Automatic Verb Valency Frames Disambiguation for Czech

Jiří Semecký
Institute of Formal and Applied Linguistics
Charles University, Prague
Czech Republic
semecky@ufal.mff.cuni.cz

Abstract

This paper deals with automatic disambiguation of verb valency frames on Czech data. Main contribution lies in determining of the most useful features for valency frame disambiguation. We experimented with diverse types of features, including morphological, syntax-based, idiomatic, animacy and WordNet-based. The considered features were classified using decision trees, rule-based learning and Naïve Bayes classifier.

On a set of 7 778 sentences we achieved accuracy of 79.86% against baseline 68.27% obtained by assigning the most frequent frame.

1 Introduction

Many recent NLP applications, including machine translation, information retrieval, and others, aiming at higher quality results need semantic analysis of language data on the sentence level. As verbs are understood as central elements of sentences, the key aspect in determination of the sentence meaning is estimation of meaning of the verb. Valency frames of verbs usually partially correspond to their meanings.

Choosing the appropriate verb frame with respect to a given frames definition could be described as a special case of word sense disambiguation. First results of verb frame disambiguation were already reported by (Erk, 2005) for German and (Lopatková et al., 2005) for Czech.

For our task we used VALEVAL (Bojar et al., 2005), a human annotated corpus of valency frames containing data selected from the Czech National Corpus (Kocek et al., 2000). VALEVAL contains frames assigned according to definitions in the VALLEX lexicon (Žabokrtský and Lopatková, 2004).

We generated a vector of features describing the contexts of a verb for each verb in our dataset. Later, we trained machine learning methods on a part of the data, and tested it on the rest. For lack of data, we employed 10-fold cross-validation.

We used three different methods, Naïve Bayes classifier, decision trees and rule-based learning. We tested five different types of features describing verb occurrences based on a context within one sentence.

This paper is divided as follows: in Section 2, we give an overview of data which we worked with, in Section 3 we describe methods which we employed in the frame disambiguation and features which we used for describing verbs in their context. In Section 4, we evaluate our results using two different metrics. In the last section, we conclude and suggest further development.

2 Data resources

2.1 Valency lexicon

For automatic assignment of valency frames we need a valency lexicon consisting of formal definitions of frames. In our experiments we used VALLEX, a manually created valency lexicon of Czech verbs, which is based on the framework of Functional Generative Description (FGD) (Sgall et al., 1986).

VALLEX is being built since 2001 and the work is still in progress. The VALLEX version 1.0 (autumn 2003), which we used in our task, defines valency for over 1,400 Czech verbs and contains

over 3,800 frames. 6000 valency frames.

The VALLEX lexicon consists of **verb entries** corresponding to particular verb lexemes, i.e. complex units consisting of the verb base lemma and its possible reflexive particle *se* or *si*. For example, the verb lexeme *dodat si* consists of a base lemma *dodat* and a reflexive particle *si*. There is also the verb *dodat* with no reflexive particle, which has other meaning.

Each verb entry consists of definitions of one or more **frames**, which roughly correspond to meanings of the verb. 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.

Each valency frame consists of a set of **frame slots** corresponding to complements of the verb. Each frame slot is described by functor, expressing the type of relation 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*).

Moreover, each frame in the lexicon is accompanied by an explanation of the meaning (using synonyms or glosses), an example sentence or phrase, and its aspectual counterpart if it exists. Some frames are assigned to semantic classes. A frame could also be marked as "idiom" if it is used idiomatically.

Figure 1 shows an example of a VALLEX entry for the verb lexeme *dodat*, containing five frames for its different senses, namely *supply*, *ship*, *mention*, *add*, and *encourage*. Each frame is described by list of frame slots (e.g. **ACT**, **ADDR**, **PAT**, **DIR** for the first frame). The superscript specify the type of the slot, and the subsript represents its surface representation (the preposition, if applicable, and the case).

dodat pf.

1 dodat$_1$ ≈ **dopravit**
–frame: **ACT**$^{obl}_{1}$ **ADDR**$^{obl}_{3}$ **PAT**$^{obl}_{4}$ ↑**DIR**^{typ}
–example: *dodat někomu zboží do domu*
–asp.counterparts: dodávat$_1$ impf.
–class: transport / exchange

2 dodat$_2$ ≈ **dopravit**
–frame: **ACT**$^{obl}_{1}$ **PAT**$^{obl}_{4}$ ↑**DIR3**^{obl} **BEN**$^{typ}_{3,pro+4}$
–example: *dodat někomu / pro někoho do domu zboží*
–asp.counterparts: dodávat$_2$ impf.
–class: transport

3 dodat$_3$ ≈ **říci; podotknout**
–frame: **ACT**$^{obl}_{1}$ **PAT**$^{opt}_{k+3}$ **EFF**$^{obl}_{4,že}$
–example: *dodal k tomu své připomínky / vše, co věděl*
–asp.counterparts: dodávat$_3$ impf.
–class: communication

4 dodat$_4$ ≈ **doplnit; připojit**
–frame: **ACT**$^{obl}_{1}$ **PAT**$^{obl}_{4}$ **EFF**$^{obl}_{k+3}$
–example: *dodal ke starému zboží nové*
–asp.counterparts: dodávat$_4$ impf.
–class: combining

5 dodat$_5$ ≈ **povzbudit** (idiom)
–frame: **ACT**$^{obl}_{1}$ **ADDR**$^{obl}_{3}$ **PAT**$^{obl}_{2,4}$
–example: *dodat někomu odvahy / odvahu*
–asp.counterparts: dodávat$_5$ impf.
–class: exchange

Figure 1: Example of VALLEX entry for verb lexeme *dodat* (meanings: *supply*, *ship*, *mention*, *add*, and *encourage*).

2.2 Training and Testing Data

For training and testing of disambiguation methods, we need data annotated according to the chosen frame definitions. There is a manually annotated corpus of frame annotations VALEVAL (Bojar et al., 2005) developed as a lexical sampling experiment using VALLEX frame definitions. It contains 109 selected base lemmas. For each base lemma, 100 sentences from the Czech National Corpus[1] (Kocek et al., 2000) were randomly selected.

For 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. For all verbs selected to be present in the VALEVAL, their aspectual counterparts including iterative forms were added too. In order to cover both "easy" and "difficult" cases, verbs were selected randomly from both ends of the difficulty spectrum. Moreover, some verbs were added on purpose to cover specific cases too.

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

2.3 Data preparation

As for input data for the frame disambiguation task, we used VALEVAL sentences where all three annotators agreed. Moreover, sentences on which

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

annotators did not agree were rechecked by another annotator, and sentences with a clear mistake were corrected and added too. This resulted in a set of 8 066 sentences.

Then, we automatically parsed the sentences using Charniak's syntactic parser (Charniak, 2000), which was trained on the Prague Dependency Treebank (Hajič, 1998). Some sentences could not have been parsed because of their length (the corpus contains sentences from fiction with length over 400 words). After excluding unparsed sentences, 7 778 sentences remained, which served as input for disambiguation methods. There were 72.0 sentences per base lemma in average, ranging from a single sentence to 100 sentences (the original amount in the VALEVAL). Figure 2 shows the distribution of number of sentences per base lemma.

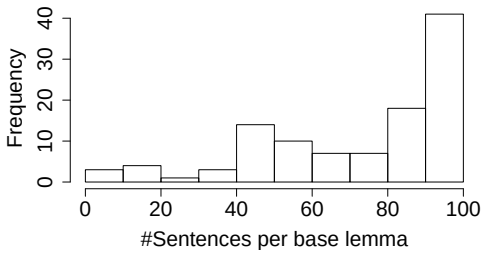

Figure 2: Distribution of the number of sentences per base lemma

3 Method

3.1 Machine Learning methods

For automatic frame disambiguation, we generated a vector of features for each instance of a verb. A detailed description of these vectors is given in Section 3.2.

Later, we trained machine learning methods for each verb separately on a part of the data, and tested it on the rest. Due to lack of annotated data, we employed 10-fold cross-validation: we divided the data into 10 parts, for each tenth we trained the algorithm on the remaining data and tested it on the selected tenth. Finally, we counted the accuracy as the average of accuracies over the ten runs.

We tested three different classification methods, namely Naïve Bayes classifier, decision trees and rule-based learning, the later two implemented in the the machine learning toolkit C5.0 (Quinlan, 2005).

Naïve Bayes classifier computes the probability that an instance belongs to a given class separately for each feature and computes the overall probability as if the features were independent.

The decision trees algorithm finds the most discriminative feature, i.e. the one that suits best for dividing the training data into two parts belonging to different classes. After the first decision, the process continues recursively in all branches resulting in a tree of decisions which indicates the features to use for division of the feature space, i.e. a *decision tree*.

The ruleset algorithm creates a set of independent rules defined as a conjunction of conditions for feature values. Conditions of individual rules may overlap, in which case the rules' predictions are aggregated using their confidence (proposed by the algorithm) to reach a verdict.

Decision trees and the rulesets are equally expressive.

3.2 Feature selection

We experimented with several types of features containing different information about the context of the verb within one sentence. The following list describes five different types of features we used.

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

The first two columns of Table 1 shows the number of features belonging to each of the groups. In the following section we give a detailed description of each group of the features.

Feature type	#Features	#Used features	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 "#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.

Table 1: Types of features.

3.2.1 Morphological features

Czech positional morphology (Hajič, 2000) uses morphological tags consisting of 12 actively used positions, each stating value of one morphological category. The morphological categories are: 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 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).

3.2.2 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.

3.2.3 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.

3.2.4 Animacy

We partially determined animacy of nouns and pronouns in the whole sentence. 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, and one integer feature stating the number of animate nouns and pronouns 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, and one integer feature stating the number of animate nouns and pronouns in the sentence. The later features can operate even in case of wrong result of syntactic parser. In cases where we could not decide, we set the feature to *false*.

Together we obtained 14 features for animacy.

We determined the animacy using several techniques.

As for nouns, the Czech lemmatizer created by Jan Hajič (Hajič, 2000) gives additional information about some lemmas. These include among others identification of first names and surnames. 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 the 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.

3.3 WordNet features

In some cases, dependency of a certain lemma or a certain type of lemma on a verb can imply its particular sense. However, as the machine learning methods which we used work with a fixed number of features, we could not have added information about individual lemmas easily. We described a lemma type in terms of belonging to WordNet (Fellbaum, 1998) classes instead.

In the first step, we used the definition of WordNet top ontology made at University of Amsterdam (Vossen et al., 1997) to obtain 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 1564 Czech lemmas associated to the WordNet top-level classes. As we worked with lemmas, and not with 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. Again, as we worked with the lemmas instead of synsets, one lemma could have been mapped to more 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 Results

4.1 Baseline for frame disambiguation

As a baseline for each base lemma we took the relative frequency of its most frequent frame using 10-fold cross validation. The baselines ranged from 24% (for base lemma *vzít* with 10 different annotated frames) to 100% for verbs with only one frame. Figure 3 shows distribution of the relative frequency of the most frequent frames.

We computed the overall baseline as weighted average of the individual baselines. The overall baseline 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. The second one better predict

	$\oslash_{data}$	$\oslash_{CNC}$
Average number of frames	4.58	5.61
10-fold baseline	68.27	60.64

$\oslash_{data}$ 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 2: Difficulty of the frame disambiguation task

Type of features	$\varnothing_{data}$ NBC	DT	RBL	$\varnothing_{CNC}$ NBC	DT	RBT
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 + A	74.52	78.76	79.22	63.5	69.77	68.63
M + S + I	73.48	78.8	78.86	63.99	68.74	69.2
M + S + W	74.32	79.16	79.47	64.94	77.25	77.41
M + A + I	72.76	74.61	74.88	61.75	63.52	64.35
M + A + W	73.23	74.23	74.29	62.26	61.16	63.84
S + A + I	73.52	78.62	78.5	63.38	70.88	70.8
S + A + W	72.96	78.89	79.16	60.81	70.71	70.9
M + S + I + W	74.19	79.43	79.36	64.91	77.38	77.55
M + S + A + I	74.51	79.05	79.27	63.5	68.6	70.6
M + S + A + W	74.63	79.81	79.41	64.69	76.94	77.04
M + S + I + A + W	74.59	79.6	79.86	64.68	76.97	77.05

Table 3: Accuracy [%] of the frame disambiguation task

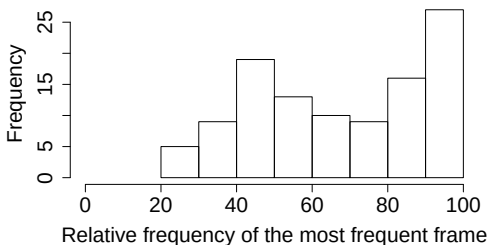

Figure 3: Distribution of the relative frequency of the most frequent frames

behaviour on real data. The difficulty of the task can be seen in the Table 2.

4.2 Evaluation

In our experiments we tested performance of automatic disambiguation classifiers based on each presented type of features separately, as well as on different combinations of feature types. Then, based on the acquired decision trees, we observed which features were most frequently used for the decisions.

Table 3 states accuracy of the word sense disambiguation task for different combinations of features. Columns corespond to different disambiguation methods – Naïve Bayes classifier (NBC), decision trees (DT), and rule-based learning (RBL). The symbol $\varnothing_{data}$ indicates the average accuracy weighted by the number of sentences in the input data, whereas the symbol $\varnothing_{CNC}$ indicates the average accuracy weighted by the relative frequency in the Czech National Corpus (CNC).

The table shows that, taken each group of features individually, the syntactic features performed best achieving accuracy 78.33% over the baseline 68.27% (using decition trees). Idiomatic features scored worst and even brought little improvement when combined with other types of features. This is mainly due to low number of idioms defined in the VALLEX lexicon, and therefore low number of idioms in the data.

Morphological features turned out to be the second best type when measured individually.

4.3 Importance of the Features

We summed the number of applications of individual features in decision trees weighted by 1 for

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
Morphological	Case of the word two possitions after the verb	6
Syntax-based	Presence of preposition *za* in accusative dependent on the verb	5.5
Syntax-based	Presence of preposition in local dependent on the verb	5.5
Syntax-based	Presence of noun or a substantive pronoun in instrumental dependent on the verb	5.5
Syntax-based	Presence of reflexive particle *si* dependent on the verb	5

Table 4: Features most often chosen in the decision trees

the features used in the root of decision trees, by 0.5 for the features applying in the first level of decision trees, by 0.25 for features applying in the second level, etc.

Over the whole data (including all 10 runs of cross-validation), 78 features were used at least once, and 345 features were not used at all. Details can be seen in Table 1.

Table 4 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.

5 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, which uses PDT-VALLEX (Hajič et al., 2003) as a frames definition, and PropBank.

We are also aiming at improving the feature set, by elaborating individual groups of features, for example by using a richer idiomatic lexicon, extending the coverage of semantic classes, or by adding other syntax-based characteristics.

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, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, November 14–15, 2003.

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. Vossen, 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.