# The University of Alicante Word Sense Disambiguation System*

**Andrés Montoyo** and **Armando Suárez**
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
{montoyo | armando}@dlsi.ua.es

## Abstract

The WSD system presented at SENSEVAL-2 uses a knowledge-based method for noun disambiguation and a corpus-based method for verbs and adjectives. The methods are, respectively, Specification Marks and Maximum Entropy probability models. So, we can say that this is a hybrid system which joins an unsupervised method with a supervised method. The whole system has been used in lexical sample english task and lexical sample spanish task.

## 1 Introduction

In this paper a Word Sense Disambiguation system based on Specification Marks (SM) and Maximum Entropy probability models (ME) is presented. SM is an unsupervised knowledge-based method and has been applied to noun disambiguation. ME belongs to the statistical approach to WSD in NLP and uses a tagged corpus in order to learn a probability model that can be used to predict the correct sense of a word. SM does not need a previously tagged corpus, it uses the semantic information stored in WordNet.

The weakness of supervised corpus-based approaches rely on availability of corpora and their dependency of the data which were used in the training phase. Knowledge-based approaches use previously acquire linguistic knowledge. This knowledge is extracted from human lexicographers experience and can be in form of electronic dictionary or lexicon. While their success seems poorest than statistical methods, they don't need neither an existing corpus nor a training phase and they can be more domain independent.

So, the University of Alicante system performs the WSD task combining unsupervised with supervised methods. The whole system has been used in lexical sample English task and lexical sample Spanish task.

## 2 Specification Marks Framework

The method we present here consists basically of the automatic sense-disambiguating of nouns that appear within the context of a sentence and whose different possible senses are quite related. Its context is the group of words that co-occur with it in the sentence and their relationship to the noun to be disambiguated. The disambiguation is resolved with the use of the WordNet lexical knowledge base.

The intuition underlying this approach is that the more similar two words are, the more informative the most specific concept that subsumes them both will be. In other words, their lowest upper bound in the taxonomy. (A "concept" here, corresponds to a Specification Mark (SM)). In other words, the more information two concepts share in common, the more similar they obviously are, and the information commonly shared by two concepts is indicated by the concept that subsumes them in the taxonomy.

The input for the WSD module will be the group of words $W = \{W_1, W_2, ..., W_n\}$. Each word wi is sought in WordNet, each one has an associated set $S_i = \{S_{i1}, S_{i2}, ..., S_{in}\}$ of possible senses. Furthermore, each sense has a set of concepts in the IS-A taxonomy (hypernymy/Hyponymy relations). First, the concept that is common to all the senses of all the words that form the context is sought. We call this concept the Initial Specification Mark (ISM), and if it does not immediately resolve the ambiguity of the word, we descend from one level

* This paper has been partially supported by the Spanish Government (CICYT) project number TIC2000-0664-C02-02.

to another through WordNet´s hierarchy, assigning new Specification Marks. The number of concepts that contain the subhierarchy will then be counted for each Specification Mark. The sense that corresponds to the Specification Mark with highest number of words will then be chosen as the sense disambiguation of the noun in question, within its given context.

At this point, we should like to point out that after having evaluated the method, we subsequently discovered that it could be improved with a set of heuristics, providing even better results in disambiguation. The set of heuristics are Heuristic of Hypernym, Heuristic of Definition, Heuristic of Common Specification Mark, Heuristic of Gloss Hypernym, Heuristic of Hyponym and Heuristic of Gloss Hyponym. Detailed explanation and evaluation of the method and heuristics can be found in (Montoyo and Palomar, 2000; Montoyo and Palomar, 2001), while its application to NLP tasks are addressed in (Montoyo et al., 2001).

## 3 Maximum Entropy Framework

Maximum Entropy(ME) modeling is a framework for integrating information from many heterogeneous information sources for classification. ME probability models were successfully applied to some NLP tasks such as POS tagging or sentence boundary detection (Ratnaparkhi, 1998).

The WSD system presented in this paper is based on conditional ME probability models (Saiz-Noeda et al., 2001). It implements a supervised learning method consisting of the building of word sense classifiers through training on a semantically tagged corpus. A classifier obtained by means of a ME technique consists of a set of parameters or coefficients estimated by means of an optimization procedure. Each coefficient is associated to one feature observed in training data. A feature is a function that gives a measure for some characteristic in a context associated to a class. The main purpose is to obtain the probability distribution that maximizes the entropy, that is, maximum ignorance is assumed and nothing apart of training data is considered. As advantages of ME framework, knowledge-poor features applying and accuracy can be mentioned; ME framework allows a virtually unrestricted ability to represent problem-specific knowledge in the form of features (Ratnaparkhi, 1998).

Let us assume a set of contexts X and a set of classes C. The function $cl : X \rightarrow C$ that performs the classification in a conditional probability model $p$ chooses the class with the highest conditional probability: $cl(x) = \arg\max_c p(c|x)$, where $x$ is a context and $c$ a class. The features have the form of (1), where $cp(x)$ is some observable characteristic[1]. The conditional probability $p(c|x)$ is defined as (2) where $\alpha_i$ are the parameters or weights of each feature, and $Z(x)$ is a constant to ensure that the sum of probabilities for each possible class in this context is equal to 1.

$$f_{c'}(x,c) = \begin{cases} 1 & \text{if } c' = c \text{ and } cp(x) = true \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$p(c|x) = \frac{1}{Z(x)} \prod_{i=1}^{K} \alpha_i^{f_i(x,c)} \tag{2}$$

## 4 The system at Senseval-2

The Spanish and English lexical sample tasks at the SENSEVAL-2 workshop had been performed by our system in three phases. The first one is a naive multi-word detection; the second one, the disambiguation of nouns by means of the SM method, and the third one, the disambiguation of verbs and adjectives by means of the ME method.

In a previous step, training and test data had been tagged with Tree-Tagger(Schmid, 1994) for English files and Conexor's FDG Parser (Tapanainen and Järvinen, ) for Spanish files in order to get the part-of-speech information and identify nouns, verbs and adjectives.

### Multi-words detection

The multi-word detection has been performed by combining the words around the target word in each sample and consulting WordNet for English (examining the training data, we conclude that this is not necessary for Spanish data). If a multi-word is found in WordNet a multi-word instance is assigned and no further single word

[1] The ME approach is not limited to binary funtions, but the optimization procedure(*Generalized Iterative Scaling*) used for the estimation of the parameters needs this kind of features.

disambiguation will be done. This kind of instances has been disambiguated with the first sense of WordNet (even if it is a polysemous one).

### Nouns with Specification Marks

The second phase consist of noun classification, and has been performed by the SM method described previously.

### Verbs and adjectives with Maximum Entropy

The third and final phase, the verbs and adjectives disambiguation, has been performed by the ME method. The SENSEVAL-2 training data has been used in order to obtain the classification functions to be applied on the test data. The set of features defined for ME training is described below and it is based on features selection made in (Ng and Lee, 1996) and (Escudero et al., 2000).

The set of features corresponds to words around the word to classify and POS labels at positions related to the target word in each sentence: $w_0$, $w_{-1}$, $w_{-2}$, $w_{-3}$, $w_{+1}$, $w_{+2}$, $w_{+3}$, $(w_{-2},w_{-1})$, $(w_{-1},w_{+1})$, $(w_{+1},w_{+2})$, $(w_{-3},w_{-2},w_{-1})$, $(w_{-2},w_{-1},w_{+1})$, $(w_{-1},w_{+1},w_{+2})$, $(w_{+1},w_{+2},w_{+3})$, $p_{-3}$, $p_{-2}$, $p_{-1}$, $p_{+1}$, $p_{+2}$, $p_{+3}$. Each $w_i$ is the lemma of the word at position $i$ in the context (in collocations, at least one of the words must be a content word). Each $p_i$ is the POS label at position $i$.

Other set of features consists of a surrounding nouns selection. This selection is doing by means of frequency information of nouns co-occurring with a sense. Nouns co-occurring with a class in a $K\%$ of examples of that class in the corpus or more are selected to build a feature for each possible class[2].

## 5 Senseval-2 results analysis

Analyzing the first evaluation results of the English lexical sample task (fine-grained scoring) reported by SENSEVAL-2 committee (*precision* = 0.421 and *recall* = 0.411) , some conclusions can be extracted from them.

The nouns disambiguation obtains the worst results (see table 1). We can mostly assure that the reason is the kind of method used: knowledge-based for nouns and corpus-based for verbs and adjectives.

| POS | precision | recall |
|---|---|---|
| Nouns | 0.299 | 0.292 |
| Verbs | 0.486 | 0.480 |
| Adjectives | 0.709 | 0.635 |

Table 1: Results of the English Lexical Sample Task (Fine-grained)

The results of the Spanish lexical sample task (fine-grained scoring) reported by SENSEVAL-2 committee are *precision* = 0.514 and *recall* = 0.503. Nevertheless, the nouns results rise to 56% of precision (table 2). It seems that the set of nouns selected for this task is easier to Specification Marks than English ones, maybe related to lexical resources used and the language itself. However, the recall of nouns is too low because a implementation error causes that the accented words had not been recognize (*corazón*, *operación* and *órgano*).

| POS | precision | recall |
|---|---|---|
| Nouns | 0.566 | 0.435 |
| Verbs | 0.511 | 0.511 |
| Adjectives | 0.687 | 0.687 |

Table 2: Results of the Spanish Lexical Sample Task (Fine-grained)

The preprocessing of the train and test data are relevant. Some errors of the POS-tagger had been detected and they affect some answer instances. Multi-words are a not resolved problem. The detection and disambiguation method is too simple and causes too much errors. More preprocessing is necessary, as well: the context information can be enriched and accuracy increased with entity recognition, full-parsing, and so on.

## 6 Conclusions

The University of Alicante system presented at SENSEVAL-2 workshop joins the two general approaches to the WSD task: knowledge-based and corpus-based methods. The Specification Marks method belongs to the first one and Maximum Entropy-based method to the second one.

[2] For example, in a set of 100 examples of sense four of the noun "interest", if "bank" is observed 10 times or more ($K$ = 10%) then a feature for each possible sense of "interest" is defined with "bank".

Specification Marks for nouns, and Maximum Entropy for verbs and adjectives had been used in order to process the test data of the English and the Spanish lexical sample tasks. The training and the test data had been used with a minimum preprocessing, just cleaning of XML-tags in order to run the Tree-Tagger. Besides, the two WSD modules had been used in the same manner as for other corpora with minor modifications: no specific changes to the algorithms used in both methods had been made for SENSEVAL-2, apart from the necessary modules to make data files available to the computer programs.

Due to the distinct approaches used in each POS, the whole system has been classified as supervised system. In the English task, the system obtains a poor score when it is compared with other supervised systems, and a great result against the unsupervised systems (we have no such information of systems for Spanish). But the truth is that our system is unsupervised for nouns but supervised for verbs and adjectives. Therefore, comparing our results with the other systems must be done separating the results of nouns, verbs and adjectives.

## 7 Future and in progress work

At this moment, the two methods presented here are being improved with new knowledge sources like full parsing information and domain categories that in order to decrease the WordNet granularity. The WSD system will be completed with other NLP software like Name Entity recognition and multi-words detection modules.

Recent work in our research group indicates that it is possible to combine the two methods in a hybrid method that assign a sense to a context combining the answers of both methods with a relevant improvement of accuracy (Suárez and Montoyo, 2001). Our intention is to extent this combination with the help of other well known WSD methods and to establish a voting method or some other manner of cooperation.

Our main objective is to develop a complete WSD system in order to help other NLP activities in our research group. The work presented here is our first attempt to participate at Senseval and we hope to get the proper conclusions in order to improve our system and compete in the next Senseval.

## References

Gerard Escudero, Lluis Màrquez, and German Rigau. 2000. Boosting applied to word sense disambiguation. In *Proceedings of the 12th Conference on Machine Learning ECML2000*, Barcelona, Spain.

A. Montoyo and M. Palomar. 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. pages 103–107.

A. Montoyo and M. Palomar. 2001. Specification Marks for Word Sense Disambiguation: New Development. In *Proceedings of 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)*, pages 182–191.

A. Montoyo, M. Palomar, and G. Rigau. 2001. WordNet Enrichment with Classification Systems. In ACL, editor, *Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, PA, USA.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In *Proceedings 34th Annual Meeting of the ACL-1996.*, San Francisco, USA.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Ph.D. thesis, University of Pennsylvania.

Maximiliano Saiz-Noeda, Armando Suárez, and Manuel Palomar. 2001. Semantic pattern learning through maximum entropy-based wsd technique. In *Proceedings of CoNLL-2001*, pages 23–29. Toulouse, France.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings International Conference on New Methods in Language Processing.*, pages 44–49, Manchester, UK.

Armando Suárez and Andrés Montoyo. 2001. Estudio de cooperación entre métodos de desambiguación léxica: Marcas de especificidad vs. máxima entropía. *Procesamiento Lenguaje Natural*, 27(1):207–214, september.

Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 64–71.