Can Subcategorization Help a Statistical Dependency Parser?

Daniel ZEMAN
Centrum komputační lingvistiky
Univerzita Karlova
Malostranské náměstí 25
Praha, Czechia, 11800
zeman@ufal.mff.cuni.cz

Abstract

Today there is a relatively large body of work on automatic acquisition of lexico-syntactical preferences (subcategorization) from corpora. Various techniques have been developed that not only produce machine-readable subcategorization dictionaries but also they are capable of weighing the various subcategorization frames probabilistically. Clearly there should be a potential to use such weighted lexical information to improve statistical parsing, though published experiments proving (or disproving) such hypothesis are comparatively rare. One experiment is described in (Carroll et al., 1998) — they use subcategorization probabilities for ranking trees generated by unification-based phrasal grammar. The present paper, on the other hand, involves a statistical *dependency* parser. Although dependency and constituency parsing are of quite a different nature, we show that a subcategorization model is of much use here as well.

Introduction

Large-scale machine-readable dictionaries are becoming available as grows the scale of corpora able to serve as source of such dictionaries and the variability of techniques for extracting lexical information from the corpora automatically. The resulting dictionaries bear (among others) data on verb subcategorization (i.e. types of *arguments* the particular verb requires, as opposed to *adjuncts* that can modify any verb). Often such data are accompanied by the relative frequencies of the alternatives. Various lexicalised grammars have been proposed (Resnik 1992, Schabes 1992, Carroll and Weir 1997) that take advantage of subcategorization information. Carroll et al. (1998) report on an experiment where verb frames were used to re-rank analyses after parsing. However, even they did not employ the subcategorization directly during parsing.

In this paper, we view the contribution of subcategorization frames to parsing from a different perspective. The model language of ours is Czech. Since Czech belongs to the part of the world where the dependency framework is popular, our parser is based on a statistical dependency model. Instead of using a grammar and reconstructing sentence structure as a side effect of the process of sentence derivation, the parser directly considers possible dependencies between words and their probabilities. Upon such information it consequently builds the optimal dependency structure (tree). The dependency parser can be lexicalized simply by modeling the word dependencies instead of dependencies of morphological tags. We argue that a "smart" lexicalization driven by subcategorization leads to far better results than the one mentioned above.

1 Background

1.1 Verb Subcategorization in Czech

Czech is a "free word-order" language. This means that the arguments of a verb do not have fixed positions and are not guaranteed to be in a particular configuration with respect to the verb.

The examples below show that while Czech has a relatively free word-order some orders are still marked. The SVO, OVS, SOV, and OSV orders in 1, 2, 3, and 4 respectively, differ in emphasis but have the same predicate-argument structure. The examples 5 and 6 can only be interpreted as a question. Such word

orders require proper intonation in speech, or a question mark in text.

The example 7 demonstrates how morphology is important in identifying the arguments of the verb. Cf. 7 with 2. The ending *-a* of *Martin* is the only difference between the two sentences. It however changes the morphological case of Martin and turns it from subject into object. Czech has 7 cases that can be distinguished morphologically.

1. *Martin otvírá soubor.*
   (SVO: "Martin opens the file.")
2. *Soubor otvírá Martin.*
   (OVS ≠ "The file opens Martin.")
3. *Martin soubor otvírá.*
4. *Soubor Martin otvírá.*
5. *# Otvírá Martin soubor.*
6. *# Otvírá soubor Martin.*
7. *Soubor otvírá Martina.*
   (= "The file opens Martin.")

As we just demonstrated, the cases distinguish between subject and object. Similarly they allow for a distinction among different sorts of objects and, most importantly, in sentences containing more than one verb the cases help a lot to resolve which argument belongs to which verb. They thus form an integral part of verb subcategorization and must be encoded in subcategorization data. Therefore the Czech frames contain elements such as N1…N7 rather than simple NP.

For extracting subcategorization frames, prepositions in Czech have to be handled carefully. In some frames, a particular preposition is required by the verb, while in others it is a class of prepositions such as locative prepositions (e.g. *in*, *on*, *behind*, …) that are required by the verb. In contrast, adjuncts can use a wider variety of prepositions. Prepositions specify the case of their noun phrase complements but some prepositions can take complements with more than one case marking with a different meaning for each case (e.g. *na mostě = on the bridge; na most = onto the bridge*). In general, verbs select not only for particular prepositions but also indicate the case marking for their noun phrase complements.

Here are some examples of the frames:

Frame	Description	Example
N4	direct object in accusative	*nese zavazadlo* "he carries luggage"
N4 N3	ditransitive — accusative and dative objects	*dal bratrovi knihu* "he gave a book to his brother"
se N2	reflexive particle *se* and a genitive object	*bojí se mě* "he is frightened of me"
VINF	infinitive (with modal verbs)	*musí odejít* "he has to leave"
R4(v)	PP with the prep. *v* governing an NP in accusative	*věří v Boha* "he believes in God"
JS(že)	clause starting with *že* "that"	*řekl, že to ví* "he said he knew it"

Published sources suggest that Czech verbs can be subcategorized into more than 130 classes (Sarkar and Zeman, 2000).

1.2 Subcategorization Acquisition

Preliminary work on subcategorization frame extraction from corpora was done by Brent (1991; 1993; 1994), Ushioda et al. (1993), and Manning (1993). They all use PoS tagged corpora and finite state shallow parsers that search the corpora for possible frames. Brent and Manning then use the hypothesis testing method to decide whether a verb-frame co occurrence was significant or only accidental. Ushioda et al. use heuristics to make the decision. They also collect frequencies of respective patterns. Manning covers the biggest number of verbs — 3104 — and in his comparison of the entries for the most common 40 with a hand-made dictionary he achieved a precision of 90% and recall of 43%. All three authors suppose the set of possible categories is relatively small and known in advance. Briscoe and Carroll (1997) also have a predefined set but it contains 160 different frames.

Gahl (1998) presents an RE-based extraction tool that creates subcorpora of the BNC containing different subcategorization frames for verbs, nouns and adjectives. The resulting subcorpora can be used to determine the (relative) frequencies of the frames.

Carroll and Rooth (1998) use an iterative approach: they start from a manually-developed

context-free grammar, train a probabilistic version of the grammar lexicalized on rule heads, then employ the EM algorithm to calculate the expected frequencies of a head word accompanied by a particular frame, and finally feed the frequencies back into the grammar and repeat the previous steps. At the same time their parser can be viewed as one of the first attempts to use the learned subcategorization for parsing but they give no figures on how the parsing accuracy improved. Chiang (2000) does use subcat information (extracted from the Penn Treebank corpus) for statistical parsing. He does not evaluate the specific contribution made by the subcat information.

All papers mentioned so far report experiments with English. Additional results have been published for Italian (Basili and Vindigni 1998), Greek (Kermanidou et al. 2001), and Czech (Sarkar and Zeman, 2000). Sarkar and Zeman compare three different statistical methods for measuring frame relevance: hypothesis testing augmented with a new subframe searching technique, the log-likelihood ratio and the t-scores. Korhonen et al. (2000) give another comparison of measuring methods, including simple relative frequency. In our experiment we reuse the Czech system of Sarkar and Zeman (2000).

1.3 Statistical Modeling of Dependencies: the Parser

Our parser is a clone of the one described in Zeman (1998) and Zeman in (Hajič, Brill et al. 1998). It uses a dependency treebank to learn dependencies between morphological (PoS) tags of words (the non-lexicalized version), aptly the lemmas or the word forms themselves (lexicalized version). Instead of using the complete Czech tag set (as described in Hajič, 1998) we use a reduced two-letter tag set (Collins in Hajič, Brill et al. 1998) which is more suitable for parsing. The total number of 1346 different tags found in the training corpus dropped to 66.

The parser maintains a table of all dependencies between words of particular morphological category (e.g. between a noun in nominative case (`N1`) and an adjective in the same case (`A1`)) together with their relative frequencies. When applied to a raw text it first determines the possible dependencies. Then it estimates the probabilities of the dependencies as their relative frequencies. And then it performs a stack search to find a tree with the highest possible product of dependency probabilities.

This simple mechanism itself is too weak to build correct dependency structures. Several constraints are defined to make the parser work. One of such constraints is projectivity.

1.3.1 Projectivity

Projectivity is a property that combines tree structure and the word order in sentence. A dependency A-B (where A is the governing node) is **projective** if and only if all the words that are placed between A and B are included in the subtree of A. If the tree is displayed so that the x-coordinates of nodes correspond to the word order, each non-projective dependency will cross at least one perpendicular from another node (it will not necessarily cross another dependency.)

Projectivity is an important attribute of the word order. It does not depend on the length of the phrases but it still allows us to distinguish "normal" sentence structures from the "wild" ones. We deliberately do not call them "incorrect" because not all non-projective dependencies are errors in Czech.[1] Nevertheless such dependencies are quite rare: Hajič et al. (1998) counted only 1.8 % of dependencies in a 19126-sentences corpus being non-projective.[2] Since the accuracy of our parser is far from 98.2 %, it is useful to require in our experiments that all the dependencies generated by the parser be projective. Unlike the Zeman's 1998 parser, our version enforces projectivity by connecting

[1] Note however that non-projective constructions cannot be described by an ordinary context-free grammar.

[2] The corpus used to measure non-projectivity was a part of the Prague Dependency Treebank (PDT, see Böhmová et al.). The average sentence length in this part is 16.3 words (dependencies). About 79.4 % of all trees have all dependencies projective. The number of trees that contain one or no crossing (non-projective) dependency is 93.8 % and the number of trees with at most two such dependencies is 98.3 %.

at each step only neighboring components (in certain configurations) in a forest.

1.3.2 Direction and distance

Finally, there are two important features that reflect the mutual positions of the two dependency members in the sentence. The first one is **direction**. The model assigns separate probabilities to a dependency where the depending node is (in the sentence) to the left of the governing node, and to the same dependency where it is to the right of the governing node. For instance, a preposition stands always to the left of its dependent.

The second thing is distance, or better, **adjacency**. The parser asks whether the two nodes are adjacent in the sentence or not. If not, it further asks whether there is at least one comma between the nodes or not.[3] Separate probability estimates are kept for all three options.

2 The Experiment

2.1 The Baseline Parser

The baseline parsing system comprises:

- A morphological tagger & lemmatizer (Hajič and Hladká, 1998). For each input word the tagger produces only the single highest-ranked tag. Mean ambiguity is roughly 2.5 tags per word, and since the alternatives often differ in case, the errors of the tagger sometimes violate parsing accuracy. Hajič and Hladká report an overall accuracy of 93% of the tagger output. We additionally measured its precision on the case of nouns to be 89%.
- A tag mapping tool from the standard set to the reduced two-letter set.
- An instance of the parser described in section 1.3, which models dependencies solely on the morphological tags.
- Training and test treebanks (of 25900 and 2600 sentences respectively) derived automatically from the Prague Dependency Treebank (PDT, Böhmová et al. 2000).

Carroll et al. (1998) test their system on a subset of the Susanne corpus that contains only sentences covered by their grammar. They argue that a fair proportion of the rest are elliptical noun or prepositional phrases, fragments from dialogue and so forth, which they do not attempt to cover. The parser we used is by nature robust enough to parse any sentence, and we give the results for all sentences in our Evaluation Section. For the sake of comparison with Carroll et al., however, we also ran the same experiment on data where "weird" sentences were not present. PDT marks elliptical constructions by the syntactic tag ExD, so any sentence containing a dependency marked this way was not used in the experiment. Moreover, we also dropped sentences with coordinations or appositions because our current dependency model is not well adjusted to such non-dependency relations. If we included those sentences, the results would be biased by errors that have nothing to do with subcategorization. Genres in the Prague Dependency Treebank are mixed only mildly. The majority is newswire text, the rest are business reports and scientific writing. The mean sentence length is 17 words.

2.2 Incorporating Acquired Subcategorization Information

Carroll et al. (1998) use subcategorization to rank the various analyses of a sentence after parsing is done. We propose several levels of incorporating it directly into the parsing method.

2.2.1 Lexicalized Dependency Model

One could argue that even a lexicalized version of the dependency model (i.e. a model of word dependencies, in contrast to tag dependencies) inherently contains a primitive subcategorization acquisition system. The problem is that subcategorization in such system has too little power to compete with other factors. Any lexical dependency model needs a very careful smoothing to ensure that it will improve the results even a little bit. To illustrate this, we include in the results the performance of the following model: Probability of each dependency is estimated as $p = \lambda p_w + (1-\lambda)p_t$, where p_w is the

[3] This comma feature is also new in our version of the Zeman's parser. It is motivated by Collins (1996; 1997).

probability of the word-word dependency and p_t is the probability of the tag-tag dependency. Empirically we found the optimal value of λ to be 0.734375. Since the value has to be driven from held-out data, and its impact on the results is very low (see the evaluation section), it is doubtful whether such plain lexicalization can help at all.

2.2.2 Selective Lexicalization

Probably the simplest shift towards subcategorization is selective lexicalization. The lexicalized model described above took into account the lexical information for both ends of a dependency, for the governor as well as the dependent, and it did so for each and every pair of words in the corpus. In contrast, a subcategorization dictionary documents that in many cases only one member of the dependency (usually the governor — verb, possibly also adjective or noun) is determined lexically, while the other is determined by its morphological class. In other cases, the dependent node has to be lexicalized as well (prepositions, subordinating conjunctions like *že* "that" etc.).

In our experiments we lexicalized all prepositions by replacing the case marker in their tag by their lemma. Thus, the R3 tag for *k* "to" (meaning that it is a preposition requiring a noun in dative) becomes Rk.[4] Similarly we lexicalize tags of subordinating conjunctions (J, for *že* "that" becomes Jže), adverbs other than derived from adjectives (Db for *více* "more" becomes Dvíce) and reflexive particles (P4 for *se* becomes Pse).

We must be careful when applying the described technique to governing verbs. The number of different verbs is too big to replace their tags by their lemmas, so we cannot simply transit from the tag VB for verbs like *give* to a "lexicalized tag" Vgive — the tag dependency probabilities would lose their ability to smooth the model. Rather we sum up the lexical (Vgive) and non-lexical (VB) probabilities.[5] This is a violation of probability laws and the resulting weight is indeed not a true probability but the real effect is positive. In fact, verb-driven dependencies are honored as their relative frequency from the corpus is counted more than once. This is exactly what we would like to do with subcategorized dependencies. And verb-driven dependencies not learned from the corpus get a weight equal to their original non-lexical probability.

The last improvement fitting into this category (although it is not a subcategorization in its very sense) involves the verb *být* "to be". Its tag is always enriched by its form from the sentence (possible negative prefix and gender ending stripped). This allows the model to distinguish constructions like *bude dělat* ("he will do", *do* is governor) from constructions like *může dělat* ("he can do", *can* is governor). It is no more necessary to include "subcategorization frames" for *to be* (and it is even undesirable because this verb has too many different functions in Czech: as an auxiliary verb, as a part of a nominal predicate, as a full-meaning verb etc.).

2.2.3 Preferring Subcategorized Dependencies

So far we have not actually used any subcategorization dictionary. The selective preferences of verbs were learned together with all other preferences in the corpus but no distinction was being made between arguments and adjuncts. Lexical acquisition systems described above make this distinction using their own statistical techniques. In this and the next sections we describe how we use an automatically acquired subcategorization dictionary.

We use a dictionary generated by a system similar to the one described by Sarkar and Zeman (2000). The dictionary is a list of verb-frame pairs (together with frame frequencies) where one verb may occur in more than one pair. A frame is a list of frame members such as N4, R3(proti) etc. The order of the frame members is not significant so they are sorted alphabetically. Our parser uses the dictionary as an additional input and before analyzing of every sentence it finds all verbs, for

[4] The number of prepositional tags raised from 8 to 77.

[5] Note that this differs from the interpolation described above. There we interpolated probabilities of whole dependencies and the parts were a word-word dependency and a tag-tag dependency. Here we alternate only the governing part of only some dependency pairs, and we use lextag (lemma) instead of word form.

each of it seeks all its frames in the dictionary and gets all the words in the sentence that match a member in at least one of the frames. The links between the verb and these words are then added to a list of "subcategorized dependencies". These are sorted according to their probabilities in the dependency model and to the probabilities of the particular frames relative to the verb (the latter is supplied by the subcategorization dictionary). Another sorting criterion is distance: an argument cannot be attached to a verb so that the dependency would skip a closer verb awaiting an argument of the same class. During the tree building phase the subcategorized dependencies are considered first. Other dependencies are considered only if the subcategorized dependency cannot be added at the given moment.

2.2.4 Taking Whole Frames into Account

The model becomes even more adequate when the subcategorized dependencies are not isolated from their frames. Some arguments occur in all frames of a particular verb while others only in one. We honour the former and penalize the latter by adding the frequency of the other frames. On the other hand, when one member of a frame is selected, the other frames for the same verb are penalized. A special case of that is multiple filling of the same frame slot — which of course is not allowed. All such constraints would easily be expressed by a CF grammar but we have to consider them separately once we build a statistical dependency model.

3 Evaluation

The standard method for evaluating dependency parsing is the (dependency) *accuracy* — the number of correctly assigned dependencies divided by the number of total dependencies (words).[6] We evaluated the accuracy of all dependencies affected by subcategorization (in test corpus they are denoted by the s-tags Sb, Obj, Adv, Pnom and AuxT). Note that our evaluation is similar to the "Grammatical Relation Evaluation" used by Carroll et al. (1998). First, we evaluated the non-selectively lexicalized model with three different lambdas to illustrate the small influence of such lexicalization. The columns correspond to different subsets of the training and test sets: "ellips" means the elliptic sentences were excluded, "coord" means sentences with coordinations or appositions were excluded as well.

λ	all	ellips	coord
0	77.0	77.7	79.9
0.734375	77.1	77.7	80.1
1	42.4	42.5	40.7

The second table shows the accuracy for the baseline model, and the subcategorization-aware model respectively:

Model	all	ellips	coord
Baseline	77.0	77.7	79.9
Subcategorized	78.7	79.3	82.1

4 Conclusion

We surveyed existing work on automatic acquisition of subcategorization from corpora. Then we described a dependency parser based on statistical dependency modeling and showed three levels of incorporating subcategorization into the model. Finally we demonstrated in an experiment that subcategorization aware lexicalization, as opposed to "naïve" lexicalization, improves significantly the performance of the parser. Current and future work includes investigating coordinated arguments and adjuncts and enriching the model to handle coordinations and appositions.

5 Acknowledgements

The research has been carried out under the project MŠMT LN00A063.

6 References

Roberto Basili, Michele Vindigni (1998): *Adapting a subcategorization lexicon to a domain.* In: "Proceedings of the ECML'98 Workshop TANLPS. Chemnitz, Germany.

Alena Böhmová, Jan Hajič, Eva Hajičová, Barbora Hladká (2000): *The Prague Dependency Treebank: Three-Level Annotation Scenario.* In: Anne Abeillé (ed.): "Treebanks: Building and Using Syntactically Annotated Corpora." Kluwer

[6] As long as the parser assigns one dependency to every word in the corpus, there is no need to distinguish precision and recall.

Academic Publishers, Dordrecht, Netherlands. See also http://ufal.mff.cuni.cz/pdt/.

Michael Brent (1991): *Automatic acquisition of subcategorization frames from untagged text.* In: ACL 1991, pp. 209-214. Berkeley, California.

Michael Brent (1993): *From grammar to lexicon: unsupervised learning of lexical syntax.* In: Computational Linguistics, 19(3):243-262.

Ted Briscoe, John Carroll (1997): *Automatic extraction of subcategorization from corpora.* In: "Proceedings of the 5th ANLP Conference", pp. 356-363. Washington, USA.

John Carroll, David Weir (1997): *Encoding frequency information in lexicalized grammars.* In: "Proceedings of the 5th ACL/SIGPARSE International Workshop on Parsing Technologies (IWPT-97)", pp. 8-17. Cambridge, Massachusetts.

John Carroll, Guido Minnen, Ted Briscoe (1998): *Can Subcategorisation Probabilities Help a Statistical Parser?* In: "Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora", Montréal, Québec.

Glenn Carroll, Mats Rooth (1998): *Valence induction with a head-lexicalized PCFG.* In: "Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP 3)". Granada, Spain.

Eugene Charniak (1997): *Statistical Techniques for Natural Language Parsing.* In: "AI Magazine, vol. 18, no. 4. American Association for Artificial Intelligence.

David Chiang (2000): *Statistical Parsing with an Automatically-extracted Tree Adjoining Grammar.* In: Proceedings of ACL. Hong Kong, China.

Michael Collins (1996): *A New Statistical Parser Based on Bigram Lexical Dependencies.* In: "Proceedings of the 36th ACL". Santa Cruz.

Michael Collins (1997): *Three Generative, Lexicalised Models for Statistical Parsing.* In: Proceedings of the 35th ACL". Madrid, Spain.

Susanne Gahl (1998): *Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus.* In: "Proc. of COLING-ACL '98". Montréal, Québec.

Jan Hajič (1998): *Building a Syntactically Annotated Corpus: The PDT.* In: "Issues of Valency and Meaning." Karolinum. Praha, Czechia.

Jan Hajič, Barbora Hladká (1998): *Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset.* In: "Proc. of COLING -ACL 98", pp. 483-490. Montréal, Québec.

Jan Hajič, Eric Brill, Michael Collins, Barbora Hladká, Douglas Jones, Cynthia Kuo, Lance Ramshaw, Oren Schwartz, Christoph Tillmann, Daniel Zeman (1998): *Workshop 98 Final Report.* http://www.clsp.jhu.edu/ws98/projects/nlp/report/. Baltimore, Maryland.

Katia-Lîda Kermanidou, Manolis Maragoudakîs, Nikos Fakôtakîs, Geôrgios Kokkinakîs (2001): *Influence of Conditional Independence Assumption on Verb Subcategorization Detection.* In: Václav Matoušek et al. (eds.): "Proceedings of TSD 2001" Springer LNAI 2166. Železná Ruda, Czechia.

Anna Korhonen (2000): *Using Semantically Motivated Estimates to Help Subcategorization Acquisition.* In: "Proc. of SIGDAT EMNLP and Very Large Corpora". Hong Kong, China.

Anna Korhonen, Genevieve Gorrell, Diana McCarthy (2000): *Statistical Filtering and Subcategorization Frame Acquisition.* In: "Proc. of SIGDAT EMNLP and Very Large Corpora". Hong Kong, China.

Christopher D. Manning (1993): *Automatic acquisition of a large subcategorization dictionary from corpora.* In: "Proceedings of the 31st Meeting of the ACL", pp. 235-242. Columbus, Ohio.

Philip Resnik (1992): *Probabilistic tree-adjoining grammar as a framework for statistical natural language processing.* In: "Proceedings of COLING-92", pp. 418-424. Nantes, France.

Anoop Sarkar, Daniel Zeman (2000): *Automatic Extraction of Subcategorization Frames for Czech.* In: "Proceedings of COLING-2000". Saarbrücken.

Yves Schabes (1992): *Stochastic lexicalized tree-adjoining grammars.* In: "Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)", pp. 426-432. Nantes, France.

Hana Skoumalová (2001): *Czech Syntactic Lexicon* (PhD Thesis). Univerzita Karlova, Praha, Czechia.

Markéta Straňáková-Lopatková, Zdeněk Žabokrtský (2002): *Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation.* In: Proceedings of LREC 2002, vol. 3, pp. 949-956. Las Palmas de Gran Canaria, Spain.

Akira Ushioda, David A. Evans, Ted Gibson, Alex Waibel (1993): *The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora.* In: Boguraev, Pustejovsky (eds.): "Proc. of the Workshop on Acquisition of Lexical Knowledge from Text". Columbus, Ohio.

Daniel Zeman (1998): *A Statistical Approach to Parsing of Czech.* In: "Prague Bulletin of Comp. Ling.", vol. 69. Univerzita Karlova, Praha.