Enriching a Valency Lexicon by Deverbative Nouns

Eva Fučíková **Jan Hajič** **Zdeňka Urešová**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{fucikova,hajic,uresova}@ufal.mff.cuni.cz

Abstract

In this paper, we present an attempt to automatically identify Czech deverbative nouns using several methods that use large corpora as well as existing lexical resources. The motivation for the task is to extend a verbal valency (i.e., predicate-argument) lexicon by adding nouns that share the valency properties with the base verb, assuming their properties can be derived (even if not trivially) from the underlying verb by deterministic grammatical rules. At the same time, even in inflective languages, not all deverbatives are simply created from their underlying base verb by regular lexical derivation processes. We have thus developed hybrid techniques that use both large parallel corpora and several standard lexical resources. Thanks to the use of parallel corpora, the resulting sets contain also synonyms, which the lexical derivation rules cannot get. For evaluation, we have manually created a gold dataset of deverbative nouns linked to 100 frequent Czech verbs since no such dataset was initially available for Czech.

1 Introduction

Valency is one of the central notions in a "deep" syntactic and semantic description of language structure. In most accounts, verbs are in the focus of any valency (or predicate-argument) theory, even if it is widely acknowledged that nouns, adjectives and even adverbs can have valency properties (Panevová, 1974; Panevová, 1994; Panevová, 1996; Hajičová and Sgall, 2003). There have been created many lexicons that contain verbs and their predicate-argument structure and/or valency, in some cases also subcategorization information or semantic preferences are included.

Creating such a lexicon is a laborious task. On top of the sheer volume of such a lexicon (to achieve good coverage of the given language), the biggest difficulty is to keep consistency among entries that describe verbs with the same or very similar behavior. The same holds for derivations; in most cases, no attempt is made to link the derivations to the base verbs in the lexicon (with NomBank (Meyers et al., 2004) being an exception, linking nouns to base verbs in the English PropBank (Kingsbury and Palmer, 2002)).

Valency information (number and function of the arguments) is shared between the base verb and its deverbatives, undergoing certain transformations in defined cases.[1] Moreover, especially in richly inflective languages, the subcategorization information (morphosyntactic surface expression of the arguments) can be derived by more or less deterministic rules from the verb, the deverbative relation and the verb's arguments' subcategorization (Kolářová, 2006; Kolářová, 2005; Kolářová, 2014). These rules, for example, transform the case of Actor (Deep subject) from nominative to genitive as the appropriate subcategorization for the deverbative noun, or delete the Actor altogether from the list of arguments in case of the derivation *teach* → *teacher* (*učit* → *učitel*).

It is thus natural to look for ways of organizing the valency or predicate-argument lexicons in such a way that they contain the links between the underlying verb and its deverbatives, which is not only

[1] Throughout the rest of the paper, we will use the term *deverbative nouns* or *deverbatives* since the term *derivations* might imply regular prefixing or suffixing processes, which we go beyond.

Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces,
pages 71–80, Osaka, Japan, December 11 2016.

natural, but if successful, would help the consistency of the grammatical properties between the verb and its deverbatives.

The goal of this study is to automatically discover deverbative nouns related to (base) verbs, using primarily parallel corpora, but also existing lexicons (mainly as an additional source and for comparison). The use of a parallel corpus should give us those deverbatives which would otherwise be hard to find using only monolingual resources. However, it is not our goal here to fully transfer the valency information from the base verb - as mentioned in the previous paragraph, that work is being done separately and we assume its results (i.e., the transfer rules) can then be applied relatively easily if we are successful in discovering and linking the appropriate nouns to the base verb.

In order to evaluate and compare the resulting automatic systems, evaluation (gold-standard) data had to be developed, due to the lack of such a resource. The language selected for this project is Czech, a richly inflectional language where derivations can be related to the word from which they are derived by regular changes (stemming with possible phonological changes, suffixing, prefixing) or - as is often the case - by more or less irregular processes.

There are many types (and definitions) of event/deverbative nouns. We are using the more general term *deverbative* throughout here, to avoid possible narrow interpretation of "event". For the purpose of our study and experiments, a deverbative noun is defined as a noun which in fact describes a state or event and can be easily paraphrased using its base verb without substantial change in meaning. For example, *Po úderu do jeho hlavy utekl.* (lit. *After hitting him in the head he ran away.*) can be paraphrased as *Poté, co ho udeřil do hlavy, utekl.* (lit: *After he hit him in the head, he ran away.*). The same noun can be used as a deverbative noun or entity-referring (referential) noun in different contexts; in Czech, however, this is rarer as the noun itself would be different for the two cases. For example, *stavba* (lit: *building*) in *Při stavbě domu jim došly peníze.* (lit: *During the building of the house, they ran out of money.*) is an event noun, while in *Tato stavba* [= budova] *se prodala levně.* (lit: *This building sold cheaply.*) it refers to an entity; here, even in Czech the same noun is used. However, another Czech derivations, *stavění* (from the same base verb, *stavět*) can only be used as event noun, and *stavení* only as a referential one. We also use the term *derivation* in a very broad sense, not only describing the very regular and productive derivation such as English *-ing* (Czech: *-ění, -a/ání, -í/ávání, -(u)tí, ...*), but also those which are much less frequent (*-ba, -nost, -ota*).

2 Related Work

Derivations, especially verbal derivations, have been studied extensively. Almost all grammars include a section on derivations, even if they use different theoretical starting points. The most recent work on Czech derivations is (Žabokrtský and Ševčíková, 2014; Ševčíková and Žabokrtský, 2014; Vidra, 2015; Vidra et al., 2015). These authors also created a resource called DeriNet (cf. Sect. 3.2). The background for their work comes from (Baranes and Sagot, 2014) and (Baayen et al., 1995). DeriNet, while keeping explicit the connection between the verb and its derivative, does not use valency as a criterion for having such a link, and therefore is broader than what we are aiming at in our study; however, we have used it as one of the starting points for the creation of the gold standard data (Sect. 4).

Event nouns, which form a major part of our definition of deverbatives, have also been studied extensively. A general approach to events and their identification in text can be found, e.g., in (Palmer et al., 2009) or (Stone et al., 2000).

NomBank (Meyers et al., 2004) is a prime resource for nominal predicate-argument structure in English. Closest to what we want to achieve here, is the paper (Meyers, 2008), where the authors also use various resources for helping to construct English NomBank; however, they do not make use of parallel resources.

For Czech, while we assume that relations between verbs and their deverbatives regarding valency structure can be described by grammatical rules (Kolářová, 2014; Kolářová, 2006; Kolářová, 2005),[2] no attempt to automatically extract deverbatives from lexicons and/or corpora has been described previously.

[2]We have also found similar work for Italian (Graffi, 1994).

3 The Data Available

3.1 Corpora

As one source of bilingual text, we have used the Prague Czech-English Dependency Treebank (PCEDT 2.0) (Hajič et al., 2012). The PCEDT is a 1-million-word bilingual corpus that is manually annotated and sentence-aligned and automatically word-aligned. In addition, it contains the predicate-argument annotation itself, where the verbs are sense-disambiguated by linking them to Czech and English valency lexicons. The English side builds on the PropBank corpus (Palmer et al., 2005), which annotates predicate-argument structure over the Penn Treebank (Marcus et al., 1993).

The associated valency lexicons for Czech - PDT-Vallex[3] (Urešová, 2011) and English - EngVallex[4] (Cinková, 2006) are also interlinked, forming a bilingual lexicon CzEngVallex (Urešová et al., 2016), which explicitly pairs verb senses and their arguments between the two languages.

The second corpus used was CzEng[5] (Bojar et al., 2011; Bojar et al., 2012; Bojar et al., 2016), a 15-million sentence parallel corpus of Czech and English texts. This corpus is automatically parsed and deep-parsed, verbs are automatically annotated by links to the same valency lexicons as in the PCEDT. The corpus is automatically sentence- and word-aligned.

The reason for using both a small high-quality annotated and a "noisy" (automatically annotated) but large corpus is to assess the ways they can contribute to the automatic identification of deverbatives, especially with regard to the amount of manual work necessary for subsequent "cleaning" of the certainly not quite perfect result (i.e., with regard to the recall/precision tradeoff).

3.2 Lexical Resources

In addition to corpora, we have also used the following lexical resources:

- DeriNet[6] (Vidra, 2015; Vidra et al., 2015; Žabokrtský and Ševčíková, 2014; Ševčíková and Žabokrtský, 2014), a large lexical network with high coverage of derivational word-formation relations in Czech. The lexical network DeriNet captures core word-formation relations on the set of around 970 thousand Czech lexemes. The network is currently limited to derivational relations because derivation is the most frequent and most productive word-formation process in Czech. This limitation is reflected in the architecture of the network: each lexeme is allowed to be linked up with just a single base word; composition as well as combined processes (composition with derivation) are thus not included. We have used version 1.1 of DeriNet.

- Morphological Dictionary of Czech called Morfflex CZ[7] (Hajič and Hlaváčová, 2016; Hajič, 2004), which is the basis for Czech morphological analyzers and taggers, such as (Straková et al., 2014). This dictionary has been used to obtain regular noun derivatives from verb, limited to suffix changes, namely for nouns ending in *-ní* or *-tí*, *-elnost* and *-ost*. The resulting mapping, which we call "Der" in the following text, contains 49,964 distinct verbs with a total of 143,556 nouns to which they are mapped (i.e., not all verbs map to all three possible derivations, but almost all do). While DeriNet subsumes most of Morflfex CZ derivations, it has proved to be sometimes too "permissive" and the deverbatives there often do not correspond to valency-preserving derivations.

- Czech WordNet version 1.9[8] (Pala et al., 2011; Pala and Smrž, 2004), from which all noun synsets with more than 1 synonym have been extracted (total of 3,432 synsets with 8,742 nouns); this set is referred to as "Syn" in the following text. Using WordNet is deemed a natural baseline for adding synonyms–in our case, to the deverbatives extracted from other sources.

[3] http://lindat.mff.cuni.cz/services/PDT-Vallex
[4] http://lindat.mff.cuni.cz/services/EngVallex
[5] http://hdl.handle.net/11234/1-1458
[6] http://hdl.handle.net/11234/1-1520
[7] http://hdl.handle.net/11234/1-1673
[8] http://hdl.handle.net/11858/00-097C-0000-0001-4880-3

4 Evaluation (Gold) Dataset Preparation

4.1 The Goal

There was no available Czech dataset for testing any particular automatic identification and extraction of deverbatives. The closest to our goals is DeriNet (Sect. 3.2), however DeriNet lists all possible derivations based on root/stem, without regard to valency (predicate-argument relations). For example, for the verb *dělit* (*divide*), DeriNet lists also *dělítko*, which is (in one rare, but possible sense) a tool for dividing things; tools used in events are not considered to share their valency, even if possible transformations are considered, as described in Sect. 1.

Two such "gold" datasets have been created: a development set, which can be used for developing the extraction rules and their optimization and tuning by both manual inspection and automatic techniques, and an evaluation set, which is used only for final "blind" evaluation of the methods developed.

An example of a set of deverbatives of the verb *klesat* (lit. *to decrease*), taken from the development dataset: *klesání, klesavost, omezování, oslabování, redukování, snižování, zmenšování* (lit. *decrease, decreaseness, limitation, weakening, reduction, lowering, diminishing*).

Each set contains 100 Czech verbs (with no overlap between the two in terms of verb senses), selected proportionally to their relative frequency in a syntactically and semantically annotated corpus, the Prague Dependency Treebank (Hajič et al., 2006), excluding verbs equivalent to *to be, to have, to do*, light and support verb senses like *close [a contract]* and all idioms (e.g. *take part*).[9]

4.2 The Annotation Process

The pre-selected sets of deverbative nouns have been extracted from several sources: PCEDT, a parallel corpus using alignments coming from an automatic MT aligner (Giza++) and the DeriNet lexicon (Sect. 3.2). To avoid bias as much as possible, these sets are intentionally much larger than we expected human annotators to create, so that the annotators would mostly be filtering out those words not corresponding to the definition of a deverbative, even if allowed to add more words as well.

Annotators had the task to amend the pre-selected list of nouns for a particular verb (actually, a verb sense, as identified by a valency frame ID taken from the Czech valency lexicon entries[10] (Urešová, 2011)) so that only deverbatives with the same or very similar meaning remain, and add those that the annotator feels are missing, based e.g. on analogies with other verb-deverbative groups and following the definition of deverbatives.

The annotation was done simply by editing a plain text file which contained, at the beginning, all the 100 verbs and for each of them, a pre-selected set of nouns, one per line. Each entry has also contained a description of the particular verb sense (meaning) used, copied form PDT-Vallex. On average, there have been pre-selected 44.1 nouns per verb. The annotators proceeded by deleting lines which contained non-deverbative nouns, and adding new ones by inserting a new line at any place in the list. The resulting average number of nouns per verb has been 6.3 per verb (in the development set).

While the development dataset has been annotated by a single annotator, the evaluation dataset has been independently annotated by three annotators, since it was expected that the agreement, as usual for such open-ended annotation, would not be very high.

4.3 Inter-Annotator Agreement (IAA)

In an annotation task where the result of each item annotation is open-ended, the classification-based measures, such as the κ (kappa) metric, cannot be sensibly used. Instead, we have used the standard F_1 measure (Eq. 1), pairwise for every pair of annotators. Precision P is the ratio of matches over the number of words annotated, and recall R is the number of matches over the other annotator's set of words.[11]

[9] This was easily done since the annotation in the Prague Dependency Treebank contains all the necessary attributes, such as verb senses and light/support/idiomatic use. Coverage of the 100-verb evaluation set is quite substantial, about 14%.

[10] `http://lindat.mff.cuni.cz/services/PDT-Vallex`

[11] While the direction of computation between the annotators matters for computing precision and recall (precision of one annotator vs. the other is equal to the recall of the opposite direction), the resulting F_1 is identical regardless of the direction, therefore we report only one F_1 number.

$$F_1 = 2PR/(P+R) \quad (1)$$

In Table 1, we list all three pairs of annotators of the evaluation dataset and their IAA.

	Annotators 1-2	Annotators 2-3	Annotators 1-3
F_1	0.5520	0.5402	0.5327

Table 1: Inter-Annotator Agreement on the Evaluation Dataset

While the pairwise F_1 scores are quite consistent, they are relatively low; again, it has to be stressed that this is an open-ended annotation task. Not surprisingly, if we only consider deletions in the pre-selected data, the agreement goes up (e.g., for Annotators 1-2, this would then be 0.6237).

To make the evaluation fair, we could not inspect the evaluation data manually, and despite using linguistically well-qualified annotators, a test proved that any attempt at adjudication would be a lengthy and costly process. We have therefore decided to use three variants of the evaluation dataset: one which contained for each verb only those nouns that appeared in the output of all three annotators (called *"intersection"*), second in which we kept also those nouns which have been annotated by two annotators (called *"majority"*) and finally a set which contained annotations from all three (called *"union"*). Such a triple would give us at least an idea about the intervals of both precision and recall which we could expect, should a careful adjudication be done in the future. We consider the "majority" set to be most likely closest to such an adjudicated dataset.

5 Extraction Methods and Experiments

5.1 Baseline

The baseline system uses only the "Der" lists (Sect. 3.2) that contain, for each verb from the Czech morphology lexicon, its basic, regularly formed event noun derivations. For example, for the verb *potisknout* (lit. *print on* [sth] *all over*) the derivations listed in "Der" are *potisknutí* (and its derivational variant *potištění*, both lit. *printing* [of sth] *all over*) and *potištěnost* (lit. *property/ration of being printed over*).

For each verb in the test set, all and only nouns listed for it in "Der" are added. The baseline experiment is used as the basis of the other methods and experiments described below.

5.2 Adding WordNet

On top of the regular derivations, synonyms of all the derivations are added, based on Czech WordNet-based "Syn" lists (Sect. 3.2). All synonyms listed for a particular noun are added; no hierarchy is assumed or attempted to extract.

5.3 Using Parallel Corpora

Using the parallel corpus is the main contribution; all the previous methods have been included for comparison only and as a baseline "sanity check". We use either the PCEDT or CzEng (Sect. 3.1), in addition to the baseline method; each of the two has different properties (PCEDT being manually annotated while CzEng is very large). For each base verb in the test set, the following steps have been taken:

1. For each occurrence of the Czech base verb, the aligned English verb (based on CzEngVallex pairings) was extracted.
2. All occurrences of that verb on the English side of the parallel corpus were identified.
3. All nouns that are aligned with any of the occurrence of the English verb were extracted from the Czech side.
4. The verb and the noun were subject to an additional filtering process, described below; if they passed, the noun was added to the baseline list of nouns associated with the base verb.

Filtering is necessary for several reasons: first, the annotation of the data is noisy, especially in the automatically analyzed CzEng corpus, and second, the alignment is also noisy (for both corpora, since it is automatic). Even if both the annotation and the alignment are correct, sometimes the noun extracted the way described above is only a part of a different syntactic construction, and not a true equivalent of the verb. In order to eliminate the noise as much as possible, two techniques have been devised.

5.3.1 Simple Prefix-based Filtering

As the first method, we have used an extremely simple method of keeping only those nouns that share the first letter with the base verb. The rationale is that the deverbatives are often (regular as well as irregular) derivations, which in Czech (as in many other languages) change the suffix(es) and ending(s), not the prefix. After some investigation, we could not find another simple and reliable method for identifying the stem or more logical part of the word, and experiments showed that on the PCEDT corpus, this was a relatively reliable method of filtering out clear mistakes (at the expense of missing some synonyms etc.).

This method is referred to in the following text and tables as "L1 filter".

5.3.2 Advanced Argument-based Filtering

As the experiments (Table 2 in Sect. 6) show, the L1 filter works well with the PCEDT corpus, but the results on CzEng are extremely bad, due to a (still) huge number of words (nouns) generated by the above method using such a noisy and large corpus.

To avoid this problem, we have devised a linguistically-motivated filter based on shared arguments of the base verb and the potential deverbative. We first extracted all arguments of the verb occurrence in a corpus, and then all dependents of the noun as found by the process described in Sect. 5.3.[12] The noun was added as a deverbative to the base verb only if at least one of the arguments (the same word/lemma) was found as a dependent at any occurrence of the noun.

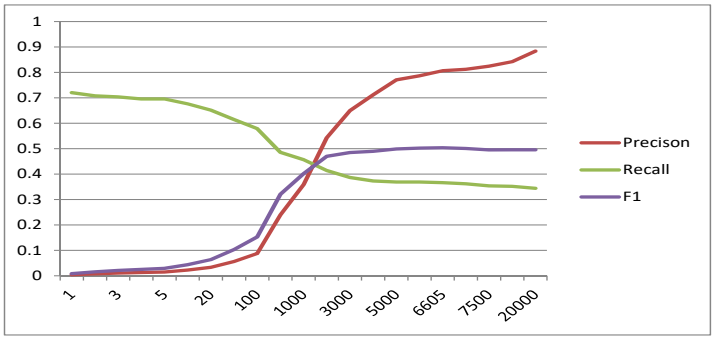

Figure 1: Recall, precision and F_1 for selected threshold values on the development dataset

However, it proved insufficient to allow the noun to be added if such sharing appeared just once - there was still too much noise, especially for CzEng. Thus we have set a threshold, which indicates how many times such a sharing should occur before we consider the noun to be a deverbative. This threshold has been (automatically) learned on the development data, and has been found to be optimal if set to 6 for the PCEDT and to 6605 for the large CzEng corpus. It has then been used in the evaluation of all the variants of the evaluation set. The effect of increasing the threshold is (as expected) that precision gradually increases (from below 1% to over 80%) while recall decreases (from slightly above 72% to below 37% at the F_1-driven optimum, Fig. 1). The F_1-optimized maximum is in fact flat and depending on the importance of recall, it could also be set at much lower point where the recall is still around 50%, which no other method came close to without dropping precision close to zero. A narrower range of precision/recall increase/decrease has been obtained on the small PCEDT corpus, with the threshold set relatively low at 6 occurrences; the highest recall (at threshold = 1) was below 53%.

[12]The deep level of annotation of the PCEDT and CzEng is used, which uses so-called tectogrammatical annotation (Mikulová et al., 2005). From this annotation, arguments and other "semantic" dependents can be easily extracted.

This filtering is referred to in the following text and tables as "shared arg" with the threshold as an index. "PCEDT" and "CzEng" indicate which corpus has been used for the primary noun extraction as described earlier in this section.

5.4 Combination with WordNet

The systems based on the parallel-corpus-based method have been also combined with the WordNet method; nouns extracted by the baseline method are always included.

6 Evaluation and Results

The measure used has been F-measure (F_1), see Eq. 1. The design of the experiments has intentionally been wide to assess how either high recall or high precision can be obtained; depending on the use of the resulting sets of deverbatives, one may prefer precision (P) or recall (R); therefore, for all experiments, we report, in addition to the standard F_1 measure, also both P and R.

All experiments have been evaluated on all three versions of the evaluation dataset (see Sect. 4 for more details on the evaluation dataset properties and the preparation process). We also report results on the development dataset, just as a sanity check. The results are summarized in Table 2.

Experiment	Measure	development dataset	intersection eval. data	union eval. data	majority eval. data
baseline	R	.3402	.3470	.1843	.3214
	P	.9151	.3720	.9640	.7920
	F_1	.4960	.3591	.3094	.4573
+WordNet	R	.3471	.3657	.1934	.3312
	P	.8519	.3415	.8815	.7108
	F_1	.4932	.3532	.3172	.4518
+parallel (PCEDT, L1 filter)	R	.4417	.4664	*.2798*	*.4302*
	P	.4909	.2090	.6120	.4431
	F_1	.4650	.2887	.3841	.4366
+parallel (CzEng, L1 filter)	R	.4801	*.5075*	*.3150*	*.4659*
	P	.0196	.0082	.0247	.0172
	F_1	.0377	.0160	.0458	.0332
+parallel (PCEDT, shared arg.$_6$)	R	.4156	*.4701*	.2492	.4107
	P	.6998	.3158	.8170	.6341
	F_1	.5215	**.3778**	.3820	**.4985**
+parallel (CzEng, shared arg.$_{6605}$)	R	.3663	.3806	.2064	.3442
	P	.8066	.3269	.8654	.6795
	F_1	.5038	.3517	.3333	.4569
+WordNet +parallel (PCEDT, shared arg.$_6$)	R	.4211	*.4813*	.2584	*.4188*
	P	.6703	.2959	.7752	.5917
	F_1	.5173	.3665	**.3876**	.4905
+WordNet +parallel (CzEng, shared arg.$_{6605}$)	R	.3731	.4067	.2194	.3604
	P	.7619	.3079	.8107	.6271
	F_1	.5009	.3505	.3454	.4577

Table 2: Summary of results of all experiments

The best F_1 scores are in **bold**, the best and second best (and close) recall scores are in *italics*.

To interpret the table, one has to take into account the ultimate goals for which the discovered deverbatives will be used. If the goal is to acquire all possible nouns which could possibly be deverbatives, and select and process them manually to extend, say, an existing noun valency / predicate argument lexicon, recall R will be more important than precision or the equal-weighted F_1 score. On the other hand, if the results are to be used, e.g., as features in downstream automatic processing or in NLP machine learning experiments, the F_1 measure, or perhaps precision P, would be preferred as the main selection criterion. It is clear that there are huge differences among the tested extraction methods, and thus all possible needs can be served by selecting the appropriate method.

Regardless of the use of the results, we can see several general trends:

- The baseline method, which used only a limited number of regular derivations of the base verb (cf. Sect. 5) and no additional lexicons or corpora, is actually quite strong and it was surpassed only by the optimized parallel corpus method(s).

- WordNet does not help much, if at all, both in the basic system where it is only combined with the baseline and in the last two systems when it adds to the results of the optimized systems. The increase in recall - which was the assumed contribution of WordNet - is small and the loss in precision substantial, even as F_1 grows.

- A manually annotated corpus, not surprisingly, gets much more precise results than a large but only automatically analyzed corpus (PCEDT vs. CzEng). The precision of the results when using CzEng alone with only simple filtering is so low that the result is beyond usefulness; however, the optimized method of filtering the results through (potentially) shared arguments between the verb and its deverbative gets surprisingly high precision even if not quite matches the PCEDT's overall F_1.

- Using a large parallel corpus (CzEng) with 100s of millions words gives us the opportunity to fine-tune the desired ratio between recall and precision by using the desired weight of recall on the F-measure definition, within a very wide range.

7 Discussion, Conclusions and Future Development

We have described and evaluated several methods for identifying and extracting deverbatives from base verbs using both lexical resources and parallel corpora. For development and evaluation, we have also created datasets, each containing 100 verbs, for further improvement of these methods and in order to allow for easy replication of our experiments.[13]

The best methods have used parallel corpora, where the translation served as a bridge to identify nouns that could possibly be deverbatives of the given base verbs through back-and-forth translation alignment. Due to the noisiness of such linking, filtering had to be applied; perhaps not surprisingly, the best method uses optimized (machine-learned) threshold for considering words shared in the deep linguistic analysis of the base verb and its potential deverbative. This simple optimization used the F_1 measure as its objective function, but any other measure could be used as well, for example F_2 if recall is to be valued twice as much as precision, etc.; this is possible thanks to the wide range of recall / precision values for the possible range of the threshold.[14]

We will further explore the argument-sharing method, adding other features, such as the semantic relation between the verb/deverbative and their arguments, in order to lower the filtering threshold and therefore to help increase recall while not hurting precision (too much). Using additional features might require new machine learning methods as well.

Finally, we will also independently check and improve our test datasets; while the "majority" voting which we have used in our experiments as the main evaluation set is an accepted practice, we would like to further improve the quality of the datasets by thoroughly checking whether the valency transformation rules as described especially in (Kolářová, 2006; Kolářová, 2005) do hold for the verb-noun pairs recorded in the datasets, amending them as necessary.

A natural continuation would be to test the methods developed on other languages, primarily English, even if the morphosyntactic transformations between a verb and a noun are not as rich as for inflective languages (such as Czech which we have used here).

We believe that for one of the intended uses of the described method, namely extending a valency lexicon of nouns with new deverbatives linked to their base verbs, the system could be used in its current state as a preprocessor suggesting such nouns for subsequent manual checking and selection; the argument sharing method optimization can be then used to balance the right ratio between desired high recall and bearable precision.

[13] The development and evaluation datasets will be freely available under the CC license, and the code will be also available as open source at `http://lindat.cz`.

[14] Upper bound for recall was at over 72% by using CzEng, see the discussion about optimization in Sect. 5.3.2.

Acknowledgments

This work has been supported by the grant No. DG16P02B048 of the Ministry of Culture of the Czech Republic. In addition, it has also been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (projects LM2010013 and LM2015071). We would like to thank the reviewers of the paper for valuable comments and suggestions.

References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Marion Baranes and Benoît Sagot. 2014. A language-independent approach to extracting derivational relations from an inflectional lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2011. Czech-English Parallel Corpus 1.0 (CzEng 1.0). LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3921–3928, İstanbul, Turkey. European Language Resources Association.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In Petr Sojka et al., editor, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238. Masaryk University, Springer International Publishing.

Silvie Cinková. 2006. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2170–2175, Genova, Italy. ELRA.

G. Graffi. 1994. *Sintassi*. Le strutture del linguaggio. Il Mulino.

Jan Hajič and Jaroslava Hlaváčová. 2016. MorfFlex CZ 160310. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th LREC 2012)*, pages 3153–3160, Istanbul, Turkey. ELRA.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum.

E. Hajičová and P. Sgall. 2003. Dependency Syntax in Functional Generative Description. *Dependenz und Valenz–Dependency and Valency*, 1:570–592.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková Razímová, and Zdeňka Urešová. 2006. *Prague Dependency Treebank 2.0*. Number LDC2006T01. LDC, Philadelphia, PA, USA.

P. Kingsbury and M. Palmer. 2002. From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. Citeseer.

Veronika Kolářová. 2005. *Valence deverbativních substantiv v češtině (PhD thesis)*. Ph.D. thesis, Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Praha, Czechia.

Veronika Kolářová. 2006. Valency of Deverbal Nouns in Czech. *The Prague Bulletin of Mathematical Linguistics*, (86):5–20.

Veronika Kolářová, 2014. *Special valency behavior of Czech deverbal nouns*, chapter 2, pages 19–60. Studies in Language Companion Series, 158. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An Interim Report. In *In Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, pages 70–77, Boston. Association for Computational Linguistics.

A. Meyers. 2008. Using Treebank, Dictionaries and GLARF to Improve NomBank Annotation. In *Proceedings of The Linguistic Annotation Workshop, LREC 2008*, Marrakesh, Morocco.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, and Lucie Kučová. 2005. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, ÚFAL MFF UK, Prague, Prague.

Karel Pala and Pavel Smrž. 2004. Building Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(2-3):79–88.

Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. 2011. Czech WordNet 1.9 PDT. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Jena D. Hwang, Susan Windisch Brown, Karin Kipper Schuler, and Arrick Lanfranchi. 2009. Leveraging lexical resources for the detection of event relations. In *Learning by Reading and Learning to Read, Papers from the 2009 AAAI Spring Symposium, Technical Report SS-09-07, Stanford, California, USA, March 23-25, 2009*, pages 81–87.

Jarmila Panevová. 1974. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Jarmila Panevová. 1994. Valency frames and the meaning of the sentence. *The Prague School of Structural and Functional Linguistics*, 41:223–243.

Jarmila Panevová. 1996. More remarks on control. *Prague Linguistic Circle Papers*, 2(1):101–120.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1087–1093, Reykjavík, Iceland. European Language Resources Association.

Matthew Stone, Tonia Bleam, Christine Doran, and Martha Palmer. 2000. Lexicalized grammar and the description of motion events *. In *TAG+5 Fifth International Workshop on Tree Adjoining Grammars and Related Formalisms*. Paris, France.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Stroudsburg, PA, USA. Johns Hopkins University, Baltimore, MD, USA, Association for Computational Linguistics.

Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.

Zdeňka Urešová. 2011. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*, volume 1 of *Studies in Computational and Theoretical Linguistics*. UFAL MFF UK, Prague, Czech Republic.

Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Milan Straka. 2015. Derinet v 1.0, http://lindat.cz.

Jonáš Vidra. 2015. Implementation of a search engine for derinet. In Jakub Yaghob, editor, *Proceedings of the 15th conference ITAT 2015: Slovenskočeský NLP workshop (SloNLP 2015)*, volume 1422 of *CEUR Workshop Proceedings*, pages 100–106, Praha, Czechia. Charles University in Prague, CreateSpace Independent Publishing Platform.

Zdeněk Žabokrtský and Magda Ševčíková. 2014. DeriNet: Lexical Network of Derivational Word-Formation Relations in Czech.