STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **XLVII**, Number 2, 2002

A WORD SENSE DISAMBIGUATION EXPERIMENT FOR ROMANIAN LANGUAGE

GABRIELA ŞERBAN AND DOINA TĂTAR

Abstract. The task of disambiguation is to determine which of the senses of an ambiguous word is invoked in a particular use of the word [5, 8]. It is known that the statistical methods produce high accuracy results for semantically tagged corpora [2]. Also, Word Net is a good source of information for WSD [3, 4]. Since for Romanian language does not exist neither a corpus nor something similar with WordNet, we make an experiment for WSD, using an algorithm for WSD [8], which requires only information that can be extracted from untagged corpus. This algorithm learns to make predictions based on local context with only a few labeled contexts and many unlabeled ones.

Keywords: Word sense disambiguation, corpus.

1. Introduction

In [9], Yarowsky observed that there are constraints between different occurrences of contextual features that can be used for disambiguation. Two such constraints are *one sense per discourse* and *one sense per collocation.* These mean that the sense of a target word is highly consistent within a given discourse (document) and the contextual features (nearby words) provide strong clues to the sense of a target word.

Notational conventions used in the following are: w is the word to be disambigued (*target word*), $s_1, \cdots, s_K$ are possible senses for w, $c_1, \cdots, c_I$ are contexts of w in a corpus, $v_1, \cdots, v_J$ are words used as contextual features for disambiguation of w. The contextual features $v_1, \cdots, v_J$ occur in a fixed position near w, in a *window* of fixed length, centered or not on w ("unrestricted collocations", in [6]).

Received by the editors: November, 20, 2002.

2000 *Mathematics Subject Classification.* 68T50, 68Q32.

1998 *CR Categories and Descriptors.* I.2.7. [**Computing Methodologies**]: Artificial Intelligence – *Natural Learning Processing*; G.3. [**Mathematics of Computing**]: Statistical Computing.

A Naive Bayes Classifier (NBC) realizes the calculus of the sense s', which for the target word w and a given context c satisfies the relation [5]: $s' = argmax_{s_k} P(s_k \mid c) = argmax_{s_k} \frac{P(c|s_k)}{P(c)} P(s_k) = argmax_{s_k} P(c \mid s_k) P(s_k)$. The Naive Bayes assumption is that the contextual features are all conditional independent. This is not generally true, but there is a large number of cases in which the algorithm works well. Concerning the probabilities $P(v_j \mid s_k)$ and $P(s_k)$, these are calculated from a labeled (annotated) corpus. In our algorithm the probabilities $P(v_j \mid s_k)$ are re-estimated until all the contexts are solved.

2. A Bootstrapping Algorithm (BA) for WSD

The BA algorithm begins by identifying a small number of training contexts. This could be accomplished by hand tagging with senses the contexts of w for which the sense of w is clear because some *seed collocations* [9, 10] occur in these contexts (for a detailed description of the BA algorithm see [8]).

The notational conventions are as above: $C = \{c_1, c_2, \cdots c_I\}$ are contexts (windows) of w, as obtained with query w and with an on-line corpus tool (at us *htdig* and a Romanian corpus). Each c_i has the form: $c_i = w_1, \cdots, w_t, w, w_{t+1}, \cdots, w_z$ where $w_1, w_2, \cdots, w_t, w_{t+1}, \cdots, w_z$ are words from the set $\{v_1, \cdots, v_J\}$ and t and z are selected by user.

Let us consider that the words $V = \{v^1, \cdots, v^l\} \subset \{v_1, \cdots, v_J\}$, where l is small (for example 2) are *surely* associated with senses for w, such that the occurrence of v^i in the context of w determines the choice of a sense s^i for w (one sense per collocation). Here $\{s^1, \cdots, s^l\}$ is a subset of $\{s_1, \cdots, s_K\}$.

These rules can be done generally as a decision list:

(1) **if** $\mathbf{v^i}$ *occurs in a context* **c** *of* **w** **then** *the sense of* **c** *is* $\mathbf{s^i}$, $s^i \in S$

So, from the set of contexts obtained as query results, some contexts can be solved.

For our algorithm, we define a relation $\delta \subset W \times P(W)$, where W is the set of all words and $P(W)$ is the power set of W . If $w \in W$ is a word and $c \in P(W)$ we say that $(w, c) \in \delta$ if $w \in c$ or, else, if exists a word $w1 \in c$ so that the words w and $w1$ have the same gramatical root (particularly c is a context).

So, a corresponding decision list has the following form:

(2) **if** $(v, c) \in \delta$ *and* **v** *has the sense* $\mathbf{s_i}$ **then** *the sense of the context* **c** *is* $\mathbf{s_i}$

3. The Application for Words Disambiguation

The application is written in Visual C++ 6.0 and its goal is to find the correct sense for a given word (the target word) in some given contexts using the algorithm described in section 2.

3.1. **Experiment.** Our aim is to use the BA algorithm for the romanian language, to disambiguate the word *poarta* in some contexts obtained with an on-line corpus tool (at us *htdig* and a Romanian corpus).

We make the following specifications:

- the target word *poarta* has, in romanian language, four possible senses (two nouns and two verbs);
- we experiment our algorithm starting with 38 contexts for the target word;
- we start with 6 words as contextual features for the disambiguation.

The input text file for our experiment is the following:

- the target word

poarta

- the senses of the target word

casa fotbal haine raspundere

- the words used as contextual features for the disambiguation and the indexes of the corresponding sense of the target word

lemn 1 casa 1 minge 2 blugi 3 raspundere 4 semnatura 4

- the contexts of the target word

- Respectivul Popa Nicolae Ioan a prezentat jandarmului de la **poarta** un buletin de identitate cu seria B.C., nr. 718609, aceasta in timp ce adevaratul Ioan Popa
- De cand s- a instalat in scaun ultimul primar, frenezia imperecherilor politice este de nestavilit. Se **poarta** negocieri secrete sau fatise, se nasc scenarii avortate dupa nici 24 de ore, se lanseaza nume alaturate te miri carei constructii politice
- hotul, natang in ce priveste alegerea modalitatii de a sustrage date de stricta confidentialitate, dar abil in a scoate pe **poarta** unei institutii, aflate in regim de paza militarizata, ditamai calculatorul

- avand rezolutia catre dl. consilier de stat Mihai Surcel, o dovedeste o alta adresa anexata la dosar, care este datata 15 aprilie 1999, **poarta** (cum se vede si in facsimilul alaturat) antetul Guvernului Romaniei, cabinetul primului-ministru, **poarta** semnatura sefei de cabinet Camelia Andrusenco si este destinata secretarului de stat Liviu Ionescu, din Ministerul de Interne
- Luptatorii SIAS s-au oprit din actiune la **poarta** unei ferme unde s-a refugiat infractorul, pe motiv ca nu aveau mandat de perchezitie
- ...

The accuracy of the BA algorithm in the proposed experiment is **60%**. We note that the *accuracy* of the disambiguation algorithm is calculated with the following formula

$$A = \frac{number\, of\, correctly\, solved\, contexts}{number\, of\, contexts} \tag{3}$$

The experiment at Hearst (1991) shows that to achieve a high precision in word sense tagging, the initial set must be large (20–30 occurrences for each sense).

We have to mention that, in our experiment, we associated a single occurrence for each sense. On the other hand, we observe that if the number of words used as contextual features for the disambiguation and the number of contexts grow, the accuracy of the BA algorithm grows, too.

3.2. **Experimental Comparison with the NBC Algorithm.** In the case of the algorithm described in section 2 (BA–Bootstrapping Algorithm), the relation δ described in Equation 2 is very important. In order to illustrate the efficiency of the BA algorithm (with an without δ), we ran at the same time the NBC algorithm for the experiment proposed in subsection 3.1. We note that "BA without relation" is the BA algorithm (Section 2), in which a decision list has the form described in Equation 1.

The comparative experimental results obtained are shown in Figure 1. In Figure 1, we give, for each algorithm, a graphical representation of accuracy/context. More exactly, for a given algorithm, for the i-th context we represent the accuracy (see Equation 3) of the algorithm for the first i contexts. From Figure 1, it is obvious that the most efficient is the BA algorithm with the relation δ (at each step, the BA algorithm's accuracy is maximum).

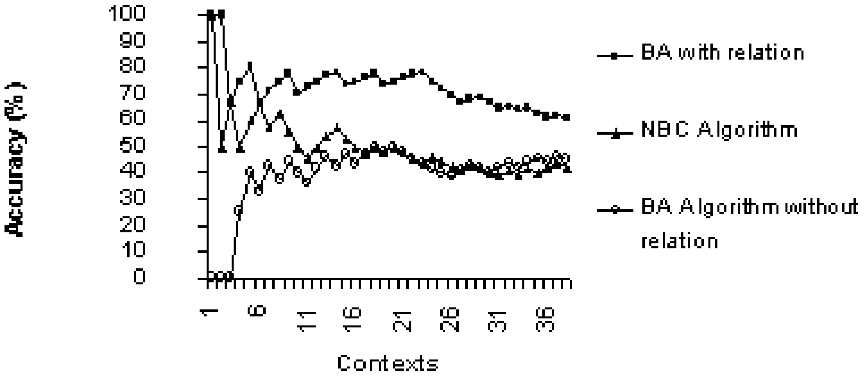

Figure 1. The comparative experimental results

4. Further Work

Further work is planned to be done in the following directions: for assuring a better efficiency of the disambiguation, we plain to retain in a database the results of the learning process. We plain to study our approach in the context of combining labeled and unlabeled data with Co-Training as in [1]. Our own goal is to solve with our method the disambiguation for a query in a future QA-system in Romanian which is now in construction.

References

[1] A. Blum, T. Mitchell: Combining Labeled and Unlabeled Data with C-Training. Proceedings of the 11th Annual Conference on Computational Learning Theory (1998) 92–100

[2] G. Escudero, L. Marquez, G. Rigau: Boosting applied to WSD. ACML,Barcelona, Spain (2000)

[3] R. Mihalcea, D. Moldovan: An iterative Approach to WSD. Proceedings of FLAIRS (2000)

[4] R. Mihalcea, D. Moldovan: A method for WSD of unrestricted text. Proceedings of the 37th Annual Meeting of the ACL, Maryland, NY (1999)

[5] C. Manning, H. Schutze: Foundation of statistical natural language processing. MIT (1999)

[6] T. Pedersen, R. Bruce: Knowledge Lean WSD. Proceedings of the Fifteenth National Conference on AI. Madison, WI (1998)

[7] P. Resnik, D. Yarowsky: Distinguishing Systems and Distinguishing sense: new evaluation methods fot WSD. Natural Language Engineering, **1** (1998)

[8] D. Tatar, G. Serban: A new algorithm for WSD. Studia Univ. Babes-Bolyai, Informatica. **2** (2001) 99–108

[9] D. Yarowsky: Hierarchical Decision Lists for WSD. Kluwer Academic Publishers (1999)

[10] David Yarowsky: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceedings of ACL'95 189–196

Babeş-Bolyai University, Cluj-Napoca, Romania
E-mail address: `gabis@cs.ubbcluj.ro, dtatar@cs.ubbcluj.ro`