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

The University of Alicante systems at SENSEVAL-3*

Sonia Vázquez, Rafael Romero
Armando Suárez and Andrés Montoyo
Dpt. of Software and Computing Systems
Universidad de Alicante, Spain
{svazquez,romero}@dlsi.ua.es
{armando,montoyo}@dlsi.ua.es

Iulia Nica and Antonia Martí †
Dpt. of General Linguistics
Universidad de Barcelona, Spain
iulia@clic.fil.ub.es
amarti@ub.edu

Abstract

The DLSI-UA team is currently working on several word sense disambiguation approaches, both supervised and unsupervised. These approaches are based on different ways to use both annotated and unannotated data, and several resources generated from or exploiting WordNet (Miller et al., 1993), WordNet Domains, EuroWordNet (EWN) and additional corpora. This paper presents a view of different system results for Word Sense Disambiguation in different tasks of SENSEVAL-3.

1 Introduction

Word Sense Disambiguation (WSD) is an open research field in Natural Language Processing (NLP). The task of WSD consists in assigning the correct sense to words in a particular context using an electronic dictionary as the source of words definitions. This is a difficult problem that is receiving a great deal of attention from the research community.

At the Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL-2, several supervised and unsupervised systems took part. The more successful systems relay on corpus-based and supervised learning methods. At SENSEVAL-2 the average level of accuracy achieved rounded 70%, which is insufficient for such other NLP tasks as information retrieval, machine translation, or question answering.

The DLSI-UA systems were applied to three SENSEVAL-3 tasks: English all-words, English lexical sample and Spanish Lexical Sample. Our systems use both corpus-based and knowledge-based approaches: Maximum Entropy(ME) (Lau et al., 1993; Berger et al., 1996; Ratnaparkhi, 1998) is a corpus-based and supervised method based on linguistic features; ME is the core of a bootstrapping algorithm that we call re-training inspired by co-training (Blum and Mitchell, 1998); Relevant Domains (RD) (Montoyo et al., 2003) is a resource built from WordNet Domains (Magnini and Cavaglia, 2000) that is used in an unsupervised method that assigns domain and sense labels; Specification Marks(SP) (Montoyo and Palomar, 2000) exploits the relations between synsets stored in WordNet (Miller et al., 1993) and does not need any training corpora; Commutative Test (CT) (Nica et al., 2003), based on the Sense Discriminators device derived from EWN (Vossen, 1998), disambiguates nouns inside their syntactic patterns, with the help of information extracted from raw corpus.

A resume of which methods and how were used in which SENSEVAL-3 tasks is shown in Table 1.

DLSI-UA Systems	Method	Combined Results
ALL-NOSU	RD	No
LS-ENG-SU	Re-t	No
LS-ENG-NOSU	RD	No
LS-SPA-SU	ME+Re-t	No
LS-SPA-NOSU	SM + ME	Nouns: SM Verbs and adj.: ME
LS-SPA-PATTERN	Pattern-Nica + ME	Nouns: SM Verbs and adj.: ME

Table 1: DLSI-UA Systems at SENSEVAL-3

Most of these methods are relatively new and our goal when participating at SENSEVAL-3 is to evaluate for the first time such approaches. At the moment of writing this paper we can conclude that these are promising contributions in order to improve current WSD systems.

In the following section each method is described briefly. Then, details of how the SENSEVAL-3 train and testing data were processed are shown. Next, the scores obtained by each system are explained. Finally, some conclusions and future work are presented.

* This paper has been partially supported by the Spanish Government (CICyT) under project number TIC-2003-7180 and the Valencia Government (OCyT) under project number CTIDIB-2002-151

2 Methods and Algorithms

In this section we describe the set of methods and techniques that we used to build the four systems that had participated in SENSEVAL-3.

2.1 Re-training and Maximum Entropy

In this section, we describe our bootstrapping method, which we call re-training. Our method is derived from the co-training method. Our re-training system is based on two different views of the data (as is also the case for co-training), defined using several groups of features from those described in Figure 1, with several filters that ensure a high confidence sense labelling.

- the target word itself
- lemmas of content-words at positions ±1, ±2, ±3
- words at positions ±1, ±2,
- words at positions ±1, ±2, ±3
- content-words at positions ±1, ±2, ±3
- POS-tags of words at positions ±1, ±2, ±3
- lemmas of collocations at positions $(-2,-1)$, $(-1,+1)$, $(+1,+2)$
- collocations at positions $(-2,-1)$, $(-1,+1)$, $(+1,+2)$
- lemmas of nouns at any position in context, occurring at least $m\%$ times with a sense
- grammatical relation of the target word
- the word that the target word depends on
- the verb that the target word depends on
- the target word belongs to a multi-word, as identified by the parser
- ANPA codes (Spanish only)
- IPTC codes (Spanish only)

Figure 1: Features Used for the Supervised Learning

These two views consist of two weak ME learners, based on different sets of linguistic features, for every possible sense of a target word. We decided to use ME as the core of our bootstrapping method because it has shown to be competitive in WSD when compared to other machine learning approaches (Suárez and Palomar, 2002; Màrquez et al., 2003).

The main difference with respect co-training is that the two views are used in parallel in order to get a consensus of what label to assign to a particular context. Additional filters will ultimately determine which contexts will then be added to the next training cycle.

Re-training performs several binary partial trainings with positive and negative examples for each sense. These classifications must be merged in a unique label for such contexts with enough evidence of being successfully classified. This "evidence" relies on values of probability assigned by the ME module to positive and negative labels, and the fact that the unlabeled example is classified as positive for a unique sense only. The set of new labeled examples feeds the training corpora of the next iteration with positive and negative examples. The stopping criteria is a certain number of iterations or the failure to obtain new examples from the unlabeled corpus.

2.2 Specification Marks

Specification Marks is an unsupervised WSD method over nouns. Its context is the group of words that co-occur with the word to be disambiguated 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 underlying hypothesis of the method we present here is that the higher the similarity between two words, the larger the amount of information shared by two concepts. In this case, the information commonly shared by two concepts is indicated by the most specific concept that subsumes them both in the taxonomy.

The input for the WSD module is a group of nouns $W = \{w_1, w_2, ..., w_n\}$ in a context. Each word w_i is sought in WordNet, each having an associated set of possible senses $S_i = \{S_{i1}, S_{i2}, ..., S_{in}\}$, and each sense having a set of concepts in the IS-A taxonomy (hypernymy/hyponymy relations). First, the common concept to all the senses of the words that form the context is gathered. This concept is marked by the initial specification mark (ISM). If this initial specification mark does not resolve the ambiguity of the word, we then descend through the WordNet hierarchy, from one level to another, assigning new specification marks. The number of concepts contained within the subhierarchy is then counted for each specification mark. The sense that corresponds to the specification mark with the highest number of words is the one chosen as the sense disambiguated within the given context

We define six heuristics for our system: Heuristic of Hypernym, Heuristic of Definition, Heuristic of Common Specification Mark, Heuristic of Gloss Hypernym, Heuristic of Hyponym and Heuristic of Gloss Hyponym.

2.3 Relevant Domains

This is an unsupervised WSD method based on the WordNet Domains lexical resource (Magnini and Cavaglia, 2000). The underlying working hypothesis is that domain labels, such as ARCHITEC-

TURE, SPORT and MEDICINE provide a natural way to establish semantic relations between word senses, that can be used during the disambiguation process. This resource has already been used on Word Sense Disambiguation (Magnini and Strapparava, 2000), but it has not made use of glosses information. So our approach make use of a new lexical resource obtained from glosses information named Relevant Domains.

First step is to obtain the Relevant Domains resource from WordNet glosses. For this task is necessary a previous part-of-speech tagging of WordNet glosses (each gloss has associated a domain label). So we extract all nouns, verbs, adjectives and adverbs from glosses and assign them their associated domain label. With this information and using the Association Ratio formula (w=word,D=domain label), in (1), we obtain the Relevant Domains resource.

$$AR(w, D) = Pr(w|D) log_2 \frac{Pr(w|D)}{Pr(w)} \quad (1)$$

The final result is for each word, a set of domain labels sorted by Association Ratio, for example, for word plant" its Relevant Domains are: genetics 0.177515, ecology 0.050065, botany 0.038544 ....

Once obtained Relevant Domains the disambiguation process is carried out. We obtain from the text source the context words that co-occur with the word to be disambiguated (context could be a sentence or a window of words). We obtain a context vector from Relevant Domains and context words (in case of repeated domain labels, they are weighted). Furthermore we need a sense vector obtained in the same way as context vector from words of glosses of each word sense. We select the correct sense using the cosine measure between context vector and sense vectors. So the selected sense is that for which the cosine with the context vector is closer to one.

2.4 Pattern-Nica

This is an unsupervised method only for Spanish nouns exploiting both EuroWordNet and corpus. In this method we adopt a different approach to WSD: the occurrence to be disambiguated is considered not separately, but integrated into a syntactic pattern, and its disambiguation is carried out in relation to this pattern. A syntactic pattern is a triplet X-R-Y, formed by two lexical content units X and Y and an eventual relational element R, which corresponds to a syntactic relation between X and Y. Examples: [X=*canal*-noun R=*de*-preposition Y=*televisión*-noun], [X=*pasaje*-noun R=Ø Y=*aéreo*-adjective]. The strategy is based on the hypothesis that syntactic patterns in which an ambiguous occurrence participates have decisive influence on its meaning. We also assume that inside a syntactic pattern a word will tend to have the same sense: the "quasi one sense per syntactic pattern" hypothesis. The method works as follows:

Step 1, the identification of the syntactic patterns of the ambiguous occurrence;

Step 2, the extraction of information related to it: from corpus and from the sentential context;

Step 3, the application of the WSD algorithm on the different information previously obtained;

Step 4, the final sense assignment by combining the partial sense proposals from step 3.

For step 1, we POS-tag the test sentence and extract the sequences that correspond to previously defined combinations of POS tags. We only kept the patterns with frequency 5 or superior.

In step 2, we use a search corpus previously POS-tagged. For every syntactic pattern of the ambiguous occurrence X, we obtain from corpus two sets of words: the substitutes of X into the pattern (S1) and the nouns that co-occur with the pattern in any sentence from the corpus (S2), In both cases, we keep only the element with frequency 5 or superior.

We perform step 3 by means of the heuristics defined by the Commutative Test (CT) algorithm applied on each set from 2. The algorithm is related to the Sense Discriminators (SD) lexical device, an adaptation of the Spanish WordNet, consisting in a set of sense discriminators for every sense of a given noun in WordNet. The Commutative Test algorithm lays on the hypothesis that if an ambiguous occurrence can be substituted in a syntactic pattern by a sense discriminator, then it can have the sense corresponding to that sense discriminator.

For step 4, we first obtain a sense assignment in relation with each syntactic pattern, by intersecting the sense proposals from the two heuristics corresponding to a pattern; then we choose the most frequent sense between those proposed by the different syntactic patterns; finally, if there are more final proposed senses, we choose the most frequent sense on the base of sense numbers in WordNet.

The method we propose for nouns requires only a large corpus, a minimal preprocessing phase (POS-tagging) and very little grammatical knowledge, so it can easily be adapted to other languages. Sense assignment is performed exploiting information extracted from corpus, thus we make an intensive use of sense untagged corpora for the disambiguation process.

3 Tasks Processing

At this point we explain for each task the systems processing. The results of each system are shown in Table2:

DLSI-UA Systems	Precision	Recall
LS-SPA-SU	84%	84%
LS-ENG-SU	82%	32%
ALL-NOSU	34%	28%
LS-ENG-NOSU	32%	20%
LS-SPA-NOSU	62%	62%
LS-SPA-PATTERN	84%	47%

Table 2: Results at SENSEVAL-3

3.1 DLSI-UA-LS-SPA-SU

Our system, based on re-training and maximum entropy methods, processed both sense labelled and unlabelled Spanish Lexical Sample data in three consecutive steps:

Step 1, analyzing the train corpus: words which most frequent sense is under 70% were selected. For each one of these words, each feature was used in a 3-fold cross-validation in order to determine the best set of features for re-training.

Step 2, feeding training corpora: for these selected words, based on the results of the previous step, each training corpus was enriched with new examples from the unlabelled data using re-training.

Step 3, classifying the test data: for the selected words, re-training was used again to obtain a first set of answers with, *a priori*, a label with a high level of confidence; the remaining contexts that re-training could not classify were processed with the ME system using a unique set of features for all words.

The lemmatization and POS information supplied into the SENSEVAL-3 Spanish data were the information used for defining the features of the system.

0ur system obtained an accuracy of 0.84 for the Spanish lexical sample task. Unfortunately, a shallow analysis of the answers revealed that the UA.5 system performed slightly worse than if only the basic ME system were used[1]. This fact means that the new examples extracted from the unlabelled data introduced too much noise into the classifiers. Because this anomalous behavior was present only on some words, a complete study of such new examples must be done. Probably, the number of iterations done by re-training over unlabelled data were too low and the enrichment of the training corpora not large enough.

[1]The ME system, without using re-training, has not competed at SENSEVAL-3: our own scoring of these set of answers reported an accuracy of 0.856

3.2 DLSI-UA-LS-ENG-SU

In the English Lexical Sample task our system goal was to prove that the re-training method ensures a high level of precision.

By means of a 3-fold cross-validation of the train data, the features were ordered from higher to lower precision. Based on this information, four executions of re-training over the test data were done with different selections of features for the two views of the method. Each execution feed the learning corpora of the next one with new examples, those that re-training considered as the most probably correct.

For this system Minipar parser (Lin, 1998)was used to properly add syntactic information to the training and testing data.

Almost 40% of the test contexts were labelled by our system, obtaining these scores (for "fine-grained" and "coarse-grained", respectively): 0.782/0.828 precision and 0.310/0.329 recall. In our opinion, such results must be interpreted as very positive because the re-training method is able to satisfy a high level of precision if the parameters of the system are correctly set.

3.3 DLSI-UA-ALL-NOSU and DLSI-UA-LS-ENG-NOSU

In the English All Words and English Lexical Sample tasks RD system was performed with information obtained from Relevant Domains resource using for the disambiguation process all the 165 domain labels.

For All Words task we used as input information all nouns, verbs, adjectives and adverbs present in a 100 words window around the word to be disambiguated. So our system obtained a 34% of precision and a reduced recall around 28%.

For Lexical Sample task we used all nouns, verbs, adjectives and adverbs present in the context of each instance obtaining around 32% precision.

We obtained a reduced precision due to we use all the domains label hierarchy. In some experiments realized on SENSEVAL-2 data, our system obtained a more high precision when grouping domains into the first three levels. Therefore we expect with reducing the number of domains labels, an improvement on precision.

3.4 DLSI-UA-LS-SPA-NOSU

We used a combined system for Spanish Lexical Sample task, using the SM method for disambiguating nouns and the ME method for disambiguating verbs and adjectives. We obtained around 62% precision and a 62% recall.

3.5 DLSI-UA-LS-SPA-PATTERN

Our goal when participating in this task was to demonstrate that the applying of syntactic patterns to WSD maintains high levels of precision.

In this task we used also a combined system for Spanish Lexical Sample task, using Pattern-Nica method for disambiguating nouns and ME method for disambiguating verbs and adjectives. We obtained around 84% precision and a 47% recall.

4 Conclusions

The supervised systems for the English and Spanish lexical sample tasks are very competitive. Although the processing of the train and test data was different for each task, both systems rely on retraining, a bootstrapping method, that uses a maximum entropy-based WSD module.

The results for the English task prove that retraining is capable of maintaining a high level of precision. Nevertheless, for the Spanish task, although the scores achieved were excellent, the system must be redesigned in order to improve the classifiers.

The re-training method is a proposal that we are trying to incorporate into text retrieval and question answering systems that could take advantage of sense disambiguation of a subset of words.

The unsupervised systems presented here does not appear to be sufficient for a stand-alone WSD solution. Wether these methods can be combined with other supervised methods to improve their results requires further investigation.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, Wisconsin, July. ACM Press.

R. Lau, R. Rosenfeld, and S. Roukos. 1993. Adaptative statistical language modeling using the maximum entropy principle. In *Proceedings of the Human Language Technology Workshop, ARPA*.

Dekang Lin. 1998. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In M. Gavrilidou, G. Crayannis, S. Markantonatu, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, pages 1413–1418, Athens, Greece.

Bernardo Magnini and C. Strapparava. 2000. Experiments in Word Domain Disambiguation for Parallel Texts. In *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, Hong Kong, China.

Lluís Màrquez, Fco. Javier Raya, John Carroll, Diana McCarthy, Eneko Agirre, David Martínez, Carlo Strapparava, and Alfio Gliozzo. 2003. Experiment A: several all-words WSD systems for English. Technical Report WP6.2, MEANING project (IST-2001-34460), http://www.lsi.upc.es/∼nlp/meaning/meaning.html.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1993. Five Papers on WordNet. *Special Issue of the International journal of lexicography*, 3(4).

Andrés Montoyo and Manuel Palomar. 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. In *Proceedings of 11th International Workshop on Database and Expert Systems Applications (DEXA 2000)*, pages 103–107, Greenwich, London, UK, September. IEEE Computer Society.

Andrés Montoyo, Sonia Vázquez, and German Rigau. 2003. Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes. *Procesamiento del Lenguaje Natural*, 30, september.

Iulia Nica, Antonia Martí, and Andrés Montoyo. 2003. Colaboración entre información paradigmática y sintagmática en la desambiguación semántica automática. *XIX Congreso de la SEPLN 2003*.

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

Armando Suárez and Manuel Palomar. 2002. A maximum entropy-based word sense disambiguation system. In Hsin-Hsi Chen and Chin-Yew Lin, editors, *Proceedings of the 19th International Conference on Computational Linguistics*, pages 960–966, Taipei, Taiwan, August. COLING 2002.

Piek Vossen. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter,* 3(1).