Semantic Annotating of Czech Corpus via WSD

Robert Král

Natural Language Processing Laboratory
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
rkral@fi.muni.cz

Abstract

We would like to describe the relationship between word sense disambiguation (WSD) and language resources (LR) working with word senses. We discuss the problem of sense division and tagging. Exploiting specific features of the inflectional languages for WSD is encouraged. We present WSD methods for Czech ambiguous nouns. The advantage of these methods consists in reducing the manual work by using a synonym training set. They could be utilized for building a semantically annotated corpus or gaining glosses for Czech WordNet.

1. Introduction

1.1. Main Problems and Solutions in WSD

WSD tries to solve the problem of assigning a sense to the words occuring in a text or speech. It is connected with two questions: how to represent word senses and in which way the correct sense of the word occurence can be determined on the contextual basis.

A lot of methodologies for WSD are being used (Ide, Véronis, 1998). The problem of the WSD methods based on machine learning consists in the manual annotation of words in a text which is needed for building a training set. The data sparseness problem occurs at the WSD solutions which utilize machine readable dictionaries or lexical databases. Finally, the WSD methods which only divide contexts into the groups in accordance with semantic similarity (Schütze, 1998) of the given ambiguous word suffer from the problem of sense interpretation.

One of our goals that we want to present in the paper is how to construct some sense-annotated data with the minimum of manual work required for tagging.

1.2. Relation between WSD and LR

The role of LRs in WSD is twofold. LRs can provide (annotated) data for the solution to WSD, but, on the other hand, WSD methods could support the creation of LRs. The main types of LRs that contain information about word senses in Czech are: 1. Czech WordNet (Pala, Ševeček, 1999) which at present does not contain any Czech glosses and 2. machine readable dictionaries, for example Dictionary of Literary Czech (2000) offering just a few examples of using senses and in our opinion with an impropriate division of word senses. This is the reason why we pay attention to the development of the WSD methods that can be used for semantic annotation of Czech corpora. It is possible to extract several sentences (documents) containing the word with the same sense from such a semantically annotated corpus or link appropriate sentences in the corpus with an item in a machine readable dictionary.

2. Word Senses

2.1. Small Sense-Tagged Corpus for Czech

By the term sense-tagged corpus we mean a corpus in which the word occurencies are tagged with their word sense. Such a corpus can give information about structure of word senses and their distribution, represents word senses in a text and could serve as a training and testing set for WSD methods. There are many sense-tagged corpora for English, such as Hector (Atkins, 1993) and others.

Since for Czech no sense-tagged corpus has been introduced so far, we have created a small sense-tagged corpus called DESEM. The textual material for the DESEM corpus has been taken from an already established Czech corpus called DESAM (Pala et. al., 1998). 1364 sense-tagged occurences of nine Czech ambiguous nouns have been manually recorded. The most difficult problem was to find the individual senses because quite fine grained classification can be used. The tags are in the following form: *sn - l - a - h*, where n is an identification number and (if it is possible) it corresponds to the number of the synset in the Czech part of EuroWordNet (Pala, Ševeček, 1999; Vossen, 1988), l is a lemma, a is a synonym and h is a hyperonym. Another 802 sense tags of the lemma *fronta* was annotated in the ESOSEM corpus whose textual material was taken from the ESO corpus. The senses of the word *fronta* together with their distributions are presented in Table 1.

We have tested our WSD methods by means of the DESEM corpus. Generally speaking, semantically annotated corpora could be used in other fields of NLP since they represent a reasonably good source of the usage of senses.

2.2. Sense Division

During the sense-tagging of DESEM there were problems with specifying the word senses, therefore, we explored this question.

We have been inspired by (Schütze, 1998) who proposed to interpret a sense as a group of the context including the common ambiguous word, therefore, the word sense in this interpretation does not exist at all. Schütze finds these groups by automatic clustering according to the similarity between two contexts. On the other hand, the word senses

sense tags	ESOSEM		DESEM	
fronta	abs. fr.	rel. fr.	abs. fr.	rel. fr.
s1-bojová-linie/front	37	4.6 %	12	16.4 %
s2-průčelí-bok/frontage	1	0.1 %	1	1.4 %
s3-sdružení-skupina/group	308	38.4 %	24	32.9 %
s4-řada-skupina/queue	92	11.5 %	18	24.6 %
s5-atmosférická-jev/atmosphere	39	4.9 %	6	8.2 %
s6-mladá-noviny/newspapers	325	40.5 %	12	16.4 %
tags total	802		73	

Table 1: The distribution of the senses for the word *fronta*.

are specified intuitively in dictionaries. We have tried to introduce the division of senses in a more formal way but with regard to human experience.

We randomly chose 50 contexts of the ambiguous word *vazba* and observed its word senses. One of the biggest problems in tagging the senses by a human is to specify a definition of a sense. We formulated an easier task: to decide if two ambiguous occurencies of the given ambiguous word *vazba* in two contexts have the same sense or not. These decisions were made by hand gradually for 126 couples of contexts. Because of transitivity of similarity, the context a and c were automatically identified as similar if we tag context a similar to b and b similar to c. Here is the algorithm which finds the partitioning $\mathcal{R}$ of n contexts according to the similarity of word senses between two contexts. This algorithm uses for this purpose the external procedure $SemSim$.

```
begin
    input C = {c_1, ..., c_n}
    for i = 1 ... n do
        [c_i] := {c_i}
    R := {[c_i]}
    for i = 1 ... n - 1 do
        if [c_i] = {c_i} then
            for j = i + 1 ... n do
                if [c_j] = {c_j} then
                    if SemSim(c_i, c_j) then
                        [c_i] := [c_i] ∪ [c_j]
                        R := R - {[c_j]} ∪ {[c_i]}
                    endif
    output R
end
```

We have got 9 classes of contexts reflecting 9 senses. There is a list of their descriptions for the most frequent among them:

1. linkage, with 19 contexts
2. detention, 17
3. feedback, 4
4. colligation, 4
5. bookbinding, 3

We compared this sense division with the division in the Dictionary of Literary Czech (2000) and found significant distinctions – some word senses are mixed together, etc. For a detailed division of word senses we would have to start, of course, from a bigger set. This experiment implies a need for constructing better and more exact LRs working with word senses.

2.3. Exploiting Relations between Inflectional Variants of Words and Senses

Each WSD system gains information given by the context in which an ambiguous word is placed. But we have anticipated that the part of the information about the sense of the word occurence is hidden in the form of the word. Of course, this approach could not be tested on the non-inflectional languages like English. But Czech has flection and we have tested our suggestion by means of DESEM corpus.

We suggest to construct a brief classifier assigning the most likely sense to a word form. These propabilities could be counted from the contingency table. Let $f_{t,s}$ be the number of co-occurencies of the word form t and the sense tag s in the training set, in our experiment given by the DESEM corpus. Then the right sense tag for an occurence of the word form t should be chosen according to the equations 1.

$$\begin{aligned} s(t) &= arg\max_{\forall s_i} P(s_i|t) \\ &= arg\max_{\forall s_i} \frac{P(s_i,t)}{P(t)} \\ &= arg\max_{\forall s_i} P(s_i,t) \\ &= arg\max_{\forall s_i} f_{t,s_i} \end{aligned} \quad (1)$$

If we suppose that the contingency table counted from DESEM represents a language model we can try to measure the precision of the rule for a word form by Equation 2, where $f_{t.}$ is the marginal frequency. In Equation 3 there is the total precision for all word forms of a given lemma.

$$\begin{aligned} prec_t &= \frac{\max_{\forall s} f_{t,s}}{f_{t.}}, \\ f_{t.} &= \sum_{\forall s} f_{t,s} \end{aligned} \quad (2)$$

$$\begin{aligned} prec &= \frac{\sum_{\forall t} \max_{\forall s} f_{t,s}}{f_{..}}, \\ f_{..} &= \sum_{\forall t} f_{t.} \end{aligned} \quad (3)$$

On the basis of the DESEM corpus we have found out, for example, that if the lemma *vazba* occurs in the context

word form⇒sense tag		expect. precision
vazbě	⇒*s8-područí/custody*	94.1 %
vazba	⇒*s5-vztah/link*	94.1 %
metru	⇒*s1-míra/measure*	92.4 %
srážek	⇒*s7-počasí/rainfall*	83.9 %
srážkách	⇒*s2-šarvátka/mellay*	83.3 %
vazbu	⇒*s5-vztah/link*	81.8 %
srážkami	⇒*s7-počasí/rainfall*	70 %

Table 2: Desambiguation rules on the word form basis.

in the locative form *vazbu*, the meaning *custody* is highly probable. If the nominative form *vazba* is used, then the meaning should be interpreted as *link*. There are more examples of the disambiguation rules with their precision in Table 2. Nevertheless, for many word forms nothing about the expected word sense can be said. Obviously, it is not appropriate to use these rules separately, however, the word form is an interesting feature which can be taken into account in conjunction with the context for WSD.

3. Some WSD methods

The task is to determine the correct sense of ambiguous (target) word in text. We can work with corpus texts thanks to the `Manatee` corpus library (Rychlý, 2000). We can also use Czech analyzer `ajka` (Sedláček, 2001) for lemmatization of word forms.

3.1. Construction of the Training Set by Using Synonyms

WSD techniques based on machine learning require a sense-tagged learning set. The problem lies with the construction of such a set because its manual construction is time-consuming. We propose substitution of the word sense with an adequate synonym. For example, the ambiguous Czech word *fronta* is synonymous with *řada/queue* or with *sdružení/association*. Thus, we suggest to extract the concordance set of synonym *řada/queue* from a common corpus where this set will serve as a learning set. For other meanings we obtain the training set the same way. However for some target words the synonyms are too ambiguous, or too rare for building a training set, or do not exist at all. In this case, we have tried to replace synonym by near synonym.

To achieve the training set no manual work is neccessary. With training set done, it is possible to apply an arbitrary learning technique. We have done experiments with an algorithm based on the vector space model, k-NN method and decision lists.

3.2. Vectors and Neighbours

The words in all concordances containing any synonym were lemmatized and some among them (included in the stoplist) were eliminated. Adapted concordances were transformed to the numeric vectors using vectors space model technique with bigram matrix (Manning, Schütze, 1999). Each vector is labelled by an appropriate synonym occuring in vector's context. We gained the training set for each sense of given target word.

The concordance containing target word (target concordance) wanted for disambiguation is transformed to (target) vector the same way. Then we choose the most similar vector from the training set labelled by a synonym and assign it to target vector. This method is called the nearest neighbour algorithm. The similarity of two vectors is measured by the angle formed by them. Thus we disambiguate the sense of target word.

We have done experiments with variant of k-NN method, where the k nearest neighbour vectors is searched. The most frequent synonym is chosen from among them.

The results depend on synonym selection and on values of many parameters, such as the length of context, number of training concordances, etc. The precision (the number of correct disambiguated occurencies divided by the total of disambiguated occurencies) ranged from 50 % to 75 % and coverage (the number of disambiguated occurencies among the total of occurencies) was 100 % (except of k-NN, where k is even).

3.3. Decision List (DL)

Using synonym training set in DL (Rivest, 1987) can de divided in two steps: obtaining the DL and the disambiguation itself.

1. In the DL there are be rules indicating that some feature of concordance determines the sense, in this case the synonym. Thanks to the training set we can compute conditional probabilities of that in concordance containing lemma w the target word has sense s. In Equatations 4 let $f_{s,w}$ be number of co-occurencies of the synonym s and lemma w and f_w is the number of occurencies of w in training set. Smoothing is also used in Equatation 5 because of rare occurencies of certain lemmata.

$$P(s|w) = \frac{P(s,w)}{P(w)} \quad (4)$$

$$\doteq \frac{f_{s,w}}{f_w}$$

$$\doteq \frac{f_{s,w}+\epsilon}{f_w+n\epsilon} \quad (5)$$

The pair (w, s) can be interpreted as a decision rule. For example *dramatický → epizoda (dramatic →episode)* conveys: if *dramatický* occurs in concordance then we predict synonym *epizoda* associated with appropriate sense. These pairs-rules sorted in descending order by probability represent the DL.

2. After lemmatization of words in concordance and elimination of some special forms included in the stoplist, each lemma is looked up in the left side of rules in DL. The right side of the rule with the highest probability score is then selected. The target word is labelled by this right side with synonym-sense. For example we disambiguate lemma *vazba* in concordance like *vzájemná vazba mezi pachateli/mutual linkage between the burglars* as a *vztah/linkage*, because the respective rule has higher score. When no rule reaches the minimal score, disambigaution will fail.

We tested this technique and one of the best results was 91 % of precision and 10 % of coverage

vazba	→sense tag-synonym
vězení/retention	→*s8-trest/penalty*
svoboda/freedom	→*s8-trest/penalty*
vzájemný/mutual	→*s5-vztah/linkage*
soud/court	→*s8-trest/penalty*
vlastnický/possessory	→*s5-vztah/linkage*
pachatel/burglar	→*s8-trest/penalty*
úzký/narrow	→*s5-vztah/linkage*
obviněný/charged	→*s8-trest/penalty*
obchodní/trading	→*s5-vztah/linkage*
sloveso/verb	→*s3-věta/sentence*
provádět/execute	→*s2-pokus/experiment*

Table 3: A part of DL for ambiguous word.

for lemma *vazba* using these synonyms: *věta/sentence, pokus/test, vztah/linkage, trest/penalty* and similar results for lemma *srážka* with synonyms: *zásah/shot, konflikt/conflict, peníze/money, počasí/weather*. Reversely, for lemma *scéna* and *fronta* the results were not good (about 50 % of precision).

4. Improving LRs via WSD

The applications of WSD to machine translation and information retrieval have been stressed in many articles. However, the WSD methods could be used for obtaining sentences containing a word of a given sense. This data may help to enrich dictionary records by adding relevant gloss to them.

Anyway, the WSD methods are going to be used for creating glosses (examples of use) for synsets in Czech Wordnet because there are not any. Obtaining such sentences from the corpus is the task which we are interested in now. For example we need a gloss for synset *strana* in the sense *stránka/page*. In the current state it is necessary to specify the synonym *stránka* manually. The occurence of target *strana* in the right sense could be automatically found. This occurence along with its context can form an appropriate gloss. For this purpose the WSD methods where sense is represented by the synonym are more suitable because synonymy is intrinsic feature of WN. For the sake of automatic gloss retrieval, precision is the most important measure. Since one or two correct glosses usually suffice enough, the coverage or recall measures are not substantial for this purpose. In the current state the tools and program libraries for automatic work with synsets in the Czech WordNet are developed. Later creating glosses could be done fully automatically. Of course, the quality of finding the right glosses depends on quality of WSD methods.

We can look on the same problem from another point of view as well. If we have a synset with at least two synonyms there is a possibility of finding a gloss containing either one or the other. We can construct two sets of concordances where appropriate synonyms occur and find the most semantically similar concordances because they will probably represent the same sense. The similarity of concordances should be measured by the similarity of word senses in concordance.

This specific task can be extended to interconnection between WordNet and corpus which would be a very useful LR. Each synset should be connected to all appropriate occurencies of a given sense in the corpus. Reversely, each occurence in corpus (except the some word classes) should be linked to the synset in WN. There would be a connection between the division of senses and their use. The advantages of hierarchical division of senses contained in WN and posibility of asking queries through the corpora managers would thus be combined.

5. Acknowledgement

This research has been supported by Czech ministry of Education, Research Program CEZ:J07/98:143300003. My thanks goes to Karel Pala, Radek Sedláček, Pavel Rychlý, Tomáš Čapek, Martin Dvořák and Miloslav Nepil for their help.

6. References

Atkins, S., 1993. Tools for computer-aided lexicography: the Hector project. *Papers in Computational Lexicography: COMPLEX'93,* Budapest.

Ide, N., Véronis, J., 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics,* Vol. 24, Num. 1.

Král, R., 2001. Three Approaches to Word Sense Disambiguation for Czech. In V. Matoušek et. al., *Text, Speech and Dialogue,* 4th International Conference, Berlin : Springer-Verlag, s. 174–179.

Manning, Ch. D., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.

Pala, K., Rychlý, P., Smrž, P., 1998. DESAM – An Annotated Corpus for Czech. *Proceedings of SOFSEM'98,* Springer.

Pala, K., Ševeček, P., 1999. Czech Lexical Database of the WordNet Type (within EuroWordNet-2). *Sborník prací filosofické fakulty brněnské university*, Masaryk University, Brno, 51–64.

Rychlý, P., 2000. *Korpusové manažery a jejich efektivní implementace.* Ph.D. thesis, Masaryk University, Brno.

Slovník spisovného jazyka českého (Dictionary of literary Czech.) Akademia, Praha, 1960, electronic version, Praha, Brno.

Rivest, R. L., 1987. Learning Decision Lists, *Machine Learning*, 2, 3, 229–246.

Sedláček, R., Smrž, P., 2001. A New Czech Morphological Analyser ajka. In V. Matoušek et. al., *Text, Speech and Dialogue,* 4th International Conference, Berlin : Springer-Verlag, s. 100–107.

Schütze, H., 1998. Automatic Word Sense Discrimination. *Computational Linguistics,* 24, 1, 97–123.

Vossen, P., 1988. *Set of Common Base Concepts in EuroWordNet-2*. Final Report, 2D001, Amsterdam.