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

UBBNBC WSD System Description

Csomai ANDRÁS
Department of Computer Science
Babes-Bolyai University
Cluj-Napoca, Romania
csomaia@personal.ro

Abstract

The Naïve Bayes classification proves to be a good performing tool in word sense disambiguation, although it has not yet been applied to the Romanian language. The aim of this paper is to present our WSD system, based on the NBC algorithm, that performed quite well in Senseval 3.

1 Introduction

According to the literature, the NBC algorithm is very efficient, in many cases it outperforms more sophisticated methods (Pedersen 1998). Therefore, this is the approach we used in our research. The word sense disambiguating process has three major steps, therefore, the application has three main components as follows:

Stemming – removal of suffixes, and the filtering out of the irrelevant information from the corpora. A simple dictionary based approach.

Learning – the training of the classifier, based on the sense tagged corpora. A database containing the number of co-occurrences is built.

Disambiguating –on the basis of the database, the correct sense of a word in a given context is estimated.

In the followings the previously mentioned three steps are described in detail.

2 Stemming

The preprocessing of the corpora is one of the most result-influential steps. The preprocessing consists of the removal of suffixes and the elimination of the irrelevant data. The removal of suffixes is performed trough a simple *dictionary based* method. For every w_i word the possible w_j candidates are selected from the dictionary containing the word stems. Then a similarity score is calculated between the word to be stemmed and the candidates, as follows:
l_i, l_j is the length of word i, respectively j.

$$score_i = \frac{2l_i}{l_i + l_j} \text{ if } l_i \le l_j \quad \text{and}$$

$$score_j = 0, \text{ otherwise.}$$

The result is the candidate with the highest score if its score is above a certain threshold, otherwise the word is leaved untouched.
In the preprocessing phase we also erase the pronouns and prepositions from the examined context. This exclusion was made upon a list of stop words.

3 Learning

The training is conducted according to the NBC algorithm. First a database is built, with the following tables:

words – contains all the words found in the corpora. Its role is to assign a sense id to every word.

wordsenses – contains all the tagged words in the corpora linked with their possible senses. One entry for a given sense and word.

nosenses - number of tagged contexts, with a given sense

nocontexts - number of tagged contexts of a given word

occurrences – number of co-occurrences of a given word with a given sense

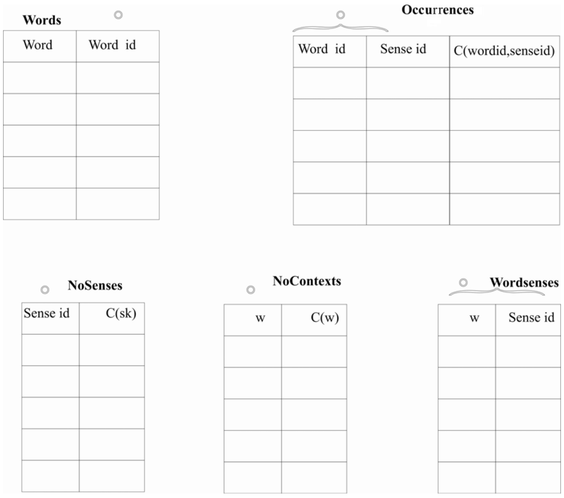

Figure1: The tables of the database

The training of the system is nothing but filling up the tables of the database.

```
fill NoSenses
fill NoContexts
fill Wordsenses
scan corpora
  $c_{akt}$=actual entry in corpora (a context)
  w=actual word in entry (the ambiguous word)
  $s_k$=actual sense of entry
    scan $c_{akt}$
      $v_j$=actual word in entry
      if $v_j$<>w then
        if $v_j$ in words then
          vi=wordid from words where w=$v_j$
        else
          add words $v_j$
        endif
      if (exists entry in occurrences where
                 wordid=vi and senseid=$s_k$) then
        increment C(wordid,senseid) in occurrences,
                 where wordid=vi and senseid=$s_k$
      else
        add occurrences(wordid, senseid, 1)
      endif
      step to next word
    endscan
step to next entry
endscan corpora
```

As it is obvious, the database is filled up (so the system is trained) only upon the training corpus provided for the Senseval3 Romanian Lexical Sample task.

4 Disambiguation

The basic *assumption* of the Naïve Bayes method is that the contextual features are *not dependent* on each other. In this particular case, we assume that the probability of co-occurrence of a word v_i with the ambiguous word *w* of sense *s* is not dependent on other co-occurrences.

The goal is to find the correct sense s', of the word *w*, for a given context. This s' sense maximizes the following equation.

$$\begin{aligned} s' &= \arg\max_{s_k} P(s_k \mid c) \\ &= \arg\max_{s_k} \frac{P(c \mid s_k)}{P(c)} P(s_k) \\ &= \arg\max_{s_k} P(c \mid s_k) P(s_k) \end{aligned}$$

At this point we make the simplifying "naïve" assumption:

$$P(c \mid s_k) = \prod_{v_j \in c} P(v_j \mid s_k)$$

The algorithm (Tătar, 2003) for estimating the correct sense of word *w* according to its *c* context is the following:

```
for every $s_k$ sense of w do
        score($s_k$)=$P(s_k)$
        for every $v_j$ from context c do
                score($s_k$)= score($s_k$)*$P(v_j \mid s_k)$
        s'= $\arg\max_{s_k}(score(s_k))$
```

where *s'* is the estimated sense, v_j is the *j*-th word of the context, s_k is the *k*-th possible sense for word *w*.

$P(s_k)$ and $P(v_j \mid s_k)$ are calculated as follows:

$$\mathrm{P}(\mathrm{s}_k) = \frac{\mathrm{C}(\mathrm{s}_k)}{C(w)}$$

$$\mathrm{P}(\mathrm{v}_j \mid \mathrm{s}_k) = \frac{\mathrm{C}(\mathrm{v}_j, \mathrm{s}_k)}{C(\mathrm{s}_k)}$$

where *C(w)* is the number of contexts for word *w*, $C(v_j, s_k)$ is the number of occurrences of word v_j in contexts tagged with sense s_k , and $C(s_k)$ is the number of contexts tagged with sense s_k

The values are obtained from the database, as follows:

$C(w)$- from **nocontexts**,

$C(v_j , s_k)$- from **occurrences**,

$C(s_k)$- from **nosenses**.

wordsenses is being used to determine the possible senses of a given word.

5 Evaluation

The described system was evaluated at Senseval 3. The output was not weighted, therefore for every ambiguous word, at most 1 solution (estimated sense) was provided. The results achieved, are the followings:

	score	correct/attempted
precision	0.710	2415 correct of 3403 attempted
recall	0.682	2415 correct of 3541 in total
attempted	96.10%	3403 attempted of 3541 in total

Figure2: Fine-grained score

	score	correct/attempted
precision	0.750	2551 correct of 3403 attempted
recall	0.720	2551 correct of 3541 in total
attempted	96.10%	3403 attempted of 3541 in total

Figure2: Coarse-grained score

A simple test was made, before the Senseval 3 evaluation. The system was trained on 90% of the Romanian Lexical Sample training corpus, and tested on the remaining 10%. The selection was random, with a uniform distribution. A coarse grained score was computed and compared to the baseline score. A baseline method consists of determining the most frequent sense for every word (based upon the training corpus) and in the evaluation phase always this sense is assigned.

	UBBNBC	Baseline
recall	0.66	0.56
precision	0.69	0.56

Figure3: baseline UBBNBC comparison

References

Ted Pedersen. 1998. *Naïve Bayes as a Satisficing Model.* Working Notes of the AAAI Spring Symposium on Satisficing Models, Palo Alto, CA

Doina Tătar. 2003. *Inteligenţă artificială - Aplicaţii în prelucrarea limbajului natural.* Editura Albastra, Cluj-Napoca, Romania.

Manning, C. D., Schütze, H. 1999. *Foundations of statistical natural language processing.* MIT Press, Cambridge, Massachusetts.