PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 31–52

Verb Valency Frames Disambiguation

Dissertation Summary

Jiří Semecký

Abstract

This is a summary of the author's PhD dissertation defended on September 17, 2007 at the Faculty of Mathematics and Physics, Charles University in Prague. Semantic analysis has become a bottleneck of many natural language applications. Machine translation, automatic question answering, dialog management, and others rely on high quality semantic analysis.

Verbs are central elements of clauses with strong influence on the realization of whole sentences. Therefore the semantic analysis of verbs plays a key role in the analysis of whole sentences. We believe that solid disambiguation of verb senses can boost the performance of many real-life applications.

In this thesis, we investigate the potential of statistical disambiguation of verb senses. Each verb occurrence can be described by diverse types of information. We investigate which information is worth considering when determining the sense of verbs. Different types of classification methods are tested with regard to the topic. In particular, we compared the Naïve Bayes classifier, decision trees, rule-based method, maximum entropy, and support vector machines. The proposed methods are thoroughly evaluated on two different Czech corpora, VALEVAL and the Prague Dependency Treebank. Significant improvement over the baseline is observed.

1. Introduction

Natural language processing (NLP) research has already grown up from the early phases of its life. Many tasks concerning the early stages of the linguistic analysis of written text, including lemmatization, morphological tagging and surface parsing, might today be considered sufficiently resolved for the mainstream NLP languages. Even if their development will probably further continue to improve, their current results are near to approaching the upper limits and they are already good enough for many practical applications.

However, the complex linguistic applications, including machine translation, question answering, dialog systems, information retrieval, and others need a deeper semantic analysis of

Please cite this article as: Jiří Semecký, Verb Valency Frames Disambiguation: Dissertation Summary. The Prague Bulletin of Mathematical Linguistics No. 88, 2007, 31–52.

text which is becoming the center of interest for current NLP research. Such an analysis tries to understand and describe not only the structure of text but also its meaning. But not all parts of speech are equally important for deep analysis.

Verbs have special roles in the analysis of text. From the syntactic point of view they are the central elements of clauses with direct influence on the presence and realization of other constituents. From the semantic point of view they are the bearers of events and their proper analysis is fundamental for a correct analysis of the rest of the sentence.

Moreover, verbs are also interesting from the linguistic perspective because they have the richest syntactic structure and also the highest level of ambiguity compared to other parts of speech.

Let us take a highly ambiguous Czech verb *dát* as an example. If we want to translate the verb into English, the most obvious translation will be *to give* as in the sentence:

*Petr **dal** Janě knihu. = Peter **gave** Jane a book.*

If we use the verb in combination with a reflexive particle *si* it changes the meaning of the sentence, and the verb needs to be translated as *put*:

*Petr **si dal** klíče do kapsy. = Peter **put** his keys in his pocket.*

Even with the same syntactic structure, we can get a completely different meaning which, again, translates differently:

*Petr si dal Guinness do půllitru. = Peter **ordered** a pint of Guinness.*

Needless to say, that when used in an idiomatic expression, the verb has a completely different translation:

*Petr **si** na tom **dal** záležet. = Peter **made** a point of it.*

*Petr **dal** na jeho slova. = Peter **took** what he said **into** account.*

*Petr **se dal** konečně dohromady. = Peter finally **got** better.*

As has been illustrated, the same Czech verb may have different English equivalents, depending on the sense in which it is used. Therefore, the correct assignment of the sense seems to be essential for the translation of the sentence. For other applications dealing with the semantic content of the text, it is naturally important, too, to take these differences into account.

Our contribution concerns the process and methods of automatic selection of the proper sense of verbs in their given contexts, i.e. verb disambiguation[1] according to a certain definition of verb senses.

[1] *to disambiguate = "to remove uncertainty of meaning from"* (Oxford Dictionary)

Czech is one of the languages which are the center of study of the world-wide computational linguistic community. A significant reason for this is the fact that there is a large amount of high-quality linguistically annotated data. As there are only ten million Czech native speakers, other languages, mainly English, Chinese, French, Spanish, and Arabic definitely receive more attention because of the far larger number of target users. However, the Czech language surely has the highest ratio of linguistically annotated tokens per native speaker[2].

In our experiments we use two Czech corpora:

First, **VALEVAL**, a small but reliable corpus, containing a few thousand running verbs in contexts annotated by three annotators in parallel. The corpus was put together as a lexical sampling experiment for an existing valency lexicon, and contains sentences randomly selected from the Czech National Corpus. Only the selected verbs are annotated in the corpus. The sentences are not selected in any larger continuous blocks except for a small context attached to each annotated unit. Only the golden part of the corpus was taken into account in our experiments. This assured highly reliable labeling which had, however, low coverage and does not respect the actual verb distribution.

Second, the tectogrammatical part of the **Prague Dependency Treebank 2.0**, a large corpus, containing almost 70,000 verb tokens[3]. The tectogrammatical annotation layer describes many linguistic characteristics, including valency which was used as an approximation of verb senses as is explained below. Each sentence of the relevant portion of the Prague Dependency Treebank was annotated on the tectogrammatical layer by one annotator only, i.e. no parallel annotations were performed. Therefore, the quality of the valency annotation is not guaranteed to be as high as for the first corpus. On the other hand, the quantity highly exceeds VALEVAL and the distribution of verbs reflects the real distribution in Czech (newspaper) text.

Our disambiguation process can be simply described by a sequence of the following steps. First, we automatically analyzed linguistically the sentences containing the annotated verbs. Second, we created a vector of features for each annotated verb in the dataset, describing its context. We experimented with a large number of different features, a great attention was paid to the comparison of individual feature types. Third, the generated features were used in machine learning algorithms. Again, we experimented with several machine learning methods, including the Naïve Bayes classifier, decision trees, rule-based learning, support vector machines, and maximal entropy model. Finally, we evaluated the obtained results. In the evaluation section, we stated the results obtained by using all types of features separately, as well as using their different combinations. Also the difference in performance of individual classification methods are evaluated, as well as several other aspects.

[2] We state here this claim without precise proof, and assuming the exclusion of dead (or nearly dead) languages where the ration is (or approaches) infinity, even with a very limited corpus.

[3] The number refers only to the portion annotated on the tectogrammatical layer.

2. Word Senses

In this section, we show that what we are going to disambiguate in this work are actually not senses of verbs but their valency frames. We explain this approximation and show that under a specific assumption it does not really matter so much.

We have worked with two different lexicons, namely VALLEX, and PDT-VALLEX.

For building a statistical word sense disambiguation system, two types of data resources are needed – a lexicon defining word senses and a corpus annotated with the senses of this lexicon.

We have decided to modify the task slightly by approximating verb senses with verb valency frames. Valency is a property of verbs which correlates with the senses to a certain extent, it is formally well defined and there are lexical resources of sufficient size available describing and using verb valency. In the following paragraphs, we point out that in our choice of valency frame lexicons, the correlation between frames and senses is relatively high.

2.1. Valency

Valency (Panevová, 1980), (Panevová, 1974), (Panevová, 1994) is the ability of a lexical item to combine with another lexical items in syntactic structures. The valency is defined for four different parts of speech — verbs, substantives, adjectives and adverbs. There is no doubt that the valency of verbs is the most differentiated and therefore the most interesting for studying. In this work we are only concerned with verb valency, leaving the valency of other parts of speech aside.

Valency is described in terms of **valency frames** which defines the ability of the given lexical item to syntactically combine with other lexical items. From a technical point of view a valency frame is usually described by a central lexical item (predicate, frame evoking element, …) and a list of participants of the frame (arguments, frame elements, …) corresponding to individual lexical items linked to the central element described by their linguistic (usually morphological and syntactic) characteristics and semantic labels. Different configurations of participants imply different valency frames. The participants are further categorized in different ways, depending on the concrete valency theory (e.g. usually distinguishing the level of obligatoriness).

2.2. Approximation of senses

The valency lexicons built at the Institute of Formal and Applied Linguistics in Prague – VALLEX and PDT-VALLEX (introduced in Section 3.1) – are, however, different from the general definition in this point: the **clearly different senses of a verb with equal valency frames are distinguished in the lexicon**. The following examples demonstrate this statement:

VALLEX:
- **Frame** 1 : **ACT**$_1$ **PAT**$_4$
  absolvovat studium
  graduate from a place

- **Frame** 2 : **ACT**$_1$ **PAT**$_4$
  absolvovat operaci
  undergo an operation

PDT-VALLEX:
- **Frame** v-w1184f1 : **ACT**$_1$ **PAT**$_4$
  chová prasata na farmě.LOC
  He breeds pigs on the farm.

⋮

- **Frame** v-w1184f4 : **ACT**$_1$ **PAT**$_4$
  chová dítě v náručí.LOC
  He cuddles the child in his arms.

When the difference in the meaning was not clear, frames did not have to be differentiated which corresponds to the uncertainty in the sense distinction.

From this perspective, **verb sense** (without any precise definition) **is a function of frames** (in VALLEX and PDT-VALLEX). The frame distinction in these lexicons is in fact driven by the combination of the valency and sense characteristics. Therefore these frames can be used as a suitable approximation of senses.

For the automatic assignment of word senses we need a lexicon containing formal definitions of senses. As already suggested above, instead of using such lexicons we are using lexicons of valency frames which take senses distinction into account.

3. Data resources

In this section, we introduce the data which we used or referred to in the experiments discussed in the thesis – two valency lexicons together with two corresponding corpora. The lexicons define the senses of verbs and the corpora use those lexicons to annotate the verbs.

3.1. VALLEX and VALEVAL

3.1.1. VALLEX

VALLEX(Žabokrtský and Lopatková, 2004) is a manually created valency lexicon of Czech verbs, which is based on the theoretical framework of Functional Generative Description.

The construction of VALLEX started in 2001 and the work is still in progress. The VALLEX

version 1.0 [4] (autumn 2003) (Lopatková et al., 2003) which we used in our task and which was published in 2003, defines valency for over 1,400 Czech verbs and contains over 3,800 frames. In 2005, the VALLEX version 1.5 was published, containing roughly 2500 verbs with more than 6000 valency frames. At the time this thesis is submitted, the new version 2.0 of the VALLEX is about to be published.

The basic structure of the VALLEX lexicon is shown in Figure 3.1.1[5]. Elements of the chart are described in more detail below.

3.1.2. VALEVAL

The manually annotated corpus VALEVAL (Bojar, Semecký, and Benešová, 2005) was created in 2005 as a lexical sampling experiment for the VALLEX lexicon. It contains frame annotations for 109 base lemmas selected from VALLEX. The term **base lemma** is used for a lemma excluding its possible reflexive particle.

For all verbs in VALEVAL, their aspectual counterparts, including iterative forms, were added, too. For each base lemma, 100 sentences from the Czech National Corpus[6] (Kocek, Kopřivová, and Kučera, 2000) (a large corpus containing over 100 million of words) were randomly selected to be present in VALEVAL. This selection resulted in an average number of frames per base lemma of 6.77 (according to VALLEX definition).

[4] http://ckl.ms.mff.cuni.cz/zabokrtsky/vallex/1.0/

[5] The rough structure of PDT-VALLEX is the same as that of VALLEX.

[6] http://ucnk.ff.cuni.cz/english/index.html

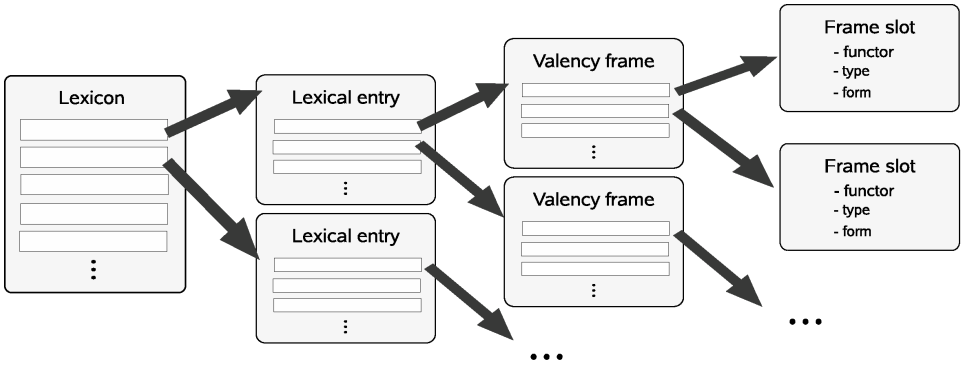


Figure 1. Structure of VALLEX and PDT-VALLEX lexicons.

3.2. Prague Dependency Treebank

The Prague Dependency Treebank (PDT) (Hajič, 2004) is a manually annotated corpus based on the theory of Functional Generative Description (FGD). Data of the PDT are part of the Czech National Corpus (Kocek, Kopřivová, and Kučera, 2000).

Data are annotated on three different layers (Hajičová, 2002), namely morphological, analytical, and tectogrammatical. This differs from the original definition of layers in the FGD.

The current version of the Prague Dependency Treebank is the version 2.0 published by the Linguistic Data Consortium in late 2006 under the number *LDC2006T01*.

Different layers contain different amounts of data. The data are organized so that each part annotated on a higher level is also annotated on all lower levels.

Moreover, the data in each section are divided into the training part, the development testing part (*dtest*), and the evaluation testing part (*etest*). The training part contains approximately 80% of the entire portion, the testing parts each contain approximately 10% of the data.

As frame annotation belongs to the tectogrammatical level, we were restricted to the tectogrammatically annotated portion of the data.

3.2.1. PDT-VALLEX

PDT-VALLEX (Hajič and Honetschläger, 2003), (Hajič et al., 2003) is a valency frames lexicon, created as a part of the PDT. It contains the definition of valency frames for four parts of speech – verbs, nouns, adjectives and adverbs. The PDT-VALLEX was created during the annotation and it contains all auto-semantic words occurring in the corpus. The lexicon was dynamically updated as the annotation went on, unlike VALLEX, described above.

4. Feature Design

To disambiguate a word or a phrase, we are looking at linguistic characteristics within its context. In our work, we look at the sentence in which the verb occurs.

The linguistic characteristics of a sentence are complex structures – trees, vectors, sets, …. On the contrary, machine learning methods can only deal with a simple description of samples, usually vectors.

The natural solution to deal with this contrast is to convert complex linguistic characteristics into simple vectors of features. As the vectors of features only describe linguistic information in a limited way, there will always be a loss of information in the feature creation process. Therefore the selection of a suitable set of features is essential for the success of the method.

4.1. Morphological features

These features are generated only from the morphological information, they are not a result of parsing.

Because syntactic parsing is computationally much more demanding than morphological tagging, those features are very simple and easy to obtain.

The morphological features are based on the Czech positional morphology (Hajič, 2000) used in the Prague Dependency Treebank. The morphological tags consist of 15 positions (characters), each stating the value of one morphological category.

In this work, we use all positions of the morphological tags, except positions 13, 14, and 15, which are not actively used.

For lemmas within a n-word window centered around the verb we used each position as a single feature.

Figure 2 shows an example of generation of morphological features for verb *odvolat – remove (from the office)*.

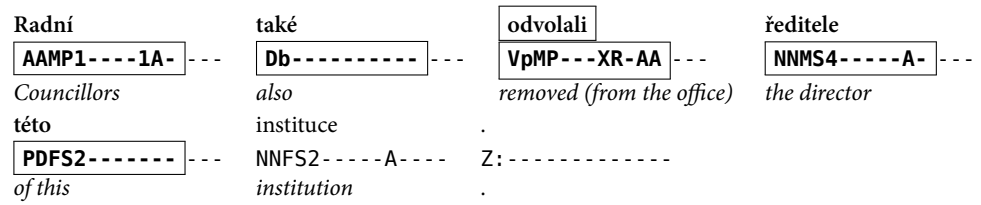


Figure 2. Generation of morphological features.

4.2. Syntax-based features

Syntax-based features, in contrast to the morphological features, are based on the result of the syntactic (analytical dependency) parser.

Syntax-based features also use morphological characteristics, but combine them with the shape of the dependency tree. As the term *syntactic features* might suggest using only syntactic information by analogy with the *morphological features* using only information about morphology, we prefer to use the term *syntax-based features*. Moreover, other types of features (idiomatic, WordNet-based, and animacy) also use the analytical syntax, however, they are in special categories because of their narrow scope.

For our experiments, we did not use a tectogrammatical parser, as we understand verb valency as a part of the tectogrammatical analysis. Therefore the tectogrammatical parsing and subsequent analysis (assignment of tectogrammatical functions) should be processed only after the valency is resolved.

We expected that syntax-based features would be very useful for the disambiguation of the valency frames as the valency frames describe the syntactic behavior of the verbs. Special care was paid to selecting the proper features. Nevertheless, since statistical parsing achieves much lower accuracy than morphological tagging, syntax-based features as opposed to morphological features can suffer much more from errors in analysis.

Based on the results of statistical syntactic parsers we extracted the following groups of features:

- Reflexive *se*
- Reflexive *si*
- Subordinate verb
- Superordinated verb
- Subordinating conjunctions
- Substantives in particular cases
- Adjectives in particular cases
- Prepositional with particular cases

A detailed description of each group follows.

4.3. Idiomatic features

Certain idiomatic expressions evoke a special (usually figurative) senses of verbs. To depict such senses, we introduced this type of features.

Each idiomatic construction (multi-word expression) described in the VALLEX lexicon was used as one boolean feature. This feature was set to *true* if this construction occurred in the raw text of the sentence containing the verb continuously. Features corresponding to non occurring idiomatic constructions were set to *false*.

In this way, we could have missed some idiomatic expressions which were in fact present in sentences but did not occur in a subsequent list of words. This could happen if the writer paraphrased the idiomatic expression. However, simply allowing the inflexion and the gaps in the multiword expression could heavily over-generate and introduce positive errors.

4.4. Animacy features

Animacy is a grammatical category of nouns and pronouns specifying whether the noun or pronoun refers to an animate object.

The introduction of the animacy features was based on an assumption that animacy can often suggest the meaning of the verb. This assumption follows from the fact that some senses of verbs can only describe a relation between (living) beings.

The main problem related to the animacy features is the difficulty of the determination of animacy. There is no simple way to determine animacy automatically, and we can only predict it for specific cases. The algorithm we used for partial animacy resolution differs for nouns and pronouns.

4.5. WordNet features

In some cases, dependency of a certain lemma or a certain type of lemma on the verb can imply a particular sense of the verb. From this perspective, it might be useful to capture the presence of each lemma among the nodes dependent on the verb. However, storing the pres-

ence for all possible lemmas would lead to a huge number of features, to a loss of generality, and possible over-fitting.

There are several possibilities of how to deal with this issue. One of them is, instead of capturing presence of each and every lemma, capturing only the "class" of the lemma. This class should generalize the meaning of each word, so words with a similar meaning should belong to the same class. This solution requires usage of some kind of ontology which maps the lemmas or meanings (disambiguated lemmas) to the classes.

WordNet (Fellbaum, 1998) seemed to be a good choise for this purpose. To define a system of coarse-grained classes of WordNet items (synsets[7]), we used the WordNet top ontology designed at the University of Amsterdam (Vossen et al., 1998). This ontology is described as a tree-based system of 64 WordNet synsets which represents the top of the WordNet hierarchy.

Using hyperonymy relation defined in WordNet we can easily determine all classes to which a given noun belongs, i.e. is related by the transitive relation of hyperonymy. This means that "the noun is type/kind of the class". Because of the transitivity of the hyperonymy relation, if a word belongs to a given class, it also belongs to all classes which are governing this class in the top-ontology.

4.5.1. Combination with Czech WordNet

For each lemma present in the synsets of the top ontology, we used the WordNet **Inter-Lingual-Index** to map the English WordNet 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 1564 Czech lemmas associated to the WordNet top-level classes.

5. Evaluation

This section summarizes the empirical results of the experiments described in this work. We ran several machine learning algorithms on two corpora using various types of features. Because of size, we used cross-validation for the VALEVAL corpus. Moreover, two different ways of counting the overall results for the VALEVAL corpus are considered. In the first one, we computed the average of the results for individual lemmas weighted by the frequencies in the corpus, but in the second one, we weighted the results by the relative frequencies measured in the Czech National Corpus relative frequencies measured in the Czech National Corpus (CNC) (Kocek, Kopřivová, and Kučera, 2000). For the Prague Dependency Treebank, we presented results for two different evaluation data sets – the development test set, and the evaluation test set. We used the development test set throughout the development period and only performed the evaluation on the evaluation data set once, for the purpose of this thesis. After that, we did not modify the methods anymore.

[7] The term **synset** is used in the WordNet for a lexicon item capturing single meaning. One lemma can belong to more synsets (suggesting different meaning of the lemma), as well as one synset can consist of more lemmas.

	VALEVAL		PDT	
	$\oslash_{data}$	$\oslash_{CNC}$	**dtest**	**etest**
Average number of frames	4.45	5.31	2.39	2.27
Baseline	68.27	60.74	73.19	71.98

$\oslash_{data}$ denotes average weighted by the number of sentences in the dataset.
$\oslash_{CNC}$ denotes average weighted by the number of sentences in the Czech National Corpus.

Table 1. Difficulty of the frame disambiguation task

As the baseline of the disambiguation task we took **the relative frequency of the most frequent frame of each lemma in the training data**. For the VALEVAL corpus, we determined the baseline using 10-fold cross validation.

For the Prague Dependency Treebank, the baseline was measured on the testing data (the dtest, and the etest section, respectively) but the most frequent frame was determined from the training data.

We computed the overall baseline as the weighted average of the individual baselines. The overall baseline for the VALEVAL corpus was 68.27% when weighted by the number of sentences in our data set and 60.74% when weighted by the relative frequency in the Czech National Corpus. The overall baseline for PDT was 73.19% for the development testing set and 71.98% for the evaluation testing set. The baseline statistics are summarized in Table 1.

5.1. Results

This section presents the evaluation results of the valency frame disambiguation using each presented type of features separately, as well as different combinations of feature types, computed by different classifiers.

Table 2 shows the results weighted by the relative frequencies in the CNC. Table 3 present the results for the Prague Dependency Treebank for evaluation testing set.

The columns of the tables correspond to different classification methods: Naïve Bayes classifier (NBC), Christian Borgelt's implementation of the decision trees (DTREE), C5 decision trees (C5-DT), and C5 rule-based learning (C5-RB), Support Vector Machines (SVM), and Maximum Entropy (ME). The rows of the table correspond to different types of features, the first five rows state the results when using each type of features separately, the following rows state the results for different combinations of the type.

The best accuracy on VALEVAL – 77.56% – was achieved by the C5 rule-based algorithm using the full set of features.

5.2. Methods Comparison

Different methods achieved different results on different data. Generally, we can claim that the C5 decision trees, C5 rulesets, Support Vector Machines and the Maximum Entropy

Corpus:	VALEVAL					
Weighting:	Relative frequencies in the Czech National Corpus					
Type of features	**NBC**	**DTREE**	**C5-DT**	**C5-RB**	**SVM**	**ME**
Baseline	60.74					
Morphological (M)	61.62	59.81	67.50	67.83	58.48	66.36
Syntactic (S)	69.98	69.34	71.01	70.43	67.90	68.51
Animacy (A)	52.87	59.86	62.32	62.67	55.12	59.60
Idiomatic (I)	60.89	60.21	61.01	61.10	60.96	62.77
WordNet (W)	45.32	53.62	58.34	59.22	50.72	54.30
M + S	63.52	60.25	69.69	69.15	63.34	64.11
M + I	61.65	59.81	67.77	68.40	58.61	63.65
S + W	59.37	60.85	71.28	70.87	60.60	61.70
S + A	63.44	61.67	70.56	70.56	63.96	63.26
S + I	69.42	69.61	70.96	70.55	68.03	69.95
M + S + I	63.52	60.25	69.27	68.54	63.43	68.76
M + S + A	63.13	58.19	69.91	69.46	64.39	64.74
M + S + W	64.80	60.28	76.61	75.08	65.27	62.62
S + A + W	60.68	61.43	70.65	71.07	58.75	65.05
S + A + I	63.32	61.67	70.95	71.31	64.04	67.22
S + I + W	59.63	60.94	71.10	71.23	61.57	65.84
M + S + I + W	64.78	60.28	76.90	77.25	65.30	63.62
M + S + A + W	64.59	58.36	76.85	77.10	62.62	67.51
S + A + I + W	60.78	61.43	71.33	71.31	58.67	64.65
M + S + A + I + W	64.58	58.36	76.97	77.56	62.64	67.45

Results are obtained by weighting individual results with the relative frequencies in the Czech National Corpus.

Table 2. Accuracy [%] of the frame disambiguation task for VALEVAL corpus.

model achieved comparably good results throughout the experiments. As has already been mentioned, we did not expect the Naïve Bayes classifier to beat other state-of-art methods. The second implementation of the decision trees algorithm (DTREE) also did not achieve results comparable with C5.

The **C5** algorithm proved to be a reliable classification method. Compared to other methods, it performed well even if the number of training samples was low. When the number of samples was higher, the Maximum Entropy models tended to outperform C5.

C5 decision trees and rule-sets are comparably powerful, sometimes one scores slightly better, sometimes the other one does. The differences are usually not significant. Still, the rule-sets seemed to work slightly better in our tasks, which corresponds to the statement of the C5's authors. On the PDT evaluation test set, both C5 algorithms achieved the same result (78.06%).

The C5 method showed some gain even with very poor feature sets (animacy or idiomatic features alone), compared to other methods which usually scored below the baseline. As a matter of fact, the C5 methods never scored worse than the baseline, which does not hold for any other method examined.

Corpus:	PDT - etest					
Weighting:	Sample counts in the corpus.					
Type of features	**NBC**	**DTREE**	**C5-DT**	**C5-RB**	**SVM**	**ME**
Baseline	71.98					
Morphological (M)	73.03	73.72	73.66	73.62	72.55	74.59
Syntactic (S)	77.84	77.89	77.47	77.35	78.63	78.60
Animacy (A)	70.23	71.05	72.37	72.37	71.99	71.44
Idiomatic (I)	72.45	72.26	72.49	72.49	72.59	72.35
WordNet (W)	68.04	70.41	72.14	72.09	70.15	70.58
M + S	75.24	75.18	77.48	77.54	76.78	78.06
M + I	73.30	73.73	73.66	73.73	72.82	74.89
S + W	74.89	76.43	77.66	77.50	76.35	76.85
S + A	76.19	74.22	77.51	77.40	77.19	77.70
S + I	78.17	78.15	77.76	77.66	78.88	78.85
M + S + I	75.18	75.22	77.71	77.80	76.89	78.10
M + S + A	75.52	75.09	77.25	77.33	75.75	78.09
M + S + W	75.72	74.97	77.60	77.75	76.46	78.17
S + A + W	75.12	73.61	77.00	76.93	75.37	76.89
S + A + I	76.45	74.38	77.75	77.61	77.42	78.04
S + I + W	74.98	76.68	77.80	77.66	76.56	76.95
M + S + I + W	75.79	75.00	78.06	78.06	76.70	64.48
M + S + A + W	75.67	75.10	77.74	77.76	75.93	78.00
S + A + I + W	75.35	73.74	77.57	77.50	75.51	77.07
M + S + A + I + W	75.51	75.13	77.91	78.04	76.10	78.26

Table 3. Accuracy [%] of the frame disambiguation task for the evaluation test set of the Prague Dependency Treebank.

Support vector machines is a popular classifier which is in general performing well. However, it requires a fine tuning of the parameters.

In our experiments, the linear kernel always scored best. This can be explained by the fact that we largely used boolean features which could be easily separated by a superspace in the linear space. Using a more sophisticated kernel adds freedom in the methods which makes the classifier more difficult to train. If there were more real-number features, the situation would probably differ. However, linguistic characteristics are rarely described by real-number features.

The support vector machines achieved the absolutely best result on both, the development and the evaluation testing dataset of the Prague Dependency Treebank.

5.3. Features Comparison

This section gives comparison of individual types of features.

Tables 2 and 3 show that the syntax-based features (see Section 4.2) clearly performed best

in all datasets. They contain most of the information which is linguistically relevant to the valency.

The morphological features turned out to be the second best. The strong difference between syntax-based and morphological features shows how much the statistical parsing helps to analyze the meaning of the verbs. The remaining feature types achieved similar results, usually in the following order: idiomatic features, animacy features, WordNet features.

When we look at the combination of syntax-based features with another type of features, the best result was achieved with the idiomatic features, while the combination with morphological features usually performed worst. In our opinion, this is because the information stored in the morphological features is already included in the syntactic features and adding it does not bring any new information. On the other hand, the other types of features contain information of a different kind, hence they help the syntactic features when combined.

5.4. Differences in Words

The success of the disambiguation task is not flat across all the verbs, it differs from one verb to another, according to the characteristics of the given verb. Most of the verbs have a single dominant sense which is assigned to the majority of the running verbs. Typical examples are the verbs *být* (the most frequent Czech verb), *říci* or *začít*. There are, however, other verbs, whose different senses are widely spread and used in the language. Typical examples are the verbs *mít* (the second most frequent Czech verb), *dát*, or *vědět*.

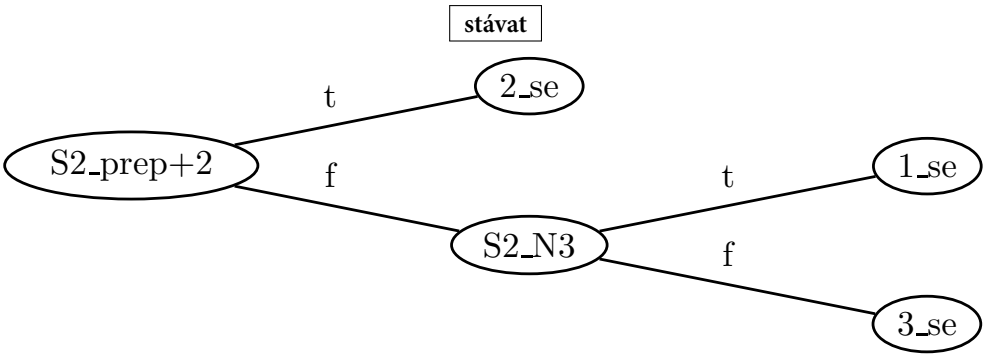
In the following sections, we present decision trees generated by the C5 algorithms. We have chosen decision trees because it is a white-box model, so they clearly show how the classifier works.

5.4.1. VALEVAL

The C5 decision trees scored worse than the baseline for eight verbs in the VALEVAL corpus. The following table lists the verbs with possible explanations of the failures:

zachytnout	(29 % loss)	low number (7) of training samples (4 frames)
spojit	(3 % loss)	high number (6) of frames
držet	(3 % loss)	high number (8) of frames
přidat	(2 % loss)	high number (7) of frames
ponechávat	(1 % loss)	
stávat	(1 % loss)	

Figure 3 shows the decision tree for the verb *stávat*, the decision trees for the other verbs from the previous list are not interesting.



S2_prep+2 …presence of a preposition in genitive dependent on the verb
S2_N3 …presence of a dative noun dependent on the verb

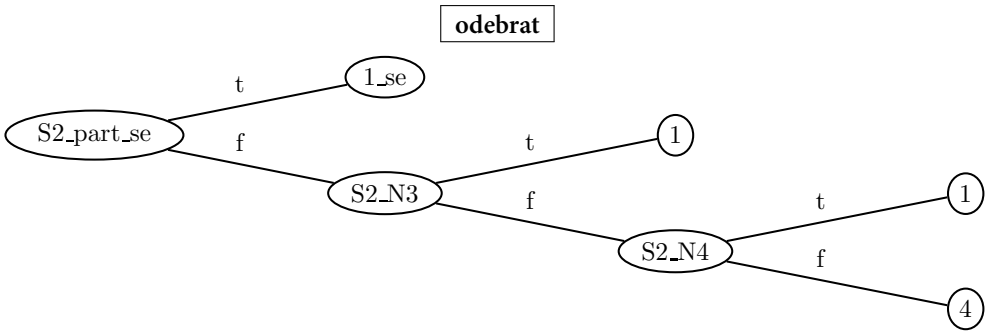
1_se	přiházet se; uskutečňovat se (Eng: *happen*) • často se mi stávalo, že jsem přišel pozdě → *lit.* it often happened to me that I came late
2_se	přeměňovat se (Eng: *become*) • pomalu se z něj stávala příšera → *lit.* slowly he became a monster
3_se	přeměňovat se v něco (Eng: *change into*) • z chlapce se stával mužem → *lit.* from a boy he changed into a man

Figure 3. Decision tree for the verb stávat from VALEVAL.

The verbs with the highest performance gain ($accuracy - baseline$) were the following:

odebrat	(48 % gain)
stát	(43 % gain)
určit	(35 % gain)
přihlížet	(33 % gain)
vyvíjet	(32 % gain)
udržovat	(31 % gain)
připadnout	(31 % gain)
orientovat	(31 % gain)
dát	(31 % gain)
umístit	(30 % gain)
vyvinout	(30 % gain)
přiznat	(30 % gain)

Figures 4 and 5 show the decision trees for the verb *odebrat* and *udržovat* respectively.



1_se	odejít; vydat se (Eng: *leave (for)*)
	• odebral se na schůzi
	→ *lit.* he left for a meeting
1	odejmout (Eng: *to take away*)
	• odebrali jí děti
	→ *lit.* they took away from her the children
4	odkoupit; převzít (Eng: *to buy, to take over*)
	• odebrali všechno objednané zboží
	→ *lit.* they bought all ordered goods

Figure 4. Decision tree for the verb odebrat from VALEVAL.

5.4.2. PDT

The C5 decision trees scored worse than the baseline for 64 verbs out of 1712. The verbs with the lowest performance were the following:

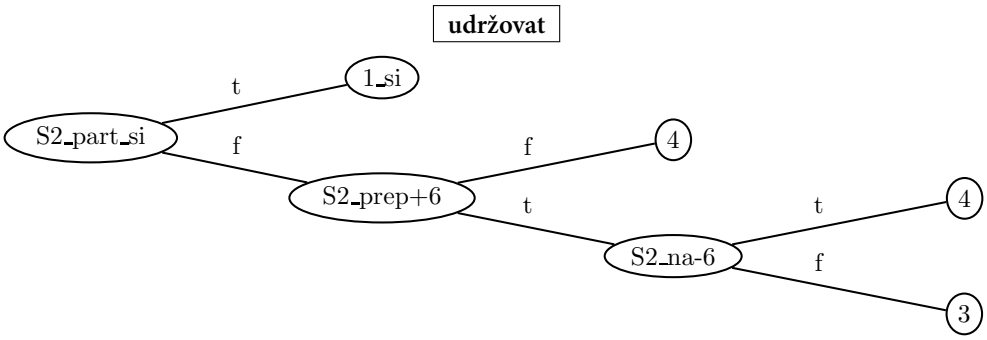
znát, *držet*, *učinit*, *přijímat*, *předpokládat*, *růst*, *fungovat*, *vyhrát*, *přinést*.

The most often reason for the fails were a low number of training data (unreliable classifier) or testing data (unreliable result), high number of frames compared to the size of training data (e.g. verb *držet* – 18 frames for 55 running verbs) and inability to distinguish two frames.

The verbs with the highest positive influence on the total performance ($accuracy-baseline$) were the following (in this order):

být, *mít*, *stát*, *dostat*, *rozhodnout*, *myslit*, *dát*.

Figures 6 and 7 show examples of decision trees for the verbs *rozhodnout* and *dělit*, respectively.



3	zachovávat v určitém stavu (Eng: *to keep in a certain state*) • udržoval byt v čistotě → *lit.* he kept the flat in good order
4	dodržet; uchránit; pečovat (Eng: *to maintain*) • udržoval kázeň / pořádek / kontakty / zahradu → *lit.* to maintain order
1_si	zachovávat (Eng: *to preserve, to keep*) • udržoval si nadhled / kondici → *lit.* he kept condition (up to the mark)

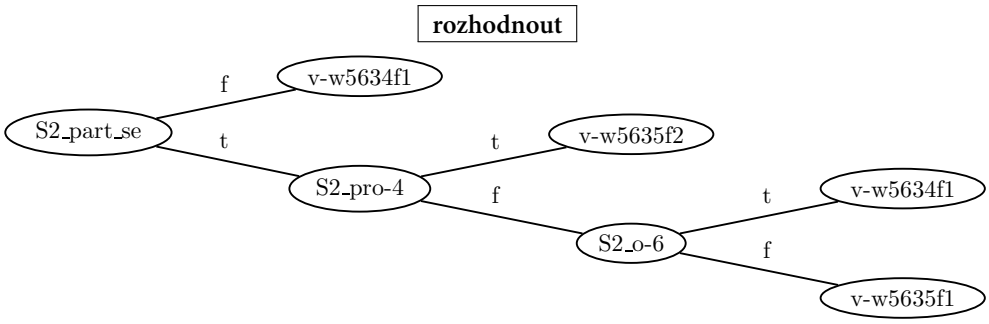
Figure 5. Decision tree for the verb udržovat from VALEVAL.

6. Conclusion

The disambiguation of verb senses in Czech has been extensively studied in this thesis. Different machine learning methods and different approaches to WSD and related tasks were introduced.

We investigated which type of information is important to consider when determining the sense of verbs. In fact, instead of senses we used the valency frames. Each verb occurrence was described by hundreds of features of five basic types. The types of the features were evaluated separately and compared to each other. The most important features turned out to be the ones using information about the surface syntax.

Experiments using different machine learning methods were performed, including the Naïve Bayes Classifier, decision trees, rule-based methods, Maximum Entropy model, and Support Vector Machines. The methods were validated on two qualitatively and quantitatively different corpora — the VALEVAL corpus and the Prague Dependency Treebank. For the smaller VALEVAL corpus, the C5 decision trees and rule-based methods turned out to be the most accurate. For the large Prague Dependency Treebank, the support vector machines and maximum entropy model performed better than other methods.



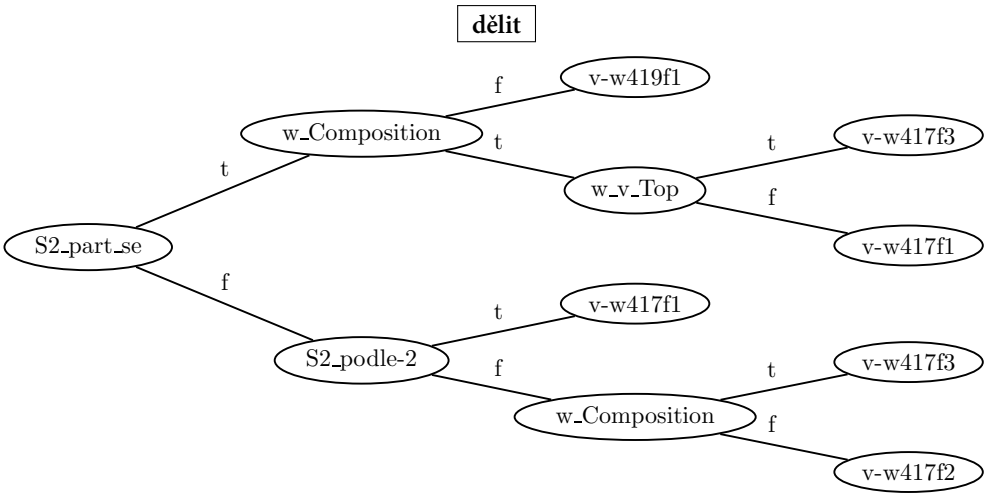
v-w5634f1	určit (Eng: *to decide*) • rychle rozhodl o jeho přijetí → *lit.* to decide on his admission • r. přijmout všechny • r., kam půjdeme
v-w5635f1 (Eng: *to decide*)	• rychle se rozhodl o dalším postupu → *lit.* to quickly decide where to go • r. se přijmout opatření • r. se, kam půjde • r. se rychle, jestli mu vydají....
v-w5635f2	volit, vybrat (Eng: *to choose*) • rozhodnout se pro Prahu mezi dvěma možnostmi → *lit.* he choose Prague as one of the two possibilities • r. se pro Karla

Figure 6. Decision tree for the verb rozhodnout from PDT.

On the VALEVAL corpus, we achieved improvement 12% absolute over the baseline. On the more challenging Prague Dependency Treebank, improvement 6.5% absolute over the baseline was measured on both the development and the evaluation testing set.

In the evaluation section we investigated the results from different perspectives giving alternative analysis and evaluations.

To summarize the thesis, different techniques of disambiguation of verb senses were proposed, implemented and thoroughly evaluated on two Czech corpora. The achieved improvement over baseline validated the correctness of the underlying ideas.



V-w417f1	členit, rozdělit, kouskovat (Eng: *to divide*) • dělit příjmení na části • d. republiku na dva státy • d. salám na poloviny • d. salám nožem v polovině • d. úkol na několik etap → *lit.* to divide the task into several phases
v-w417f2	odloučit Eng: to separate • minuta dělila kajakářku od medaile
v-w417f3	rozdělit, dát, podělit (Eng: *to distribute*) • dělit archívy mezi republiky • dělit dětem dárky → *lit.* to distribute presents among children • d. mezi děti dárky • d. aktivity na střediska, do středisek, střediskům • d. peníze do rozpočtu obcí
v-w419f1	rozdělit se (Eng: *to go share with a person*) • dělil se s příbuznými o majetek • ODS se dělí s ČSSD o politickou moc

Figure 7. Decision tree for the verb dělit from PDT.

Further perspectives Even though this work deals with the disambiguation task, extensively discussing many alternatives, there still remain several directions for the potential extension of the work.

In our opinion, more attention given to the tuning of parameters of non-linear SVM kernels might bring some improvement in performance.

The problem with low number of training samples can be partially avoided by merging aspectual counterparts which often share the valency behavior. However, this might not be applicable for all verbs, and it would require a further exploration. We would also need the mapping of aspectual pairs which is part of the VALLEX lexicon but is missing in the PDT-VALLEX.

The proposed methods might also be further adapted to other languages. However, for languages with limited morphology, e.g. English, a revision of features should be considered, as the current feature set is heavily based on information resulting from morphology.

Acknowledgement This research has been supported in part or in full by the following grants: Grant Agency of the Czech Republic GA405/06/0589 and Ministry of Education of the Czech Republic projects ME 838 and ME 752.

Bibliography

Bojar, O., J. Semecký, and V. Benešová. 2005. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Hajič, Jan. 2000. Morphological Tagging: Data vs. Dictionaries. In *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pages 94–101, Seattle, Washington.

Hajič, Jan. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Hajič, Jan and Václav Honetschläger. 2003. Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. *Prague Bulletin of Mathematical Linguistics*, (79–80):61–86.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová-Řezníčková, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In J. Nivre and E. Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Hajičová, Eva. 2002. Theoretical description of language as a basis of corpus annotation: The case of Prague Dependency Treebank. *Prague Linguistic Circle Papers*, 4:111–127.

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

Lopatková, Markéta, Zdeněk Žabokrtský, Karolína Skwarska, and Václava Benešová. 2003. Vallex 1.0 valency lexicon of czech verbs. Technical report, ÚFAL MFF UK.

Pala, Karel and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2):pp. 79–88.

Panevová, Jarmila. 1974. On verbal frames in Functional generative description I. *Prague Bulletin of Mathematical Linguistics*, (22):3–40.

Panevová, Jarmila. 1980. *Formy a funkce ve stavbě české věty*. Prague:Academia.

Panevová, Jarmila. 1994. Valency frames and the meaning of the sentence. *The Prague School of Structural and Functional Linguistics*, pages 223–243.

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. The eurowordnet base concepts and top ontology. Technical report, Centre National de la Recherche Scientifique, Paris, France, France.

Žabokrtský, Zdeněk and Markéta Lopatková. 2004. Valency Frames of Czech Verbs in VALLEX 1.0. In Adam Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 70—77, Boston. Association for Computational Linguistics.