Verbal Valency Frame Detection and Selection in Czech and English

Ondřej Dušek, Jan Hajič and **Zdeňka Urešová**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Prague 1, Czech Republic
{odusek,hajic,uresova}@ufal.mff.cuni.cz

Abstract

We present a supervised learning method for verbal valency frame detection and selection, i.e., a specific kind of word sense disambiguation for verbs based on subcategorization information, which amounts to detecting mentions of events in text. We use the rich dependency annotation present in the Prague Dependency Treebanks for Czech and English, taking advantage of several analysis tools (taggers, parsers) developed on these datasets previously. The frame selection is based on manually created lexicons accompanying these treebanks, namely on PDT-Vallex for Czech and EngVallex for English. The results show that verbal predicate detection is easier for Czech, but in the subsequent frame selection task, better results have been achieved for English.

1 Introduction

Valency frames are a detailed semantic and syntactic description of individual predicate senses.[1] As such, they represent different event types. We present a system for automatic detection and selection of verbal valency frames in Czech and English, which corresponds to detecting and disambiguating mentions of events in text. This is an important step toward event instance identification, which should help greatly in linking the mentions of a single event. We took advantage of the fact that the Prague family of dependency treebanks contains comparable valency frame annotation for Czech and English (cf. Section 2). Thus the feature templates used in frame selection are the same and the features initially considered differ only in their instantiation (cf. Section 3).

While somewhat similar to the CoNLL 2009 Shared Task (Hajič et al., 2009) in the predicate detection part, our task differs from the semantic role labeling task in that the whole frame has to be detected, not only individual arguments, and is therefore more difficult not only in terms of scoring, but also in the selection part: several verbal frames might share the same syntactic features, making them virtually indistinguishable unless semantics is taken into account, combined with a detailed grammatical and morphological context.

2 Valency in the tectogrammatical description

The annotation scheme of the Prague Dependency Treebank (Bejček et al., 2012, PDT) and the Prague Czech-English Dependency Treebank (Hajič et al., 2012, PCEDT) is based on the formal framework of the Functional Generative Description (Sgall, 1967; Sgall et al., 1986, FGD), developed within the Prague School of Linguistics. The FGD is dependency-oriented with a "stratificational" (layered) approach to a systematic description of a language. The notion of valency in the FGD is one of the core concepts operating on the layer of linguistic meaning (*tectogrammatical layer, t-layer*).

2.1 Valency frames

The FGD uses syntactic as well as semantic criteria to identify verbal complements. It is assumed that all semantic verbs – and, potentially, nouns, adjectives, and adverbs – have subcategorization requirements, which can be specified in the *valency frame*.

Verbal valency modifications are specified along two axes: The first axis concerns the (general) opposition between inner participants (*arguments*) and free modifications (*adjuncts*). This dis-

[1] Valency can be observed for verbs, nouns, adjectives and in certain theories, also for other parts of speech; however, we focus on verbal valency only, as it is most common and sufficiently described in theory and annotated in treebanks.

Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 6–11,
Baltimore, Maryland, USA, June 22-27, 2014. ©2014 Association for Computational Linguistics

tinction is based on criteria relating to:

(a) the possibility of the same type of complement appearing multiple times with the same verb (arguments cannot), and

(b) the possibility of the occurrence of the given complements (in principle) with any verb (typical for adjuncts).

The other axis relates to the distinction between (semantically) *obligatory* and *optional* complements of the word, which again is based on certain operational criteria expressed as the *dialogue test* (Panevová, 1974). Five arguments are distinguished: *Actor* (ACT), *Patient* (PAT), *Addressee* (ADDR), *Origin* (ORIG), and *Effect* (EFF). The set of free modifications is much larger than that of arguments; about 50 types of adjuncts are distinguished based on semantic criteria. Their set can be divided into several subclasses: temporal (e.g., TWHEN, TSIN), local (e.g., LOC, DIR3), causal (such as CAUS, CRIT), and other free modifications (e.g., MANN for general *Manner*, ACMP for *Accompaniment*, EXT for *Extent* etc.).

All arguments (obligatory or optional) and obligatory adjuncts are considered to be part of the valency frame.

2.2 Tectogrammatical annotation

The PDT is a project for FGD-based manual annotation of Czech texts, started in 1996 at the Institute of Formal and Applied Linguistics, Charles University in Prague. It serves two main purposes:

1. to test and validate the FGD linguistic theory,
2. to apply and test machine learning methods for part-of-speech and morphological tagging, dependency parsing, semantic role labeling, coreference resolution, discourse annotation, natural language generation, machine translation and other natural language processing tasks.

The language data in the PDT are non-abbreviated articles from Czech newspapers and journals.

The PCEDT contains English sentences from the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993, PTB-WSJ) and their Czech translations, all annotated using the same theoretical framework as the PDT.

The annotation of the PDT and the PCEDT is very rich in linguistic information. Following the stratificational approach of the FGD, the texts are annotated at different but interlinked layers. There are four such layers, two linear and two structured:

- the word layer (*w-layer*) – tokenized but otherwise unanalyzed original text,
- the morphological layer (*m-layer*) with parts-of-speech, morphology and lemmatization,
- analytical layer (*a-layer*) – surface dependency syntax trees,
- tectogrammatical layer (*t-layer*) – "deep syntax" trees according to the FGD theory.

While the PDT has all the layers annotated manually, the PCEDT English annotation on the *a-layer* has been created by automatic conversion from the original Penn Treebank, including the usual head assignment; morphology and the tectogrammatical layer are annotated manually, even if not as richly as for Czech.[2]

Valency is a core ingredient on the t-layer. Since valency frames guide, i.a., the labeling of arguments, valency lexicons with sense-distinguished entries for both languages have been created to ensure consistent annotation.

2.3 Valency Lexicons for Czech and English in the FGD Framework

PDT-Vallex (Hajič et al., 2003; Urešová, 2011) is a valency lexicon of Czech verbs, nouns, and adjectives, created in a bottom-up way during the annotation of the PDT. This approach made it possible to confront the pre-existing valency theory with the real usage of the language.

Each entry in the lexicon contains a headword, according to which the valency frames are grouped, indexed, and sorted. Each valency frame includes the frame's "valency" (number of arguments, or frame members) and the following information for each argument:

- its label (see Section 2.1),
- its (semantic) obligatoriness according to Panevová (1974)'s dialogue test,
- its required surface form (or several alternative forms) typically using morphological, lexical and syntactic constraints.

Most valency frames are further accompanied by a note or an example which explains their meaning and usage. The version of PDT-Vallex used here contains 9,191 valency frames for 5,510 verbs.

EngVallex (Cinková, 2006) is a valency lexicon of English verbs based on the FGD framework, created by an automatic conversion from

[2] Attributes such as tense are annotated automatically, and most advanced information such as topic and focus annotation is not present.

PropBank frame files (Palmer et al., 2005) and by subsequent manual refinement.[3] EngVallex was used for the tectogrammatical annotation of the English part of the PCEDT. Currently, it contains 7,699 valency frames for 4,337 verbs.

3 Automatic frame selection

Building on the modules for Czech and English automatic tectogrammatical annotation used in the TectoMT translation engine (Žabokrtský et al., 2008) and the CzEng 1.0 corpus (Bojar et al., 2012),[4] we have implemented a system for automatic valency frame selection within the Treex NLP Framework (Popel and Žabokrtský, 2010).

The frame selection system is based on logistic regression from the LibLINEAR package (Fan et al., 2008). We use separate classification models for each verbal lemma showing multiple valency frames in the training data. Due to identical annotation schemata in both languages, our models use nearly the same feature set,[5] consisting of:

- the surface word form of the lexical verb and all its auxiliaries,
- their morphological attributes, such as part-of-speech and grammatical categories,
- formemes – compact symbolic morphosyntactic labels (e.g., `v:fin` for a finite verb, `v:because+fin` for a finite verb governed by a subordinating conjunction, `v:in+ger` for a gerund governed by a preposition),[6]
- syntactic labels given by the dependency parser,
- all of the above properties found in the topological and syntactic neighborhood of the verbal node on the *t-layer* (parent, children, siblings, nodes adjacent in the word order).

We experimented with various classifier settings (regularization type and cost C, termination criterion E) and feature selection techniques (these involve adding a subset of features according to a metric against the target class).[7]

[3] This process resulted in the interlinkage of both lexicons, with additional links to VerbNet (Schuler, 2005) where available. Due to the refinement, the mapping is often not 1:1.

[4] Note that annotation used in TectoMT and CzEng does not contain all attributes found in corpora manually annotated on the tectogrammatical layer. Valency frame IDs are an example of an attribute that is missing from the automatic annotation of CzEng 1.0.

[5] The only differences are due to the differences of part-of-speech tagsets used.

[6] See (Dušek et al., 2012; Rosa et al., 2012) for a detailed description of formemes.

[7] The metrics used include the Anova F-score, minimum Redundancy-Maximum Relevance (mRMR) (Peng et al., 2005), ReliefF (Kononenko, 1994), mutual information (MI), symmetric uncertainty (Witten and Frank, 2005, p. 291f.), and an average of the ranks given by mRMR and MI.

4 Experiments

We evaluated the system described in Section 3 on PDT 2.5 for Czech and on the English part of PCEDT 2.0 for English. From PCEDT 2.0, whose division follows the PTB-WSJ, we used Sections 02-21 as training data, Section 24 as development data, and Section 23 as evaluation data. Since the system is intended to be used in a fully automatic annotation scenario, we use automatically parsed sentences with projected gold-standard valency frames to train the classifiers.

The results of our system in the best setting for both languages are given in Table 1.[8] The *unlabeled* figures measure the ability of the system to detect that a valency frame should be filled for a given node. The *labeled* figures show the overall system performance, including selecting the correct frame. The *frame selection accuracy* value shows only the percentage of frames selected correctly, disregarding misplaced frames. The accuracy for *ambiguous verbs* further disregards frames of lemmas where only one frame is possible. Here we include a comparison of our trained classifier with a baseline that always selects the most frequent frame seen in the training data.[9] Our results using the classifier for both languages have been confirmed by pairwise bootstrap resampling (Koehn, 2004) to be significantly better than the baseline at 99% level.

We can see that the system is more successful in Czech in determining whether a valency frame should be filled for a given node. This is most probably given by the fact that the most Czech verbs are easily recognizable by their morphological endings, whereas English verbs are more prone to be misrepresented as nouns or adjectives.

The English system is better at selecting the correct valency frame. This is probably caused by a more fine-grained word sense resolution in the Czech valency lexicon, where more figurative uses and idioms are included. For example, over 16%

[8] The best setting for Czech uses L1-regularization and 10% best features according to Anova, with other parameters tuned on the development set for each lemma. The best setting for English uses L2-regularization with best feature subsets tuned on the development set and fixed parameters $C = 0.1$, $E = 0.01$.

[9] All other parts of the system, up to the identification of the frame to be filled in, are identical with the baseline.

		Czech	English
Unlabeled precision		99.09	96.03
Unlabeled recall		94.81	93.07
Unlabeled F-1		96.90	94.53
Labeled precision		78.38	81.58
Labeled recall		74.99	79.06
Labeled F-1		76.65	80.30
Frame selection accuracy		79.10	84.95
Ambiguous verbs	baseline	66.68	68.44
	classifier	72.41	80.03

Table 1: Experimental results

of errors in the Czech evaluation data were caused just by idioms or light verb constructions not being recognized by our system. In Czech, additional 15% of errors occurred for verbs where two or more valency frames share the same number of arguments and their labels, but these verb senses are considered different (because they have different meaning), compared to only 9% in English.

5 Related Work

As mentioned previously, the task of detecting and selecting valency frames overlaps with semantic role labeling (Hajič et al., 2009). However, there are substantial differences: we have focused only on verbs (as opposed to all words with some semantic relation marked in the data), and evaluated on the exact frame assigned to the occurrence of the verb in the treebank. On the other hand, we are also evaluating predicate identification as in Surdeanu et al. (2008), which Hajič et al. (2009) do not. Tagging and parsing have been automatic, but not performed jointly with the frame selection task. This also explains that while the best results reported for the CoNLL 2009 Shared task (Björkelund et al., 2009) are 85.41% labeled F-1 for Czech and 85.63% for English, they are not comparable due to several reasons, the main being that SRL evaluates each argument separately, while for a frame to be counted as correct in our task, the whole frame (by means of the reference ID) must be selected correctly, which is substantially harder (if only for verbs). Moreover, we have used the latest version of the PDT (the PDT 2.5), and EngVallex-annotated verbs in the PCEDT, while the English CoNLL 2009 Shared Task is PropBank-based.[10]

Selecting valency frames is also very similar to Word Sense Disambiguation (WSD), see e.g. (Edmonds and Cotton, 2001; Chen and Palmer, 2005). The WSD however does not consider subcategorization/valency information explicitly.

Previous works on the PDT include a rule-based tool of Honetschläger (2003) and experiments by Semecký (2007) using machine learning. Both of them, unlike our work, used gold-standard annotation with just the frame ID removed.

6 Conclusions

We have presented a method of detecting mentions of events in the form of verbal valency frame selection for Czech and English. This method is based on logistic regression with morphological and syntactic features, trained on treebanks with a comparable annotation scheme. We believe that these results are first for this task on the granularity of the lexicons (PDT-Vallex for Czech and EngVallex for English), and they seem to be encouraging given that the most frequent verbs like *to be* and *to have* have tens of possible frames, heavily weighing down the resulting scores.

We plan to extend this work to use additional features and lexical clustering, as well as to see if the distinctions in the lexicons are justified, i.e. if humans can effectively distinguish them in the first place, similar to the work of Cinková et al. (2012). A natural extension is to combine this work with argument labeling to match or improve on the "perfect proposition" score of Surdeanu et al. (2008) while still keeping the sense distinctions on top of it. We could also compare this to other languages for which similar valency lexicons exist, such as SALSA for German (Burchardt et al., 2006) or Chinese PropBank (Xue, 2008).

Acknowledgments

This work was supported by the Grant No. GPP406/13/03351P of the Grant Agency of the Czech Republic, the project LH12093 of the Ministry of Education, Youth and Sports of the Czech Republic and the Charles University SVV project 260 104. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

[10] Please recall that EngVallex is a manually refined PropBank with different labeling scheme and generally $m : n$ mapping between PropBank and EngVallex frames.

References

E. Bejček, J. Panevová, J. Popelka, P. Straňák, M. Ševčíková, J. Štěpánek, and Z. Žabokrtský. 2012. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of COLING 2012: Technical Papers*, Mumbai.

A. Björkelund, L. Hafdell, and P. Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, United States, June.

O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna. 2012. The joy of parallelism with CzEng 1.0. In *LREC*, page 3921–3928, Istanbul.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*.

J. Chen and M. Palmer. 2005. Towards robust high performance word sense disambiguation of English verbs using rich linguistic features. In *Natural Language Processing–IJCNLP 2005*, pages 933–944. Springer.

S. Cinková, M. Holub, and V. Kríž. 2012. Managing uncertainty in semantic tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 840–850. Association for Computational Linguistics.

S. Cinková. 2006. From PropBank to EngValLex: adapting the PropBank-Lexicon to the valency theory of the functional generative description. In *Proceedings of the fifth International conference on Language Resources and Evaluation (LREC 2006), Genova, Italy*.

O. Dušek, Z. Žabokrtský, M. Popel, M. Majliš, M. Novák, and D. Mareček. 2012. Formemes in English-Czech deep syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 267–274.

P. Edmonds and S. Cotton. 2001. Senseval-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL '01, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. E Fan, K. W Chang, C. J Hsieh, X. R Wang, and C. J Lin. 2008. LIBLINEAR: a library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA.

J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, and Z. Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC*, pages 3153–3160, Istanbul.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolářová, and P. Pajas. 2003. PDT-VALLEX: creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, page 57–68.

V. Honetschläger. 2003. Using a Czech valency lexicon for annotation support. In *Text, Speech and Dialogue*, pages 120–125. Springer.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Empirical Methods in Natural Language Processing*, pages 388–395.

I. Kononenko. 1994. Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, page 171–182.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):330.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

J. Panevová. 1974. On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

H. Peng, F. Long, and C. Ding. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, page 1226–1238.

M. Popel and Z. Žabokrtský. 2010. TectoMT: modular NLP framework. *Advances in Natural Language Processing*, pages 293–304.

R. Rosa, D. Mareček, and O. Dušek. 2012. DEPFIX: a system for automatic correction of Czech MT outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 362–368. Association for Computational Linguistics.

K. K. Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Univ. of Pennsylvania.

J. Semecký. 2007. Verb valency frames disambiguation. *The Prague Bulletin of Mathematical Linguistics*, (88):31–52.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.

P. Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Praha.

M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England, August. Coling 2008 Organizing Committee.

Z. Urešová. 2011. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, ISBN 978-80-904571-1-9, 375 pp.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2nd edition.

N. Xue. 2008. Labeling Chinese predicates with semantic roles. *Computational linguistics*, 34(2):225–255.

Z. Žabokrtský, J. Ptáček, and P. Pajas. 2008. TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, page 167–170. Association for Computational Linguistics.