SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

# UBB system at Senseval3

**Gabriela Serban**
Department of Computer Science
University "Babes-Bolyai"
Romania
gabis@cs.ubbcluj.ro

**Doina Tatar**
Department of Computer Science
University "Babes-Bolyai"
Romania
dtatar@cs.ubbcluj.ro

## Abstract

It is known that whenever a system's actions depend on the meaning of the text being processed, disambiguation is beneficial or even necessary. The contest Senseval is an international frame where the research in this important field is validated in an hierarchical manner. In this paper we present our system participating for the first time at Senseval 3 contest on WSD, contest developed in March-April 2004. We present also our intentions on improving our system, intentions occurred from the study of results.

## 1 Introduction

Word Sense Disambiguation (WSD) is the process of identifying the correct meanings of words in particular contexts (Manning and Schutze, 1999). It is only an intermediate task in NLP, like POS tagging or parsing. Examples of final tasks are Machine Translation, Information Extraction or Dialogue systems. WSD has been a research area in NLP for almost the beginning of this field due to the phenomenon of *polysemy* that means multiple related meanings with a single word (Widdows, 2003). The most important robust methods in WSD are: machine learning methods and dictionary based methods. While for English exist some machine readable dictionaries, the most known being WordNet (Christiane Fellbaum, 1998), for Romanian until now does not exist any. Therefore for our application we used the machine learning approach.

## 2 Machine learning approach in WSD

Our system falls in the supervised learning approach category. It was trained to learn a classifier that can be used to assign a yet unseen example to one or two of a fixed number of senses. We had a *trained* corpus (a number of annotated contexts), from where the system learned the classifier, and a *test* corpus which the system will annotate.

In our system we used the Vector Space Model: a context $c$ was represented as a vector $\vec{c}$ of some features which we will present bellow. By a context we mean the same definition as in Senseval denotation: the content between ¡context¿ and ¡/context¿.

The notations used to explain our method are (Manning and Schutze, 1999):

- $w$ - the word to be disambiguate;
- $s_1, \cdots, s_{Ns}$ the senses for $w$;
- $c_1, \cdots, c_{Nc}$ the contexts for $w$;
- $v_1, \cdots, v_{Nf}$ the features selected.

As we treated each word $w$ to be disambiguated separately, let us explain the method for a single word. The features selected was the set of ALL words used in the trained corpus (nouns, verbs, prepositions, etc) , so we used the *cooccurrence* paradigm (Dagan, Lee and Pereira , 1994).

The vector of a context $c$ of the target word $w$ is defined as:

- $\vec{c} = (w_1, \cdots, w_{|W|})$ where $w_i$ is the number of occurences of the word $v_i$ in the context $c$ and $v_i$ is a word from the entire trained corpus of $\mid W \mid$ words.

The similarity between two contexts $c_a, c_b$ is the *normalised cosine* between the vectors $\vec{c_a}$ and $\vec{c_b}$ (Jurafsky and Martin, 2000):

$$cos(\vec{c_a}, \vec{c_b}) = \frac{\sum_{j=1}^{m} w_{a,j} \times w_{b,j}}{\sqrt{\sum_{j=1}^{m} w_{a,j}^2 \times \sum_{j=1}^{m} w_{b,j}^2}}$$

and $sim(\vec{c_a}, \vec{c_b}) = cos(\vec{c_a}, \vec{c_b})$.

The number $w_i$ is the weight of the feature $v_i$. This can be the frequency of the feature $v_i$ (term frequency or $tf$), or "inverse document frequency ", denoted by $idf$. In our system we considered all the words from the entire corpus, so both these aspects are satisfied.

## 3 k-NN or memory based learning

At training time, our k-NN model memorizes all the contexts in the training set by their associated features. Later, when proceeds a new context $c_{new}$, the classifier first selects $k$ contexts in the training set that are closest to $c_{new}$, then pick the best sense (senses) for $c_{new}$ (Jackson and Moulinier, 2002).

- TRAINING: Calculate $\vec{c}$ for each context $c$.
- TEST: Calculate
  Step1.

  $$A = \{\vec{c} \mid sim(\vec{c_{new}}, \vec{c})\, is\, maxim, \mid A \mid = k\}$$

  that means $A$ is the set of the $k$ nearest neighbors contexts of $\vec{c_{new}}$.
  Step2.

  $$Score(c_{new}, s_j) = \sum_{\vec{c_i} \in A} (sim(\vec{c_{new}}, \vec{c_i}) \times a_{ij})$$

  where $a_{ij}$ is 1 if $\vec{c_i}$ has the sense $s_j$ and $a_{ij}$ is 0 otherwise.
  Step3. Finally,

  $$s' = argmax_j Score(c_{new}, s_j).$$

We used the value of $k$ set to 3 after some experimental verifications.

A major problem with supervised approaches is the need for a large sense tagged training set. The bootstrapping methods use a small number of contexts labeled with senses having a high degree of confidence.

These labeled contexts are used as seeds to train an initial classifier. This is then used to extract a larger training set from the remaining untagged contexts. Repeating this process, the number of training contexts grows and the number of untagged contexts reduces. We will stop when the remaining unannotated corpus is empty or any new context can't be annotated. In (Tatar and Serban, 2001), (Serban and Tatar, 2003) we presented an algorithm which falls in this category. The algorithm is based on the two principles of Yarowsky (Resnik and Yarowsky, 1999):

- *One sense per discourse*: the sense of a target word is highly consistent within a given discourse (document);
- *One sense per collocation*: the contextual features ( nearby words) provide strong clues to the sense of a target word.

Also, for each iteration, the algorithm uses a NBC classifier. We intend to present a second system based on this algorithm at the next Senseval contest.

## 4 Implementation details

Our disambiguation system is written in JDK 1.4.

In order to improve the performance of the disambiguation algorithm, we made the following refinements in the above k-NN algorithm. First one is to substitute the lack of an efficient tool for stemming words in Romanian.

1. We defined a relation between words as $\delta : W \times W$, where $W$ is the set of words. If $w1 \in W$ and $w2 \in W$ are two words, we say that $(w1, w2) \in \delta$ if $w1$ and $w2$ have the same grammatical root. Therefore, if $w$ is a word and $C$ is a context, we say that $w$ **occurs** in $C$ iff exists a word $w2 \in C$ so that $(w, w2) \in \delta$. In other words, we replaced the stemming step with collecting all the words with the same root in a single class. This collection is made considering the rules for romanian morphology;
2. The step 3 of the algorithm for choosing the appropriate sense (senses) of a polysemic word $w$ in a given context $C$ (in fact the sense that maximizes the set $S = \{Score(C, s_j) \mid j = 1, \cdots Ns\}$ of scores for $C$) is divided in three sub-steps:
   - If there is a single sense $s$ that maximizes $S$, then $s$ is reported as the appropriate sense for $C$;
   - If there are two senses $s1$ and $s2$ that maximize $S$, then $s1$ and $s2$ are reported as the appropriate senses for $C$;
   - Consider that $Max1$ and $Max2$ are the first two maximum values from $S$ where $(Max1 > Max2)$. If $Max1$ is obtained for a sense $s1$ and if $Max2$ is obtained for a sense $s2$ and if

     $$Max1 - Max2 \leq P$$

where $P = \frac{Max1-Min}{(Ns-1)}$ and $Min$ is the minimum score from $S$, then $s1$ and $s2$ are reported as the appropriate senses for $C$.

Experimentally, we proved that the above improvements grow the precision of the disambiguation process.

## 5 Conclusions after the evaluation

Coarse-grained score for our system UBB using key "EVAL/RomanianLS.test.key" was:

precision: 0.722 (2555.00 correct of 3541.00 attempted)

recall: 0.722 (2555.00 correct of 3541.00 in total)

attempted: 100.00

Fine-grained score was:

precision: 0.671 (2376.50 correct of 3541.00 attempted)

recall: 0.671 (2376.50 correct of 3541.00 in total)

attempted: 100.00

Considering as baseline procedure the majority sense (all contexts are solved with the most frequent sense in the training corpus), for the word *nucleu* (noun) is obtained a precision of 0,78 while our procedure obtained 0,81. Also, for the word *desena* (verb) the baseline procedure of the majority sense obtains precision 0,81 while our procedure obtained 0,85.

At this stage our system has not as a goal to label with $U$ (unknown) a context, every time choosing one or two from the best scored senses. Annotating with the label $U$ is one of our coming improving. This can be done simply by adding as a new sense for each word the sense $U$. A simple experiment reported a number of right annotated contexts.

Another direction to improve our system is to exploit better the senses as they are done in training corpus: our system simply consider the first sense.

## References


I. Dagan, L. Lee and F. C. N. Pereira. 1994. Similarity-based Estimation of Word Cooccurences Probabilities. *Meeting of the Association for Computational Linguistics, 272–278.*

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database.* The MIT Press.

P. Jackson and I. Moulinier. 2002. *Natural Language Processing for Online Applications.* John Benjamin Publ. Company.

D. Jurafsky and J. Martin. 2000. *Speech and language processing.* Prentice-Hall, NJ.

C. Manning and H. Schutze. 1999. *Foundation of statistical natural language processing.* The MIT Press.

Ruslan Mitkov,editor. 2002 *The Oxford Handbook of Computational Linguistics.* Oxford University Press.

P. Resnik and D. Yarowsky. 1999. Distinguishing Systems and Distinguishing sense: new evaluation methods for WSD. *Natural Language Engineering, 5(2):113-134.*

G. Serban and D. Tatar. 2003. Word Sense Disambiguation for Untagged Corpus: Application to Romanian Language. *CICLing-2003, LNCS 2588, 270–275.*

D. Tatar and G. Serban. 2001. A new algorithm for WSD. *Studia Univ. "Babes-Bolyai", Informatica, 2 99–108.*

D. Widdows. 2003. A mathematical model for context and word meaning. *Fourth International Conference on Modeling and using context, Stanford, California, June 23-25.*