Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-153

Hybrid Approach for Automatic Identification of Multi-Word Expressions in Lithuanian

Justina MANDRAVICKAITĖ [a,1], Erika RIMKUTĖ [b] and Tomas KRILAVIČIUS [c]

[a] *Baltic Insitute of Advanced Technology, Vilnius, Lithuania*
[b] *Vytautas Magnus University, Kaunas, Lithuania*
[c] *Vytautas Magnus University, Kaunas, Lithuania and Baltic Institute of Advanced Technology, Vilnius, Lithuania*

Abstract Identification of MultiWord Expressions (MWE) is one of the most challenging problems in Computer Linguistic and Natural Language Processing. A number of techniques are used to solve this problem in different language, mostly English. However not all techniques and approaches can be directly transferred to Lithuanian. Hence, in this paper we experiment with automatic identification of bi-gram MWEs for Lithuanian, which is considered to be under-resourced in terms of lexical resources and availability or accuracy of special lexical tools (e.g., POS-taggers, parsers). We use a raw corpus and combination of lexical association measures and supervised machine learning, which was shown to perform well for English and some other languages. Using this approach we have reached 70.4% precision for identification of typical MWEs, 77.1% precision for non-typical MWEs as well as 60.0% and 81.6% precision for typical adjective + noun and noun + noun MWEs respectively.

Keywords. Multiword Expressions, Lithuania language, MWE identification

1. Introduction

In this paper we report our experiment with automatic detection of *multi-word expressions* (MWEs) in Lithuanian by combining *lexical association measures* (LAMs) and *supervised machine learning* (ML). An MWE is *a sequence of at least two words frequently used together* [10]. MWE, among other features, "acts as a single unit at some level of linguistic analysis" [4]. Accurate identification and processing of MWEs is one of the main problems for the development of large-scale, effective and precise NLP technologies with applications such as foreign language acquisition, machine translation, text analytics & retrieval.

[1] Corresponding Author: Justina Mandravickaitė, Baltic Institute of Advanced Technology, Saulėtekio 15, Vilnius; E-mail: justina@bpti.lt.

Lithuanian is a synthetic language, thus simple statistical approaches for identification of MWEs do not provide satisfactory results due to the morphological richness resulting in lexical sparseness. However, combining LAMs and ML could help. LAMs compute an association score for each collocation candidate assessing the degree of association between its components. ML allows various properties of text to be encoded in feature vectors (lexical, morphological, syntactic, semantic, etc.) associated with output classes, as well as identifying complex non-linear relations. It permits capturing features in languages with complex morphology.

Certainly, ML only could provide good results as well, however, for this method to work in automatic identification of MWEs, annotated (morphologically, syntactically, etc.) data is neccessary for training. There is no such freely available "gold standard" for Lithuanian at the moment. Therefore values of LAMs with evaluation against reference list were chosen for training as this approach allowed to get some results even without sophisticated lexical resources and/or tools.

Combination of LAMs and ML has been investigated by several authors, e.g. [17] used such approach for extraction and evaluation of MWEs in English (reached 67.7% precision); extraction of nominal MWEs in French is reported by [7] (reached 60-75% precision for various models). Combining LAMs helps in the collocation extraction task [11,12]. Improvement can be achieved by combining a relatively small number of measures. So far there is no universal combination of LAMs that works best, since collocation extraction depends on the data, language and type/notion of MWEs.

2. Method

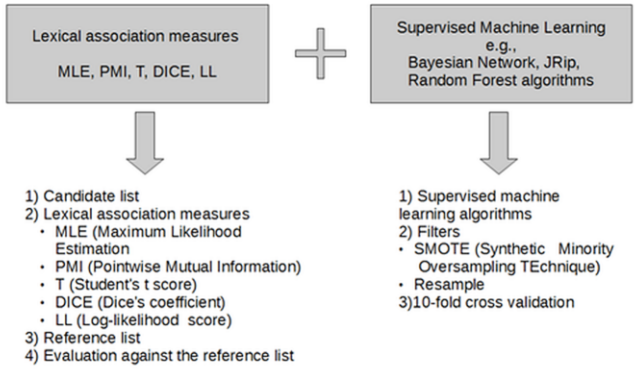


Figure 1. Scheme of the used methods

We use LAMs combined with ML algorithms in this research. Getting values of LAMs was performed using *mwetoolkit* [15] and for application of ML algorithms for MWEs candidates with LAMs values WEKA [8] was applied. Firstly, the candidate MWE bi-grams were extracted from the raw text. Then values of

5 LAMs (Maximum Likelihood Estimation, Dice's coefficient, Pointwise Mutual Information, Student's t score and Log-likelihood score) [15] were calculated and evaluated against the reference list of bi-gram MWEs. LAMs (MLE, PMI, T, DICE and LL) were combined in terms of vector containing values of association of each and every LAM for each and every MWE candidate. Combination of LAMs values for MWE candidates presented better results then taking values of separate LAMs or combination of values of lesser number of aforementioned LAMs.

Then selected algorithms, namely JRip (rule-based classifier) [6], Bayesian Network [16] and Random Forest [3] were applied. Classifiers for ML task were chosen according to their features (e.g., higher robustness in terms of noise is more characterictic to Random Forest than to other popular algoritms used in similar ML tasks [3]) and reported success by other authors, e.g. JRip was used successfully by [17], Bayesian Network - by [7]. Two filters were used due to sparseness: SMOTE (it re-samples a dataset with the Synthetic Minority Oversampling TEchnique) [5] and Resample (it produces a random subsample of a dataset using either sampling with replacement or without replacement) [8]. Precision, Recall and F-measure [13, 14] were used to evaluate the results. See the whole scheme in Figure 1.

We chose not to use lemmatiser because without extensive morphosyntactic information it is not possible to produce well-formed MWE lemmata [2]. Besides, as Lithuanian lexical resources are limited and linguistic tools are not freely available or need to be improved, we decided to use raw text with minimal preprocessing (lowercasing and tokenizing one-sentance-per-line only). Finally, using lemmata with LAMs resulted in zero values in majority of our initial experiments.

3. Data: Corpus & Reference List for Evaluation of MWE Candidates

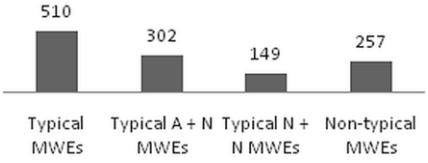


Figure 2. Summary of reference lists

Corpus of transcribed Lithuanian parliamentary speeches, containing speeches of members of the Lithuanian Parliament (MPs) from March 1990 to December 2013 was used. The size of the whole corpus is 23,908,302 words [9].

To evaluate MWE candidates with calculated LAMs, we combined the aforementioned list of MWE candidates with the highest LAMs scores. It was reviewed by the linguist in order to remove non-MWEs. We selected bi-grams, as statistical methods were generally reported to be more successful with shorter n-grams [1].

As most LAMs are designed for bi-gram, e.g. log-likelihood ratio, and in order for them to work for longer sequences they need to be adapted, we plan to explore LAMs+ML approach for longer MWE in the future.

MWEs for compilation of reference lists for different types of MWEs were identified and classified by a linguist who was native speaker of Lithuanian via review of candidate MWEs produced by *mwetoolkit*. The aforementioned linguist mostly works with various types of Lithuanian MWEs and thus MWEs among candidate MWEs were identified based on linguistic intuition as well as taking into consideration their grammatical structure (e.g., sequences *noun + conjunction* or *preposition*, *conjunction + verb* were not considered as MWEs because sequences of the latter structures in Lithuanian are regarded as incomplete) and meaning. Even though lists of candidate MWEs were reviewed by only one linguist, there was hardly a possibility that additional liguist(s) would not regard sequences kept in the reference lists as MWEs in this research. More discussions could rise about classification of MWEs, however, it should not have more significant impact on MWEs identification in this case.

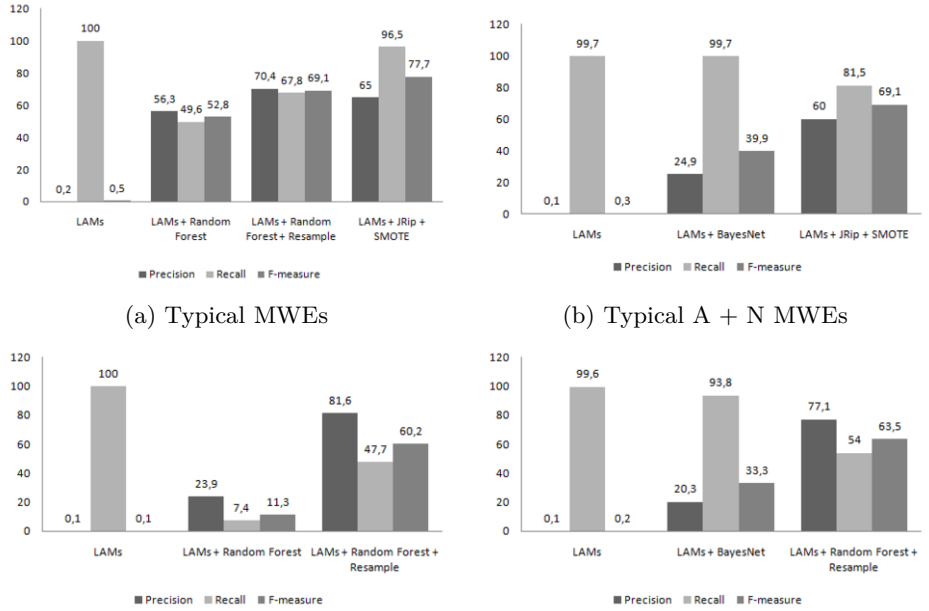
The majority of candidate MWEs kept for reference were typical collocations, i.e., sequences which are grammatically and semantically well-formed and used frequently, e.g., *atominė elektrinė* (nuclear plant). Another group contained non-typical collocations, i.e., grammatical collocations (composed of inflective or uninflected parts of speech that form semantically and syntactically unified, non-compositional unit with one syntactic function, e.g., multi-word adverbs (*taip pat* (also, too)), prepositions (*iki pat* (to, until), etc.), foreign words (e.g., *financial times*), coligations (*atrodo, kad* (seems that)), combinations of dependent verbs (*būtų priimtas* (would be accepted)), pronoun and noun compounds. The smallest group were collocations which are not in their typical form because of certain word order in the text, e.g., *apskaitai kompiuterizuoti* (for accounting to computerize): common word order would be *kompiuterizuoti apskaitą* (to computerize the accounting), i.e., verb governs the noun in accusative form but if it turns into final adjunct, then word order in the sentence changes and accusative becomes dative.

We used separate lists of these MWEs for evaluation of MWE candidates with calculated LAMs values for different experiments: automatic identification of typical MWEs, non-typical MWEs combined with MWEs not in their typical form and typical MWEs with certain morphological patterns: *Adjective (A) + Noun (N)* and *Noun (N) + Noun (N)*. Summary of reference lists is presented in Figure 2.

4. Experiments and Discussion

Using only LAMs from the *mwetoolkit* combined with the reference list for evaluation gave almost perfect Recall but extremely low Precision (see Figure 3). Thus it seems that almost any candidate MWE out of the 218 240 was identified as an MWE. Thus, association measures did not suffice for the successful extraction of MWEs for Lithuanian.

LAMs and ML algorithms were combined in 3 ways: (i) without any filter, (ii) with the SMOTE filter and (iii) with the Resample filter. All the models



(a) Typical MWEs

(b) Typical A + N MWEs

(c) Typical N + N MWEs

(d) Non-typical MWEs

Figure 3. Identification Results

were tested using standard 10-fold cross-validation. Summary of all experimental results are presented in Figure 3.

The highest Precision (70.4%) for typical MWEs was reached with **Random Forest classifier and the Resample filter**. For typical A + N MWEs we got the highest Precision (60.0%) with **JRip classifier and the SMOTE filter**. Resample filter did not manage to improve results significantly with any classifier we tried for this pattern. The best Precision (81.6%) for typical N + N MWEs was obtained with **Random Forest classifier and the Resample filter**. The same configuration gave the highest Precision (77.1%) for non-typical MWEs. For the latter two experimental settings SMOTE filter was less successful than Resample with any classifier we tried and therefore was excluded from the results.

In this stage of research we identified that in non-typical MWEs requires different recipe, i.e., different configuration in experimental setup, then the typical ones. The best results for automatic identificaton of typical MWEs were received with configuration of LAMs+JRip+SMOTE, while the best results for non-typical MWEs were achieved with experimental setup of LAMS+Random Forest+Resample. We plan more extensive research regarding non-typical (word order, gaps in between MWE components) MWEs in the future.

5. Conclusions

We report our experiments for extraction of bi-gram MWEs for Lithuanian by combining lexical association measures and supervised machine learning. This

experimental setup improved our results in comparison with just using association measures or machine learning only. Also, the reported experiments showed that different "recipes" (i.e., different configuration of lexical association measures, supervised machine learning algoritms, filters, etc.) are needed for diferent types/notions of MWEs in Lithuanian. Thus our future plans include experiments with automatic extraction of different types of MWEs and a greater diversity of MWEs.

Acknowledgements This research was funded by a grant (No. LIP-027/2016) from the Research Council of Lithuania.

References

[1] Sabine Bartsch and Stefan Evert. Towards a Firthian notion of collocation. *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Rep. of the Academic Network Internet Lexicography, OPAL–Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache, Mannheim*, 2014.

[2] Loïc Boizou, Jolanta Kovalevskaitė, and Erika Rimkutė. Automatic lemmatisation of lithuanian mwes. In *Proc. of the 20th Nordic Conference of Computational Linguistics, NODALIDA*, number 109, pages 41–49. Linköping Univ., 2015.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Nicoletta Calzolari, Charles J Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. Towards best practice for multiword expressions in computational lexicons. In *LREC*, 2002.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Jrnl. of Artificial Intelligence Research*, 16:321–357, 2002.

[6] William W Cohen. Fast effective rule induction. In *Proc. of the 12th Int. Conf. on Machine Learning*, pages 115–123, 1995.

[7] Marie Dubremetz and Joakim Nivre. Extraction of nominal multiword expressions in french. In *Proc. of the 10th Wksp on Multiword Expressions (MWE)*, pages 72–76, 2014.

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[9] Jurgita Kapočiūtė-Dzikienė, Andrius Utka, and Ligita Šarkutė. Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. *Linguistics: Germanic & Romance Studies*, 66, 2014.

[10] Rūta Marcinkevičienė. Tradicinė frazeologija ir kiti stabilūs žodžių junginiai. *Lituanistica*, 4(48):81–98, 2001.

[11] Pavel Pecina. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158, 2010.

[12] Pavel Pecina and Pavel Schlesinger. Combining association measures for collocation extraction. In *Proc. of the COLING/ACL*, pages 651–658. ACL, 2006.

[13] James W Perry, Allen Kent, and Madeline M Berry. Machine literature searching X. machine language; factors underlying its design and development. *American Documentation*, 6(4):242–254, 1955.

[14] David M. W. Powers. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders Univ., Adelaide, Australia, 2007.

[15] Carlos Ramisch. A generic framework for Multiword Expressions treatment: from acquisition to applications. In *Proc. of ACL 2012 Student Research Workshop*, pages 61–66. ACL, 2012.

[16] Jiang Su, Harry Zhang, Charles X Ling, and Stan Matwin. Discriminative parameter learning for Bayesian networks. In *Proc. of the 25th Int. Conf. on Machine Learning*, pages 1016–1023. ACM, 2008.

[17] Leonardo Zilio, Luiz Svoboda, Luiz Henrique Longhi Rossi, and Rafael Martins Feitosa. Automatic extraction and evaluation of MWE. In *8th Brazilian Symp. in Information and Human Language Technology*, pages 214–218, 2011.