Word Sense Discrimination for Czech

Robert Král

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

Abstract. This paper deals with the automatic discrimination of contexts of Czech ambiguous words. The Schütze's methodology was used, modified and transformed for the Czech language. This algorithm is based on word space and clustering. The semantic discrimination could be understood as a subtask of word sense disambiguation. In this approach, the sense of word is defined as the cluster of contexts of ambiguous word. We show that Schütze's method is transportable into Czech. Our results are not so good as his because we have experimented with a highly ambiguous word.

1 Introduction

The problem of word sense disambiguation is defined usually as "the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word" [3]. The solution consists in two steps: the summarizing of target word senses and associating the correct meaning with the word occurrence. The area of wsd problem particularly covers the task of machine learning, information retrieval and text processing as well as others.

In this area many approaches have recently appeared: some of them are based on thesaurus (Wilks and Stevenson [9]), other take advantage of WordNet (Sussna [7]) or bilingual corpora (Brown [1]).

The sense definition and representation is one of the biggest problems in wsd. Schütze [6] proposed an approach which evades this question because he just diversifies and unites the contexts of given word according to whether they are semantically similar or not. For this purpose he speaks about the context group discrimination based on clustering. The semantically similar contexts share a common cluster. Even if the contexts are not labelled by senses this labelling can be done if the correspondence between the senses and clusters exists. The sense discrimination is a subtask of the word sense disambiguation.

In this paper we would like present approach which is based on Schütze's research. His method has been simplified and applied to the Czech language.

In addition, we think that this approach could be applied not only to the word sense disambiguation but it could also be used for gaining comment examples for Czech WordNet synsets.

2 Algorithm

The task is to classify the contexts of target words according to semantic similarity. The target word is a Czech noun which has more senses. In the following text we would cling to Schütze's terminology.

The idea of the algorithm consists in creating the word space in which each context is represented by a vector. All the context vectors are grouped in a few clusters. The vectors in the same cluster have a similar angle, that is they are semantically similar. Clusters could represent the senses of words.

The algorithm of word sense discrimination could be described in the three steps:

1. The construction of bigram matrix.
2. The calculation of the context vector for each context.
3. Grouping the context vectors into the clusters.

In the first step, we construct bigram matrix so that in the left margin of the matrix there are words which are called features and the words in the top margin dimensions. The number in the cell of the matrix expresses the number of times for which both words share the same context. Usually, the number of features words is much higher than that of dimensions words. The overcoming of data sparseness is hidden above this because each word in the context will next be represented by the vector of a few dimensions. The words in the matrix are most frequent and sensefull lemmas from nouns, adjectives and verbs in contexts.

In the next step the context is lemmatized by a morphological analyzer `ajka` [5] and the word vectors are gained for any lemma from the adequate row of matrix. Indeed, the scalar in vector expresses how strongly the lemma from the context corelates with the corresponding dimension. If the word form in the context has more lemmas, then the word vector is created by means of vectors. If the lemma is not in the matrix, it is ignored. Then, we get the context vector by computing means (arithmetic average of scalars) of word vectors.

	otrávit/to poison	*cesta/journey*
jet/go	0	4
jed/poison	2	0
kolo/bicycle	1	5

Table 1. Illustration example for two-dimensional word space.

In the sentence *Jedu na kole/I am riding a bicycle*, for example, the first word has two lemmas *jet, jed* and the word vector is (1, 2). The third word in the sentence is represented by the vector (1,5). The context vector for the sentence is thus (1, 3.5). It means that the sentence is closer to the second dimension.

The clustering of the context vectors into groups is the last stage of the algorithm. We chose the non-hierarchical clustering algorithm which is called k-means. In the beginning, k centroids in the word space are selected. Then, each object (context vector) is assigned to the closest centroid. The clusters are created by vectors assigned to the same centroid. After that, the new centroids are calculated as a means of the cluster's members. This computation is repeated till the clusters are stabilized. The distance is measured by the angle between two vectors and describes the similarity of contexts.

The peculiarity of these clustering techniques is the initialization of centroids. They are usually selected at random. However, we could influence the cluster results with an appropriate initialization. In the task of gaining semantically similar contexts we could select centers so that they would correspond to the sense definition in a dictionary.

3 Experiment

We have tested the previous algorithm using the Czech National Corpus [2], or, to be more precise, using the two thousands contexts of the target *srážka*. The contexts had the length of 20 positions or shorter if they overlapped the document border.

The bigram matrix had the dimensions of 300 rows (features words) by 40 columns (dimensions word). We decided to find four clusters (k = 4) because the word *srážka* has four basic meanings. The senses of the word target *srážka* are *clash, discount, precipitation* and *collision* or *crash*. The centroids were initialized manually so that they corresponded to sense division. The settings could be done automatically from an electronic dictionary such as SSJČ [10] or EuroWordNet [8].

Cluster:	ozbrojený armed	daň tax	mzda wage	teplota temperature	oblačnost cloudiness	zahynout deaden	voják soldier	vlak train
1	0.07	0.01	0.01	0.01	0.00	0.02	0.04	0.01
2	0.01	0.16	0.20	0.01	0.00	0.01	0.01	0.00
3	0.02	0.01	0.01	0.12	0.08	0.01	0.01	0.00
4	0.03	0.01	0.01	0.01	0.00	0.08	0.02	0.04

Table 2. The founded centroids in selected dimensions for *srážka*.

The algorithm divided two thousand contexts into four groups. These contained 959, 288, 342 and 284 contexts. 127 contexts were not clustered because there was not enough context information within. The precision rate was that of 74 % and was measured for 200 contexts. Schütze claims the precision rate between 76 % and 83 % for a natural ambiguous word. But there is a distinction in the comparison because we have tried to disambiguate the highly ambiguous word.

4 Conclusions and Future Work

In this paper we have presented an approach that is based on the Schütze's research which uses the word spaces and the clustering. The contexts of target word and the senses are represented by the vectors in this word space. We enrich his approach with the solution to more lemma variants in the context. By using the k-means clustering algorithm we attempt to show that there are many ways of initializing cluster centers. Considering that variability of algorithm, it is possible to change the number of senses and their definitions according to individuals neccesities. On the other hand, some simplification of the algorithm was done - we did not use the singular value decomposition and χ^2 test.

We have conducted small experiments with a highly ambiguous word. The advantage of gaining senses automatically was confirmed and it turned out to be neccessary to assess the words according to their discriminating potential (the log document frequency known from an information retrieval).

In the future work we would like to focus on enlarging the bigram matrix and testing the algorithm using more czech lemmas. We hope it will be possible to incorporate this algorithm into the sophisticated wsd system used for the sense tagging of nouns in the corpus.

References

1. Brown, P. F. et all: Word Sense Disambiguation using statistical methods, In Proc. of the 29th Annual Meeting, Berkeley, p. 264-270, 1991.
2. Čermák F.: Czech National Corpus: Its Character, Goal and Background, Proc. of Workshop on TSD'98, Springer, Pilsen, 1999.
3. Ide, N., Véronis, J.: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, Computational Linguistics, Vol. 24, Num. 1, 1998.
4. Manning Ch.D., Schütze H.: Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, 1999.
5. Sedláček R., Smrž P.: A New Czech Morphological Analyser `ajka`, In proceedings of the 4th Workshop on Text, Speech and Dialogue - TSD 2001, Berlin, 2001.
6. Schütze H.: Automatic Word Sense Discrimination, [3], p. 97–123.
7. Sussna M.: Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, Proc. of the 2nd International Conference on Information and Knowledge Management, Arlington, 1993.
8. Vossen, P., et al.: Set of Common Base Concepts in EuroWordNet-2, Final Report, 2D001, Amsterdam, October 1988.
9. Wilks Y., Stevenson M.: Sense Tagging: Semantic Tagging with a Lexicon, Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?, Washington, D.C., 1997.
10. Slovník spisovného jazyka českého (Dictionary of literary Czech), Akademia, Praha, 1960, electronic version, Praha, Brno, 2000.