Towards Partial Word Sense Disambiguation Tools for Czech

Tomáš Čapek and Pavel Šmerk

Faculty of Informatics, Masaryk University, Brno, Czech Republic
xcapek1@aurora.fi.muni.cz, smerk@fi.muni.cz

Abstract. Complex applications in natural language processing such as syntactic analysis, semantic annotation, machine translation and especially word sense disambiguation consist of several relatively simple independent tasks. Czech, belonging among Slavonic languages with many inflectional features, requires more effort for such tasks, in comparison with other languages. In this article we present two software tools to tackle morphological disambiguation and multi-word expression recognition for Czech in a cost saving and time-efficient way.

1 Introduction

Word sense disambiguation (WSD) is the most fundamental task in NLP. In past decades much effort has been made to develop tools to resolve WSD in its entirety, i.e. to correctly disambiguate all words in all contexts. This issue is discussed in detail in [3] as largely unsuccessful. Numerous authors are cited to discuss the reasons behind this and conclude that:

- Many WSD systems assign sense labels from pre-established lexical resources (sense inventories) such as traditional dictionaries and are therefore relative to the sense inventory used, content of which may be at each instance subject to interpretation and might ultimately be unsuitable for some applications. Quality of sense inventories is however not the focus of this paper.
- Whether a sense inventory is used or not, the focus is too often set on division of senses that is too fine-grained even for a human user to distinguish. The effort to encompass as much exceptions and rare sense occurrences can lead to needless complexity of the WSD system whereas NLP can provide useful results by relying on far less.
- Computational WSD should reorient itself to tasks it can easily perform with high accuracy even if they only provide partial results compared to full WSD.

By combining partial solutions for WSD we can make our results more accurate in overall – in ideal case, each step in the text processing might be able to filter out some of the potential variants of the particular word. Below we present two software tools that allow us to partially disambiguate words or collocations in Czech texts. In Section 2 we introduce *Desamb*, a hybrid morphological tagger and Section 3 deals with multi-word expression recogniser called *mrec*.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 103–108, 2010. © Tribun EU 2010

2 Desamb – Morphological Tagger

Morphological analysis is a process which assigns all possible pairs of a lemma and a morphological tag to an analysed word form. A morphological guesser does the same for words unknown to the morphological analyzer. Morphological disambiguation, also called tagging, is a process to determine which lemma and morphological tag is correct with respect to a particular context of the analysed word form in a sentence.

Desamb is an experimental hybrid tagger for Czech, in which rule-based and statistical algorithms are combined [9]. The disambiguation process consists of several independent tools whose inputs and outputs are managed by additional scripts. These tools include morphological analyser, morphological guesser, chunk parser and a tagger based on hidden Markov models (HMM).

As an input, *Desamb* accepts vertical file where the text is stored in the format of one word form or punctuation token per line. The first step of the disambiguation process is a simple detection of sentence boundaries. Then each word form is assigned all its possible pairs of a lemma and a morphological tag by morphological analyzer *ajka* [6]. Similarly, morphological guesser for Czech then computes the same information for word forms that are not covered by *ajka* dictionary [10]. This step concludes the necessary preprocessing of the input text.

Next phase performs the actual disambiguation and can be divided into two steps. In the first step, various lexical filters and morpho-syntactic rules are applied on the data to remove obviously incorrect tags from the list of potential ones for each word form. The filters can be context-independent, for example pronouns *si* or *mi*, which are very frequent in texts, are also recognized by *ajka* as solmization syllables, i.e. nouns. In real texts there are virtually no occurences of this alternative so it can be omitted altogether without causing any measurable inaccuracy in the results. Other filters use simple context information, for example *Se* at the beginning of a sentence can never be a pronoun but is always a preposition. More complex rules are part of a partial syntactic analyser DIS (also called a chunk parser), which recognizes noun, prepositional, and verb phrases in sentences [11]. In Czech, these phrases regularly demostrate certain agreements among several grammatical categories which allows us to remove such tags that do not correspond to given phrase pattern described by a rule.

These filters and rules are designed to be as accurate as possible even if they perform with low recall. Their ambition is always to make the resulting ambiguity lower and to never remove a correct tag. In the end it is still possible for some word forms to have more than one tag attached to them, so the disambiguation at this stage is only partial. On the other hand, these filters and rules are highly reliable and can also be used independently on other tools discussed in this paper.

Finally, a statistical trigram algorithm is used to prune the remaining tags that still need to be disambiguated. An HMM tagger represents the implementation of this approach. At this step each word form is left with exactly one morphological tag.

Because the guesser cannot deal properly with foreign words, especially names, the overall precision of *Desamb*, i.e. the portion of word forms with the correct morphological tag, is 91.0%. When we exclude the foreign words (or, if we can assume flawless morphological analysis of the input text) the precision increases to 95.3%. These results are however quite preliminary, as they were experimentally computed on just 2000 tokens originating in newspaper articles where names occur relatively frequently.

In overall, the main advantage of our approach is its modularity. The individual components of *Desamb* can be replaced or left out. For example, if we use just the HMM tagger, we get fully disambiguated results with relatively low accuracy while using only filters and rules yield highly accurate but partial disambiguation.

3 mrec – Multi-Word Expression Recogniser

Multi-word expression (MWE) recognition is one of the important tasks in NLP. For many applications we need to process MWEs (collocations) as standalone lexical entities for the purpose of lemmatization or parsing. By *MWE* or *collocation* we understand a lexical unit whose meanings can't be inferred from the meanings of the words that make it up, i.e. set phrases, compound words and idioms, rather than any statistically significant word group occurring in a large volume of texts, such as a corpus. To be more specific, we count the following categories among collocations – all can be inflected in Czech:

- multi-word named entities such as toponyms, geographical, proper and other names (e.g. Mediterranean Sea, Julius Caesar, Creative Commons),
- general collocations and set phrases (e.g. carnivorous plant, elementary school, red tape),
- multi-word abbreviations (e.g. a. s., před n. l.),
- Czech reflexive verbs (e.g. kolíbat se, usnadnit si) — these, along with phrasal verbs constitute vast majority of Czech multi-word expressions among verbs. In English, this majority is represented by phrasal verbs alone (e.g. catch on, take off).

One motivation to develop a MWE recogniser closely relates to WSD. Lexical units intrinsically possess a feature of having exactly one sense if they consist of more than one word as it has been verified by D. Yarowsky in [12]. By exploiting this feature we can basically get a partial disambiguation of any text "for free"[1]

Statistical techniques in MWE recognition provide rather approximate results and are more suitable for discovering general multi-word regularities that are outside the scope of our definition of a collocation. For reliable semantic classification of collocations we prefer to utilize rule-based methods and to have a large MWE database at our disposal.

[1] We can also observe similar feature of lexical units if we consider how many senses of a unit we get within one discourse. It has been verified with high accuracy that it is one sense per discourse [2].

We have developed a large Czech MWE database which at the moment contains 160,470 lexical units. It was compiled mostly semi-automatically from various resources such as encyclopedias and dictionaries, public databases of proper nouns and toponyms, collocations obtained from Czech WordNet [4], botanical and zoological terminologies and others. It was originally built for a question answering system UIO which inevitably influenced its composition [7]. The current data can serve as a metalexicon because for each entry the reference to a real dictionary or similar resource is available – this increases the quality of our data in comparison with its previous version [5].

In Table 1 we present basic statistics of the MWE types in the database and their frequencies in SYN2000 corpus which contains 114,363,813 tokens and is a part of Czech National Corpus (CNC) [1].

- column # *MWEs* contains a numbers of the MWEs in our database for each given domain.
- column # *Occs* presents a number of MWE occurrences of each given domain in the SYN2000.
- # *Unique* is a number of individual MWEs from a given domain which occur in the SYN2000 at least once.
- *% of all* represents percent of MWEs occurring in the SYN2000 in comparison with all MWEs from a given domain.
- # *HL* denotes *hapax legomena*, i.e. MWEs with only one occurrence in the SYN2000.
- # *not in corpus* is a number of the MWEs, which did not occur in the SYN2000.

Table 1. Statistics of Czech MWE Database.

Domain	# MWEs	# Occs	# Unique	% of all	# HL	# not in corpus
Botanics, zoology	63,153	13,707	3,279	5.1	1,538	59,874
Culture	6,828	30,279	2,042	29.9	505	4,786
Toponyms	14,561	102,683	2,554	17.5	652	12,007
Proper names (people)	61,152	289,794	15,092	24.7	3,851	46,060
Unsorted	14,776	656,971	7,628	51.6	774	7,148
Total	160,470	1,093,434	30,595	19.1	7,320	129,875

Mrec itself consists of one script that accepts the output from *Desamb* in a form of vertical text file with each line looking as follows:

```
word_form <l>word_form_lemma <c>morphological_tag
```

The vertical file represents the preprocessed corpus data – word order and sentences are preserved. Due to complex inflection in Czech, each line in the *mrec* MWE database uses the following syntax:

$$\mathtt{c}_1\ \mathtt{c}_2\ \ldots\ \mathtt{c}_n\#\mathtt{l}_1\ \mathtt{l}_2\ \ldots\ \mathtt{l}_k\text{\`{}}\mathtt{l}_{k+1}\ \mathtt{l}_{k+2}\ \ldots\ \mathtt{l}_{k+m}$$

In this format, variables c_x denote component words of a collocation while variables l_x represent lemmata of the respective component words. Character “`” is used as a mark that divides the collocation in two parts; the component words to the left from the mark may be inflected while other component words to the right from the mark have their word forms fixed. For example, there is a collocation in Czech *běžný účet platební bilance (current account of balance of payments)* which would have the mark right in the middle. Both left and right part of the lemmatized form of the collocation may be empty, i.e. the whole collocation may be either fully inflectable or fully fixed.

The purpose of the mark is to help filter out collocations which have some of its component words incorrectly inflected. It also increases performance of the collocation recognition process as it is unnecessary to check fixed component words for any inflection they could otherwise demonstrate. If a collocation exists as a sub-collocation of another and both are recognized in the database, only the longer is by default returned to output.

Mrec is designed as a lightweight tool to be used as a component in a larger system. High processing speed was a major issue in its development. We can report that *mrec* processes as much as 6,000 input tokens per second.

4 Future Work

In the future work we intend to optimize the performance of *Desamb*, specifically the chunk parser and processing speed of the morpho-syntactic rules. The implementation is done in Prolog programming language which is not optimal for tagging large volumes of corpora texts.

Our primary application for the tools described in this paper is to improve semantic annotation of free text with WordNet database serving as the main sense inventory. One of the long-term tasks to do this is to enrich the WordNet database with collocations from our MWE database that are missing in the semantic network and thus can’t be used for the annotation. Another way to improve the results is to decrease the sense granularity in the WordNet lexical data with the help of specially designed heuristic tests [8].

Desamb can also be exploited as a part of text processing during the annotation. In this task only information about part-of-speech is necessary to get about the words on input as every lexical unit in the sense inventory is stored as a lemma. By simply adding *Desamb* to the process we get partial WSD “for free" as obviously incorrect morphological tags get filtered out even before we get to recognize collocations. The primary goal of this approach in general is, at each step, get at least slightly less ambiguous form of input text.

5 Conclusion

We have presented two software tools that provide us with partial disambiguation of Czech texts at two different levels – morphological tags and multi-word expression recognition. By combining them with other techniques such as discourse boundary spotting or word sense labeling we predict can get reasonable and useful WSD results without developing a standalone, monolithic and complex system dedicated to the problem specifically.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and by the Czech Science Foundation under the project 407/07/0679.

References

1. *Český národní korpus – SYN2000*. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Praha, Czech Republic, 2000.
2. W.A. Gale, K.W. Church, and D. Yarowsky. One Sense per Discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics, 1992.
3. N. Ide and Y. Wilks. Making Sense About Sense. *Word Sense Disambiguation*, pages 47–73, 2006.
4. K. Pala and P. Smrž. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7:79–88, 2004.
5. K. Pala, L. Svoboda, and P. Šmerk. Czech MWE Database. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, pages 1–5. European Language Resources Association (ELRA), 2008.
6. R. Sedláček. *Morphemic Analyser for Czech*. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2005.
7. L. Svoboda. Processing of Natural Language Multiword Expressions. In V. Snášel, editor, *Proceedings of Znalosti 2004*, Ostrava, Czech Republic, 2004. Technical University Ostrava.
8. T. Čapek. Semantic Network Integrity Maintenance via Heuristic Semi-Automatic Tests. In *Proceedings of the RASLAN Workshop 2009*, pages 63–67. Masaryk University, Brno, Czech Republic, 2009.
9. P. Šmerk. *K morfologické desambiguaci češtiny (On Morphological Desambiguation of Czech)*. Ph.D. thesis proposal, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2007.
10. P. Šmerk. Towards Czech Morphological Guesser. In P. Sojka and A. Horák, editors, *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*, Brno, 2008. Masarykova univerzita.
11. E. Žáčková. *Parciální syntaktická analýza češtiny (Partial Syntactic Analysis of Czech)*. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2002.
12. D. Yarowsky. One Sense per Collocation. In *Proceedings of the workshop on Human Language Technology*, pages 266–271. Association for Computational Linguistics, 1993.