Valency Frames of Polysemous and Homonymous Verbs in VerbaLex

Vít Baisa

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic
xbaisa@fi.muni.cz

Abstract. Verb valency lexicon VerbaLex is one of few language resources which take a local context into account. For each Czech verb with its sense, VerbaLex contains appropriate valency frames (patterns with morphological and syntactical information) in which the verb can appear. This, and the fact that a context of a word is crucial for determining sense of the word, makes VerbaLex suitable for disambiguation of polysemous and homonymous verbs. This paper tries to manifest this via investigation of valency frames of polysemous and homonymous verbs. Some related aspects of VerbaLex and its contents are discussed too.

1 Introduction

Polysemy and homonymy are phenomena which cause many problems in natural language processing. It is easier to solve the latter as was shown by [1] and [2]: accuracy about 95% can be attained in disambiguation of homonyms.

Homonymy is accidental phenomenon and that is why two homonymous words usually differ a lot in their behaviour and contexts. E.g. two homonymous Czech verbs *sladit*. Their meanings are a. *to sweeten* and b. *to coordinate*. Obviously, their usual contexts differ a lot.

Polysemy is far harder task to deal with. Accuracy of a solution depends strongly on granularity of senses in a reference dictionary. In the case of a fine-grained sense distinction even human annotators may not agree each other on particular disambiguation.

Since supervised methods perform better than other approaches [3, p. 56] it is reasonable to set eyes on data sources with semantic annotations. VerbaLex falls into this class. In this article we will discuss its possible contribution to *verb sense disambiguation* (VSD).

2 VerbaLex

VerbaLex is valency lexicon of Czech verbs. Each verb lemma, together with a number of its sense, represents a *literal*. Synonymic literals are grouped into *synsets* – basic elements of both WordNet [4] and VerbaLex. Each synset in VerbaLex has list of *frames* valid for particular literals from the synset. VerbaLex

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 9–14, 2010. © Tribun EU 2010

nowadays contains about 6,300 verb synsets, 21,000 literals, 10,500 verb lemmas and 20,000 frames. For more detailed description of VerbaLex structure see [5]. Here we will describe only its most important component – *frame*.

2.1 VerbaLex Frame

A frame includes combination of semantical, syntactical and morphological information about context of a particular verb. The frame is represented by a list of *semantic roles* with important additional information. There are 29 roles such as AG (agens), LOC (location), PAT (patient) etc. Verbs themselves have label VERB in frames.

Additional information is

- obligation: if a role is obligatory or optional in the frame,
- semantic class: a set of possible words on a position of the role represented by WordNet literal and
- morphological and syntactical constraints of the role, e.g. direct or prepositional case, animality etc.

For more detailed description of additional information, again, see [5]. Example of valency frame for verb *mačkat* follows:

$$\mathsf{AG}^{kdo1}_{person:1} + \mathsf{VERB} + \mathsf{OBJ}^{co4}_{object:1} + (\mathsf{PART}^{v\ \check{c}em6}_{hand:1}).$$

Optional roles are surrounded by parentheses and additional information is in superscripts and subscripts. The frame is formal representation of usual verb behaviour. A realisation of the frame may be e.g. *Honza mačkal bankovky v ruce.* (*John was pressing bank-notes in his hand.*). All constraints are met: agens *Honza* is in nominative and is animate (*kdo*1, i.e. *who* in animate nominative), object *bankovky* is in accusative and is inanimate (*co*4, i.e. *what* in inanimate accusative) etc.

Frames are core of VerbaLex. They describe the majority of possible contexts of verbs which might be used in disambiguation of polysemous and homonymous verbs.

We assume that each meaning of a verb has its own specific context. Since local contexts are represented by frames in VerbaLex, we can evaluate disambiguational potential of VerbaLex by checking uniqueness of these frames.

3 Frames of Polysemous and Homonymous Verbs

3.1 VerbaLex as Python Data Structure

VerbaLex comes in two formats: XML and text format. Our simple procedure goes through VerbaLex and converts it into a dictionary of frames. Keys of the dictionary are lemmas and its value is a list of couples: (number of a meaning of lemma, a list of valid frames for an appropriate literal). Frames are lists of semantic roles and a semantic role bears mentioned additional information.

3.2 Comparison of Frames

After the converting, another procedure compares all frames of all pairs of polysemous and homonymous verbs (lemmas). Thanks to described structure of the dictionary this step is quite straightforward.
The number of all possible pairs is expressed by the following formula:

$$\sum_{v \in V} \sum_{i,j \in S_v, i<j} min \left(|\mathrm{F}_v^i|, |\mathrm{F}_v^j| \right) .$$

Structure of the formula corresponds (to a certain extent) to steps of the latter procedure: for each homonymous or polysemous verb (lemma) v in VerbaLex (V) and for each pair of meanings i and j from a set of meanings of v (S_v), the procedure compares all combinations of pairs of frames. F_v^i represents a set of frames for a given lemma v and its meaning i. Cardinality of this set is $|\mathrm{F}_v^i|$.

The formula expresses the highest number of possible identical frames between all pairs of meanings. In the case of VerbaLex it equals to 190,795. The second procedure looks for frames which are identical.

4 Results

The number of identical frames depends on several criteria. At first we looked for absolute identities. I.e. the case when two frames are identical in all information they bear: semantic roles and all additional information. It yielded 891 pairs of literals with at least one identical frame.

Then we checked less strict identity: two frames were considered as identical if they consisted of same obligatory semantic roles (whereas the previous identity takes into account obligatory as well as optional semantic roles) together with appropriate additional information. In that case there were 2,203 identical frames.

The third option was to compare only names of semantic roles, i.e. to omit additional information. This option yielded 3,343 frames.

Since substantial number of matches corresponded to perfective and imperfective variants of verbs which share same frames, we manually removed all imperfective variants and obtained 605 pairs of literals as the fourth option.

Other synonymic variants of verbs (two forms of infinitive – *pomoci* and *pomoct*) also share frames so number of really identical frames in Verbalex is even smaller.

Results are summarized in Table 1 on the following page.

4.1 Examination of Identical Frames

Results shown above are very promising. Out of 21,000 literals, only 605 share a frame. Moreover, if we take a look at concrete pairs of literals with identical frames we will discover that utter majority of them are rather annotation inconsistencies than real identities between frames. We can classify these identities into 3 groups.

Table 1. Number of identical frames according to various criteria

quantity	‰	**identity criterion**
891	4.67	absolute identity
2,203	11.55	identity only for obligatory semantic roles
3,343	17.52	identity of semantic role names
605	3.17	absolute identity without perfective and imperfective variants
190,795	10^3	all possible identities

Invalid Verb in Subsynset Since frames are usually not valid for all literals in a synset, they are assigned to subsynsets. A lexicographer must decide for which verbs in a synset a frame holds and then to create an appropriate subsynset for the frame.

Frame shared between literal *mrkat*:2 (*to wink*) and *mrkat*:3 (*to watch something with interest intermittently*) is

$$\mathsf{AG}^{kdo1}_{person:1} + \mathsf{VERB} + \mathsf{PAT}^{na\ koho4}_{person:1}.$$

The subsynset which contains the literal *mrkat*:2 contains also literals *mrknout*:2 and *zamrkat*:1. It is all right since the two literals are just perfective variants of *mrkat*:2. The problem is with the subsynset which contains the literal *mrkat*:3. It also contains literals *mrknout*:3 (perfective variant of *mrkat*:3), *pokukovat*:2 and *pomrkávat*:1. In the case of this subsynset, the first two literals (including *mrkat*:3) are not valid for this frame and should be removed from the subsynset.

Insufficient Distinction by Semantic Class Some verbs whose senses are distinguished very finely differ in details. If the only distinction between them is on lexical level and there is not suitable semantic class in WordNet, they can not be distinguished by valency frame itself. Literals *hořet*:1 and *hořet*:4 may serve as example. Their share frame

$$\mathsf{SUBS}^{co1}_{substance:1} + \mathsf{VERB}.$$

The first literal stands for *undergo combustion* and the second for *glow*. It is hard to imagine some substance which is burning and at the same time is not glowing. If we wanted to distinguish these cases we would need semantic classes for substances which are glowing whilst burning and for substances which can burn without emitting any light.

There are also other annotation errors in frames, especially in additional information: wrong animality, prepositional case etc. These mistakes should also be corrected.

Too Fine-Grained Distinction between Verb Senses In some cases there are very fine differences between senses which can not be distinguished by frames.

Frame

$$\mathsf{AG}^{kdo1}_{person:1} + \mathsf{VERB} + \mathsf{ART}^{co4}_{artifact:1} + (\mathsf{SUBS}^{\check{c}\acute{\imath}m7}_{material:1}).$$

is shared between literals *pokreslit*:1 and *pokreslit*:2. Both verbs have meaning *to deface*. The former meaning is rather neutral – *to cover with paintings*, whereas the latter is more negative – *to deface and to depreciate something by that*. This distinction can not be expressed in VerbaLex using only frames and additional information.

5 Conclusion

The second and the third group point to general problem of fine-grained word senses in natural language processing. We are able to distinguish tens of senses per word but then we are not able to distinguish between them in real applications automatically: the more senses we have the worse are results of word sense disambiguation. The question is whether we need to have all these fine-grained senses in our dictionaries at all.

Nevertheless, the experiment proved that VerbaLex could be very useful for *verb sense disambiguation* of polysemous and homonymous verbs. If we recognised a frame of a polysemous verb in a sentence, we would attain high precision in VSD. And since many synsets in VerbaLex are linked to appropriate synsets in English WordNet, this VSD could be used directly in machine translation from Czech to English.

6 Future Work

Our goal is to check all pairs of verb senses with identical frames manually and, if possible, to correct annotation errors. The procedure should be tuned to be fast enough and not to enable importation of new errors by an annotator.

The second goal is to check all frames in VerbaLex for soundness. The main endeavour is to find out whether all frames in VerbaLex are well-founded using a Czech corpus. Then we plan to discover which frames are useless or, on the contrary, which frames must be added to increase coverage of VerbaLex [6].

Both should improve consistency and quality of language data in VerbaLex.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536. I want also to thank people involved in building VerbaLex for their respectable work on the advanced data source.

References

1. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (Cambridge, MA). 189–196. 1995.

2. Stevenson, M., Wilks, Y. The interaction of knowledge sources in word sense disambiguation. Computational Linguistics 27, 3, 321—349. 2001.
3. Navigli, R. *Word Sense Disambiguation: A Survey*. ACM Computing Surveys, 41 (2), 2009, pp. 1–69.
4. Fellbaum, C. *WordNet: An electronic lexical database*. The MIT Press, 1998.
5. Hlaváčková, D., Horák, A. *VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech*. In Computer Treatment of Slavic and East European Languages. Bratislava, Slovakia: Slovenský národný korpus, 2006. pp. 107–115.
6. Jakubíček, M., Kovář, V., Horák, A. *Measuring Coverage of a Valency Lexicon using Full Syntactic Analysis*. In Sojka, P., Horák, A. (Eds.): RASLAN 2009: Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2009. pp. 75–79.