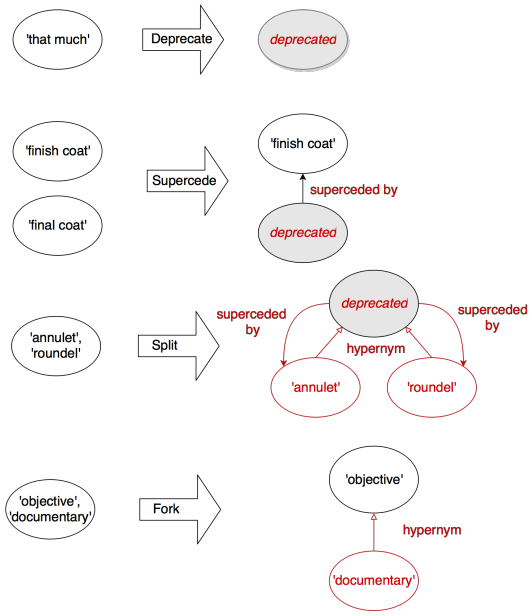


Figure 1: Examples of changes that can be made to the CILI

case, two new concepts need to be created and linked to the original one, which should be deprecated.

- The remaining 737 mappings were changes of part-of-speech between satellite and non-satellite adjectives. These require no changes to the ILI as it does not mark part-of-speech.

In summary, these changes should be the most common changes of the CILI as it develops.

Deprecate A synset may be flagged as deprecated, meaning that we no longer consider it a true lexical concept. This is primarily the case when a compound term has been introduced by mistake. The synset identifier is not removed from the ILI.

Supercede If a duplicate is detected we would choose one of the synsets to remain, and the second synset identifier is marked as deprecated and a link is introduced to the superceding synset, but this second synset is not removed from the CILI.

Split If a synset is considered to generalize two distinct concepts we split it into two new synsets and add these as hyponyms, which are marked as supercedents of the original synset. The original synset is marked as deprecated but not removed from the ILI.

Fork Alternatively if the original synset is still considered valid it is kept undeprecated and a new more specific and closely related synset is added.

Note, that a new wordnet version on its own does not give enough information to decide when a concept should be deprecated or superceded. The platform must therefore allow projects to suggest this as a separate operation.

7 The Open Dutch Wordnet

The Open Dutch Wordnet (ODWN, Postma et al. (2016)) was created from PWN through a mixture of expand and merge methods. PWN synset identifiers and relations have been re-used as much as possible. However, new concepts that originate from the Referentie Bestand Nederlands (RBN: Van der Vliet, 2007)) and have no equivalence relation to PWN synsets have been added. Table 1 shows the distribution of synsets with mappings to PWN (Dutch PWN synsets) and synsets without (Dutch ODWN synsets). To maintain the PWN hierarchy, the wordnet includes hypernym synsets from PWN even if they do not have any Dutch synonyms (English PWN synsets).

Table 1: Overview of the Open Dutch Wordnet

Open Dutch Wordnet	Total	Nouns	Verbs
Word forms	57,602	50,255	7,347
Lexical Units	94,140	78,612	15,528
Dutch ODWN synsets	21,636	15,992	5,644
Dutch PWN synsets	19,980	15,706	4,274
English PWN synsets	75,376	66,409	8,967
Total	116,992	98,107	18,885

In all cases that we could use a PWN synset, we could also map the concept to a CILI record. All synsets with an ODWN identifier were not mapped to the CILI. Consider the word *bierbuik* which is ambiguous between two senses: one for *a big belly because of drinking too much beer* and the second referring to *a person with such a belly*. Neither sense is currently in PWN3.0, and thus new CILI concepts would need to be created. The first sense is lexicalized in English (*beer belly, beer gut*, but has not yet been added to PWN. Both the concepts are linked (as hyponyms) to existing synsets in PWN, therefore to add the concepts to the CILI, the ODWN project would just need to write English glosses.

In total there are 21,636 synsets without a mapping to a PWN synset and therefore without a

Dutch	English
perzikhuid	peach skin
kalfskotelet	veal chop
natuurramp	natural disaster
verwachtingspatroon	expectations
sluikreclame	product placement

Figure 2: ODWN entries not in PWN

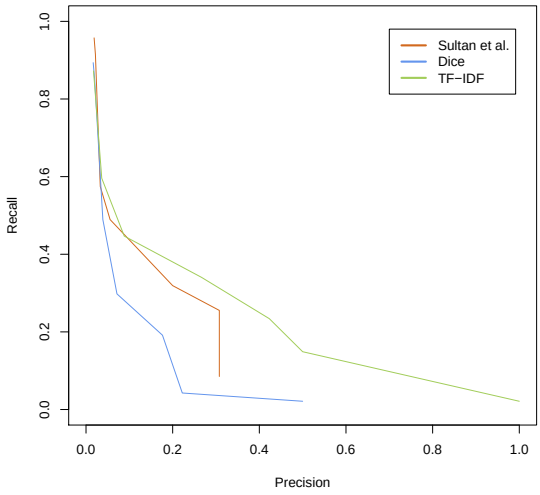


Figure 3: Precision-Recall curve of gloss matching between ODWN and PWN

mapping to the CILI. From these, there are about 5,067 synsets in which case the lemma as translated by Google Translate is not an entry in PWN and another 4,479 synsets for which the translations have a low similarity according to the PWN hierarchy, using the method described by Leacock and Chodorow (1998). We show some more examples of translations not in PWN in Figure 2.

We consider these 9,546 synsets potential new CILI concepts. To validate these as new, we need to ensure that they do not match an existing English gloss. This task is complicated by two main issues: firstly, semantic textual similarity is still a difficult task and secondly, we are using machine translations of the definitions, which introduces further error into the process. To investigate whether automatic methods would solve this task we translated the definitions of Dutch synsets which were already aligned to synsets in PWN and attempted to see if we can distinguish this gloss from similar glosses, in particular glosses of synsets that were up to 3 hyperonym/hyponym links from the target synset. We tried three similarity metrics, namely, the Dice co-efficient, the cosine of TF-IDF vectors of the glosses and an alignment method (Sultan et al., 2014), which had the strongest performance for the Semantic Textual Similarity Task at SemEval-2014. For each of these methods, we varied the acceptance threshold and calculated precision and recall in the usual manner and the results are presented in figure 3. A random baseline has an expected precision of 3.0% and the highest F-Measure was 35.5%: a strong improvement.

However, the performance of the semantic matching is still low, and while high recall can be achieved, which would allow us to select a list of potential duplicates this is only at very low precision, meaning that annotators may have to work through a very long list of candidates. We believe this in part due to the relatively short glosses in ODWN, for example, for 'afweersystem' ('immune system'), the gloss is only 'afweer tegen ziektes' ('defense against diseases') where as the PWN gloss is 27 words long: "a system (including the thymus and bone marrow and lymphoid tissues) that protects the body from foreign substances and pathogenic organisms by producing the immune response". As such, automatic systems can aid in the detection of duplicates in the CILI but must be considered along with guidelines that glosses must be submitted in English and not automatically translated and that glosses must conform to quality guidelines.

Our impression overall so far is that many of the ODWN synsets are already in PWN, although it is very difficult in some cases to find them. The best candidates for new concepts are actually translations of synonyms that could not be found as entries in PWN (5,067 in total). However, even these need critical review and their glosses should have zero scores compared with a wide range of candidate synsets. Concluding, we can say that extending the CILI with new concepts should be done conservatively and with great care.

8 Future work and conclusions

We presented the implementation of the Global Wordnet Grid, which has been pending for many years. We described the data structures and data points as well as the main principles, the protocols and the motivation. The success of the platform will depend on the community. We described two use cases. They made it clear that the process is not trivial and we still will need to discuss many

details. We welcome any further suggestions and contributions of wordnet builders and users.

Finally, the position of the Princeton WordNet in the Grid is essential. Reference to concepts, words, word senses and versions of resources is essential. We hope that future version of PWN will support the GWG and make reference to the ILI just as other wordnets should do.

References

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Coling 2004 Workshop on Multilingual Linguistic Resources*, pages 101–108. Geneva.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: The collaborative interlingual index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. (this volume).

Lus Morgado da Costa and Francis Bond. 2015. OMWEdit - the integrated open multilingual wordnet editing system. In *ACL-2015 System Demonstrations*.

Christiane Fellbaum and Piek Vossen. 2010. Connecting the universal to the specific: Towards the global grid. *Lecture Note in Computer Science (LNCS) Vol.4568*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Fellbaum (1998), chapter 11, pages 265–283.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch wordnet. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. (this volume).

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Hennie Van der Vliet. 2007. The referentiebestand Nederlands as a multi-purpose lexical database. *International Journal of Lexicography*, 3(20):239–257.

P. Vossen, E. Agirre, G. Rigau, and A. Soroa. 2013a. *New Trends of Research in Ontologies and Lexical Resources*, chapter KYOTO: a knowledge-rich approach to the interoperable mining of events from text, pages 65–90. Springer Verlag, Heidelberg.

P. Vossen, G. Rigau, L. Serafini, P. Stouten, F. Irving, and W. Van Hage. 2014. Newsreader: recording history from daily news streams. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Piek Vossen and Christiane Fellbaum. 2011. *Multilingual FrameNets in Computational Lexicography, Methods and Applications*, chapter Universals and Idiosyncracies in Multilingual WordNets, pages 319–346. Mouton de Gruyter, Berlin.

Piek Vossen, Claudia Soria, and Monica Monachini. 2013b. LMF - lexical markup framework. In Gil Francopoulo, editor, *LMF - Lexical Markup Framework*, chapter 4. ISTE Ltd + John Wiley & sons, Inc.

This Table is Different: A WordNet-Based Approach to Identifying References to Document Entities

Shomir Wilson
Carnegie Mellon University
shomir@cs.cmu.edu

Alan W Black
Carnegie Mellon University
awb@cs.cmu.edu

Jon Oberlander
University of Edinburgh
j.oberlander@ed.ac.uk

Abstract

Writing intended to inform frequently contains references to document entities (DEs), a mixed class that includes orthographically structured items (e.g., illustrations, sections, lists) and discourse entities (arguments, suggestions, points). Such references are vital to the interpretation of documents, but they often eschew identifiers such as "Figure 1" for inexplicit phrases like "in this figure" or "from these premises". We examine inexplicit references to DEs, termed *DE references*, and recast the problem of their automatic detection into the determination of relevant word senses. We then show the feasibility of machine learning for the detection of DE-relevant word senses, using a corpus of human-labeled synsets from WordNet. We test cross-domain performance by gathering lemmas and synsets from three corpora: website privacy policies, Wikipedia articles, and Wikibooks textbooks. Identifying DE references will enable language technologies to use the information encoded by them, permitting the automatic generation of finely-tuned descriptions of DEs and the presentation of richly-structured information to readers.

1 Introduction

It is rare that communication in a written document is a simple linear endeavor. Writers make use of orthographic, paralinguistic, and discursive structures to augment and enhance what they write. These structures commonly include figures, tables, sections, subsections, extended quotations, examples, arguments, summaries, and other means of organizing the communication channel. Such document entities (*DEs*, for brevity) may be linguistic or pictorial, and they may be well-delineated or vaguely bounded. Additionally, they may be entirely distinct from the prose or embedded in it.

DEs are necessarily connected to the text that they appear with (or subsume) in a document. Although the relationship may be implicit, a referring expression is often used to make a local connection. When style permits, these referring expressions may use identifiers for DEs such as "Table 4" or "Problem #3". However, phrases like "this table" or "this section" are also used, with the assumption that the reader can decode them. Consider the following sentences:

(1) This table shows the augmented performance statistics.
(2) The ideas in this section are new.

Notably, the referents of *table* and *section* in the above examples differ from those below:

(3) This table should be moved to the kitchen.
(4) The shelves in this section are unfinished.

To understand (1) or (2) (in contexts with referents), the reader must realize that *table* and *section* refer to DEs rather than entities in another class of referents, as in (3) or (4). The presence or absence of potential referents may help; however, the (1)/(3) and (2)/(4) distinctions are clear even out of context. This suggests that differing word senses are responsible.

References to DEs (*DE references*, for brevity) are frequent in text written to inform, and they profoundly affect the referential structure and practical value of passages that contain them. Entity linking and coreference resolution address similar phenomena, but systems for those tasks are unsuitable for DE references (as explained in Section 3). Little has been done to empirically understand DE references or automatically iden-

tify them in text, which would allow language technologies to exploit links between DEs and discourse context. This would enable the tagging of DEs with precise descriptive information from referring text, enabling (for example) relevance-based caption generation for DEs, automatic document layout generation, and tools to help readers quickly skim documents for specific resources or explanations of those resources.

This paper presents results on developing a method to automatically label noun word senses that represent references to DEs. This was done using logistic regression and a selection of features from synsets in the English WordNet (Fellbaum, 1998), from which word senses were sampled. To give the task a practical focus, word senses were selected for words in deictic phrases from three corpora: the set of featured textbooks from Wikibooks, a random selection of articles from Wikipedia, and a selection of privacy policies from popular websites. Wikibooks was selected because prior work has noted a high density of DE references. Wikipedia was selected for the informative value of its text, which differs in style and purpose from Wikibooks. The domain of privacy policies was chosen as a strong contrast with the other two domains, and for the potential benefits of downstream research to reduce reader confusion (Reidenberg et al., 2014). The diversity of these corpora also provided an opportunity for cross-domain evaluation.

The contributions of this work are threefold:

- The first evaluation results for using machine learning to discriminate between DE-referential and non-DE referential word senses, establishing a baseline for the task;
- A corpus of word senses (synsets) labeled for DE-referential capacity, with a rich diversity of DEs identified by them; and
- A procedure for extracting strong candidates for DE reference from a document along with the DE structure of the document.

Although we do not identify instances of DE reference in text, the results of this work create a bridge to existing work on word sense disambiguation, making feasible the goal of DE reference detection. This goal is also supported by the domain flexibility of the results. The corpus of word senses was labeled in a domain-agnostic fashion, and the use of WordNet enables easy labeling of additional word senses not covered by the present work (e.g., for new corpora).

The remainder of this paper is structured as follows. Section 2 summarizes a prior study of DE reference, with examples of the phenomenon and differences from the present work. Several related topics are reviewed in Section 3. Section 4 details the collection of word senses and the manual annotation process. In Sections 5 and 6, the procedure for the automatic labeling of synsets is presented, along with results for intra-domain and cross-domain labeling. We conclude with a discussion of the significance of these results and some directions for future work.

Category	Examples
Structural	Many of the resources listed elsewhere in **this section** have…
	In **this chapter**, we will show you how to draw…
Illustrative	Consider **these sentences**: [followed by example sentences]
	[following a source code fragment] …the first time the computer sees **this statement**, 'a' is zero, so it is less than 10.
Discourse	Utilizing **this idea**, subunit analogies were invented…
	In **this case**, you've narrowed the topic down to "Badges."
Non-DE Reference	Devices similar to resistors turn **this energy** into light, motion…
	What type of things does a person in **that career field** know?

Table 1. Examples of candidate instances from the prior study. Bold text denotes the determiner and head noun in each instance.

2 Background

The present work builds upon findings from a prior study of word senses relevant to DE reference (Wilson & Oberlander, 2014). There, the set of 122 English Wikibooks[1] textbooks with printable versions was selected as a corpus. The set contained eleven subject areas, including computing, humanities, sciences, and languages. This corpus was chosen for several reasons. Among the alternatives, it provided the largest volume of text with a reuse-friendly license. It addressed a diverse set of topics with text written to inform, thus implying a diverse set of DEs. Additionally, the corpus represented the collaboration of a large number of writers.

Phrase templates were used to gather candidate instances of DE reference. These templates consisted of noun phrases beginning with the demonstratives *this*, *that*, *these*, and *those*. A subset of the candidates was read and annotated

[1] http://en.wikibooks.org/

with categories, shown in Table 1. Three varieties of DE reference emerged: structural (i.e., reference to divisions of a document or the document in its entirety), illustrative (to DEs that present information in non-prose form), and discourse (to DEs embedded in the prose). The researchers estimated that 48% of candidate phrases were examples of DE reference.

Directly labeling large numbers of candidate instances proved to be time-consuming, and instead work focused on labeling the word senses (from WordNet) of the 27 most frequent nouns in candidate instances. These senses were manually labeled by reading their definitions to judge their ability to refer to DEs. By fitting the labeled DE senses into the WordNet ontology, observations became possible on the kinds of entities that served as DEs. For example, DEs were more likely to be abstractions than physical entities.

The word sense annotations from the prior study showed that, for 15 of the 27 examined nouns, the first (most common) word sense of the noun was able to refer to a DE. They also illustrated a permeable boundary between DEs thought of as discourse entities and DEs that reside outside of the prose. For example, *a question raised for consideration or solution* (the definition of *problem.n.02*) could refer to a question embedded in informative prose or an orthographically-distinct exercise in a problem set.

3 Related Work

Prior studies showed the communicative value of multiple representations and their tight integration, motivating the present work. Mayer (2009) presented the cognitive theory of multimedia learning and explored how pictorial DEs augment and enhance textual artifacts. Similarly, Ayres and Sweller (2005) argued that learning materials should be presented so that "disparate sources of information are physically and temporally integrated". Power, et al. (2003) argued for "abstract document structure as a separate descriptive level in the analysis and generation of written texts", further motivating our work.

The aggregation of word senses discussed in the present work has a precedent in supersense tagging (Ciaramita & Johnson, 2003), especially for Wikipedia text (Chang, Tsai, & Chang, 2009). Notably, one of WordNet's lexicographer files is *noun.communication*, which contains "nouns denoting communicative processes and contents" ("WordNet 3.0 Reference Manual", 2012). However, the set of senses in this file is a poor match for current purposes, as it includes many senses that do not fit a written or document-oriented context (for example, a word sense for *airwave* is included in the file). The present work also identifies several DE senses outside of this lexicographer file. Overall, the meta-communicative focus of the present work is novel compared to prior efforts.

The task of automatically identifying instances of DE reference bears some similarity to coreference resolution. However, coreference resolvers are not suited for the present task; those tried by the researchers include CoreNLP (Recasens, de Marneffe, & Potts, 2013), ArkRef (O'Connor & Heilman, 2013) and the work of Bengtson and Roth (2008). One problem is that many DEs are partly pictorial or are not recognized by NLP tools as cohesive entities. Many DEs are distinguished by their non-linguistic aspects (i.e., diagrams) or stylistic markup (bulleted lists, quotations delimited by quote marks).

The task at hand also has commonalities with entity linking (Hachey et al., 2013) and Wikification, the process of linking named entities in text with corresponding Wikipedia pages (Cheng & Roth, 2013). However, DEs differ markedly from named entities. DEs vary widely in their representation and they often reside in the same communication medium as references to them. References to DEs often incorporate pragmatic information: for example, the referent of "this figure" may be the closest figure or the one most recently referred to. The potentially non-textual nature of DEs also separates them from mentioned language (Wilson, 2012), although the phenomena share a metalinguistic quality.

Shell nouns are nouns used anaphorically to refer to complex concepts such as points, assumptions, acts, or feelings (Schmid, 2000). Their referents intersect with DEs, although neither set subsumes the other: Schmid's taxonomy of shell nouns does not include typical DE-referential nouns like *section*, *figure*, or *list*, yet it does include non-DEs like *fury*, *miracle*, and *pride*. Kolhatkar and Hirst (2014) have automatically detected referents of some shell nouns, but their methods share the limitations of coreference resolvers, as described above.

4 Synset Collection and Labeling

The prior study of DE senses provided groundwork for the study of DE reference, but the dataset it created lacked the size and diversity for appreciable machine learning results. This sec-

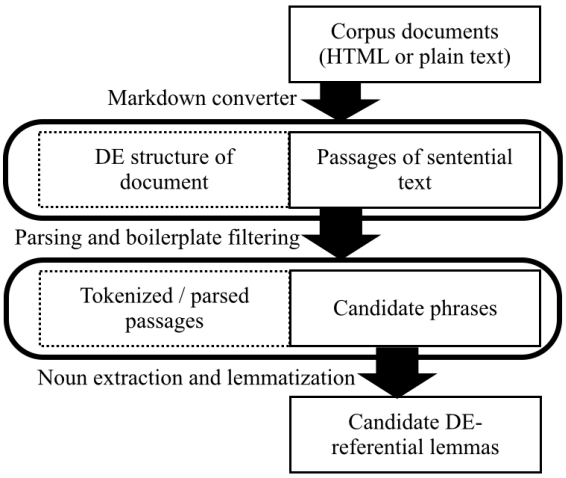


Figure 1. Pipeline used to process the corpora.

tion describes a procedure used to collect and label more word senses. A processing pipeline collected promising lemmas from three corpora, and a manual labeling procedure resulted in synset labels agreed upon by multiple annotators.

4.1 Processing Pipeline

An eventual goal of this research is to link DE references with their referents, and a processing pipeline was constructed to retain document features to enable that task. Although DE reference-referent linking is not a contribution of this paper, we present a pipeline that enables DE inventorying for two reasons. First, it illuminates our procedure for collecting lemmas for sense labeling. Second, it shows a method for preserving valuable information on orthographically-structured DEs in web documents. Such information is generally discarded by text processing pipelines. This pipeline shares some motivation with work by Poesio et al. (2011) on document structure, but the present work retains structure inline with contents, simplifying analysis.

Figure 1 shows the pipeline stages. The input consists of corpus documents in HTML format (or if HTML is unavailable, plaintext). Documents are first converted to Markdown (Gruber & Swartz, 2006), which preserves the orthographic organization of the text while simplifying the document to the extent that it can (if desired) be read as plaintext. Items such as titles, sections, lists, tables, and block quotations are shown in the output of the Markdown converter using ASCII symbols (e.g., asterisks for bullet points, hashes around section headers), but all HTML is removed. Inventorying the orthographically-structured DEs then becomes a simple matter of parsing Markdown syntax and recording the character indices where each DE begins and ends. This approach avoids the need for a complex parser to directly handle the variability and complexity of DEs represented in HTML.

Statistic	Privacy Policies	Wikipedia	Wikibooks
Documents	1010	500	149
Words	2646864	720013	5429978
Cand. Phrases	34181	2371	47546

Table 2. Statistics on each of the three corpora.

After conversion to Markdown, boilerplate text is discarded[2], and the remaining passages are part-of-speech tagged and parsed with Stanford CoreNLP (Socher et al, 2013). Candidate phrases for DE reference are then gathered using dependency templates. These identify noun phrases beginning with demonstratives *this*, *that, these*, and *those*; such phrases were productive for gathering DE references in previous work. Two new templates were added for noun phrases containing *above* and *below*. These captured additional relevant phrases, such as "the above notation" and "the examples below". DE-referential nouns were gathered from candidate phrases, lemmatized, and ranked by frequency.

The prior study noted an informal correlation between lemma frequency in candidate phrases and fertility for DE reference. Also, it was unclear if less frequent DE-referential senses have different qualities. For those reasons, and because labeling word senses for *all* candidate lemmas was infeasible, two methods were used to sample lemmas from each corpus. The first was a "high-rank" sampling of the most frequent lemmas, continuing down the ranks until selections were collectively responsible for at least 200 synsets. The second was a smaller "broad rank" random sampling of 25% of the 100 most frequent lemmas, which included some in the long tail of the distribution. Care was taken to avoid any overlap between the broad rank and high rank lemma sets. [3]

Table 2 shows descriptive statistics for the three corpora, which consisted of:

- **Privacy Policies (PP)**: a corpus collected by Liu et al. (2014) to reflect Alexa Internet's assessment of the internet's most popular sites

[2] Sentences in each corpus were discarded if they appeared verbatim in ten or more corpus documents.

[3] The procedure differed slightly for Wikibooks. Its high rank sample consisted of the 27 most frequent lemmas, whose 200 synsets were labeled by the prior study. Those labels are reused in the present work.

Privacy Policies		Wikibooks		Wikipedia	
Lemma	Freq.	Lemma	Freq.	Lemma	Freq.
policy	5945	case	790	page	535
information	3862	license	687	article	168
site	2151	book	686	time	67
website	1233	page	574	year	27
statement	859	example	515	period	21
party	852	section	486	list	18
company	720	way	385	case	15
cookie	638	type	363	section	15
service	585	point	344	issue	15
page	462	equation	337	game	15

Table 3. The ten most frequent lemmas in candidate phrases in each of the three corpora.

```
For each synset's definition, perform
the following:

Imagine instantiating the type represent-
ed by the definition. Judge its suitabil-
ity for the following statements.

(1) [an instantiation of the type] is
intended to communicate.

(2) [an instantiation of the type] can
be produced in a document or as a doc-
ument to convey information.

If both of the above statements are
coherent, mark 'y' for the definition.
Otherwise, mark 'n'.
```

Figure 2. Labeling rubric for the synsets.

```
y: table.n.01: a set of data arranged in
rows and columns

n: table.n.02: a piece of furniture hav-
ing a smooth flat top that is usually
supported by one or more vertical legs

n: table.n.03: a piece of furniture with
tableware for a meal laid out on it
```

Figure 3. Examples of synset labels.

Set Name	PP	WB	WP
High Rank	205 (35/170)	200 (62/138)	200 (28/172)
Broad Rank	57 (21/36)	93 (16/77)	136 (26/110)

Table 4. Sizes of the sets of synsets, along with their label compositions (positive/negative).

- **Wikibooks (WB)**: all English books with printable versions
- **Wikipedia (WP)**: random English articles, excluding disambiguation and stub pages

Table 3 shows the most frequent lemmas in candidate phrases, illustrating topical differences between corpora. The frequency distribution for Wikibooks showed a "heavier tail", as the text in its candidate phrases was more varied. It was hypothesized that this was not a reflection of a greater diversity of DEs, but instead showed a larger variety of references to non-DE entities fitting the phrase templates. The results of synset labeling appeared to validate this hypothesis.

4.2 Manual Annotation of Synsets

Using WordNet, all word senses were collected for all high rank lemmas. For broad rank lemmas, word senses were collected only if they were not present in the union of the sets of synsets gathered for the high rank lemmas. The total union of these collections was a set of 723 unique synsets. 200 of them were labeled in the prior study, and the researchers used a similar procedure (Figure 2) to label those remaining. Figure 3 shows some example labels. One annotator produced labels for all 523 new synsets, and two annotators respectively labeled new synsets in the high rank and broad rank samples. Thus, each new synset was labeled twice. Annotators worked independently and met to resolve differences. To promote domain-independent results, annotators were unaware which corpus (or corpora) triggered the inclusion of each synset.

Kappa values between the annotators who labeled the high rank set and the broad rank set were 0.60 and 0.72, respectively. Although kappa is an imperfect agreement metric (Carletta, 1996), these values are generally regarded as moderate to substantial (Viera & Garrett, 2005). The contrast in kappa values mostly arose from differing interpretations of the DE status of psychological entities. All annotators agreed that it was challenging to determine the degree of their presence in a document and thus their DE status.

Table 4 summarizes the results of labeling, with positive and negative representing "y" and "n" marks respectively. The numbers do not sum to 723 (the total number of unique synsets labeled) due to redundancies among the sets of synsets. Since the broad rank sets did not include any synsets in the union of the high rank sets, the sizes of the broad rank sets reflect differing vocabulary diversity. Lemmas from Wikipedia diverged furthest from the vocabulary of the other corpora, producing a much larger broad rank set.

5 Automatic Labeling of Synsets

The present work substantially increased the number of DE-labeled synsets available, but the intensity of the labeling task still constrained the volume of new labels generated. This limitation partly shaped the experimental procedure, and it

Name (Type)	Description
ss_rank (numeric)	Rank of synset for its namesake lemma (e.g., *2* for *section.n.02*)
ss_depth (numeric)	Length of shortest hypernym chain from the instance-synset to the noun root synset
hyper_*synset* (binary)	Presence of *synset* in the shortest hypernym chain from the instance-synset to the root noun synset
gloss-self_*word* (binary)	Presence of *word* in the instance-synset's definition
gloss-hypo_*word* (binary)	Presence of *word* in the definitions of the instance-synset's hyponyms

Table 5. Features used to classify synsets.

also reinforced the motivation for automatic, domain-independent labeling of DE synsets.

5.1 Classifier and Feature Set

Preliminary experiments with the labeled data from the prior study compared the advantages of various supervised learning algorithms and feature sets. A diverse sample of classifiers was tried using Weka (Hall et al., 2009), which led to the selection of its implementation of logistic regression. Other classifiers showed substantially lower precision and recall, regardless of parameter adjustments. SMO (Keerthi et al., 2001) was the runner-up for selection, with a potentially insignificant difference in F-score for most runs.

Table 5 describes features extracted for each instance (i.e., for each labeled synset). A total of 3607 features were generated. *ss_rank* and *ss_depth* characterize the vicinity of a synset in the ontology but are agnostic to its semantic properties. The *gloss-self_*word and *gloss-hypo_*word feature families were intended to exploit words used often to describe DEs (*writing*, *message*, etc.) or their hyponyms[4]. Finally, the *hyper_*synset feature family exploited varying concentrations of DE senses in the ontology.

Two additional binary feature families were considered. These were *hypo_*synset (presence of *synset* in the hyponym closure of the instance-synset) and *gloss-hyper_*word (presence of *word* in the definitions of the immediate hypernyms of the instance-synset). However, these had negligible effects on classifier performance.

[4] Incidentally, the annotators found that hyponyms of DE senses were not assured to be DE senses as well. This was partly due to vagueness in synset definitions. We also recognize that the ontology cannot reflect all use cases (such as ours) with equal precision.

5.2 Evaluation Protocol

Evaluation was devised to answer four questions:

(Q1) How difficult is it to automatically label DE senses if the classifier is trained with data from the same corpus?

(Q2) How difficult is the above task when using training data from a different corpus?

(Q3) For intra-corpus training and testing, are there differences in classifier performance between corpora?

(Q4) Are correct labels harder to predict for the broad rank set than for the high rank set?

To answer these questions, the classifier was run on a total of 33 different train-test set pairs or configurations. The limited quantity of labeled data posed a challenge to evaluation, and it was partly mitigated by performing all the aforementioned preliminary experiments on the Wikibooks high rank set (i.e., the data obtained from the prior study). Also, the broad rank synsets for all corpora were segregated from the rest of the labeled data and unexamined prior to evaluation.

The following classifier trials were performed, addressing the questions as indicated:

(T1) Leave-one-out cross validation (LOOCV) on each high rank set (Q1, Q3)

(T2) Training on a corpus' high rank set and testing on its broad rank set (Q1, Q3, Q4)

(T3) Training on 1 or 2 high rank sets and testing on the remaining high rank set(s) (Q2)

(T4) Training on 1 or 2 high rank sets and testing on the broad rank set(s) for the other corpus or corpora (Q2, Q4)

It was noted that, for each corpus, the positive/negative ratio for the high rank set differed from the ratio in the broad rank set. Accordingly the broad rank sets were resampled prior to T2 and T4 to contain equivalent ratios to their high rank counterparts. Additionally, some duplication of contents was observed between the high rank sets, complicating T3. Having an intersection between the train and test sets accurately reflected corpus composition, but it also biased the classifier. Thus, we generated performance statistics twice for each T3, with the intersection included and excluded from the test set.

6 Results

We first discuss the results of the classifier trials, and then add observations on a potential performance ceiling and the most valuable features.

		LOOCV	Cross-Train (1)			Cross-Train (2)		
			PP	WB	WP	PP/WB	PP/WP	WB/WP
Evaluation Set	PP	.53/.89/.67	-	.55/.86/.67	.94/.43/.59	-	-	.61/.89/.72
				.41/.77/.53	.91/.33/.49			.46/.81/.59
	WB	.68/.77/.72	.90/.60/.72	-	.96/.36/.52	-	.85/.79/.82	-
			.86/.49/.62		.92/.23/.37		.77/.70/.73	
	WP	.44/.79/.56	.80/.43/.56	.57/.86/.69	-	.67/.86/.75	-	-
			.70/.30/.42	.44/.78/.56		.52/.77/.62		

Table 6. Performance statistics (precision/recall/f-score) for the logistic regression classifier when trained and evaluated on high rank sets. Shaded cells show intersection-included performances.

		Same Corpus (High Rank)	Cross-Train (1)			Cross-Train (2)		
			PP	WB	WP	PP/WB	PP/WP	WB/WP
Eval. Set	PP	.33/.57/.42	-	.36/.71/.48	.55/.86/.67	-	-	.33/.57/.42
	WB	.61/.69/.65	.60/.56/.58	-	.34/.61/.44	-	.56/.56/.56	-
	WP	.34/.61/.44	.34/.72/.46	.43/.67/.52	-	.43/.72/.54	-	-

Table 7. Performance statistics (precision/recall/f-score) for the logistic regression classifier when training on the indicated high rank sets and predicting labels for the broad rank sets.

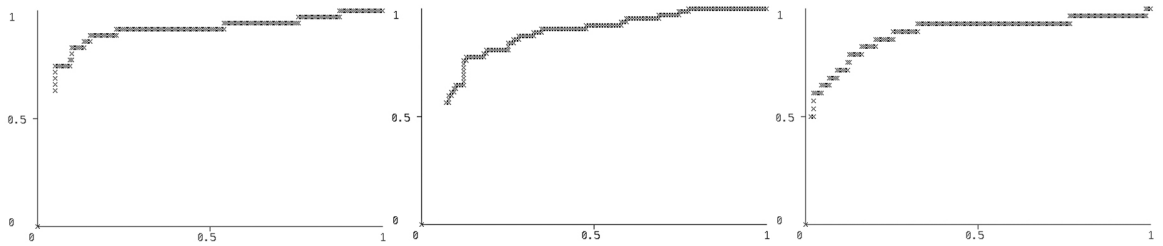


Figure 4. ROC curves (false positive rate on the horizontal axis and true positive rate on the vertical axis) for the logistic regression classifier with LOOCV on the high-rank sets.

Privacy Policies		Wikibooks		Wikipedia	
Info. Gain	Feature	Info. Gain	Feature	Info. Gain	Feature
.28284	hyper_communication.n.02	.18307	hyper_communication.n.02	.05860	hyper_part.n.01
.11949	hyper_written_communication.n.01	.08880	gloss-self_written	.05860	gloss-hypo_issue
.10539	gloss-self_written	.07950	gloss-hypo_written	.05860	gloss-hypo_author
.09347	hyper_abstraction.n.06	.07077	hyper_written_communication.n.01	.05529	gloss-hypo_newspaper
.07786	hyper_writing.n.02	.06694	hyper_writing.n.02	.05529	hyper_creation.n.02
.07226	hyper_message.n.02	.05398	ss_rank	.04794	hyper_communication.n.02
.07138	gloss-hypo_written	.05219	gloss-hypo_page	.04550	gloss-hypo_year
.06612	hyper_object.n.01	.04513	hyper_message.n.02	.04358	gloss-hypo_bill
.06440	gloss-hypo_document	.04328	gloss-hypo_question	.04358	gloss-hypo_publication
.06089	hyper_physical_entity.n.01	.04328	gloss-hypo_statement	.04150	hyper_product.n.02

Table 8. The highest-ranked features by information gain for the three high-rank sets.

6.1 Task Performance

Table 6 shows performance statistics for the trials that trained and evaluated with high rank sets (T1 and T3). In this table (and in Table 7) columns specify training sets and rows specify evaluation sets. F-scores for overlap-excluded runs varied from .37 (training on Wikipedia and evaluating on Wikibooks) to .73 (training on privacy policies/Wikipedia and testing on Wikibooks). For perspective, these figures are similar to the state of the art for overall labeling of discourse relations (Lin, Ng, & Kan, 2014) or discourse mentions (Recasens et al., 2013). The performance figures shown in Tables 6 and 7 are for the positive class only; overall weighted accuracy figures were generally .8 or higher.

The precision-recall gap was largest for runs trained on Wikipedia and tested on the other two sets. Manual inspection of errors from those two runs showed that the model made correct predictions for DE senses that closely resembled those in Wikibooks and Wikipedia but missed a variety of more esoteric DE senses. It appeared that non-DE suggestive lemmas had a relatively strong presence in Wikipedia's high rank sample, leading to impoverished training. This was reflected

by the relatively low ratio of positive labels in Wikipedia's high rank set. In contrast, Wikibooks' diverse positive instances led to higher recall when its high rank set was used as training.

High rank cross-training results varied widely: some exceeded LOOCV performance and some fell below it. It appeared that training on two corpora produced better results than training on one, which validates intuitions on the advantages of a diverse (and larger) training set. Also as expected, intersection-inclusive performances were superior to their exclusive counterparts.

Table 7 shows performance statistics for the trials that were trained using the high rank sets and evaluated with the broad rank sets (T2 and T4). Resampling of the high rank sets (described in 5.2) meant that there were few positive instances in them, with 7, 16, and 18 respectively for privacy policies, Wikibooks, and Wikipedia. Lower performances were a consistent trend in comparison to T1 and T3. It appeared that many (if not most) of the prediction errors involved entities that were close to the conceptual border between discourse DEs and non-DE psychological entities. This aligns with the researchers' observations on manual labeling agreement, suggesting that a practical ceiling exists for classifier performance on the task as currently conceived.

6.2 Additional Analysis

Figure 4 shows receiver operating characteristic (ROC) curves for the LOOCV high rank runs (T1). All three show a drawback of achieving high recall for the task: many DE synsets resist correct classification without a high tolerance for false positives. ROC curves for cross-training runs were similar. These observations resemble prior results on *mentioned language*, a related metalinguistic phenomenon for which many positive instances appear to lack reliable predictive features (Wilson, 2013). On the other hand, labeling a small "core" group of positive instances with high precision seems possible.

Finally, information gain was used to rank the utility of features for T1, and Table 8 shows the results. The *hyper*_synset and *gloss*_hypo feature families dominated the top features for all corpora. The strength of *hyper*_synset was expected, given prior observations of DE "neighborhoods" in the ontology. The strength of *gloss*_hypo (and the relative absence of *gloss*_self) was not expected, though an intuitive explanation for it exists: the aggregated vocabulary of multiple hyponyms' definitions provides more robust evidence for a synset's DE status than its own definition.

7 Discussion

The difficulty in identifying DE synsets is substantial; specifically, recall poses a challenge for the current prediction scheme. However, training on one corpus' high rank set and testing on a different corpus' set produced results that were not consistently better or worse than LOOCV, which suggests that labeling synsets gathered for a new domain (or all of WordNet) is no less feasible. These observations answer Q1 and Q2.

Toward Q3, some variation seemed to exist: for intra-corpus runs (T1 and T2), Wikibooks synsets produced the highest score and Wikipedia synsets produced the lowest. However, this ordering may be the result of differing positive-negative label ratios, and it did not hold for cross-training. The answer to Q3 may be a nominal affirmation: the label ratio, which varies by corpus, naturally affects classifier performance.

Finally, Q4 is simpler to answer: evaluating on broad rank sets generally produced worse performances than evaluating on high rank sets. The greater prevalence of discourse and psychological entities in broad rank sets seemed to be responsible. Excluding discourse entities from the class of DEs may appear to be an effective *ad hoc* solution, but it causes a new problem: many DEs appear interchangeably as orthographic or prose-embedded entities (e.g., lists, which may appear in bullet form or in a sentence). Since phrases that refer to DEs do not distinguish between the two, the exclusion of discourse entities would create further artificial distinctions.

8 Conclusion and Future Work

In this paper we presented a method for automatically identifying word senses that refer to document entities. Evidence suggests that identifying non-discourse DE senses was attainable with high precision and recall, but the ambiguities of discourse DEs—which were in some ways inseparable—poses a problem. We also introduced a corpus of DE-labeled word senses from three domains and a method for extracting orthographically-structured DEs from web documents. These contributions enable future work on the automatic detection of DE reference and the development of associated applications.

The use of these results toward DE supersense tagging and referent identification is a clear next step. The researchers have experimented with a prototype DE reference tagger, and preliminary results suggest that integrating tagging and referent identification may be advantageous. A low-

precision high-recall DE reference tagger will produce many false positives, but the availability of (or lack of) referents for each instance may serve as a sieve to eliminate those false positives.

Acknowledgements

This research was supported in part by the National Science Foundation under grants OISE 11-59236 (Metalanguage Identification for Interactive Language Technologies) and CNS 13-30596 (Towards Effective Web Privacy Notice & Choice: A Multi-Disciplinary Perspective).

References

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press.

Bengtson, E., & Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proc. EMNLP* (pp. 294–303). Stroudsburg, PA, USA: Association for Computational Linguistics.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2), 249–254.

Chang, J., Tzong-Han Tsai, R., & S. Chang, J. (2009). WikiSense: Supersense tagging of Wikipedia named entities based WordNet. In *Proc. PACLIC*.

Cheng, X., & Roth, D. (2013). Relational inference for wikification. *Urbana*, *51*.

Ciaramita, M., & Johnson, M. (2003). Supersense tagging of unknown nouns in WordNet. In *Proc. EMNLP* (pp. 168–175). Stroudsburg, PA, USA: Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Gruber, J., & Swartz, A. (2006). *Markdown*. http://daringfireball.net/projects/markdown/syntax.

Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating Entity Linking with Wikipedia. *Artificial Intelligence*, *194*, 130–150.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, *13*, 637–649.

Kolhatkar, V., & Hirst, G. (2014). Resolving shell nouns. In *Proc. EMNLP*, pp. 499–510.

Lin, Z., Ng, H. T., & Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 151–184.

Liu, F., Ramanath, R., Sadeh, N. M., & Smith, N. A. (2014). A step towards usable privacy policy: Automatic alignment of privacy Ssatements. In *Proc. COLING*.

Mayer, R. E. (2009). *Multimedia Learning*. Cambridge University Press.

O'Connor, B., & Heilman, M. (2013). ARKref: a rule-based coreference resolution system. *arXiv:1310.1975 [cs]*. Retrieved from http://arxiv.org/abs/1310.1975

Poesio, M., Barbu, E., Stemle, E. W., & Girardi, C. (2011). Structure-preserving pipelines for digital libraries. In *Proc. ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 54–62). Stroudsburg, PA, USA: Association for Computational Linguistics.

Power, R., Scott, D., & Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, *29*(2), 211–260.

Recasens, M., de Marneffe, M., & Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Proc. NAACL HLT*.

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J. T., … Schaub, F. (2014). *Disagreeable privacy policies: Mismatches between meaning and users' understanding*. SSRN Scholarly Paper No. ID 241829. Rochester, NY: Social Science Research Network.

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Walter de Gruyter.

Socher, R., Bauer, J., Manning, C. D., & Andrew Y., N. (2013). Parsing with compositional vector grammars. In *Proc. ACL* (pp. 455–465).

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wilson, S. (2012). The Creation of a Corpus of English Metalanguage. In *Proc. ACL* (pp. 638–646).

Wilson, S. (2013). Toward automatic processing of English metalanguage. In *Proc. IJCNLP* (pp. 760–766).

Wilson, S., & Oberlander, J. (2014). Determiner-Established deixis to communicative artifacts in pedagogical text. In *Proc. ACL*.

WordNet 3.0 Reference Manual. (2012). Cognitive Science Laboratory, Princeton University. https://wordnet.princeton.edu/wordnet/documentation/.

WordNet and beyond : the case of lexical access

Michael Zock
AMU, LIF, UMR 7279
163, Avenue de Luminy
13288 Marseille / France
zock@free.fr

Didier Schwab
Univ. Grenoble Alpes
LIG - GETALP
Campus de Grenoble / France
didier.schwab@imag.fr

Abstract

For humans the main functions of a dictionary is to store information concerning words and to reveal it when needed. While *readers* are interested in the meaning of words, *writers* look for answers concerning usage, spelling, grammar or word forms (lemma). We will focus here on this latter task : help authors to find the word they are looking for, word they may know but whose form is eluding them. Put differently, we try to build a resource helping authors to overcome the tip-of-the-tongue problem (ToT).

Obviously, in order to access a word, it must be stored somewhere (brain, resource). Yet this is by no means sufficient. We will illustrate this here by comparing WordNet (WN) to an equivalent lexical resource bootstrapped from Wikipedia (WiPi). Both may contain a given word, but ease and success of access may be different depending on other factors like quality of the query, proximity, type of connections, etc. Next we will show under what conditions WN is suitable for word access, and finally we will present a roadmap showing the obstacles to be overcome to build a resource allowing the text producer to find the word s/he is looking for.

1 Introduction

When speaking or writing we encounter basically either of the following two situations: one where everything works automatically (Segalowitz, 2000), somehow like magic, words popping up one after another as in a fountain spring, leading to a discourse where everything flows like in a quiet river (Levelt et al. 1999; Rapp and Goldrick, 2006) The other situation is much less peaceful : discourse being hampered by hesitations, the author being blocked somewhere along the road, forcing him to look deliberately and often painstakingly for a specific, possibly known word (Zock et al. 2010; Abrams et al. 2007; Schwartz, 2002; Brown, 1991).

We will be concerned here with this latter situation. More specifically, we are concerned here with authors using an electronic dictionary to look for a word. While there are many kind of dictionaries, most of them are not very useful for the language producer. The great majority of dictionaries are semasiological, that is, words are organized alphabetically. Alas, this kind of organisation does not fit well the language producer whose *starting points* (input) are generally meanings [1], and only the *end point* (outputs) the corresponding target word. While it is true that most dictionaries have been built with the reader in mind, one must admit though that attemps have been made to assist also the writer. The best known example is probably Roget's Thesaurus (Roget, 1852), but as we will see, there is also WordNet (Miller, 1990; Fellbaum, 1998) [2], a very special kind of resource integrating in a single place information 'normally' spread over different dictionaries. Rather than creating different volumes for different tasks (allowing the user to find a definition, synonyms, antonyms, etc.), WordNet (WN) has integrated all these functions into a a single resource. As its spirit is closest to what we have in mind, we will focus

[1] More or less well specified thoughts (concepts, elements of the word's definition), or somehow related elements : collocations, i.e. associations (elephant: tusk, trunk, Africa).

[2] For other pointers to onomasiological dictionaries, see (Zock et al. 2010).

on it in this paper, commenting on its strenghts and weaknesses with respect to word access.

This paper is organized as follows. We start by providing evidence that storage does not guarantee access. That this holds for humans has been shown already 50 years ago (Tulving and Pearlstone, 1966), in particular via Brown and Mc Neill's (1966) seminal work devoted to the *tip-of-the-tongue problem* (henceforth, ToT) [3]. We will show here that this can also hold for machines. The assumption that what is stored can also be accessed (anytime), is simply wrong. To illustrate our claim we will compare an extended version of WN (Mihalcea and Moldovan, 2001) to an equivalent resource based on Wikipedia.

Next, we will discuss under what conditions WN is adequate for word access, and finally, we will sketch a roadmap describing the steps to be performed in order to go beyond the Princeton resource. The goal is to build an index (association network) and navigational tools (categorial tree) to help authors to find the word they are looking for when being in the ToT state.

2 Storage does not guarantee access

To test this claim we ran a small experiment, comparing an extended version of WN and *Wikipedia*, which we converted into a lexical resource. Our goal was not so much to check the quality of WN or any of its extensions as to show, firstly, that storage does not guaranteee access and, secondly, that access depends on a number of factors like (a) quality of the resource within which the search takes place (organisation, completeness), (b) index, and (c) type of the query (proximity to the target) [4]. Having two resources built with different foci, our goal was to check the efficiency of each one of them with respect to word access. For practical reasons we considered only direct neighbors. Hence, we defined a function called *direct neighborhood*, which, once applied to a given window (sentence/ paragraph [5], produces all its co-occurences. Of course, what holds for *direct associations* (our case here), holds also for indirectly related words, that is, words whose distance >1 (mediated associations).

2.1 Examples and comparisons of the two resources

The table here below shows the results produced by *eXtended* WN and WiPi for the following, randomly given inputs : 'wine', 'harvest' or their combination 'wine + harvest'.

Input:	*Output* : eXtended **WN**	*Output* : **WiPi**
wine	**488 hits** grape, sweet, serve, France, small, fruit, dry, bottle, produce, red, bread, hold...	**3045 hits** name, lord characteristics, christian, grape, France, ... vintage (81st), ...
harvest	**30 hits** month, fish, grape, revolutionary, calendar, festival, butterfish, dollar, person, make, wine, first,...	**4583 hits** agriculture, spirituality, liberate, production, producing, ..., vintage (112th), ...
wine + harvest	**6 hits** make, grape, fish, someone, commemorate, person, ...	**353 hits** grape, France, vintage (3d), ...

Table 1: Comparing two corpora with various inputs

Our goal was to find the word 'vintage'. As the results show, 'harvest' is a better query term than 'wine' (488 vs 30 hits), and their combination is better than either of them (6 hits). What is more interesting though is the fact that none of these terms allows us to access the target, eventhough it is contained in the database of *x*WN, which clearly supports our claim that storage does not guarantee access. Things are quite

[3] The ToT problem is characterized by the fact that the author has only partial access to the word form s/he is looking for. The typically lacking parts are phonological (Aitchison, 2003). The ToT problem is a bit like an incompleted puzzle, containing everything apart from some minor small parts (typically, syllables, phonemes). Alas, not knowing what the complete picture (target, puzzle) looks like, we cannot determine the lacking part(s). Indeed, we cannot assume to know the target, and claim at the same time to look for it or any of its elements. Actually, if we knew the target (word) there wouldn't be a search problem to begin with, we would simply spell out the form.

[4] To show the relative efficiency of a query, we have developed a website in Java as a servlet which will soon be released on our respective homepages. Usage is quite straightforward: people add or delete a word from the current list, and the system produces some output. The output is an ordered list of words, whose order depends on the overall score (i.e. the number of co-occurrences between the input, i.e. 'source word' (S_w) and the directly associated words, called 'potential target word' (PT_w)). For example, if the S_w 'bunch' co-occured five times with 'wine' and eight times with 'harvest', we would get an overall score or weight of 13: ((wine, harvest), bunch, 13). Weights can be used for ranking (i.e. prioritizing words) and the selection of words to be presented, both of which may be desirable when the list becomes long.

[5] Optimal size is an empirical question, which may vary with the text type (encyclopedia vs. raw text).

different for an index built on the basis of information contained in WiPi. The same input, 'wine' evokes many more words (3045 as opposed to 488, with '*vintage'* in the 81st position). For 'harvest' we get 4583 hits instead of 30, 'vintage' occurring in position 112. Combining the two yields 353 hits, which pushes the target word to the third position, which is not bad at all.

We hope that this example is clear enough to convince the reader that it makes sense to use real text (ideally, a well-balanced corpus) to extract from it the information needed (associations) in order to build an index allowing users to find the elusive word.

One may wonder why we failed to access information contained in WN and why WiPi performed so much better. We believe that the relative failure of WN is mainly due to the following two facts: the size of the corpus (114,000 words as opposed to 3,550,000 for WiPi), and the number of syntagmatic links, both of which are fairly small compared to WiPi. Obviously, being an encyclopedia, WiPi contains many more syntagmatic links than WN. Of course, one could object that we did not use the latest release of WN (version 3.0) which contains many more words (147,278 words, clustered into 117,659 synsets). True as it is, this would nevertheless not affect our line of reasoning or our conclusion. Even in a larger lexical resource we may fail to find what we are looking for because of the lack of *syntagmatic links*. As mentioned already, the weak point is not so much the quantity of the data, as the quality of the index (the relative sparsity of links). Yet, in order to be fair towards WN, one must admit that, had we built our resource differently, for example, by including in the list of related terms, not only the directly evoked words, i.e. potential target words, but all the words containing the source-word (wine) in their definition (Bordeaux, Retsina, Tokay), then we would get 'vintage', as the term 'wine' is contained in its definition ('vintage': a season's yield of 'wine' from a vineyard). Note that in such cases even Google works often quite well, but see also (Bilac et al. 2004, El-Kahlout and Oflazer, 2004; Dutoit and Nugues, 2002).

Another noteworthy point is the fact that success may vary quite dramatically, depending on the input (quality of the query). As Table 2 shows, WN outperforms WiPi for the words 'ball', 'racket' and 'tennis'. Yet, WiPi does not lag much behind; additionally, it contains many other words possibly leading to the target words ("player, racket, court", ranked, respectively as numbers 12, 18 and 20).

Input:	*Output* : eXtended **WN**	*Output* : **WiPi**
ball	**346 hits** game, racket, player, court, volley, Wimbledon, championships, inflammation, ... , tennis (15th), ...	**4891 words** sport, league, football, hand, food, foot, win, run, game, ..., tennis (27th), ...
racket	**114 hits** break, headquarter, gangster, lieutenant, rival, kill, die, ambush, tennis (38th), ...	**2543 words** death, kill, illegal, business, corrupt, ..., tennis (72nd), ...
ball + racket	**11 hits** game, tennis, (2nd), ...	**528 hits** sport, strike, tennis (3d), ...

Table 2: Comparing two corpora with various inputs

Not being an encyclopedia, WN lacks most of them, though surprisingly, it contains named entities like 'Seles' and 'Graf', two great female tennis players of the past. Given the respective qualities of WN and WiPi one may well consider integrating the two by relying on a resource like *BabelNet* (Navigli and Ponzetto, 2012) [6]. This could be done in the future. In the meantime let us take a closer look at WN and its qualities with respect to word look up.

3 Under what condition is WN really good for consultation ?

Many people know that WN is based on psycholinguistic principles. What is less known though is the fact, that despite its psycholinguistic origines, it has never been built for consultation. It has been primarily conceived for usage by machines: "WordNet is an online lexical database designed for use under program control." (Miller, 1995, p. 39). This being said, WN can nevertheless be used for consultation, all the more as it is quite good at it under certains circumstances.

Remains the question under what conditions WN is able to reveal the elusive target word. We believe that it can do so perfectly well provided that the following three conditions are met :

(a) the *author knows* the *link* holding between the source word (input, say 'dog') and the target, e.g.
([dog]+*synonym* = [?] → [bitch]);
([dog]+*hypernym* = [?] → [canine]);

[6] http://lcl.uniroma1.it/babelnet/

(b) the *input* (source word) and the *target* are *direct neighbors* in the resource. For example,
[seat]-[leg] (*meronym*);
[talk]-[whisper] (*troponym*), ...

(c) the *link* is *part* of WN's database, e.g.
'hyponym/hypernym', 'meronym', ...

4 The framework of a navigational tool for the dictionary of the future

To access a word means basically to reduce the entire set of words stored in the resource (lexicon), to one (target). Obviously, this kind of reduction should be performed quickly and naturally, requiring as little time and effort (minimal number of steps) as possible on the users' side. Note that this process is knowledge based, meaning that the user may have stored the word and, if he cannot find it, he may nevertheless be aware of some other word(s) somehow connected to the target. This is a very important aspect, as we will start from that.

When we wrote that WN is quite successful with regard to word look-up under certain circumstances, we also meant to say that it is not so good when these conditions are not met. More precisely, this is likely to occur when :

(a) the source (input) and the target are only *indirectly* related, the distance between the two being greater than 1. This would be the case when the target ('Steffi Graf') cannot be found directly in reponse to some input ('tennis player'), but only via an additional step, say, 'tennis pro' : ([tennis player] → [tennis pro]); given as input at the next cycle, it will definitely reveal the target [7].

(b) the input ('play') and the target ('tennis') belong to different parts of speech (see 'tennis problem', Fellbaum, 1998);

(c) the prime and the target are linked via a *syntagmatic association* ('smoke'-'cigar'). Since the majority of relations used by WN connect words from the same part of speech, word access is difficult if the output (target) belongs to a different part of speech than the input (prime) [8];

(d) the user ignores the link, he cannot name it, or the link is not part of WN's repertory [9]. Actually this holds true (at least) for nearly all syntagmatic associations;

Let us see how to go beyond this. To this end we present here briefly the principles of the resource within which search takes place, as well as the required navigational aid (categorial tree) to allow authors to find quickly the word they are looking for. Yet, before doing so, let us clarify some differences between hierarchically structured dictionaries and our approach.

While lexical ontologists (LO) try to integrate all words of a language into a neat subsumption hierarchy, we try to group them only in terms of direct neighborhood, not mentioning at all the type of the link. Words are grouped later on by category (see, figure 1). This yields a quite different network than WN. Our graph is fully connected and, not being concerned with exhaustivity, we try to reveal only the words typically evoked by some input. This being so, our graph (or, any equivalent association network) will yield different results than WN for the same input (see table 3).

WN : *hypernym*: solid; *part_holonym*: nutrient; **hyponyms :** leftovers**,** fresh_food**,** convenience_food, chocolate, baked_goods, loaf, meat, pasta, health_food, junk_food, breakfast_food, green_goods, green_groceries, coconut, coconut_meat, dika_bread, fish, seafood, butter, yoghourt, cheese, slop
E.A.T : at, drink, good, thought, dinner, eating, hunger, salad, again, apple, baby, bacon, bread, breakfast, case, cheese, consumption, cook, firm, fish, France, goo, great, hungry, indian, kitchen, lamb, loot, meal, meat, mix, mouth, noah, nosy, of, pig, please, poison, rotten, sausage, steak, stomach, storage, store, stuff, time, water, yoghurt, yum

Table 3: The respective outputs produced by a *lexical ontology* (here WN) as opposed to an *association network* (here, the E.A.T).

Suppose we started from a broad term like 'food'. A LO like WN would produce the entire list of objects referring to 'food' (hyponyms), while an association network would only reveal typically evoked words {food, bread, noodles, rice, fish, meat, cook, eat, buy, starving, good, expensive, fork, chopsticks....}. This list contains, of course, a subset of the terms found in a LO (terms referring to 'food'), but also syntag-

[7] Note that the situation described is a potential problem for any association network. Note also that, eventhough Named Entities (NEs) are generally not contained in a lexicon, some of them have made it into WN. This is the case for some famous tennis players, like Steffi Graf. Anyhow, since NEs are also words, the point we are trying to make holds for both. Hence, both can be organized as networks, and whether access is direct or indirect depends on the relative proximity of the input (prime) with respect to the target word.

[8] This being said, WN does have cross-POS relations, i.e. "morphosemantic" links holding among semantically similar words : observe (V), observant (Adj) observation (N).

[9] For example : 'well-known_for', 'winner_of', ...

matically related words (*origine* : France; *state* : hungry, ...). Compare the results obtained by WN and the Edinburgh Association Thesaurus [10].

By taking a look at this second list one can see that it contains not only hyponyms, that is, specific kinds of food (meat, cheese, ...), but also syntagmatically related words (cook, good, France, ...), i.e. words typically co-occurring with the term 'food'. Note that our list may lack items like 'bagles', 'cheese' or 'olives. This is quite normal, if ever these words are not strongly associated with our input (food), which does not imply, of course, that we cannot activate or find them. Had we given 'wine' or 'oil' 'green' and 'Greece' as input, chances are that 'cheese' and 'olives' would pop up immediately, while they are burried deep down in the long list of food produced by a LO.

Let us return to the problem of word access. Just as orientation in real world requires tools (map, compass) we need something equivalent. While the *semantic map* defines the territory within which search takes place, the *lexical compass* guides the user, helping her or him to reach the goal (target word). Obviously, the terms map and compass are but metaphors, as there are important differences between world maps and lexical graphs (see below) on one hand, and compasses sailors use and the tool an information seeker is relying on (human brain) on the other. The map we have in mind is basically an association network. It is a fully connected graph encoding all directly associated words given some input. This kind of graph has many redundancies, and the links are not labeled. In this respect it is very different from WN and even more so from the maps we use when traveling in real world. Also, when using a world map the user generally knows more or less precisely the destination or relative location of the place he is looking for, for example, south of Florence. He may also be able to deduce its approximate location, eventhough she is not able to produce its name (Rome). This does not hold in the case of a user resorting to a lexical resource (map) based on associations. While the user may know the starting point (knowledge available when trying to find the target, the elusive word), he cannot name the destination (target), as if he could, there would be no search problem to begin with. The user either knows the word (in which case the problem is solved), or he does not. In this latter case all he can do is to rely on available knowledge concerning the target, an assumption we make here. Knowledge is fragmentary. Yet, incomplete as it may be, this kind of information may allow us to lead him to the target, guiding him in a reduced, clearly marked search space (details here below).

To get back to navigation in real world. In the case of spatial navigation it suffices to know that 'Rome' is south of 'Florence', which is part of 'Lazio', and that it can be reached by car in about 2 hours. Having this kind of knowledge we could initiate search in the area of 'Lazio', since 'Lazio' is an area south of 'Tuscany', the area containing 'Florence'. While this strategy works fine in the case of spatial navigation, it will not work with lexical graphs. In this kind of network terms are related in many ways and their strength may vary considerably. Hence, it is reasonable to show a term only if it is above a certain threshold. For example, a term A (Espresso) being connected to term B (coffee) may be shown only if it is sufficiently often evoked by B. Note that eventhough words are organized in terms of neighborhood, the link between them (explicited or not) may be of many other kinds than a spatial relation. In sum, the links connecting words in an associative network are much more diverse than the ones typically found in a lexical ontology.

As mentioned already, humans using world maps usually know the name of their destination, whereas people being in the ToT state do not. Yet, even if they did, they would not be able to locate it on the map. Lexical graphs are simply too big to be shown entirely on a small screen [11]. In sum, we need a different approach : search must be performed stepwise, taking place in a very confined space, composed of the input and the direct neighbors (directly associated words). It is like a small window moved by the user from one part of the graph to the next. If there are differences between world maps and association networks (lexical graphs), there are also important differences between a conventional compass and our navigational tool. While the former automatically points to the north, letting the user compute the path between his current location and the desired goal (destination, target), the latter (brain) assumes the user to know, the

[10] http://www.eat.rl.ac.uk

[11] Associative networks contain many redundancies and are potentially endless, since they contain loops. For example, an input, say 'Rome' may well appear to be the direct neighbor of one of its outputs, 'Italy' : ([Rome] → {[capital], [Italy], [city]}); ([Italy] → {[country], [France], [Rome]}).

goal, i.e. target word [12], or its direction (even if one does not know its precise location). While the user cannot name the goal —he has only passive knowledge of it,— the system cannot guess it. However it can make valuable suggestions. In other words, eventhough the system can only make suggestions concerning the target or the directions to go (which word to use as input for the next cycle), it is the user who finally decides whether the list contains the target or not, and if so, in what direction to go. He is the only one to know which suggestion corresponds best to the target (the word he has in mind) or which one of them is the most closely connected to it. Of course, the user may go wrong, but as experience shows his intuitions are generally quite good.

Let us now see quickly how to make this idea work. Imagine an author wishing to convey the name of a beverage commonly found in coffee shops (target : 'mocha'). Failing to do so, he reaches for a lexicon. Since dictionaries are too huge to be scanned from cover (letter A) to cover (Z), we suggest a dialog between the user and the computer to reduce incrementally the search space. The user provides the input [13], — word coming to his/her mind, generally a word more or less directly related to the target,— and the system makes a set of proposals (list of words), trying to guide the user on the basis of her input.

Suppose that the target were 'gull'. In such a case one might ask : 'do you know the name of a bird able to swim', having yellow feet, and a long beak [14]? To simplify matters and to convey as simply as possible the rationale underlying our approach (see figure 1, next page), let us assume that the input is a single word. The process consists basically in the following steps : (a) user input (query), (b) system output (answer), (c) user's choices concerning the target (does the list contain it?), or, choice of the word to continue search with. Concretely speaking this leads to the following kind of dialogue. The user starts by providing her input, that is, any word coming to her mind, word somehow connected to the target (step-1, figure 1) [15]. The system presents then in a clustered and labeled form (categorial tree) all direct associates (step-2, figure 1) [16]. The user navigates in this tree, deciding on the category within which to look for the target, and if he cannot find it in any of them, in what direction to go. If he could find the target, search stops, otherwise the user will pick one of the associated terms or provides an entirely new word and the whole process iterates. The system will come up with a new set of proposals.

As one can see, this method is quite straightforward, reducing considerably time and space needed for navigation and search. Suppose that you had to locate a word in a resource of 50.000 words. If your input triggered 100 direct associates, one of them being the target, then we would have reduced in a single step the search space by 99,8%, limiting navigation and search to a very small list. Suppose that our hundred words were evenly spread over 5 groups, than search would consist in spotting the target in a list of 25 items: 5 being category names and 20 being words within the chosen group.

A small note concerning the 2nd step. Step-2 yields a tree whose leaves are *potential target words* and whose nodes are *categories*, which while being also words are not at all the goal of the search. They are only the means to reach the goal. Put differently, their function is orientational, guide the user during his search.

[12] It has been shown over and over again that people being in the ToT state are able to identify immediately, and without making any mistakes the target word if it is shown to them, eventhough they could not name it. This is passive knowledge.

[13] This latter can be a single word —'coffee' in the case of target 'mocha'— or a set of words, which in a normal communicative setting would yield a sentence, where the information seeker asks someone else to help him to find the elusive word.

[14] This kind of wording can be generalized to a pattern for asking the following question: "What is the word for '[X] that [Y]?", where [X] is usually a hypernym and [Y] a stereotypical, possibly partial functional/relational/case description (action) of the target word. A similar pattern could be used for namefinding. For example, asking "What is the name of the <conqueror> of <empire>?" could yield 'Pizarro' or 'Cortés', depending on the value of the empire (Inca/Aztec). As one can see, the processes underlying wordfinding and namefinding are not very different.

[15] Note, that in order to determine properly the initial search space (step-1), we must have already well understood the input [$mouse_1$/$mouse_2$ (rodent/device)], as otherwise our list will contain a lot of noise, presenting 'cat, cheese' together with 'computer, mouse pad' {cat, cheese, computer, mouse pad}, which is not quite what we want, since some of these candidates are irrelevant, i.e. beyond the scope of the user's goal.

[16] This labeling is obligatory to allow for realistic navigation, as the list produced in response to the input may be very long and the words being of the same kind may be far apart from each other in the list. Hence it makes sense to structure words into groups by giving them appropriate (i.e. understandable) names so that the user, rather than looking up the entire list of words, searches only within a specific bag labeled by a category.

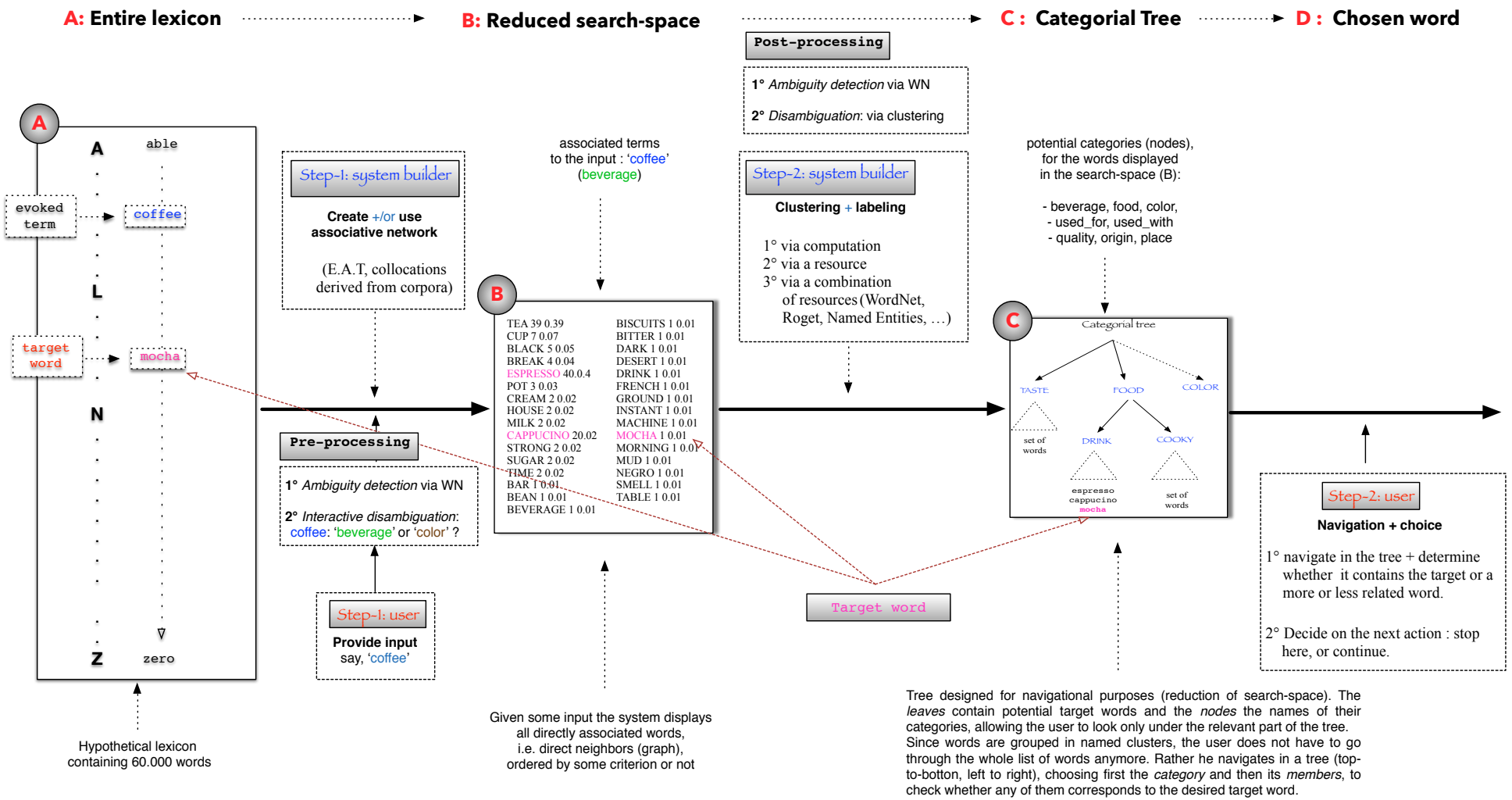


Figure 1 : Lexical access as a two-step dialogue

Words at the leave-level are potential target words, while the ones at the intermediate level (category names; preterminal nodes) are meant to reduce the number of words among which to perform search, and to help the user to decide on the direction to go. Hence, category names are reductionist and orientational (signposts), grouping terminal nodes into a bag, signaling via their name not only the bag's content, but also the direction to go. While the system knows the content of a bag, it is only the user who can decide which of the bags is likely to contain the elusive word. Because, eventhough he cannot name the target, he is the only one to know the target, be it only passively and in fairly abstract terms. This is where the categoy names have their role to play. In sum, it is not the system that decides on the direction to go next, but the user. Seeing the names of the categories she can make reasonable guesses concerning their content.

In sum, categories act somehow like signposts signaling the user the kind of words he is likely to find going one way or another. Indeed, knowing the name of a category (fruit, animal), the user can guess the kind of words contained in each bag (kiwi vs. crocodile). Assuming that the user knows the category of the searched word [17], she should be able to look in the right bag and take the best turn. Navigating in a categorial tree, the user can search at a fairly high level (class) rather than at the level of words (instances). This reduces not only the cognitive load, but it increases also chances of finding the target, while speeding up search, i.e. the time needed to find a word.

While step-1 is mainly a matter of 'relatedness' ('wine' and 'red' being different in nature, they are nevertheless somehow related), step-2 deals with 'similarity' : there are more commonalities between 'dogs' and 'cats' than between 'dogs' and 'trees'. Put differently, the first two terms are more similar in kind than the last two. The solution of the second step is certainly more of a challenge than the one of step-1 which is largely solved (eventhough there is an issue of relevance : not all co-occurences are really useful) [18]. To put words into clusters is one thing, to give them names an ordinary dictionary user can understand is quite another [19]. Yet, arguably building this categorial tree is a crucial step, as it allows the user to navigate on this basis. Of course, one could question the very need of labels, and perhaps this is not too much of an issue if we have only say, 3-4 categories. We are nevertheless strongly convinced that the problem is real, as soon as the number of categories (hence the words to be classified) grows.

To conclude, we think it is fair to say that the 1st stage seems to within reach, while the automatic construction of the categorical tree remains a true challenge despite some existing tools (word2vec) and the vast literature devoted to this topic or to strongly related problems (Zhang et al., 2012; Biemann, 2012; Everitt et al., 2011).

5 Conclusion

We have started the paper by pointing out the fact that word access is still a problem for dictionary builders and users (see also Thumb, 2004), in particular humans being in the production mode (Zock, 2015). Next, we showed that the fact that an item is stored in a lexical resource does not guarantee its access. We continued then to discuss why even a psycholinguistically motivated resource like WN often fails to reveal the word authors are looking for.

Finally, we presented a roadmap to overcome this problem. The idea is to build a resource guiding a human user allowing him to find the word he is looking. Given some input (user's knowledge concerning the target word), the system would provide the direct neighbors in a clustered and labeled form (output) to allow the user to check whether this tree contains the elusive word. While the system's task with respect to the user's input (step-1) is to reduce search space, the function of the second step is to support navigation. Just as it is unreasonable to perform search in the entire lexicon, is it cumbersome to drill down huge lists. This is why we suggested to cluster and label the outputs produced in response to the query. After all, we want users to find the target quickly and naturally, rather than drown them under a huge, unstructured (or poorly structured) list of words.

[17] A fact which has been systematically observed for people being in the ToT state who may tell the listener that they are looking for the name of a ”fruit typically found in a<PLACE>”, say, New Zealand, in order to get ’kiwi’.

[18] Take for example the Wikipedia page devoted to 'Panda', and check which of the co-occurrences are those typically evoked when looking for the word 'Panda'.

[19] For example, while the sequence of hypernyms listed by WN for *horse* captures much of the phylogenetic detail a biologist would want to see recorded (horse → equine → odd-toed ungulate → ungulate → placental mammal → mammal → vertebrate → chordate → animal → organism → entity), most of these terms mean next to nothing to an ordinary dictionary user.

Reference

Abrams, L., Trunk, D. L., and Margolin, S. J. (2007). *Resolving tip-of-the-tongue states in young and older adults: The role of phonology*. In L. O. Randal (Ed.), Aging and the Elderly: Psychology, Sociology, and Health (pp. 1-41). Hauppauge, NY: Nova Science Publishers, Inc.

Aitchison, J. (2003). Words in the Mind: an Introduction to the Mental Lexicon. Oxford, Blackwell.

Biemann, C. (2012). *Structure discovery in natural language*. Springer.

Bilac, S., Watanabe, W., Hashimoto, T., Tokunaga, T. and Tanaka, H. (2004). *Dictionary search based on the target word description*. In: Proc. of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004), pages 556-559.

Brown, R and Mc Neill, D. (1966). *The tip of the tongue phenomenon.* In: Journal of Verbal Learning and Verbal Behaviour, 5:325-337.

Brown A.S (1991), *The tip of the tongue experience A review and evaluation*. Psychological Bulletin, 10, 204-223

Dutoit, D. and P. Nugues (2002): *A lexical network and an algorithm to find words from definitions*. In Frank van Harmelen (ed.): ECAI2002, Proceedings of the 15th European Conference on Artificial Intelligence, Lyon, pp.450-454, IOS Press, Amsterdam.

El-Kahlout I. D. and K. Oflazer. (2004). *Use of Wordnet for Retrieving Words from Their Meanings*. 2nd Global WordNet Conference, Brno

Fellbaum, C. editor. (1998). *WordNet: An electronic lexical database and some of its applications*. MIT Press.

Levelt W., Roelofs A. and A. Meyer. (1999). *A theory of lexical access in speech production*. Behavioral and Brain Sciences, 22, 1-75.

Mihalcea, R. and D. Moldovan, (2001): Extended WordNet: progress report. In NAACL 2001 - Workshop on WordNet and Other Lexical Resources, Pittsburgh, USA.

Miller, G. A. (1995). *WordNet : A lexical database for English*. Communications of the ACM, 38 (11), 39–41.

Miller, G.A. (ed.) (1990): WordNet: An On-Line Lexical Database. International Journal of Lexicography, 3(4), 235-244.

Navigli, R. and Ponzetto, S. (2012), BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, pp. 217-250

Rapp, B. and Goldrick, M. (2006). Speaking words: Contributions of cognitive neuropsychological research. Cogntive Neuropsychology, 23 (1), 39-73

Roget, P. (1852). Thesaurus of English Words and Phrases. Longman, London.

Segalowitz, N. (2000). *Automaticity and attentional skill in fluent performance*. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 200-219). Ann Arbor, MI: University of Michigan Press.

Schwartz, B. L. (2002). *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical*. Mahwah. New Jersey: Lawrence Erlbaum Associates,

Thumb, J. (2004). Dictionary Look-up Strategies and the Bilingualised Learner's Dictionary. A Think-aloud Study. Tübingen: Max Niemeyer Verlag.

Tulving, E., and Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. Journal of Verbal Learning and Verbal Behavior, 5, 381-391

Zhang, Z., Gentile, A. and Ciravegna, F. (2012*). Recent advances in methods of lexical semantic relatedness – a survey*. Journal of Natural Language Engineering, Cambridge Universtiy Press, 19(4):411–479.

Zock, M., Ferret, O. and Schwab, D. (2010) Deliberate word access : an intuition, a roadmap and some preliminary empirical results, In A. Neustein (Ed.) *'International Journal of Speech Technology'*, Springer Verlag, 13(4):107-117.

Zock, M. (2015) 'Errare humanum est'. Refusing to 'appreciate' this fact could be a big *mistake* ! In Adda, G., Adda-Decker, M., Mariani, J., Mititelu, V., Tufis, D., Vasilescu , I. (eds). "Errors by Humans and Machines in multimedia, multimodal and multilingual data processing. Proceedings of ER-RARE 2015". Romanian Academy Publishing House

Author index