Automatic Methods to Supplement Broad-Coverage Subcategorization Lexicons

Michael Schiehlen, Kristina Spranger

Institute for Computational Linguistics
Universität Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany
{mike, sprangka}@ims.uni-stuttgart.de

Abstract

The paper describes a system for extracting subcategorization frames of verbs not found in existing broad-coverage valency lexicons. The system uses two parameters: the results of a finite-state parser and the predictions of a set of automatically learned rules which transfer subcategorization frames from cognate verbs. An in-depth evaluation quantified the contribution of the individual parameters.

1. Introduction

This paper describes ongoing research in the area of subcategorization acquisition. Information on subcategorization is urgently needed in many Human Language Technology (HLT) applications and potentially welcome in lexicography. At least two research strands endeavour to address this need: On the one hand, large lists recording the subcategorization behaviour of thousands of words have been made available for many European languages (e.g. “COMLEX Syntax” with 6,000 English verbs (Grishman et al., 1994), Eckle’s subcategorization lexicon with 14,000 German verbs (Eckle-Kohler, 1999)). This approach aims at high precision and relies on semi-automatic extraction and manual checking. On the other hand, much effort has gone into the fully automatic acquisition of subcategorization frames from large corpora, usually by using wide-coverage parsers (e.g. a finite-state parser (Manning, 1993), a probabilistic LR parser (Briscoe and Carroll, 1997), a lexicalized probabilistic context-free grammar (Carroll and Rooth, 1996; Schulte im Walde, 2002)). All approaches of the latter kind have attempted to learn a subcategorization lexicon from scratch. Usually the most important task for applications in both HLT and lexicography is to add new words to existent subcategorization lexicons. Since such new words tend to be infrequent, standard statistical techniques run into problems, and more emphasis needs to be laid on heuristics and linguistic features.

The paper describes experiments which we conducted in German. As a basis, a version of Eckle-Kohler’s lexicon (called EKL hereafter) was used, comprising 16,630 verbs. EKL provides a fine-grained distinction in subcategorization frames amounting to 1,580 different frames. (This number is rather high compared with the 19 frame types distinguished in (Manning, 1993), 160 in (Briscoe and Carroll, 1997), and 38[1] in (Schulte im Walde, 2002).) EKL differentiates not only between 36 different types of prepositions, but also between 5 clause types, as well as correlates and reflexives. Automatic disambiguation is hard or impossible in cases such as argument–adjunct distinction, and the determination of semantically empty (i.e. inherently reflexive or correlative) arguments.

The paper is organized as follows. Section 2. describes the architecture of the subcategorization frame extraction system. It discusses three options for disambiguation among subcategorization frames. Section 3. presents a heuristical method to transfer subcategorization frames between cognate verbs via rules and gives three conditions that arguably should restrict the process of rule formation. Section 4. describes a novel approach to evaluation of subcategorization frame extraction, and presents the results of our system in such an evaluation. Section 5. concludes.

2. Experimental Setup

For acquisition, 36.2 million tokens of newspaper text were processed in several steps. The corpus was tagged by a POS tagger (the Tree Tagger), named entities were determined, and finally the corpus was parsed with a cascaded finite state parser (Schiehlen, 2003). Since the parser integrates information from EKL, it was built to distinguish all 1,580 subcategorization frames. However, only a subset of these frames (1,256) really crops up in the corpus. After parsing, possibly ambiguous case frames were extracted for all lemmas not in EKL and tagged as verbs or adjectives. We tried to eliminate tagging errors by only considering those verbs recognized by a morphological analyzer (Lezius et al., 2000). In the entire corpus, we found 3,278 verbs of this kind (1,845 hapax legomena), with an average frequency of 2.68 under a standard deviation of 4.88.

We went beyond existing approaches by also inspecting attributive present and past participles, gerundives and “-bar” adjectives (of the form ‘auflösend’, ‘aufgelöst’, ‘aufzulösend’, ‘auflösbar’, respectively). In these constructions, the subject argument is easy to determine (it is the head noun). Furthermore, gerundives, “-bar” adjectives, and most past participles express the passive of verbs with accusative object and possibly further arguments.

The parse results were fed into a *patternset extractor* (Briscoe and Carroll, 1997). The extracted subcategorization patterns included the syntactic categories and head lemmas of constituents for all relevant verbs and adjectives (cf. (1)). This phase encompasses a transformation from passive to active voice. Passive participles governed by the auxiliary *sein* are systematically ambiguous between passive (for transitive verbs) and active (for ergative verbs). Therefore, we had to consider both analyses. We also inte-

[1] (Schulte im Walde, 2002) distinguishes 579 frames in the version where she takes different types of prepositions into account.

grated the heuristic assumption that PPs and accusative NPs headed by temporal nouns have adjunct status. For this purpose we made use of a hand-compiled list of 79 temporal nouns (Spranger, 2002). The result of this phase is a list of corpus examples for the respective verbs and adjectives and their patternsets as illustrated in (1).

(1) Während die schlampige Clownsfrau die Tätowierungen eines Zuschauers bewunderte, marschierte ihr gepflegter Kompagnon durch die Reihen und wedelte Hüte, Röcke, oder Schuhe des Publikums mit einem Rasierpinsel ab.
While the slovenly clowness was admiring the tattoos of a spectator, her well-groomed partner walked through the rows and dusted hats, skirts or shoes of the audience with a shaving brush.

```
ab#wedeln VVFIN |Nom,Gen|Nom,Akk|Nom,Akk,PP/mit:D|
      |Gen|Akk|Akk|             Hut|Rock|Schuh
      |ADJ|ADJ|PP/mit:Dat| Rasierpinsel
```

The example shows the only instance in our corpus for the verb *abwedeln* (i.e. *dust off*). In the patternset, the lemma for *abwedeln* (*ab#wedeln*, since *ab* is a separable prefix) is followed by the POS tag of the token (finite verb) and a list of the subcategorization frames that the parser assigned. Finally, the head lemmas filling the argument slots are listed.

As illustrated in (1), the parser could sometimes only determine ambiguous subcategorization frames. In such cases, a disambiguation routine is required (*patternset evaluator* (Briscoe and Carroll, 1997)). We investigated three disambiguation strategies: In the first option, **longest-match**, subcategorization frames are ordered by arity and complexity, preferring longer over shorter frames (cf. (Briscoe and Carroll, 1995)) as well as frames incorporating inherently reflexive or correlative arguments over genuine arguments. In the second option, **global frequency**, frames are ordered by their frequency in the parsed 36.2 million word corpus (Carroll and Rooth, 1996). Thus, it is assumed that the probability distribution of subcategorization frames does not change between less frequent and more frequent verbs. In the third option, **local frequency**, the order is based on frame frequencies calculated from the corpus examples retrieved by the patternset extractor. Hence, it is assumed that there is a probability distribution of subcategorization frames which is special to rare verbs.

3. Inferring Frames from Cognate Verbs

Further information relevant for disambiguation can be gleaned from the subcategorization lexicon, as morphologically related words are usually also linked in their subcategorization behaviour. In particular, there are correlations between the subcategorization behaviour of prefixed verbs and that of their stems (Aldinger, 2004). In the corpus, 2,155 of the unknown verbs were prefix verbs with stems already listed in EKL. To handle these verbs, correspondence rules were learned from the prefix occurrences in the database and applied to yield predictions for unknown prefix verbs.

We extracted rules from EKL mapping the frames of stem verbs to frames of prefix verbs for individual prefixes. Every combination of a frame f_p of a prefix verb v_p and a frame f_s of v_p's stem could trigger a rule subject to the following conditions:

1. f_p extends f_s, so that all arguments of f_s occur in f_p.
2. There have to be at least two other prefix verbs with the same prefix as v_p and the frame f_p so that their stems have the frame f_s.
3. v_p's stem has no frame f'_s different from f_s which fulfils the conditions with respect to v_p's prefix and f_p.

Condition (1) can be motivated by the following train of thought: Semantically transparent prefix verbs imply their stems, hence all semantically obligatory arguments of the stem verb also need to be expressed in the prefix verb. If a functional mapping between semantic and syntactic arguments can be assumed (Levin, 1993), these arguments will be realized in the same syntactic form in both stem and prefix verb. The prefix verb may feature additional arguments (e.g. *get* vs. *get out* which may subcategorize for *of*).

Condition (2) derives from the fact that semantically opaque verbs are idiomatic in the sense that the meaning of the prefix verb cannot be entailed from the meaning of the prefix and the meaning of the stem. Only in semantically transparent prefix verbs the prefix has an independent meaning. Thus it is a characteristic of semantically transparent prefix verbs that they occur in groups centered around some meaning of the prefix. Condition (2) follows on two assumptions, viz. that verb meaning can be modelled by subcategorization behaviour (Levin, 1993) and that EKL already includes enough verbs to decide whether verbs form a group. On these assumptions, we can infer that combinations of prefix, prefix verb frame and stem frame do not describe independent prefix readings if they occur rarely in the lexicon. We exclude such combinations.

Condition (3) expresses the assumption that the prefix verb never is derived from more than one reading of the stem verb. Again we have to hypothesize that semantic distinctions manifest themselves in syntactic realizations.

Each rule which met the described conditions was weighted by the relative frequency of f_p given f_s and v_p's prefix. When an unknown prefix verb with a known stem verb was encountered, the rules were applied to all subcategorization frames of the stem verb. Weights for the resulting subcategorization frames for the prefix verbs were determined by the sum of the weights of the rules by which they could be derived. Finally, for every prefix verb with known stem verb, the subcategorization frame proposed by the parser that had maximal weight was chosen.

4. Evaluation

Due to the nature of the task which involves low-frequency words but fine-grained subcategorization frames, the standard evaluation technique (comparison with an independent machine readable dictionary) (Schulte im Walde, 2002) is not applicable. Even large published dictionaries often[2] do not contain entries for the words, and

[2] A cursory countercheck showed that 12.7% of the verbs investigated are missing in published dictionaries.

	Total			Hapax Legomena		
	F-value	Precision	Recall	F-value	Precision	Recall
baseline	40.14			39.95		
longest-match	35.07	35.66	34.51	34.39	35.09	33.73
global freq	51.59	55.50	53.71	52.61	53.67	51.59
local freq	42.09	42.79	41.41	34.27	34.96	33.61
longest-match with prefix	38.43	39.07	37.81	37.15	37.90	36.43
global freq with prefix	55.05	55.97	54.16	53.33	54.40	52.29
local freq with prefix	44.53	45.27	43.81	37.87	38.63	37.13
upper limit	67.29			66.98		

Table 1: Evaluation Results

even if they do, subcategorization frames are not described in sufficient detail. Hence a standard evaluation is exposed to a large number of false negatives.

We also tried to cut down on false positives, i.e. cases where the system outputs correct results, but derives them from data that do not validate these results. We manually disambiguated and corrected a sample of the output of the patternset evaluator (cf. (1)) so that we could measure the success of the patternset evaluator as the percentage of examples it got correct. In this evaluation regime, the system is expected to find the correct subcategorization frame for each verb token. In this respect, our evaluation contrasts with standard evaluation, where systems are only expected to determine correct subcategorization frames for each verb type. Note also that a simple list of subcategorization frames for verbs is not much use to a lexicographer who depends on corpus evidence, i.e. corpus examples correctly annotated with proposed subcategorization frames.

4.1. Annotation Guidelines

For manual annotation, we set up the guidelines in (2). In view of our target group, lexicographers and developers of NLP tools, who both are in need of fine-grained subcategorization information, we opted for semantically rather than strictly syntactically motivated principles.

(2) a. Syntactically obligatory arguments are subcategorized.

b. PPs (and other adverbials) that occur with almost all verbs are assumed to be adjuncts. Truly subcategorized PPs are much more selective.

c. Potential complements that can be moved into an unambiguous complement position by different kinds of alternations (e.g., reflexivization, morphologically triggered alternations like “bar”-adjectivization and nominalizations) are subcategorized. We exclude instrumental PPs from this rule due to the fact that they can be combined with almost all verbs and that the instrumental slot can be filled by several arguments at once.

d. PPs introduced by a preposition that is synonymous with the prefix of the subcategorizing verb have argument status.

e. A PP introduced by a preposition P1 that can be replaced by a preposition P2 with contrary meaning is not subcategorized.

In total, 1,333 examples were annotated by the authors, involving 971 verbs and 70 frames. All data were checked at least twice, interannotator agreement gave a kappa value of 80.9%. This relatively high kappa value reflects the quality of the annotation guidelines in (2). Among the annotated data, there were 851 hapax legomena (i.e. verbs that only occur once), on average every verb occurred in 1.37 examples under a standard deviation of 1.71.

4.2. Discussion

Table 1 lists the evaluation results for the individual approaches to disambiguation (cf. section 2.). The baseline approach consists in always choosing the most frequent subcategorization frame, i.e. transitive. The upper limit was computed as the percentage of examples in which the parser found the correct frame (possibly among others).

The results show that longest-match is the worst strategy which performs below the baseline. Local-frequency is also below the baseline with hapax legomena, but slightly surpasses the baseline in the total set. Global-frequency outperforms the other two disambiguation strategies. The prefix heuristics yields an improvement in all cases. (Carroll and Rooth, 1996) advocate the use of the Expectation Maximization algorithm, but we refrained from applying further EM iterations since they had no impact with global-frequency and even a deteriorating effect with local-frequency. All results shown here were determined without EM iterations. The figures in Table 1 compare badly with those presented in the literature, but it should be borne in mind that the task evaluated here is harder.

Like (Manning, 1993), we checked the token recall of our system. To this purpose, we inspected 367 occurrences of unknown verbs drawn randomly from the corpus. Out of these cases, 239 had a subcategorization frame discovered by the system, 81 were tagging errors, and 1 case was a typo. Thus, the system achieves a token recall of 83.85%.

4.3. Impact of Prefix Rule Conditions

The figures in Table 1 were determined with prefix rules that conformed to all three conditions on frame selection for prefix verbs (cf. section 3.). We went on to investigate the contribution of the individual conditions; results are displayed in Table 2. With the global-frequency strategy, omission of one, two or all three of the conditions has no significant effects. With the longest-match strategy, however, there was a steep rise of 9% in F-value when all or some of the conditions were left out. Condition (3) generally performed somewhat worse than the other two.

	F values	
	longest-match	global frequency
No Condition	47.50	55.20
Condition 1	47.05	55.43
Condition 2	47.12	55.20
Condition 3	45.14	55.20
Condition 1,2	46.74	55.36
Condition 1,3	47.05	55.43
Condition 2,3	45.14	54.82
Condition 1,2,3	38.43	55.05

Table 2: Impact of the Prefix Rule Conditions

The fact that the conditions did not lead to improvements can be explained by the following considerations.

Different kinds of exceptions to Condition (1) can be found. Transitive *umranken* (i.e. *twine itself round*) has a stem (*ranken*) which takes an inherently reflexive and a PP argument (introduced by *um*). Semantically, the prepositional object of *ranken* is the same argument as the accusative object of *umranken* in violation of (Levin, 1993)'s hypothesis that the mapping from semantic arguments to syntactic roles is independent of the verb involved. Another counterexample, intransitive *umher#suchen* (i.e. *search around*), lacks the prepositional object of its stem verb *suchen* entirely. In semantics, the argument slot is either filled anaphorically or quantified over by the spatial adverb *umher*. In a third example, *aneinander#reiben* (i.e. *rub on each other*), the prefix itself fills the missing argument slot.

Condition (2) is arguably thwarted by data sparseness in EKL.

The semantic foundation of Condition (3) is impeccable, but the Condition also rests on (Levin, 1993)'s hypothesis. Condition (3) prevents prefix rules from applying to verbs like transitive *ein#lernen* (i.e. *show the ropes*): The accusative object of the stem verb may be left implicit so that *lernen* has a transitive and intransitive frame, which do not, however, correspond to different readings.

5. Conclusion

The stated results are preliminary for several reasons. First, a large part of the errors was due to tagging errors and consequential parsing errors. Better tagging and parsing quality will improve performance, also for the present task. Second, our system essentially depends on three knowledge sources, viz. the parse output, global frequencies of subcategorization frames, and the learned prefix rules. The combination of these factors might be suboptimal and improved by machine learning. In particular, the fact that parse results are always given preference imposes a rather low upper limit on performance. A system more skeptical of quality of parse results might perform better.

We presented a system extracting subcategorization frames for infrequent verbs and adjectives based on a finite-state parser. Several methods to disambiguate proposed subcategorization frames were discussed and compared by evaluation. Furthermore, we integrated rules cross-relating subcategorization frames for morphologically cognate verbs, which were learned automatically from a large lexicon. We showed that integrating these rules leads to performance gains. We performed a thorough evaluation of the system, using 1,333 manually annotated and double-checked corpus examples as a gold standard. A concise set of guidelines on the argument–adjunct distinction ensured a high degree of interannotator agreement. The usefulness of our tool for lexicographic purposes is underscored by the fact that amazingly many verbs (12.7% of all verbs) for which we were able to extract fine-grained linguistic information are missing in even the most comprehensive published dictionaries of German.

In sum, we found extraction of rare subcategorization frames a challenging and interesting topic. One of the authors plans to pursue this topic in her further research.

6. References

Aldinger, Nadine, 2004. Towards a dynamic lexicon: Predicting the syntactic argument structure of complex verbs. In LREC '04. Lissabon.

Briscoe, Ted and John Carroll, 1995. Towards Automatic Extraction of Argument Structure from Corpora. ACQUILEX II Working Paper.

Briscoe, Ted and John Carroll, 1997. Automatic extraction of subcategorization from corpora. In ANLP'97. Washington, DC.

Carroll, Glenn and Mats Rooth, 1996. Valence Induction with a Head-Lexicalized PCFG. In EMNLP 3.

Eckle-Kohler, Judith, 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textkorpora*. Berlin: Logos-Verlag.

Grishman, Ralph, Catherine Macleod, and Adam Meyers, 1994. COMLEX Syntax: Building a Computational Lexicon. In COLING '94. Kyoto, Japan.

Levin, Beth, 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Lezius, Wolfgang, Arne Fitschen, and Stefanie Dipper, 2000. IMSLex — representing morphological and syntactical information in a relational database. In EURALEX'00. Stuttgart.

Manning, Christopher D., 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In ACL'93. Columbus, OH.

Schiehlen, Michael, 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the Research Note Sessions of EACL'03*. Budapest, Hungary.

Schulte im Walde, Sabine, 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In LREC '02. Las Palmas.

Spranger, Kristina, 2002. *Lexically Informed Chunking Analyses as a Starting Point for the Extraction of Linguistic Information from Dutch Text*. Master's thesis, IMS Stuttgart.