Extensive Study on Automatic Verb Sense Disambiguation in Czech

Jiří Semecký, Petr Podveský

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

Abstract. In this paper we compare automatic methods for disambiguation of verb senses, in particular we investigate Naïve Bayes classifier, decision trees, and a rule-based method. Different types of features are proposed, including morphological, syntax-based, idiomatic, animacy, and WordNet-based features. We evaluate the methods together with individual feature types on two essentially different Czech corpora, VALEVAL and the Prague Dependency Treebank. The best performing methods and features are discussed.

1 Introduction

Verb sense disambiguation (VSD) is an interesting and challenging problem of assigning the right sense to a given verb according to context. VSD aims at selecting the right sense using surrounding words or, perhaps, a thorough analysis of larger context. Verbs are usually central elements of sentences, therefore, the key aspect in determining the meaning of the whole sentence is a proper analysis of the verb sense. A verb can have several senses, for example in Czech the verb *dodat* can mean *to supply* or *to add*. VSD can also help in improving other NLP tasks, such as machine translation, information retrieval, etc.

Previous experiments on VSD have been already reported in the literature, e.g. [1] and [2] studied English VSD; initial experiments on Czech VSD have been also published [3]. Related problems are studied in the Curpus Pattern Analysis project http://nlp.fi.muni.cz/projekty/cpa/ and in [4].

In this paper we focus on automatic VSD methods. We propose novel elaborate features and employ them in standard automatic classifiers. We evaluate our approach on two corpora.

The paper is divided as follows. Section 2 introduces the corpora and lexicons that we used in our experiments. Section 3 describes the proposed features in detail. Section 4 covers the machine learning methods which we used for VSD. In Section 5, we summarize and evaluate achieved results.

2 Data

In this section we describe corpora which were used throughout our experiments. We worked with two corpora VALEVAL and the Prague Dependency Treebank

	# unique verbs	# annotated running verbs	Ørunning verbs per verb	Øsenses per running verb
VALEVAL	109	7,779	71.4	4.58
PDT	1,636	67,015	41.0	14.8

Table 1. Corpora statistics after parsing and cleaning.

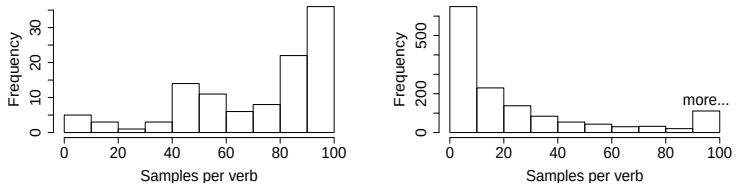
2.0. Verb senses are not directly annotated in the corpora, instead, the verbs are annotated with valency frames. The valency lexicon which was used for annotation of VALEVAL was VALLEX version 1.0 [3]. The valency frame annotation of PDT corpus was done according to valency lexicon PDT-VALLEX. Verb valency frames are closely related to verb senses. In addition, in the valency lexicons, different verb senses even with the same configuration of syntactical constituents are labeled with two different frames. For example the verb *chovat* with accusative object have in Czech two different meanings: *cuddle*, and *breed*. In both valency lexicons, the two meanings are described by two different frames. As there is no straightforward procedure to determine the verb reflexivity, Verbs with reflexive particles are assumed to be variants of the main verb.

VALEVAL. VALEVAL contains randomly selected sentences from the Czech National Corpus [5]. 109 representative verbs were chosen to form VALEVAL. For each verb, 100 sentences were selected from the Czech National Corpus to constitute VALEVAL. For more details about the verb selection, see [6].

The corpus was independently annotated by three annotators. The inter-annotator agreement of all three annotators was 66.8%, the average pairwise match was 74.8%. Sentences on which the three annotators disagreed were double-checked by an expert who determined the correct annotation. Sentences with an obvious mistake were corrected.

To prepare the data for subsequent feature extraction, we automatically parsed the sentences using Charniak's syntactic parser [7]. The parser was trained on the Prague Dependency Treebank [8]. Some sentences could not be parsed due to their enormous length. Such long sentences were excluded from our corpus yielding the total number of 7,779 parsed sentences. In the parsed corpus, a verb occured 71.4 times in average, ranging from a single occurrence to 100 occurrences. The average number of senses per verb was 4.58, the average was computed over the corpus.

Prague Dependency Treebank 2.0 (PDT). PDT is a large corpus of manually annotated Czech data with linguistically rich information. PDT is based on the theory of Functional Generative Description [9]. It contains three layers of annotation – morphological, analytical, and tectogrammatical. We worked only with the tectogrammatically annotated part of the corpus. It contains about 800 thousand words. The verb frame annotation was done according to the PDT-VALLEX lexicon.



VALEVAL PDT

Fig. 1. Distributions of the number of samples per lemma.

We automatically parsed PDT using the MST parser [10] trained on PDT using deleted interpolation. The tectogrammatical annotations were done only by one annotator, therefore the PDT data may be more biased than VALEVAL corpus. We excluded verbs which were only present either in the training set or in the testing set. This resulted in 67,015 annotated verbs occurences. For training we used the *train* portion of PDT which was comprised of 58,304 sentences. For testing we used so-called *dtest* portion which had 8,711 sentences. The number of unique verbs was 1,636. There were 41.0 occurrences of verb in average, ranging from two occurrences (one in each part of data) to 11,345 occurrences (for the verb *být*).

Table 1 summarizes the basic statistics of the corpora. Figure 1 shows distribution of the number of verb occurences in VALEVAL and PDT corpora respectively.

3 Features

Features are essential to any automatic classification method. Each occurrence of a verb in a context is described by a vector of features. Based on this feature vector, a verb sense is assigned. Features reflect various information about the context of a verb. We worked only with features with context confined to the actual sentence. No information behind sentence boundary was considered. We experimented with five types of features, namely morphological features, syntax-based features, idiomatic features, animacy features, and WordNet-based features. In the following paragraphs, we thoroughly describe each group of features.

3.1 Morphological features

Morphological features are reliably estimated, and easy to obtain. Czech positional morphology [11] uses tags with 15 positions, out of which we used first 12 positions. Each position expresses one morphological category: part of speech, detailed part of speech, gender, number, case, possessor's gender, possessor's

number, person, tense, grade, negation and voice. Categories which are not relevant for a given word are assigned a special void value.

We introduced one feature for each possition of the current verb tag. Moreover, we added tag features for two preceding words, and two following words. Thus we obtained 60 morphological features (5 words times 12 features).

3.2 Syntax-based features

We believe that syntax can capture deeper relation crucial to sense disambiguation, therefore we added the following features based on syntax:

- 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, each for one subordinating conjunction defined in the VALLEX lexicon (*aby*, *ať*, *až*, *jak*, *že* and *zda*) stating whether this subordinating conjunction depends 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.
- 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.

  All together we proposed 100 syntax-based features.

3.3 Idiomatic features

Idiomatic constructions can alter verb sense. 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 idiomatic expressions which did not occure in the sentence were set to *false*. In total we obtained 118 idiomatic features.

3.4 Animacy

We partially determined animacy of nouns and pronouns in the sentence using information from lemmatization and morphological analysis. 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 stating the same information for animate nouns and pronouns anywhere in the sentence. Together we obtained 14 features for animacy.

Feature type	#Features	VALEVAL		PDT	
		#Features used	Relative weight [%]	#Features used	Relative weight [%]
Morphological	60	27	35.92	44	45.37
Syntax-based	100	23	46.28	39	30.76
Idiomatic	118	3	0.85	16	1.20
Animacy	14	8	5.25	9	3.12
WordNet	128	44	11.70	92	19.55
Total	**420**	**105**	**100**	**200**	**100**

Table 2. Types of features. The column "#Used features" 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.

3.5 WordNet features

Dependency of a certain lemma or a certain type of lemma on a verb can imply its particular sense. We described the type of a lemma in terms of WordNet [12] classes.

In the first step, we used the definition of WordNet top ontology [13] to obtain a tree-like hierarchy of 64 classes. Then, for each lemma captured in the definition of the top ontology, we used the WordNet **Inter-Lingual-Index** to map English lemmas to the Czech EuroWordNet [14], extracting all Czech lemmas belonging to the top level classes. We ended up with 1,564 Czech lemmas associated to the WordNet top-level classes. Moreover, if a lemma was mapped to a class, it belonged also to all the predecessors of the class.

In the second step, we used the relation of **hyperonymy** in the Czech EuroWordNet 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. As we worked with the lemmas instead of synsets, one lemma could be mapped to many top-level classes.

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 Methods

To disambiguate verb senses, we tried several machine learning methods: Naïve Bayes classifier, decision trees, and a rule-based method.

Naïve Bayes is a straightforward probabilistic classifier based on an assumption that features are independent of each other. We did not expect this classifier to perform very well but rather use it for a direct comparison.

Decision tree is an algorithm based on the *divide and conquer* principle. It finds the most discriminative feature, and divides the training data into groups according to feature's possible values. The procedure is applied recursively for each group which results in a tree of decisions. Nodes of the tree represent tests on feature values.The decision tree divides the feature space into disjunctive parts. We tried two different implementations of the decision tree algorithms, namely Christian Borgelt's implementation of decision trees [15] (using information gain ratio as the attribute selection measure), and the comercial toolkit C5.0 [16], which implements an improved version of the C4.5 algorithm.

The rule-based classifier generates a set of independent *if-then* rules. Conditions of the rules may overlap, in which case the rules have to compete to reach the verdict. We used rule-based classifier implemented in the C5.0 toolkit which constructs the rules from the decision trees. Therefore, the results of the two methods are strongly correlated. However, the final classifier might differ from the C5.0 decision tree classifier.

5 Results

5.1 Baseline of frame disambiguation

As a baseline we chose the most common frame according to the relative frequency. The baseline is computed individually for each verb. For VALEVAL cor-

Type of features	**VALEVAL**				**PDT**			
	NBC	**dtree**	**C5-DT**	**C5-RB**	**NBC**	**dtree**	**C5-DT**	**C5-RB**
baseline	60.7				73.2			
Morphological (M)	61.62	59.81	65.66	66.48	74.42	75.26	75.73	75.86
Syntax-based (S)	69.98	69.34	70.70	70.68	78.64	78.76	79.08	79.04
Animacy (A)	52.87	59.86	62.62	62.49	71.61	72.82	73.50	73.53
Idiomatic (I)	60.89	60.21	60.86	61.03	73.77	73.71	73.55	73.54
WordNet (W)	45.32	53.62	60.95	59.67	68.97	71.53	72.52	72.68
M + S	63.52	60.25	68.81	68.97	76.16	76.13	78.85	78.96
M + I	61.65	59.81	67.66	67.96	74.39	75.31	76.09	76.23
M + W	62.03	59.87	67.58	66.26	74.70	75.15	74.92	75.34
S + W	59.37	60.85	70.94	70.86	76.00	77.41	78.10	78.28
M + S + I	63.52	60.25	68.00	70.07	76.40	76.23	79.19	79.28
M + S + A	63.13	58.19	70.64	69.37	76.21	75.94	78.92	79.08
M + S + W	64.80	60.28	76.69	77.03	76.44	76.01	78.37	78.91
M + S + I + W	64.78	60.28	76.86	77.16	76.55	76.10	78.82	79.20
M + S + A + W	64.59	58.36	76.35	77.03	76.25	75.93	78.21	78.72
M + S + A + I + W	64.58	58.36	77.06	77.21	76.47	76.02	78.58	79.08

Table 3. Accuracy [%] of the frame disambiguation task for PDT corpus. Columns in the table corespond to individual disambiguation methods – Naïve Bayes classifier (NBC), Borgelt's implementation of decision trees (dtree), C5 decision trees (C5-DT), and C5 rule-based classifier (C5-RB).

Feature type	Feature description	Weight
Syntax-based	Presence of reflexive particle *se* dependent on the verb	291.0
Syntax-based	Presence of noun or a subst. pron. in dative dep. on the verb	64.6
Syntax-based	Presence of reflexive particle *si* dependent on the verb	61.7
Morphological	Detailed part of speech of the word following the verb	56.8
Syntax-based	Presence of preposition *do* with genitive dependent on the verb	39.8
Morphological	Detailed part of speech of the word two possitions after the verb	36.6
Syntax-based	Presence of noun or a subst. pron. in accusative dep. on the verb	36.3
Syntax-based	Presence of preposition in dative dependent on the verb	35.1
Syntax-based	Presence of noun or a subst. pron. in nominative dep. on the verb	34.6
Syntax-based	Presence of preposition *na* with accusative dependent on the verb	32.1

Table 4. Features most often chosen in the decision trees on PDT.

pus, we computed the baseline using 10-fold cross-validation. Then, we weighed baselines of individual verbs by the relative frequency as observed in the Czech National Corpus. The weighted baseline was 60.7%.

For PDT, we acquired the most common frame from the training data and measured the baseline on the testing data. The baseline was 73.2%.

5.2 Evaluation

We tested performance of classifiers on both corpora using each feature type separately. We also experimented with different combinations of feature types. Table 3 states accuracy for VALEVAL and PDT corpus, respectively. The table shows that the syntactic features performed best among individual feature types. Morphological features turned out to be the second best. On VALEVAL, we achieved the best with the full feature set. On PDT, the best accuracy was achieved using combination of morphological, syntax-based and idiomatic features.

For VALEVAL it was 77.06% over the baseline of 60.7%. For PDT it was 79.28% over the baseline of 73.2%.

To compare individual features, we computed scores which represent importance of the features in the constructed decision trees. We summed the number of applications of features weighted by the 0.5-based exponent of the level in which they occurred (i.e. 1 for root, 0.5 for the first level, 0.25 for the second level, ...). Table 4 shows the features with the highest weights on PDT corpus. Syntax-based features were used most often for important decisions. From the total amount of 420 features, 105 features were used in VALEVAL corpus, and 200 features were used in PDT corpus. Details can be seen in Table 2.

6 Conclusion

We have compared performance of different machine learning methods for automatic verb sense disambiguation on two qualitatively and quantitatively different

corpora. We have investigated performance of various feature types describing local context of annotated verbs. Syntax-based features have shown to be the most effective of all feature types.

7 Acknowledgement

The research reported in this paper has been partially supported by the project of Information Society No. 1ET101470416, the grants of the Grant Agency of the Charles University No. 372/2005/A-INF/MFF and 375/2005/A-INF/MFF, and the grant MSM0021620838.

References

1. Dang, H.T., Palmer, M.: The Role of Semantic Roles in Disambiguating Verb Senses. In: Proceedings of ACL, Ann Arbor MI (2005)
2. Ye, P.: Selectional Preferenced Based Verb Sense Disambiguation Using WordNet. In: Australasian Language Technology Workshop 2004, Australia (2004) pp. 155–162
3. Lopatková, M., Bojar, O., Semecký, J., Benešová, V., Žabokrtský, Z.: Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In: 8th International Conference on TSD. (2005) pp. 99–106
4. Král, R.: Jaký to má význam? PhD thesis, Masaryk University (2004)
5. Kocek, J., Kopřivová, M., Kučera, K., eds.: Czech National Corpus - introduction and user handbook (in Czech). FF UK - ÚČNK, Prague (2000)
6. Bojar, O., Semecký, J., Benešová, V.: VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. Prague Bulletin of Mathematical Linguistics **83** (2005)
7. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of NAACL-2000, Seattle, Washington, USA (2000) pp. 132–139
8. Hajič, J.: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. Issues of Valency and Meaning (1998) pp. 106–132
9. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in its Semantic and Pragmatic Aspects. Academia, Prague, Czech Republic/Reidel Publishing Company, Dordrecht, Netherlands (1986)
10. McDonald, R., Pereira, F., Ribarov, K., Hajic, J.: Non-Projective Dependency Parsing using Spanning Tree Algorithms. In: Proceedings of HLT Conference and Conference on EMNLP, Vancouver, Canada, ACL (2005) pp. 523–530
11. Hajič, J.: Morphological Tagging: Data vs. Dictionaries. In: Proceedings of ANLP-NAACL Conference, Seattle, Washington, USA (2000) pp. 94–101
12. Fellbaum, C.: WordNet An Electronic Lexical Database. The MIT Press (1998)
13. Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., Peters, W.: The EuroWordNet Base Concepts and Top Ontology. Technical report (1997)
14. Pala, K., Smrž, P.: Building Czech Wordnet. Romanian Journal of Information Science and Technology **7**(1-2) (2004) pp. 79–88
15. Borgelt, C.: A Decision Tree Plug-In for DataEngine. In: Proceedings of 2nd Data Analysis Symposium, Aachen, Germany, MIT GmbH (1998)
16. Quinlan, J.R.: Data Mining Tools See5 and C5.0 (2005) http://www.rulequest.com/see5-info.html.