What makes sense?
Searching for strong WSD predictors in Croatian

Nikola Bakarić
Department of Information Sciences,
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: nbakaric@ffzg.hr

Jasmina Njavro
Department of Information Sciences,
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: jnjavro@ffzg.hr

Nikola Ljubešić
Department of Information Sciences,
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: nljubesi@ffzg.hr

Summary

The goal of this research was to investigate and determine position of strong predictors for word sense disambiguation of Croatian nouns. Research was conducted using supervised learning methods and a corpus of around 70 million words. We have concluded that words in the immediate vicinity of an observed lexeme (1-5 words left and right) have the highest discriminative power. We have also measured the applicability and accuracy of the one-sense-per-discourse method and found it to be very successful as well as the impact of sentence boundaries which proved not to be a good criterion for selecting strong predictors.

Key Words: word sense disambiguation, Croatian language, strong predictors

Introduction

Multi-sensed words have presented a problem in computer processing of natural languages since its beginnings. These words carry more than one sense and therefore present a problem in many high-level NLP tasks like information retrieval, automated indexing and machine translation. There are two general approaches in dealing with such problems: stochastic and deterministic [8].

In recent decades the stochastic approach has gained popularity due to increase of processing power and its high efficiency. Again, there are two approaches in stochastic methods - supervised which uses labeled data to build statistical models and unsupervised which uses clustering algorithms without having any labeled data on hand. Supervised algorithms, as expected, achieve much better results than unsupervised [9]. In this research we use a supervised method.

Determining sense of a word is often a complex task, even for humans. The inter-annotator agreement between annotators that prepare data for the SENSEVAL competition is around 60% [2]. In the SENSEVAL competition annotators focus on fine-grained sense distinctions. This research deals with lexemes that have related, but distinctly different senses (strong polysemy) because we believe that there is no point in trying to distinguish fine nuances of meaning which are often unclear to human evaluators. The approach we use is gaining popularity in the NLP community [10]. When trying to determine the sense of a particular lexeme, humans rely on the information given through the context [11]. This research focuses solely on the context of observed lexemes as we try to determine the relationship between position of a word regarding the observed lexeme and its discriminative power in WSD.

Preparing the data

The corpus on which the research was conducted consists of on-line articles of Vjesnik daily paper from May 30th 1999 to December 31st 2006 and it is not POS tagged or lemmatized [6]. The main identifier of the article is the URL and the structure is as following: title, subtitle, text.

• Two separate lists were put together. Each list consisted of articles extracted from the corpus in which lexemes "miš" ("mouse" – the first list) and "stanica" ("cell" – the second list) appear. The lists were then randomly divided into ten sets used in 10-fold cross-validation. They were verticalised and sentence boundaries were marked.

• Next step was to determine possible word senses present in the lists and then to manually annotate the sense of very occurrence of the observed lexemes. Around 1000 occurrences were evaluated. The occurrences in the first 60 percent of the lists were annotated by both annotators together to determine the sense inventory. The remaining 40 percent was annotated separately so as to determine the inter-annotator agreement. The lexeme "miš" was annotated with eight different senses while "stanica" was annotated with six different senses.

Naïve Bayes classifier

Naive Bayes is the simplest probabilistic learning method of all supervised corpus-based methods for word sense disambiguation [5]. The main idea of this classifier is that it calculates in the training corpus the conditional probability of an event (in our case a token in a specific window around the

observed lexeme) regarding a specific sense of the lexeme. Each token contributes potentially useful information about the sense of the ambiguous word present. The classifier does no feature selection – all types are features – it uses all tokens as bag-of-words around the observed lexeme [3]. It is possible to use some feature selection method as the chi-square or mutual information [4], but at this point our primary interest lies in the relative difference in accuracy concerning the size of the window and its distance from the observed lexeme. The greatest disadvantage of such simple classifier is the fact it assumes that the variables given are independent. In spite of this naiv e design and apparently over-simplified assumptions, naiv e Bayes classifier often works better than some other, more complex classifiers. Due to its simplicity, this classifier is robust enough not to be affected by the curse of dimensionality. Like all probabilistic classifiers under the maximum a posteriori decision rule, it arrives at the correct classification as long as the correct class is more probable than any other class; hence class probabilities do not have to be estimated very well [5].

Results

The one-sense-per-discourse hypothesis assumes that in one discourse a polysemous lexeme is used in only one sense. Yarowsky measures that phenomenon as applicable to English and uses it effectively in his semi-supervised approach to WSD [7]. Since we annotated all occurrences of chosen lexemes in selected documents, it was possible to measure the applicability and accuracy of this hypothesis in our corpus. The method has proven to be applicable in almost one third of cases as well as quite accurate as can be seen in Table 1.

	Applicability	Accuracy
“miš”	28,92%	88,98%
“stanica”	26,31%	97,10%

Table 1 - one-sense-per-discourse applicability and accuracy

We trained and tested the Naive Bayes classifier by using 10-fold cross-validation. Since there was no additional parameter estimation, we did not need a validation set. The experiments were performed with the varying window size and the varying window distance with window size one. Results are shown in figures 1 to 4.

The results of varying window size show that accuracy decreases as the window size increases. In the case of the lexeme "stanica", it decreases constantly while the highest accuracy for the lexeme "miš" is obtained with window size 3. The results for varying window distance show that in case of both lexemes best

sense predictors lie in the first five positions from the observed lexeme and that the discriminative power of more distant tokens is quite constant.
The difference between the lexemes “miš” and “stanica“ lies in the fact that the lexeme “stanica” mostly makes strong NP collocation (“matične stanice“, “autobusna stanica“). That is in our belief the reason why the lexeme “stanica” has it's accuracy peak with a window distance and size of one. The lexeme “miš” needs, as stated before, a window distance or size of three to achieve peak accuracy.

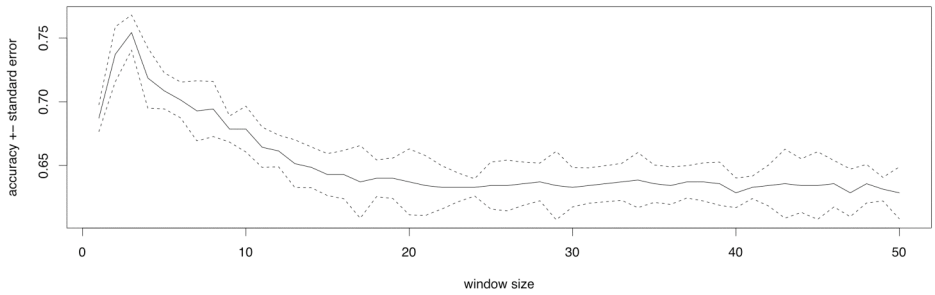

Figure 1 – window size/accuracy ratio for “miš”

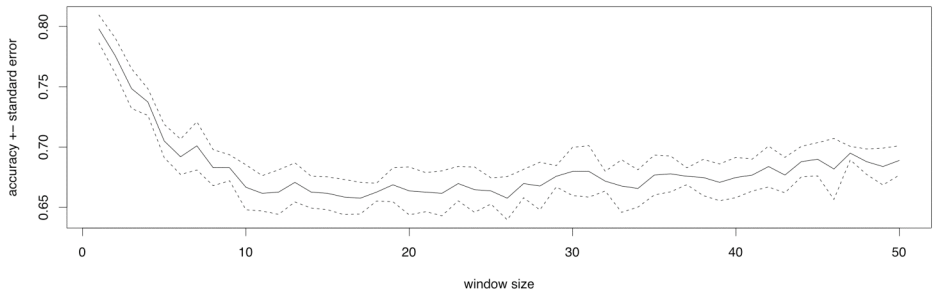

Figure 2 - window size/accuracy ratio for "stanica"

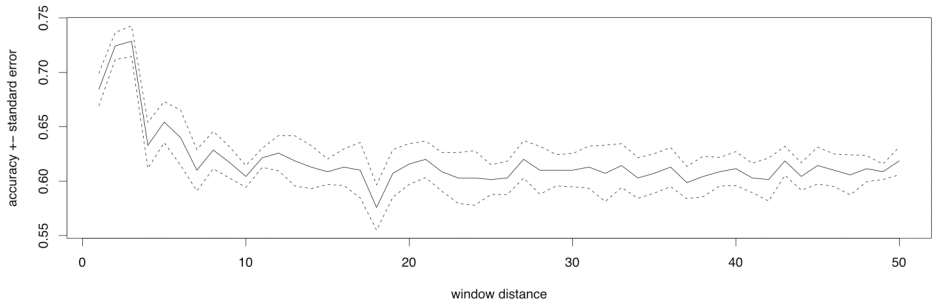

Figure 3 - window distance/accuracy ratio for "miš"

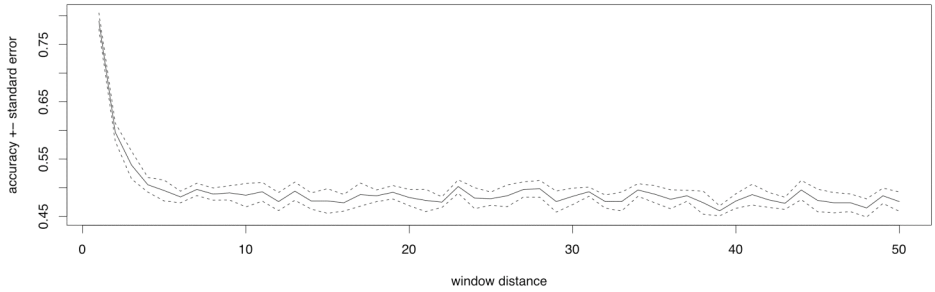

Figure 4 - window distance/accuracy ratio for "stanica"

Furthermore, we experimented with the importance of sentence endings for WSD. We trained one classifier with three first and last tokens in the sentence in which the lexeme occurs and one classifier with three last and three first tokens in neighbouring sentences. The accuracy difference between these two classifiers is shown in table 2. While tokens in the sentence of the observed lexeme are better sense predictors, the difference is rather small and it remains unclear to what extent it is the result of the smaller distance from the observed lexeme in comparison to its possibly bigger discriminative power.

	Before sentence boundary	After sentence boundary
"miš"	68,00%+1,52% (SE)	64,14%+2,09%
"stanica"	57,37%+1,75%	57,27%+1,18%

Table 2 - accuracy with standard error in relation to 3 tokens before/after observed lexeme sentence boundary

Conclusion

Applications of sense disambiguation systems are many. Apart from machine translation; information retrieval, information extraction and text mining could also benefit from a working word sense disambiguation system as well as lexicography [1]. The main goal of the research was to examine the connection between the distance of a token to the ambiguous lexeme and its' discriminative power for WSD. Our main conclusion is that best predictors of a sense of the observed lexemes are situated near that lexeme, usually from 1 to 5 places to the left or right. The one-sense-per-discourse is proven to be applicable in a third of cases and is quite accurate. Since this method is applicable only when the lexeme is mentioned more than once in a discourse, its possible application in a WSD system is limited, but it can still strongly affect the final results, especially in unsupervised and semi-supervised approaches. The sentence limits have not proven to be any significant border of strong WSD predictors.

Since we are not aware of any research of WSD for Croatian, we believe that conclusions drawn in this paper represent a stepping stone for further research in WSD and natural language processing of Croatian language.

References

[1] Agirre, E.; Edmonds, P., editors. Word Sense Disambiguation: Algorithms and Applications. // *Text, Speech and language Technology.* Vol. 33 (2006), pages 10-11

[2] Edmonds, P. SENSEVAL: The evaluation of word sense disambiguation systems. // *ELRA Newsletter*. Vol. 7 (2002), No. 3

[3] Manning, C.D.; Schütze, H. Foundations of Statistical Natural Language Processing. London: MIT Press, 1999, pages 237-239

[4] Manning, C.D.; Schütze, H. Foundations of Statistical Natural Language Processing. London: MIT Press, 1999, pages 169-172

[5] Marquez, L.; Escudero G.; Martinez D.; Rigau G. Supervised Corpus-Based Methods for WSD // Word Sense Disambiguation: Algorithms and Applications. // *Text, Speech and language Technology.* Vol. 33 (2006), pages 185-186

[6] Vjesnik d.d., On-line edition from 30 May 1999 to 31 Dec 2006. http://www.vjesnik.hr (25 Mar 2007)

[7] Yarowski, D. Unsupervised word sense disambiguation rivaling supervised methods. // *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Cambridge, USA, 1995, pages 189-196

[8] Wikipedia, Natural Language Processing. http://en.wikipedia.org/wiki/Natural_language_processing (25 Mar 2007)

[9] Agirre, E.; Edmonds, P., editors. Word Sense Disambiguation: Algorithms and Applications. // *Text, Speech and language Technology.* Vol. 33 (2006), page 7

[10] Ide, N.; Wilks, Y. Making Sense About Sense // Word Sense Disambiguation: Algorithms and Applications. // *Text, Speech and language Technology*. Vol. 33 (2006), pages 48-49

[11] Wikipedia, Distributional Hypothesis. http://en.wikipedia.org/wiki/Distributional_hypothesis (25 Mar 2007)