On Automatic Assignment of Verb Valency Frames in Czech

Jiří Semecký

Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Prague, Czech Republic
jiri.semecky@mff.cuni.cz

Abstract

Many recent NLP applications, including machine translation and information retrieval, could benefit from semantic analysis of language data on the sentence level. This paper presents a method for automatic disambiguation of verb valency frames on Czech data. For each verb occurrence, we extracted features describing its local context. We experimented with diverse types of features, including morphological, syntax-based, idiomatic, animacy and WordNet-based features. The main contribution of the paper lies in determining which ones are most useful for the disambiguation task. The considered features were classified using decision trees, rule-based learning and a Naïve Bayes classifier. We evaluated the methods using 10-fold cross-validation on VALEVAL, a manually annotated corpus of frame annotations containing 7,778 sentences. Syntax-based features have shown to be the most effective. When we used the full set of features, we achieved an accuracy of 80.55% against the baseline 67.87% obtained by assigning the most frequent frame.

1. Introduction

Since verbs are understood as central elements of sentences, the key aspect in the determination of sentence meaning lies in estimation of meaning of the verb. Verbs valency frames usually partially correspond to verbs' different meanings. As there is no exact definition of word meaning, we consider frames to reflect the verb meanings sufficiently. Initial results of verb frame disambiguation were already reported in (Erk, 2005) for German and (Lopatková et al., 2005) for Czech.

The paper is divided as follows. In Section 2., we give an overview of the VALEVAL corpus, which was used for the task. In Section 3., we describe features extracted from the data. In Section 4., we mention methods which we used for disambiguation. In Section 5., we evaluate our results using two different metrics. And finally, in the last section, we conclude and suggest directions for further development.

2. Data resources

Our frame definition is taken over from **VALLEX** (Žabokrtský and Lopatková, 2004), a manually-created valency lexicon of Czech verbs, based on the framework of Functional Generative Description (FGD) (Sgall et al., 1986). VALLEX version 1.0, which we used for our task, defines valency for over 1,400 Czech verbs and contains over 3,800 frames.

The lexicon consists of **verb entries** corresponding to particular verb lexemes, i.e. complex units consisting of the verb base lemma and its eventual reflexive particle (*se* or *si*). Each verb entry consists of definitions of one or more **frames** which roughly correspond to verb meanings. Each frame is composed of several **slots** corresponding to complements of the verb. Each frame slot is described by a functor, expressing the relationship between the verb and the complement (e.g. *Actor*, *Patient*, *Addressee*), list of possible morphological forms in which the frame slot might be expressed, and type of the slot (*obligatory*, *optional* or *typical*). Verbs with identical slots and functors but different meanings received different frames in VALLEX.0 The average number of frames per verb lexeme in VALLEX is 2.7 and the average number of frames per base lemma is 3.9.

As data for disambiguation methods, we used the manually annotated corpus of frame annotations **VALEVAL** (Bojar et al., 2005), which uses VALLEX frame definitions. It contains 109 selected base lemmas. For each base lemma, 100 sentences were randomly selected from the Czech National Corpus (Kocek et al., 2000). For detail on verb selection see (Bojar et al., 2005).

For the purpose of the VALEVAL corpus, reflexivity of verbs (expressed by a separate reflexive particle) was disregarded, as there is no automatic procedure to determine it.

VALEVAL was concurrently annotated by three annotators looking at the sentence containing the verb and three preceding sentences. Annotators had the option of selecting no frame if the corresponding frame was missing in the lexicon or if the decision could not be made due to an incorrect morphological analysis. The inter-annotator agreement of all three annotators was 66.8%, the average pairwise match was 74.8%.

For our experiment we used only VALEVAL sentences where all three annotators agreed. Moreover, sentences on which annotators did not agree were rechecked by another annotator, and sentences with a clear mistake were corrected and included as well. This resulted in a set of 8,066 sentences.

Then, we automatically parsed the sentences using Charniak's syntactic parser (Charniak, 2000) trained on the Prague Dependency Treebank (Hajič, 1998). Some sentences could not be parsed because of their enormous length resulting from bad segmentation in the Czech National Corpus. After excluding unparsed sentences, 7,778 sentences remained. There were 61.2 sentences per base lemma in average, ranging from a single sentence to 100 sentences (the original amount in the VALEVAL).

Note that the number of sentences for different lemmas is not related to the real data distribution.

3. Verb description

For automatic verb frame disambiguation, we generated a vector of features describing each instance of a verb. We experimented with several types of features containing dif-

Feature type	#Features	#Features used	Relative weight
Morphological	60	21	24.28%
Syntax-based	103	22	58.40%
Idiomatic	118	1	0.82%
Animacy	14	9	5.76%
WordNet	128	25	10.74%
Total	**423**	**78**	**100.00%**

The column "#Features used" indicates the number of features used in the decision trees.
The column "Relative weight" indicates the weight based on the feature occurrences in the decision trees.

Table 1: Types of features.

ferent information about the context of the verb within one sentence. The following list describes the five different types of features we used.

- **Morphological**: purely morphological information about lemmas in a small window centered around the verb.
- **Syntax-based**: information based on the result of the syntactic parser (including mainly morphological and lexical characteristics).
- **Idiomatic**: occurrence of idiomatic expressions in the sentence according to the VALLEX lexicon.
- **Animacy**: information about animacy of nouns and pronouns syntactically dependent on the verb and occurring anywhere in the sentence.
- **WordNet**: information based on the WordNet top-ontology classes of the lemmas both syntactically dependent on the verb and occurring anywhere in the sentence.

Detailed description of each feature group follows.

3.1. Morphological features

Czech positional morphology (Hajič, 2000) uses morphological tags consisting of 12 actively used positions, each stating the value of one morphological category. Categories which are not relevant for a given lemma (e.g. tense for nouns) are assigned a special value.
For lemmas within a five-word window centered around the verb (two preceding lemmas, the verb itself, and two following lemmas) we used each position as a single feature. Hence we obtained 60 morphological features (5 lemmas, 12 features for each).

3.2. Syntax-based features

Based on the result of an automatic syntactic parser we extracted the following features:

- Two boolean features stating whether there is a pronoun *se* or *si* dependent on the verb.
- One boolean feature stating whether the verb depends on another verb.
- One boolean feature stating whether there is a subordinate verb dependent on the verb.
- Six boolean features, one for each subordinating conjunction defined in the VALLEX lexicon (*aby*, *ať*, *až*, *jak*, *že* and *zda*), stating whether this subordinating conjunction occurs dependently on the verb.
- Seven boolean features, one for each case, stating whether there is a noun or a substantive pronoun in the given case directly dependent on the verb.
- Seven boolean features, one for each case, stating whether there is an adjective or an adjective pronoun in the given case directly dependent on the verb.
- Three boolean features, one for each degree of comparison (positive, comparative, superlative), stating whether there is a lemma in the given degree directly dependent on the verb.
- Seven boolean features, one for each case, stating whether there is a prepositional phrase in this case dependent on the verb.
- 69 boolean features, one for each possible combination of preposition and case, stating whether there is the given preposition in the given case directly dependent on the verb.

Together, we used 103 syntax-based features.

3.3. Idiomatic features

We extracted a single boolean feature for each idiomatic expression defined in the VALLEX lexicon. We set the value of the corresponding feature to *true* if all words of the idiomatic expression occurred anywhere in the sentence contiguously. Features corresponding to not occurring idiomatic constructions were set to *false*.
Together, we obtained 118 idiomatic features.

3.4. Animacy

We partially determined animacy of all nouns and pronouns in the sentence (method described below). Then, we introduced seven boolean features, one for each case, stating whether there is an animate noun or pronoun in this case syntactically dependent on the verb. Moreover, we introduced another seven boolean features, one for each case, stating whether there is an animate noun or pronoun in this case anywhere in the sentence.
Together we obtained 14 features for animacy.

We determined the animacy using several techniques.
For nouns, the Czech lemmatizer (Hajič, 2000) gives additional information about some lemmas. This includes identification of first names and surnames, among others. In cases where the lemmatizer marked a lemma as a name we set the animacy to *true*. We also used the fact that the morphological category *gender* distinguishes between masculine animate and masculine inanimate in some cases, as Czech masculines behave differently for animate and inanimate nouns. However, for common feminine and neutrum nouns we could not determine the animacy.
As for pronouns, the morphological category *detailed part of speech* gives us information about the type of the pronoun. Some types of pronoun imply animacy. Again, not all cases can be determined in this way.
In cases where we could not determine the animacy, we set the feature to *false*.

3.5. WordNet features

In some cases, the dependence of a certain lemma type on a verb can imply its particular sense. We described lemmas in terms of belonging to WordNet (Fellbaum, 1998) classes. In the first step, we used the definition of WordNet top ontology made at University of Amsterdam (Vossenm et al., 1997) which defines a tree-based hierarchy of 64 classes.
Then for each lemma present in the definition of the top ontology, we used the WordNet **Inter-Lingual-Index** to map English lemmas to the Czech EuroWordNet (Pala and Smrž, 2004), extracting all Czech lemmas belonging to the top level classes. After this step we ended up with 1,564 Czech lemmas associated to the WordNet top-level classes. As we worked with lemmas, instead of synsets, one lemma could have been mapped to more top-level classes. Moreover, if a lemma is mapped to a class, it belongs also to all the predecessors of the class.
In the second step, we used the relation of **hyperonymy** in the Czech WordNet to determine the top-level class for other nouns as well. We followed the relation of hyperonymy transitively until we reached a lemma assigned in the first step.
For each top level class we created one feature telling whether a noun belonging to this class is directly dependent on the verb, and one feature telling whether such noun is present anywhere in the sentence.
This resulted into 128 WordNet class features.

4. Disambiguation Methods

We trained machine learning methods for each verb separately using 10-fold cross-validation.
We tested three different classification methods, namely Naïve Bayes classifier, decision trees and rule-based learning, the latter two implemented in the machine learning toolkit C5.0 (Quinlan, 2005). The results of decision trees and rule-based methods are strongly correlated as C5.0 derives the rules from decision trees. Still, the rule-based methods are different classifiers and could perform differently, according to the author's statement.
As a baseline for each base lemma we chose the most frequent frame according to the relative frequency using 10-fold cross-validation.

5. Evaluation

The overall baseline computed as the weighted average of the individual lemma baselines was 68.27% when weighting by the number of sentences in our dataset and 60.64% when weighting by the relative frequency in the Czech National Corpus (CNC)[1].
We tested performance of automatic disambiguation classifiers based on each presented type of features separately, as well as on different combinations of feature types.
Table 2 presents the achieved accuracy for different combinations of features. Columns correspond to different disambiguation methods – Naïve Bayes classifier (NBC), decision trees (DT), and rule-based learning (RBL). The symbol $\emptyset_{data}$ indicates the average accuracy weighted by the number of sentences in the input data, whereas the symbol $\emptyset_{CNC}$ indicates the average accuracy weighted by the relative frequency in the CNC.
Syntactic features appeared to perform best, achieving accuracy 70.65% over the baseline 60.64% (using rule-based learning and CNC-weighting). Morphological features turned out to be the second best type (accuracy 66.26%). Idiomatic features scored worst. They brought little improvement when combined with other types of features. We achieved the best accuracy of 77.05% using the full set of features for CNC weighting.

5.1. Feature Importance

We summed the number of applications of individual features in decision trees weighted by the 0.5-based exponent of the level in which they occurred (1 for the root, 0.5 for first level, 0.25 for second level, ...) over the whole data (over all 10 runs of cross-validation). 78 features were used at least once, and 345 features were not used at all. Table 1 displays the number of features for each feature type, the number of them that have been used in the decision trees and the summed weighted number of applications collected from the decision trees for the full feature set.
Table 3 shows the features which resulted as the most important ones, and their respective relative weights. Syntax-based features were used most often for important decisions.

6. Conclusion

We have performed automatic disambiguation of verb valency frames using machine learning techniques. We have tried various types of features describing context of verbs. Syntax-based features have shown to be most effective.
Currently we are working on applying the methods on larger lexical resources, namely the tectogrammatically annotated part of the Prague Dependency Treebank, and the PropBank.
We are also aiming to improve the feature set by elaborating individual groups of features, using a richer idiomatic lexicon, and extending the coverage of semantic classes.

[1]Weighting by number of sentences in the dataset shows the performance of the methods, however weighting by the relative frequencies in the Czech National Corpus depicts how the methods performs on real data.

Type of features	$\oslash_{data}$ NBC	DT	RBL	$\oslash_{CNC}$ NBC	DT	RBT
Baseline		68.27			60.64	
Morphological	71.88	73.83	74.25	62.06	66.26	65.33
Syntax-based	77.05	78.33	78.23	70.46	70.65	70.77
Idiomatic	68.31	68.37	68.31	60.97	60.93	60.73
Animacy	65.89	70.77	70.76	52.84	62.58	62.46
WordNet	63.01	70.64	70.59	45.4	60.21	60.04
M + S	73.51	78.9	78.7	63.98	69.48	68.97
M + W	72.69	73.85	73.9	62.08	66.07	66.47
S + A	73.51	78.58	78.48	63.51	70.69	71.19
S + I	77.14	78.29	78.32	69.87	70.69	71.06
S + W	73.8	78.49	78.86	59.87	71.15	71.28
M + S + I + A + W	74.59	79.6	79.86	64.68	76.97	77.05

Table 2: Accuracy [%] of the frame disambiguation task

Feature type	Feature description	Weight
Syntax-based	Presence of reflexive particle *se* dependent on the verb	51.5
Syntax-based	Presence of preposition in accusative dependent on the verb	26
Morphological	Gender of the word following the verb	17.5
Syntax-based	Presence of a noun or a nominal pronoun in dative dependent on the verb	13.5
Morphological	Part of speech of the word following the verb	8
Morphological	Gender of the verb	7.5
Syntax-based	Presence of preposition *z* in genitive dependent on the verb	7
Morphological	Voice of the verb	6.25
Syntax-based	Presence of preposition in dative dependent on the verb	6.125
Syntax-based	Presence of a verb (in infinitive) dependent on the verb	6

Table 3: Features most often chosen in the decision trees

7. Acknowledgement.

The research reported in this paper has been partially supported by the project of Information Society No. 1ET101470416, the grant of Grant Agency of the Czech Republic GA405/06/0589, and the grant of the Grant Agency of the Charles University No. 372/2005/A-INF/MFF.

8. References

Ondřej Bojar, Jiří Semecký, and Václava Benešová. 2005. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*, pages 132–139, Seattle, Washington, USA, April.

Katrin Erk. 2005. Frame assignment as word sense disambiguation. In *Sixth International Workshop on Computational Semantics (IWCS)*, Tilburg.

Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database*. The MIT Press.

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning*, pages 106–132.

Jan Hajič. 2000. Morphological Tagging: Data vs. Dictionaries. In *Proceedings of ANLP-NAACL Conference*, pages 94–101, Seattle, Washington, USA.

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. 2000. *Czech National Corpus - introduction and user handbook (in Czech)*. FF UK - ÚČNK, Prague.

Markéta Lopatková, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdeněk Žabokrtský. 2005. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In *8th International Conference on Text, Speech and Dialogue*.

Karel Pala and Pavel Smrž. 2004. Building czech wordnet. *Romanian Journal of Information Science and Technology*, pages 79–88.

J. R. Quinlan. 2005. Data mining tools see5 and c5.0. http://www.rulequest.com/see5-info.html.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

P. Vossenm, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. 1997. The eurowordnet base concepts and top ontology. Technical report.

Zdeněk Žabokrtský and Markéta Lopatková. 2004. Valency Frames of Czech Verbs in VALLEX 1.0. In *proceedings of the Workshop of the HLT/NAACL Conference*, pages 70–77, May 6, 2004.