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 R2D2 Team at SENSEVAL-3*

Sonia Vázquez, Rafael Romero
Armando Suárez and Andrés Montoyo
Dpto. de Lenguajes y Sistemas. Informáticos
Universidad de Alicante, Spain
{svazquez,romero}@dlsi.ua.es
{armando,montoyo}@dlsi.ua.es

Manuel García, M. Teresa Martín †
M. Ángel García and L. Alfonso Ureña
Dpto. de Informática
Universidad de Jaén, Spain
{mgarcia,maite}@ujaen.es
{magc,laurena}@ujaen.es

Davide Buscaldi, Paolo Rosso ‡
Antonio Molina, Ferrán Plá and Encarna Segarra
Dpto. de Sistemas Informáticos y Computación
Univ. Polit. de Valencia, Spain
{dbuscaldi,prosso}@dsic.upv.es
{amolina,fpla,esegarra}@dsic.upv.es

Abstract

The R2D2 systems for the English All-Words and Lexical Sample tasks at SENSEVAL-3 are based on several supervised and unsupervised methods combined by means of a voting procedure. Main goal was to take advantage of training data when available, and getting maximum coverage with the help of methods that not need such learning examples. The results reported in this paper show that supervised and unsupervised methods working in parallel, and a simple sequence of preferences when comparing the answers of such methods, is a feasible method...

The whole system is, in fact, a cascade of decisions of what label to assign to a concrete instance based on the agreement of pairs of systems, when it is possible, or selecting the available answer from one of them. In this way, supervised are preferred to unsupervised methods, but these last ones are able to tag such words that not have available training data.

1 Introduction

Designing a system for Natural Language Processing (NLP) requires a large knowledge on language structure, morphology, syntax, semantics and pragmatic nuances. All of these different linguistic knowledge forms, however, have a common associated problem, their many ambiguities, which are difficult to resolve.

In this paper we concentrate on the resolution of the lexical ambiguity that appears when a given word in a context has several different meanings. This specific task is commonly referred as Word Sense Disambiguation (WSD). This is a difficult problem that is receiving a great deal of attention from the research community because its resolution can help other NLP applications as Machine Translation (MT), Information Retrieval (IR), Text Processing, Grammatical Analysis, Information Extraction (IE), hypertext navigation and so on.

The R2D2 Team has participated in two tasks: English all-words and lexical sample. We use several different systems both supervised and unsupervised. The supervised methods are based on Maximum Entropy (ME) (Lau et al., 1993; Berger et al., 1996; Ratnaparkhi, 1998), neural network using the Learning Vector Quantization algorithm (Kohonen, 1995) and Specialized Hidden Markov Models (Pla, 2000). The unsupervised methods are Relevant Domains (RD) (Montoyo et al., 2003) and the CIAOSENSO WSD system which is based on Conceptual Density (Agirre and Rigau, 1995), frequency of WordNet (Miller et al., 1993a) senses and WordNet Domains (Magnini and Cavaglia, 2000).

In the following section we will show a more complete description of the systems. Next, how such methods were combined in two voting systems, and the results obtained in SENSEVAL-3. Finally, some conclusions will be presented.

2 Systems description

In this section the systems that have participated at SENSEVAL-3 will be described.

2.1 Maximum Entropy

ME modeling provides a framework for integrating information for classification from many heterogeneous information sources (Manning and Schütze, 1999). ME probability models have been success-

* 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

fully applied to some NLP tasks, such as POS tagging or sentence boundary detection (Ratnaparkhi, 1998). ME have been also applied to WSD (van Halteren et al., 2001; Montoyo and Suárez, 2001; Suárez and Palomar, 2002), and as meta-learner in (Ilhan et al., 2001).

Our ME-based system has been shown competitive (Màrquez et al., 2003) when compared to other supervised systems such as Decision Lists, Support Vector Machines, and AdaBoost. The features that were defined to train the system are those described in Figure 1.

- 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

Figure 1: Features Used for the Supervised Learning of the ME system

Because the ME system needs annotated data for the training, Semcor (Miller et al., 1993b) was used for the English All-Words task, the system was trained using Semcor (Miller et al., 1993b), and parsed by Minipar (Lin, 1998). Only those words that have 10 examples or more in Semcor were processed in order to obtain a ME classifier.

For the Spanish Lexical Sample task, the training data from SENSEVAL-3 was the source of labeled examples. We did not use any parser, just the lemmatization and POS-tagging information supplied into the training data itself.

2.2 UPV-SHMM-AW

The upv-shmm-aw WSD system is a supervised approach based on Specialized Hidden Markov Models (SHMM).

Basically, a SHMM consists of changing the topology of a Hidden Markov Model in order to get a more accurate model which includes more information. This is done by means of an initial step previous to the learning process. It consists of the redefinition of the input vocabulary and the output tags. This redefinition is done by means of two processes which transform the training set: the selection process chooses which input features (words, lemmas, part-of-speech tags, ...) are relevant to the task, and the specialization process redefines the output tags by adding information from the input. This specialization produces some changes in the model topology, in order to allow the model to better capture some contextual restrictions and to get a more accurate model.

We used as training data the part of the SemCor corpus that is semantically annotated and supervised for nouns, verbs, adjectives and adverbs, and the test data set provided by SENSEVAL-2.

We used 10% of the training corpus as a development data set in order to determine the best selection and specialization criteria.

In the experiments, we used WordNet1.6 (Miller et al., 1993a) as a dictionary that supplies all the possible semantic senses for a given word. Our system disambiguated all the polysemic lemmas, that is, the coverage of our system was 100%. For unknown words (words that did not appear in the training data set), we assigned the first sense in WordNet.

2.3 Relevant Domains

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

The LVQ-JAÉN-ELS system (García-Vega et al., 2003) is based on a supervised learning algorithm for WSD. The method trains a neural network using the Learning Vector Quantization (LVQ) algorithm (Kohonen, 1995), integrating Semcor and several semantic relations of WordNet.

The Vector Space Model (VSM) is used as an information representation model. Each sense of a word is represented as a vector in an n-dimensional space where n is the number of words in all its contexts.

We use the LVQ algorithm to adjust the word weights. The input vector weights are calculated as shown by (Salton and McGill, 1983) with the standard $(tf \cdot idf)$. They are presented to the LVQ network and, after training, the output vectors are obtained, containing the adjusted weights for all senses of each word.

Any word to disambiguate is represented with a vector in the same way. This representation must be compared with all the trained sense vectors of the word by applying the cosine similarity rule:

$$sim(w_k, x_i) = \frac{w_k \cdot x_i}{\mid w_k \mid \cdot \mid x_i \mid} \qquad (2)$$

The sense corresponding to the vector of highest similarity is selected as the disambiguated sense.

To train the neural network we have integrated semantic information from two linguistic resources: SemCor1.6 corpus and WordNet1.7.1 lexical database. From Semcor1.6 we used the paragraph as a contextual semantic unit and each context was included in the training vector set. From WordNet1.7.1 some semantic relations were considered, specifically, synonymy, antonymy, hyponymy, homonymy, hyperonymy, meronymy, and coordinate terms. This information was introduced to the training set through the creation of artificial paragraphs with the words of each relation. So, for a word with 7 senses, 7 artificial paragraphs with the synonyms of the 7 senses were added, 7 more with all its hyponyms, and so on.

The learning algorithm is very simple. First, the learning rate and the codebook vectors are initialized. Then, the following procedure is repeated for all the training input vectors until a stopping criterion is satisfied:

- Select a training input pattern, x, with class d, and present it to the network

- Calculate the Euclidean distance between the input vector and each codebook vector $\mid\mid x - w_k \mid\mid$

- Select the codebook vector, w_c, that is closest to the input vector, x, like the winner sense.

- The winner neuron updates its weights according the learning equation:

$$w_c(t+1) = w_c(t) + s \cdot \alpha(t) \cdot [x(t) - w_c(t)] \quad (3)$$

where $s = 0$, if $k \neq c$; $s = 1$, if $x(t)$ and $w_c(t)$ belong to the same class $(c = d)$; and $s = -1$, if they do not $(c \neq d)$. $\alpha(t)$ is the learning rate, and $0 < \alpha(t) < 1$ is a monotically decreasing function of time. It is recommended that $\alpha(t)$ should already initially be rather small, say, smaller than 0.1 (Kohonen, 1995) and $\alpha(t)$ continues decreasing to a given threshold, u, very close to 0.

2.5 CIAOSENSO

The CIAOSENSO WSD system is an unsupervised system based on Conceptual Density, the frequency of WordNet sense, and WordNet Domains. Conceptual Density is a measure of the correlation among the sense of a given word and its context. The noun sense disambiguation is performed by means of a formula combining the Conceptual Density with WordNet sense frequency (Rosso et al., 2003). The context window used in both the English all-words and lexical sample tasks is of 4 nouns. Additional weights are assigned to those senses having the same domain as the context nouns' senses. Each weight is proportional to the frequency of such senses, and is calculated as $MDW(f, i) = 1/f \cdot 1/i$ where f is an integer representing the frequency of the sense of the word to be disambiguated and i gives the same information for the context word. Example: If the word to be disambiguated is doctor, the domains for senses 1 and 4 are, respectively, Medicine and School. Therefore, if one of the context words is university, the resulting weight for *doctor(4)* and *university(3)* is $1/4 * 1/3$.

The sense disambiguation of an adjective is performed only on the basis of the above weights.

Given one of its senses, we extract the synsets obtained by the *similar_to*, *pertainym* and *attribute* relationships. For each of them, we calculate the MDW with respect to the senses of the context noun. The weight assigned to the adjective sense is the average between these MDWs. The selected sense is the one having the maximum average weight.

The sense disambiguation of a verb is done nearly in the same way, but taking into consideration only the MDWs with the context words. In the all-words task the context words are the noun before and after the verb, whereas in the lexical sample task the context words are four (two before and two after the verb), without regard to their morphological category. This has been done in order to improve the recall in the latter task, for which the test corpus is made up mostly by verbs.

The sense disambiguation of adverbs (in both tasks) is carried out in the same way of the disambiguation of verbs for the lexical sample task.

3 Tasks Processing

We have selected several combinations of such systems described before for two voting systems, one for the Lexical-Sample task and the other for the All-Words task.

3.1 English Lexical Sample Task

At the English Lexical Sample task we combined the answers of four systems: Relevant Domains, CIAOSENSO, LVQ-JAÉN-ELS and Maximum Entropy.

The four methods worked in parallel and their sets of answers were the input of a majority voting procedure. This procedure selected those answers with more systems agreements. In case of tie we gave priority to supervised systems.

With this voting system we obtained around a 63% precision and a 52% recall.

3.2 English All Words Task

For this task we used a voting system combining the results of Relevant Domains, Maximum Entropy, CIAOSENSO and UPV-SHMM-AW. So we obtained the final results after 10 steps.

Step 1, we selected those answers with agreement between ME and UPV-SHMM-AW (supervised systems).

Step 2, from no agreement in step 1 we selected those answers with agreement between ME and Relevant Domains.

Step 3, from no agreement in step 2 we selected those answers with agreement between ME and CIAOSENSO.

Step 4, from no agreement in step 3 we selected those answers with agreement between CIAOSENSO and UPV-SHMM-AW.

Step 5, from no agreement in step 4 we selected those answers with agreement between UPV-SHMM-AW and Relevant Domains.

Step 6, from no agreement in step 5 we selected those answers with agreement between Relevant Domains and CIAOSENSO.

Step 7, from no agreement in step 6 we selected Maximum Entropy answers.

Step 8, from the remaining unlabeled instances we selected UPV-SHMM-AW answers.

Step 9, from the remaining unlabeled instances we selected Relevant Domains answers.

Step 10, from the remaining unlabeled instances we selected CIAOSENSO answers.

Last step was labeling with the most frequent sense in WordNet those instances that had been not tagged by any system, but in view of the final results only two instances had not answer and we didn't find them in WordNet.

With this voting system preference was given to supervised systems over unsupervised systems.

We obtained around a 63% precision and a 63% recall.

4 Conclusions

This paper presents the main characteristics of the Maximum Entropy, LVQ-JAEN-ELS, UPV-SHMM-AW, Relevant Domains and CIAOSENSO systems within the framework of SENSEVAL-3 English Lexical Sample and All Words tasks. These systems are combined with a voting technique obtaining a promising results for English All Words and English Lexical Sample tasks.

References

Eneko Agirre and German Rigau. 1995. A proposal for word sense disambiguation using Conceptual Distance. In *Proceedings of the International Conference "Recent Advances in Natural Language Processing" (RANLP95)*.

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.

Manuel García-Vega, María Teresa Martín-Valdivia, and Luis Alfonso Ureña. 2003. Aprendizaje competitivo lvq para la desambiguación léxica. *Revista de la Sociedad Espaola para el Procesamiento del Lenguaje Natural*, 31:125–132.

H. Tolga Ilhan, Sepandar D. Kamvar, Dan Klein, Christopher D. Manning, and Kristina Toutanova. 2001. Combining Heterogeneous Classifiers for Word-Sense Disambiguation. In Judita Preiss and David Yarowsky, editors, *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (*SENSEVAL-2*)*, pages 87–90, Toulouse, France, July. ACL-SIGLEX.

T. Kohonen. 1995. Self-organization and associative memory. *2nd Ed. Springer Verlag, Berlin*.

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.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

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. 1993a. Five Papers on WordNet. *Special Issue of the International journal of lexicography*, 3(4).

George A. Miller, C. Leacock, R. Tengi, and T. Bunker. 1993b. A Semantic Concordance. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, New Jersey.

Andrés Montoyo and Armando Suárez. 2001. The University of Alicante word sense disambiguation system. In Judita Preiss and David Yarowsky, editors, *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (*SENSEVAL-2*)*, pages 131–134, Toulouse, France, July. ACL-SIGLEX.

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.

F. Pla. 2000. *Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos*. Tesis doctoral, Departamento de Sistemas Informáticos y Computación. Universidad de Politécnica de Valencia, Septiembre.

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

P. Rosso, F. Masulli, D. Buscaldi, F. Pla, and A. Molina. 2003. Automatic noun disambiguation. *LNCS, Springer Verlag*, 2588:273–276.

G. Salton and M.J. McGill. 1983. Introduction to modern information retrieval. *McGraw-Hill, New York*.

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.

H. van Halteren, J. Zavrel, and W. Daelemans. 2001. Improving accuracy in wordclass tagging through combination of machine learning systems. *Computational Linguistics*, 27(2):199–230.