Combining Supervised-Unsupervised Methods for Word Sense Disambiguation*

Andrés Montoyo, Armando Suárez and Manuel Palomar

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
{montoyo, armando, mpalomar}@dlsi.ua.es

Abstract. This paper presents a method to combine two unsupervised methods (Specification Marks, Conceptual Density) and one supervised (Maximum Entropy) for the automatic resolution of lexical ambiguity of nouns in English texts. The main objective is to improved the accuracy of knowledge-based methods with statistical information supplied by the corpus-based method. We explore a way of combining the classification results of the three methods: "voting" is the way we have chosen to combine the three methods in one unique decision.
These three methods have been applied both individually as in a combined way to disambiguate a set of polysemous words. Our results show that a combination of different knowledge-based methods and the addition of statistical information from a corpus-based method might eventually lead to improve accuracy of first ones.

1 Introduction

In this paper we concentrate on the resolution of the lexical ambiguity that arises when a given word has several different meanings. This specific task is commonly referred to as Word Sense Disambiguation (WSD). In general terms, WSD involves assigning a definition to a given word, in either a text or a discourse, that endows it with a meaning that distinguishes it from all of the other possible meanings that the word might have in other contexts.

Currently, two main tendencies can be found in this research area: *knowledge-based* methods and *corpus-based* methods.

The first group of methods rely on previously acquired linguistic knowledge, and work disambiguating of words by matching the context in which they appear with information from an external knowledge source. To accomplish this task, the two knowledge-based methods (Specification Marks Method [6, 8] and Conceptual Density [1, 2]) used in this paper, chose WordNet as it combines the features of both dictionaries and thesauruses, and also includes other links among words by means of several semantic relations, (Hyponymy, hypernymy,

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

meronymy, etc). In other words, WordNet provides definitions for the different senses that a given word might have (as a dictionary does) and defines groups of synonymous words by means of "Synsets", which represent distinct lexical concepts, and organises them into a conceptual hierarchy (as a thesaurus does).

The second one use techniques from statistics and machine learning to induce models of language usage from large samples of text [11]. These last methods can perform supervised or unsupervised learning, that is, the corpus is previously tagged with correct answers or not.

Usually, supervised learning methods represents linguistic information in the form of features. Each feature informs of the occurrence of certain attribute in a context that contains a linguistic ambiguity. That context is the text surrounding this ambiguitiy and relevant to the disambiguation process. The features used can be of distinct nature: word collocations, part-of-speech labels, keywords, topics and domain information, etc.

A WSD method using supervised learning tries to classify a context containing an ambiguous word or compound word in one of its possible senses by means of a classification function. This function is obtained after a training process on a sense tagged corpus. The information source for this training is the set of results of the features evaluation on each context, that is, each context has its vector of feature values. The supervised learning WSD method (Maximum Entropy) used in this paper to do such analysis is based on Maximum Entropy probability models (ME) [13].

This paper is organized as follows. After this short introduction, section 2 shows the methods we have applied. Section 3 describes the test sets and shows the results. With this results, the contribution of each method to the disambiguation process is analyzed. Finally, some conclusions and future and in progress work will be presented.

2 Methods WSD for Combining

2.1 Specification Marks Framework

The method we present here [7, 6] 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 $W = \{W_1, ..., W_n\}$. Each word wi is sought in WordNet, each one has an associated set $S_i = \{S_{i1}, ..., S_{in}\}$ 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 to another through WordNets 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 of the method and heuristics can be found in [8], while its application to NLP tasks are addressed in [9, 10].

2.2 Conceptual Density Framework

Conceptual distance tries to provide a basis for determining closeness in meaning among words, taking as reference a structured hierarchical net. The measure of conceptual distance among concepts we are looking for should be sensitive to:

- the length of the shortest path that connects the concepts involved.
- the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.
- the measure should be independent of the number of concepts we are measuring.

We are working with the Agirre-Rigau Conceptual Density formula [2] shown in the formula 1, which compares areas of subhierarchies.

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c} \quad (1)$$

The numerator expresses the expected area for a subhierarchy containing m senses of the words to be disambiguated, while the divisor is the actual area, and is given by the formula 2:

$$descendants_c = \sum_{i=0}^{h-1} nhyp^i \quad (2)$$

2.3 Maximum Entropy Framework

Maximum Entropy(ME) modeling is a framework for integrating information from many heterogeneous information sources for classification [4]. ME probability models were successfully applied to some NLP tasks such as part-of-speech(POS) tagging or sentence boundary detection [12].

The WSD method used in this paper is based on conditional ME probability models [13]. It implements a supervised learning method consisting of building 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 the training data. 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 the ME framework, knowledge-poor features applying and accuracy can be mentioned; The ME framework allows a virtually unrestricted ability to represent problem-specific knowledge in the form of features [12].

Let us assume a set of contexts X and a set of classes C. The function $cl : X \rightarrow C$ chooses the class c with the highest conditional probability in the context x: $cl(x) = \arg\max_c p(c|x)$. Each feature is calculated by a function that is associated to a specific class c' and it has the form (3), where $cp(x)$ is some observable characteristic in the context [1]. The conditional probability $p(c|x)$ is defined as (4) where α_i is the parameter or weights of the feature i, K the number of features defined, and $Z(x)$ a constant to ensure that the sum of all conditional probabilities for this context is equal to 1.

$$f(x,c) = \begin{cases} 1 \text{ if } c' = c \text{ and } cp(x) = true \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

$$p(c|x) = \frac{1}{Z(x)} \prod_{i=1}^{K} \alpha_i^{f_i(x,c)} \tag{4}$$

3 Experiments and Results

it is to prove the effectiveness of the three applied methods in an individual way and in a combined way.

The main objective of these experiments is to check the effectiveness of the three methods, applied in an individual or combined way, on oneself group of examples. The individual evaluation to each method has been conducted on the SemCor collection [5], a set of 171 documents where all content words are annotated with the most appropiate WordNet sense. However, the evaluation in a combined way has been carried out on 18 documents of SemCor. In order to

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

evaluate each previously described method and their combination, we selected a set of nouns at random: account, age, art, car, child, church, cost, duty, head, interest, line, member, people, term, test, and work.

3.1 Experiments on Specification Marks

In this experiment, all the sentences were selected when some of the previously selected nouns appeared in the whole corpus Semcor. For each one of these sentences the nouns were obtained, forming the context of the word to be disambiguated. This context is introduced to the method of WSD, and it returns the sense corresponding of WordNet automatically for each one of the nouns. An important advantage of the method we present here consists basically of the automatic sense-disambiguating of nouns that appear within the context of a sentence. Therefore, it does not require any sort of training process, no hand-coding of lexical entries, nor the hand-tagging of texts. However, an inconvenience found in the experiments carried out with the Semcor is that the method relies on the semantics relations (Hypernymy/Hyponymy) and the hierarchical organization of WordNet used for disambiguate the sense of the words. For this reason, when the method of ME is applied on the selected nouns, there are words that have a percentage of desambiguacin so low. As it is shown in the table 1, i.e., the word "test" obtains a low percentage of disambiguation, because the other nouns of the context are not related semantically by WordNet.

nombre	#	P	R	A
account	21	**0,048**	0,048	1,000
age	86	**0,523**	0,523	1,000
art	64	**0,333**	0,328	0,984
car	65	**0,734**	0,723	0,985
child	180	**0,622**	0,594	0,956
church	107	**0,539**	0,514	0,953
cost	76	**0,289**	0,289	1,000
duty	23	**0,348**	0,348	1,000
head	168	**0,204**	0,190	0,935
interest	126	**0,444**	0,444	1,000
line	118	**0,209**	0,203	0,975
member	68	**0,515**	0,515	1,000
people	244	**0,531**	0,520	0,980
term	45	**0,156**	0,156	1,000
test	34	**0,088**	0,088	1,000
work	190	**0,255**	0,253	0,989
TOTAL	1615	**0,404**	0,395	0,978

Table 1. Results of Specification Marks Method in SemCor

3.2 Experiments on Conceptual Density

In this experiment, all the sentences were selected when some of the previously selected nouns appeared in the whole corpus Semcor. For each one of these sentences the nouns were obtained, forming the context of the word to disambiguate. This context is introduced to the Conceptual Density Method, and it computes the Conceptual Density of each concept in WordNet according to the senses it contains in its subhierarchy. It selects the concept with highest Conceptual Density and selects the senses below it as the correct senses for the respective words. Besides completely disambiguating a word or failing to do so, in some cases the disambiguation algorithm returns several possible senses for a word. In this experiment we considered these partial outcomes as failure to disambiguate. In the table 2 is shown the results of each words.

nombre	P	R	A
account	**0,000**	0,000	1,000
age	**0,333**	0,333	1,000
art	**0,121**	0,088	0,733
car	**1,000**	1,000	1,000
child	**0,352**	0,352	1,000
church	**0,500**	0,464	0,928
cost	**1,000**	1,000	1,000
duty	**0,500**	0,500	1,000
head	**0,000**	0,000	1,000
interest	**0,277**	0,263	0,947
line	**0,000**	0,000	1,000
member	**0,166**	0,166	1,000
people	**0,454**	0,396	0,873
term	**0,250**	0,250	1,000
test	**0,333**	0,333	1,000
work	**1,000**	0,500	0,500
TOTAL	**0,393**	0,353	0,936

Table 2. Results of Conceptual Density in SemCor

3.3 Experiments on Maximum Entropy

Some evaluation results over a few terms of the aforementioned corpus are presented in Table 3. The system was trained with features that inform of content words in the sentence context (w_{-1}, w_{-2}, w_{-3}, w_{+1}, w_{+2}, w_{+3}), collocations ((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})), and POS tags (p_{-1}, p_{-2}, p_{-3}, p_{+1}, p_{+2}, p_{+3}).

For each word, the training set is divided in 10 folds, 9 for training and 1 for evaluation; ten tests were accomplished using a different fold for evaluation in each one (10-fold cross-validation). The accuracy results are the average accuracy on the ten tests for a word.

Table 3. Results of Maximum Entropy Method in SemCor

noun	#	P	R	A
account	2,7	**0,285**	0,263	0,872
age	10,3	**0,313**	0,143	0,438
art	7,3	**0,596**	0,575	0,966
car	6,9	**0,959**	0,959	1,000
child	19,1	**0,957**	0,169	0,189
church	12,7	**0,558**	0,543	0,967
cost	8,4	**0,883**	0,851	0,962
duty	2,5	**0,778**	0,685	0,870
head	16,6	**0,600**	0,582	0,961
interest	13,7	**0,485**	0,454	0,932
line	12,2	**0,070**	0,067	0,946
member	7,3	**0,874**	0,874	1,000
people	27,1	**0,626**	0,359	0,530
term	5,2	**0,445**	0,430	0,951
test	3,6	**0,258**	0,252	0,938
work	20,3	**0,405**	0,392	0,962
TOTAL		**0,586**	0,473	0,805

Some low results can be explained by the corpus itself. There has not been made any selection of articles and fiction and non-fiction ones had been processed. Moreover, the number of examples of the selected words is very low too.

3.4 Experiments on Voting

Two experiments had been done: *voting* and *"quality" voting*. The first one consists on considering only those contexts where at least two methods classify it as the same sense. The second one consists on assigning a "quality" vote to ME method, that is, if none of the method agrees with other, then the response of ME is the sense in which the context is classified.

In order to obtain the results shown in table 4, 18 articles of Semcor had been selected. All methods work on this set classifying the selected words previously mentioned. Context by context, classification results of every context are compared and they take a vote on each context to decide its sense.

Table 4. Results comparison

method	precision	recall	attempted
SM	0.361	0.330	0.914
CD	0.358	0.327	0.891
ME	0.638	0.614	0.963
Voting	0.514	0.345	0.670
QVoting	0.517	0.517	1.000

4 Discussion

The main objective of this work is to enforce the knowledge methods and raise their accuracy but maintaining their virtues: no corpus dependence. In order to get this, two strategies had been defined: adding more knowledge-based methods and adding statistical information too.

Voting is the kind of cooperation chosen and, for those contexts which doesn't reach the enough number of votes, statistical information of a corpus-based method is supplied to resolve the ambiguity, finally.

ME, the corpus-based method, obtains better results than the knowledge-based methods, SM and CD, when applied on the evaluation set of articles, but, we have no security on what happens when the domain changes [3].

Due to this, we consider a good result the gain in precision obtained when voting is applied. The low recall of pure voting is resolve when the ME method uses its quality vote. It is usual, in circumstances like that described here, to assign the most frequent sense (in the corpus or sense one in WordNet) but this is a statistical information too. Therefore, more elaborated statistical information has been preferred; moreover, ME also applies MFS rule when the context has no enough information.

These results are indicative of a promising approach: the combination of several WSD methods in order to improve accuracy. More complex cooperation formulas can be explored too.

5 Conclusions

A study of cooperation between different WSD methods has been shown. Two knowledge-based methods, Specification Marks and Conceptual Density, and a corpus-based method, Maximum Entropy probability models, had been used in a voting strategy of sense classification.

Two voting methods had been performed. The first one only considers those context in which at least two methods agree in the classification sense. The second one, for those contexts in which there is not a minimum agreement, the ME method decides which is the sense of them.

The analysis of the results presented in this paper shows that the knowledge-based methods can obtain a considerably gain in accuracy when used jointly

and combining statistical information of a corpus-based method. Approximately a 15% of precision gain is achieved in both voting methods and the number classified contexts rise to 100% when corpus-based method uses its quality vote.

As future and in progress work, more WSD methods are being studied and more complex cooperation strategies are being developed.

References

1. Eneko Agirre and German Rigau. A proposal for Word Sense Disambiguation using Conceptual Distance. In *Proceedings of the International Conference "Recent Advances in Natural Language Processing" (RANLP95)*, 1995.
2. Eneko Agirre and German Rigau. Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistic (COLING96*, Copenhagen, Denmark, 1996.
3. Gerard Escudero, Lluis Màrquez, and German Rigau. On the portability and tuning of supervised word sense disambiguation systems. In Hinrich Schütze and Keh-Yih Su, editors, *Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, China, 2000.
4. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
5. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.
6. Andrés Montoyo and Manuel Palomar. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. In *Proceedings of 11th International Workshop on Database and Expert Systems Applications (DEXA 2000). 11th International Workshop on Database and Expert Systems Applications*, pages 103–107, Greenwich, London, UK, September 2000. IEEE Computer Society.
7. Andrés Montoyo and Manuel Palomar. WSD Algorithm Applied to a NLP System . In Mokrane Bouzeghoub, Zoubida Kedad, and Elisabeth Mtais, editors, *Proceedings of 5th International conference on Applications of Natural Language to Information Systems (NLDB-2000). Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, pages 54–65, Versailles, France, June 2000. Springer-Verlag.
8. Andrés Montoyo and Manuel Palomar. Specification Marks for Word Sense Disambiguation: New Development. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 182–191, Mexico City, February 2001. Springer-Verlag.
9. Andrés Montoyo, Manuel Palomar, and German Rigau. Wordnet enrichment with classification systems. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customisations Workshop. The Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, pages 101–106. Carnegie Mellon University. Pittsburgh, PA, USA, 2001.
10. M. Palomar, M. Saiz-Noeda, R. Muñoz, A. Suárez, P. Martínez-Barco, and A. Montoyo. PHORA: A NLP aystem for Spanish. In A. Gelbukh, editor, *Proceedings of 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 126–139, Mexico City, February 2001. Springer-Verlag.

11. Ted Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In ACL, editor, *Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, PA, USA, 2001.
12. Adwait Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, 1998.
13. Maximiliano Saiz-Noeda, Armando Suárez, and Manuel Palomar. Semantic pattern learning through maximum entropy-based wsd technique. In *Proceedings of CoNLL-2001*, pages 23–29. Toulouse, France, 2001.