Automatic Identification of Valency Frames in Free Text

Martin Wörgötter

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

Abstract. In this paper I present a versatile tool for automatic labelling of Czech verbs in free text with VerbaLex valency frames. The effective implementation can process one sentence in 0.03 seconds on average. I provide an overview of the algorithm and its evaluation.

Key words: VerbaLex, WordNet, valency frame, verb, annotation, tagger

1 Introduction

Valency lexicons are important lexical resources, which make it possible to disambiguate morphological, syntactic as well as semantic aspects of language. One such resource is *VerbaLex* [1] developed at the *Natural Language Processing Center at the Faculty of Informatics*. A significant feature of *VerbaLex* is its interconnection with the semantic lexical database *WordNet* [2].

Another notable verb valency lexicon for Czech is *VALLEX* [3]. A machine learning algorithm for matching verbs with corresponding valency frames of *VALLEX* was proposed in [4].

Assigning appropriate *VerbaLex* valency frames to verbs in a text is challenging. The difficulty of the task lies in discriminating between multiple valency frames. The algorithm for solving this task must necessarily encompass morphological and syntactic analysis.

In the following sections, I describe the implemented rule-based algorithm and the evaluation on five manually prepared test sets. Due to the exploitation of verb valency features specific to *VerbaLex*, a generalization of the algorithm for use with other verb valency lexicons is quite limited.

2 Implementation

The tool processes the output of the syntactic parser *SET* [5] given the options `-preserve-xml-tags` and `-long-phrases`. The second option ensures that morphological tagging is output together with the syntactic tree.

Syntactic analysis segments the input into clauses, which are the main scope for the algorithm. In each clause, each verb is looked up in *VerbaLex* for a list of

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 91–96, 2016. © Tribun EU 2016

possible candidate valency frames, i.e. those which allow the given lemmatized form of the verb. A set of tests, which is depending on the frame specification, is evaluated on each candidate. Only if all tests succeed, the candidate is accepted. Each verb triggers the following two tests.

– Principal verb—auxiliary verbs are discovered using the syntactic structure of the clause and are discarded.
– Reflexivity—the verb has to be reflexive/irreflexive according to the frame specification. This is verified by searching for a reflexive particle.

Valency frames in *VerbaLex* have the structure of a list of participants, which are either obligatory or facultative. Participants can capture semantic information via subcategorization features represented by *WordNet* literals. Surface grammar constraints are encoded using the properties listed bellow. Multiple possible values for each constraint are supported.

Depending on its specification, each obligatory participant imposes some of the following tests.

– Subcategorization features require a constituent which falls into the set of hyponyms of the second level semantic role.
– Surface grammar constraints are the following:
  • Case: same case number
  • Category of personality: a heuristic by which the grammatical gender of a *masculine noun* has to match the specified category
  • Prepositional lemma: it has to be found
  • Adverb: it has to match a word
  • Infinitive: a verb in infinitive has to be found
  • Subordinating conjunction lemma: has to be found in a subordinate clause

The described algorithm heavily relies on database lookups of *VerbaLex* and *WordNet*. To achieve the required performance for tagging large corpora, a command line option is available which employs a caching procedure during initialization to overcome this problem.

3 Preparation of a gold standard

Currently, no collection of sentences, manually annotated with *VerbaLex* valency frames is available, thus it was necessary to prepare the annotated data.

A simple web application, depicted in Figure 1 served this purpose. The annotators were five students with a specialization in computational linguistics.

The sentences for annotation were taken from the *czTenTen* [6] corpus using the *Sketch Engine* corpus interface [7] to filter out sentences not containing verbs and sort them by their *GDEX* score [8], a value expressing suitability for being used as a dictionary example. From the resulting list the top 900 sentences were extracted. 150 sentences were put aside and the rest was divided into five

Fig. 1: The web application interface for annotating verbs.

sets. The 150 separate sentences were then randomly intermingled into each set raising their size to 300. The intersection between sets was later used to assess the inter-annotator agreement.

Each annotator was required to choose exactly one of the following options for each verb.

1. *No allowed frame*: a preselected option in case the verb is not recorded in *VerbaLex*
2. *Match*: a suitable frame was found
3. *No match*: no appropriate frame is available in *VerbaLex*
4. *Not a verb*: the word was erroneously recognized as a verb
5. *Auxiliary*: the verb is auxiliary
6. *Infinitive*: an infinite verb.

Only in case of a match, the annotator continued by choosing one or more appropriate valency frames which were listed with respect to the verb lemma.

4 Evaluation

The difficulty of the task is underlined by the obtained inter-annotator agreement. Only in 17.5%, all five annotators agreed on the same set of valency frames. To alleviate this problem, four gold standards, according to the number of agreements between annotators, were established, see Table 1.

For each gold standard the third column presents the total number of verbs for which at least 2-5 annotators (depending on the gold standard) agreed to assign at least one frame (corresponds to the option *match* described above). The last column shows the number of verbs for which the annotators agreed on the exact set of assigned frames, with respect to the third column.

The evaluation of the implementation is presented in Table 2. According to the results nearly every third analysed verb is correctly assigned the exact set of appropriate valency frames.

Table 1: Gold standards according to the number of agreements between annotators.

name	agreements	agreed to assign	full agreement (%)
GS1	at least 2	160	70.0
GS2	at least 3	119	43.7
GS3	at least 4	81	34.6
GS4	at least 5	40	17.5

Table 2: Evaluation results.

	precision (%)	recall (%)	F-score (%)
GS1	13.8	8.0	10.1
GS2	21.2	13.4	16.4
GS3	31.5	21.4	25.5
GS4	25.0	14.2	18.1

5 Error analysis

Figure 2 gives an example of both a successful and unsuccessful assignment of frames for verbs *přát* and *jet*.

The only frame accepted by the algorithm for verb *přát* is plausible. A problem occurs in the subordinate clause, as the *agens* does not match the semantic role *machine:1*. The algorithm assigned two frames, the first one being incorrect due to missing anaphora resolution.

The algorithm is very sensitive in processing the results of syntactical and morphological analysis and cannot cope with errors in the input data, which is

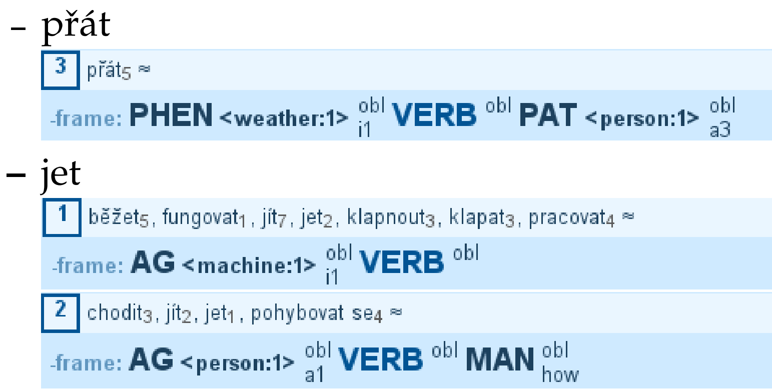

Fig. 2: Example of erroneous result: "Počasí nám zatím nepřálo a tak jsme rychle jeli dál." (The weather was not good so far so we quickly went away.)

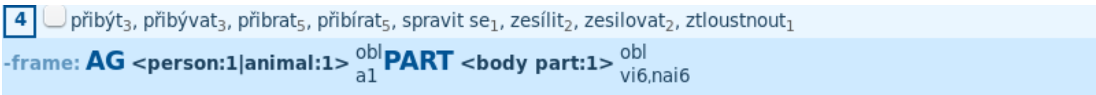

Fig. 3: Example of erroneous result: "Pokud se to ignoruje a pokračuje se v sestupu, bolest v uších stále zesiluje." (If it is ignored and the descent continues, the pain in the ears keeps increasing.)

demonstrated in Figure 3. An irrelevant frame is assigned to verb *zesílit* in the third clause. According to the syntactical analysis, this clause has a zero subject and the noun *bolest* is syntactically an object and any subcategorization features on a zero subject succeed.

6 Conclusions

In this paper I presented a tool, which is able to assign appropriate *VerbaLex* valency frames to verbs in free text.

The evaluation has proven the difficulty of the task, especially considering inter-annotator agreement. On the other hand, by relaxing the conditions for matching the gold standard, better results could be achieved.

The provided implementation could be used to enhance *VerbaLex* by semi-automatically adding corpus examples to the respective valency frames.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and the Specific University Research.

References

1. Hlaváčková, D.: Databáze slovesných valenčních rámců VerbaLex. Disertační práce, Masarykova univerzita, Filozofická fakulta, Brno (2008)

2. Fellbaum, C.: WordNet: An Electronic Lexical Database. Language, speech, and communication. MIT Press, Cambridge (1998)
3. Žabokrtský, Z., Lopatková, M.: Valency information in VALLEX 2.0: Logical structure of the lexicon. The Prague Bulletin of Mathematical Linguistics (87) (2007) 41–60
4. Lopatková, M., Bojar, O., Semecký, J., Benešová, V., Žabokrtský, Z.: Valency lexicon of czech verbs vallex: Recent experiments with frame disambiguation. In: Text, Speech and Dialogue. Volume 3658., Karlovy Vary, Česká republika, Springer (2005) 99–106
5. Kovář, V., Horák, A., Jakubíček, M.: Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In: Human Language Technology. Challenges for Computer Science and Linguistics, Berlin/Heidelberg, Springer (2011) 161–171
6. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In: 7th International Corpus Linguistics Conference CL. (2013) 125–127
7. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1** (2014)
8. Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M., McAdam, K.: GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the XIII EURALEX International Congress, Barcelona, Institut Universitari de Lingüística Aplicada (2008) 425–432