Word Sense vs. Word Domain Disambiguation: a Maximum Entropy approach*

Armando Suárez and Manuel Palomar

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

Abstract. In this paper, a supervised learning system of word sense disambiguation is presented. It is based on *maximum entropy conditional probability models*. This system acquires the linguistic knowledge from an annotated corpus and this knowledge is represented in the form of features. The system were evaluated both using WordNet's senses and domains as the sets of classes of each word. Domain labels are obtained from the enrichment of WordNet with subject field codes which produces a polysemy reduction. Several types of features has been analyzed for a few words selected from the DSO corpus. Currently, the system implementation does not support any smoothing technique or complex pre-processing but its accuracy of the system is good when it is compared with, for example, the systems at SENSEVAL-2. Using the domain enrichment of WordNet, a 14% of accuracy improvement is achieved.

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 using an electronic dictionary as the source of word definitions. This is a hard problem that is receiving a great deal of attention from the research community.

Currently, there are two main methodological approaches in this research area: *knowledge-based* methods and *corpus-based* methods. The former approach relies on previously acquired linguistic knowledge, and the latter uses techniques from statistics and machine learning to induce models of language usage from large samples of text [1]. These last methods can perform supervised or unsupervised learning. With supervised learning, the actual status (here, sense label) for each piece of data in the training example is known, whereas with unsupervised learning the classification of the data in the training example is not known [2].

At SENSEVAL-2, researchers showed the latest contributions to WSD. Some supervised systems competed in the English lexical sample tasks [3]. The Johns

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

Hopkins University system combines several WSD subsystems based on different methods: decision lists [4]; transformation-based, error-driven learning [5] [6]; cosine-based vector models; decision stumps; and two feature-enhanced naive Bayes systems. The Southern Methodist University system is an instance-based learning method but also uses word-word relation patterns obtained from WordNet1.7 and Semcor, as described in [7]. Boosting [9] is based on the AdaBoost.MH algorithm.

[8] proposes a baseline methodology for WSD that relies on decision tree learning and Naive Bayesian classifiers, using simple lexical features. Several systems combining different classifiers based on distinct sets of features competed at SENSEVAL-2, both in the English and Spanish lexical sample tasks.

This paper presents a system that implements a corpus-based method of WSD. The method used to perform the learning over a set of sense-disambiguated examples is that of maximum entropy probability models (ME). Linguistic information is represented in the form of feature vectors, which identify the occurrence of certain attributes that appear in contexts containing linguistic ambiguities. The context is the text surrounding an ambiguity that is relevant to the disambiguation process. The features used may be of a distinct nature: word collocations, part-of-speech labels, keywords, topic and domain information, grammatical relationships, and so on.

[10] do machine translation tasks using ME to perform some kinds of semantic classification, but they also rely on another statistical training procedure to define word classes. In addition, we are aware of a few sites on the Internet which describe attempts to apply ME to WSD, but to our knowledge, these results have not yet been published.

Word Domain Disambiguation (WDD) is a variant of WSD where words in a text are tagged with a domain label in place of a sense label, and an enrichment of WordNet is proposed using subject field codes [11]. On the one hand, labeling with such information causes a synsets clustering and then a polysemy reduction. Therefore, WDD must be more accurate than WSD. On the other hand, several researches argue that applications like Information Retrieval (IR) and Question Answering (QA) will be better improved with domain disambiguation than with sense disambiguation. Another proposal of enrichment of WordNet that uses IPTC[1] subject codes can be seen in [12]

In the following discussion, the ME framework and the features implementation will be described. Then, the complete set of feature definitions used in this work will be detailed. Next, evaluation results using several combinations of these features for a few words will be shown. Finally, some conclusions will be presented, along with a brief discussion of work in progress and future work planned.

[1] The IPTC Subject Reference System has been developed to allow Information Providers access to a universal language independent coding system for indicating the subject content of news items. http://www.iptc.org

2 The Maximum Entropy Framework

ME modeling provides a framework for integrating information for classification from many heterogeneous information sources [2]. ME probability models have been successfully applied to some NLP tasks, such as part-of-speech (POS) tagging or sentence boundary detection [13].

The WSD method used in this paper is based on conditional ME probability models. It has been implemented using a supervised learning method that consists of building word-sense classifiers using a semantically tagged corpus. A classifier obtained by means of an ME technique consists of a set of parameters or coefficients which are estimated using an optimization procedure. Each coefficient is associated with 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 from the training data is considered. Some advantages of using the ME framework are that even knowledge-poor features may be applied accurately; the ME framework thus allows a virtually unrestricted ability to represent problem-specific knowledge in the form of features [13].

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 takes the form of equation (1), where $cp(x)$ is some observable characteristic in the context[2]. The conditional probability $p(c|x)$ is defined by equation (2), where α_i is the parameter or weight of the feature i, K is the number of features defined, and $Z(x)$ is a constant to ensure that the sum of all conditional probabilities for this context is equal to 1.

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

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

The implementation of the ME-based WSD system was done in C++ and the features used to test its accuracy are described in the following section. A complete description of the system and some of the features mentioned in the following section can be found in [14].

A usual definition of features would substitute $CP(x)$ in equation (1) with an expression like $info(x,i) = a$, where $info(x,i)$ informs of a property that can be found at position i in a context x, and a is a predefined value. For example, if we consider that 0 is the position of the word to be learned and that i is related to 0, then *word(x,-1) = "best"*. In the following, we will refer to this type of features as "non-relaxed features".

[2] 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, uses this feature.

Other expressions, such as $info(x,i) \in W_{(c',i)}$, may be substituted for the term $CP(x)$, as a way to reduce the number of possible features. In the expression above, $W_{(c',i)}$ is the set of attributes present in the learning examples at position i. So this kind of function reduces the number of features to one per each sense at position i. In the following, we will refer to this type of features as "relaxed features".

3 Evaluation

In this section we present the results of our evaluation. Some polysemous nouns and verbs have been selected and evaluated using the DSO sense-tagged English corpus [15]. This corpus is structured in files containing tagged examples of several nouns and verbs. Tags correspond to senses in WordNet 1.5 [16]. All examples in each file (both those from the Brown Corpus and those from the Wall Street Journal) have been processed. This corpus has been parsed using MiniPar [17].

The set of features defined for the training of the system is described below (figure 1). Features are automatically defined as explained earlier and depend on the data in the training corpus. These features are based on words, collocations, part-of-speech (POS) tags, and grammatical properties in the local context.

Fig. 1. List of types of features

- *Non-relaxed*
  - ***0* features**: ambiguous-word shape
  - s **features**: words in positions ±1, ±2, ±3
  - p **features**: POS-tags of words in positions ±1, ±2, ±3
  - km **features**: lemmas of nouns at any position in context, occurring at least m% times with a sense
  - r **features**: grammatical relation of the ambiguous word
  - d **features**: the word that the ambiguous word depends on
  - m **features**: the ambiguous word belongs to a multi-word, as
- *Relaxed*
  - L **features**: lemmas of content-words in positions ±1, ±2, ±3
  - W **features**: content-words in positions ±1, ±2, ±3
  - S **features**: words in positions ±1, ±2, ±3
  - B **features**: lemmas of collocations in positions $(-2,-1)$, $(-1,+1)$, $(+1,+2)$
  - C **features**: collocations in positions $(-2,-1)$, $(-1,+1)$, $(+1,+2)$
  - P **features**: POS-tags of words in positions ±1, ±2, ±3
  - D **features**: the word that the ambiguous word depends on
  - M **features**: the ambiguous word belongs to a multi-word, as identified by the parser

Table 1 shows the best results obtained using a 10-fold cross-validation evaluation method. Several feature combinations have been tested in order to find the best set for each selected word. The main goal was to achieve the most relevant information from the corpus for each feature rather than applying the same combination of features to all words.

Table 1. 10-fold cross-validation best results on DSO files

	Senses	Examples	Features	Functions	Accur	MFS
age.N	3	491	0CsprDMk5	1587	73.8	11.71
art.N	4	393	0sprdm	1594	65.2	18.49
car.N	2	1363	s	3036	97.1	1.97
child.N	2	1057	sp	2731	90.5	9.63
church.N	3	367	0rDMCk3	228	67.9	6.81
cost.N	2	1456	0WrDM	62	90.0	2.67
head.N	7	844	sprdm	2911	80.8	43.95
interest.N	6	1479	0sprDM	4059	70.1	25.03
line.N	22	1320	0LSBCrdm	1542	54.7	32.77
work.N	6	1419	0sprdm	4784	53.2	21.45
fall.V	6	1341	LSBCrdm	503	84.9	14.82
know.V	6	1425	0rDMCk10	230	47.9	13.02
set.V	11	1246	BsprDMk5	4569	57.3	20.43
speak.V	5	510	0sp	1667	74.5	5.40
take.V	19	794	LWBCsrDMk10	3706	43.0	7.45
Averages	**7**	**1034**		**2214**	**70.1**	**15.71**
Nouns	6	1019		2253	74.3	17.45
Verbs	9	1063		2135	61.5	12.22
All words			**0sprdm**	**3411**	**68.8**	**14.4**
Nouns				3013	73.5	16.6
Verbs				4208	59.4	10.1

In order to perform the ten tests on each word, some preprocessing of the corpus was done. For each word file in DSO, all senses were uniformly distributed in the ten folds (each fold contains one tenth of examples of each sense, except for the tenth fold, which contains the remaining examples). Those senses that had fewer than ten examples in the original corpus file were rejected and not processed; therefore, the "Senses" column shows the number of senses effectively learned.

Senses is the number of distinct senses in the corpus, *Features* the feature selection with the best result, *Functions* the number of functions generated from features, and *Accur* (for "accuracy") the number of correctly classified contexts divided by the total number of contexts. Column *MFS* is the gain in accuracy of our ME method against the most-frequent-sense classification.

The data summarized in table 1 reveal that all types of features, relaxed and non-relaxed ones, are useful. Moreover, each word has its own best-feature-

selection. If such strategy of selection is assumed, for this fifteen words an average of 70.1% of contexts are correctly classified. This result means a gain in accuracy of 15.7% against the most-frequent-sense classification. Nouns are better classified than verbs.

Applying a fixed set of features to all words, the best result is obtained with the "0*sprdm*" one, a 68.8% of accuracy (14.4% more than MFS). Again, nouns obtain better results than verbs.

The results of the ME method were also compared with the train and test data from SENSEVAL-2 for the Spanish lexical sample task. The results reported in SENSEVAL-2 for thirteen systems range from 71.2% to 51.4% for accuracy. The ME method obtained an accuracy rate of 65.03% (4th place) using $0LWSBCQ$ features for all words, and 64.04% (4th place too) using $OLBK5$. In this comparison, we used the Conexor FDG Parser [18]. We intend to next test for all features.

Table 2 shows the evaluation results when the sets of classes include domain labels instead of sense labels. The first consequence of using domains instead synsets is the reduction of the number of classes, and then the gain in accuracy of the method. Obviously, those words with the same number of domains than senses do not contribute to a gain in accuracy. Currently, domain labels are assigned to nouns only. In order to perform the comparison between WDD and WSD results, 14 nouns were selected.

Table 2. Results with a fixed feature selection

	Doms	Ex	Features	Accur	MFS	Sens	Features	Accur	MFS	**Gain**
action.N	4	1049	sprdm	59.4	46.7	5	0sprdm	52.5	46.7	**+6.80**
activity.N	2	786	0sprdm	86.9	85.7	3	0sprdm	71.3	68.8	**+15.61**
art.N	2	393	SP	97.5	97.5	4	0sprdm	65.2	48.0	**+32.32**
body.N	2	390	0LSsBCprdm	86.3	77.9	4	0LSBCprdm	68.6	60.5	**+17.69**
book.N	3	615	0sprdm	84.1	80.6	4	0LSBCprdm	68.8	65.0	**+15.29**
business.N	6	1483	0sprdm	64.4	50.3	7	0sprdm	64.1	50.3	**+0.34**
case.N	3	1419	0LSsBCprdm	74.0	66.8	9	0SQ	53.2	32.5	**+20.76**
center.N	3	546	0LSsBCprdm	80.9	58.3	6	0SP	66.9	58.3	**+14.02**
church.N	2	367	0sprdm	69.6	67.1	3	0CrDMk3	67.9	62.0	**+1.73**
condition.N	2	624	sprdm	87.6	84.6	3	0LB	81.0	79.6	**+6.60**
course.N	4	337	0LSsBCprdm	77.8	49.4	5	0SP	65.0	42.3	**+12.88**
interest.N	5	1476	0LSsBCprdm	70.9	45.9	6	0sprDM	70.1	45.9	**+0.80**
line.N	14	1320	LWBCp	73.4	42.5	22	sprdm	55.9	22.7	**+17.41**
work.N	3	1419	0LSBCprdm	80.6	71.7	6	0sprdm	53.2	32.8	**+27.48**
Averages	**4**	**873**		**78.1**	**66.1**	**6**		**64.6**	**51.11**	**+13.55**
All words			**0sprdm**	**76.9**			**0sprdm**	**63.2**		**13.7**

The left half of the table is related to WDD and the right half to WSD. *Doms* is the number of distinct domains in the corpus, and *Sens* the number of

senses, *Features* the feature selection with the best result, and *Accur* (for "accuracy") the number of correctly classified contexts divided by the total number of contexts. Columns *MFS* are the accuracy of our ME method when the most-frequent-domain or sense is selected. Finally, *Gain* is the gain in accuracy of WDD against WSD. The last row is the average results when a fixed set of features (*0sprdm*) is applied to all words.

As a direct consequence of the polysemy reduction, an average gain in accuracy of 13% had been achieved. Because the majority of nouns in DSO reduce their polysemy with domains rather than senses, this is a promising result[3]. Currently, we are doing a complete evaluation using all nouns of the corpus.

4 Conclusions

A WSD system based on maximum entropy conditional probability models has been presented. It is a supervised learning method that needs a corpus previously annotated with sense labels, or domain labels.

For a few words selected from the DSO corpus, several combinations of features were analyzed in order to identify which were the best. The WSD evaluation results of the system were compared with the Spanish lexical sample task at SENSEVAL-2.

Several researches criticize the excessive polysemy of WordNet, specially for IR and QA applications, and propose a clustering of synsets to achieve more efficiency. WordNet Domains [11] is a proposal that assigns a subject field code to each synset reducing the polysemy degree, currently for nouns only. In order to evaluate the accuracy of the method when the set of classes is formed by domain labels instead of sense labels, 14 nouns were selected. A gain of a 14% of accuracy of WDD against WSD were obtained.

Future research will incorporate domain information as an additional information source for the system in order to improve WSD and WDD. These attributes will be incorporated into the learning of the system in the same way that features were incorporated, as described above.

As we work to improve the ME method, we are also working to develop a cooperative strategy between several other methods as well, both knowledge-based and corpus-based.

References

1. Pedersen, T.: A decision tree of bigrams is an accurate predictor of word sense. [19]
2. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)

[3] The DSO corpus has 121 nouns and 938 senses: 100 nouns reduce its polysemy to 629 subject field codes (an average reduction of 3.09 classes)

3. SENSEVAL-2: Second international workshop on evaluating word sense disambiguation systems: system descriptions. http://www.sle.sharp.co.uk/senseval2/ (2001)
4. Yarowsky, D.: Hierarchical decision lists for word sense disambiguation. Computers and the Humanities **34** (2000)
5. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics **21** (1995) 543–565
6. Florian, R., Ngai, G.: Multidimensional transformation-based learning. In Daelemans, W., Zajac, R., eds.: Proceedings of CoNLL-2001, Toulouse, France (2001) 1–8
7. Mihalcea, R., Moldovan, D.: An iterative approach to word sens disambiguation. In: Proceedings of FLAIRS-2000, Orlando, FL (2000) 219–223
8. Pedersen, T.: A baseline methodology for word sense disambiguation. [20] 126–135
9. Escudero, G., Màrquez, L., Rigau, G.: Boosting applied to word sense disambiguation. In: Proceedings of the 12th Conference on Machine Learning ECML2000, Barcelona, Spain (2000)
10. García-Varea, I., Och, F.J., Ney, H., Casacuberta, F.: Refined lexicon models for statistical machine translation using a maximum entropy approach. In: Proceedings of 39th Annual Meeting of the Association for Computational Linguistics. (2001) 204–211
11. Magnini, B., Strapparava, C.: Experiments in Word Domain Disambiguation for Parallel Texts. In: Proceedings of the ACL Workshop on Word Senses and Multilinguality, Hong Kong, China (2000)
12. Montoyo, A., Palomar, M., Rigau, G.: WordNet Enrichment with Classification Systems. [19]
13. Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD thesis, University of Pennsylvania (1998)
14. Suárez, A., Palomar, M.: Feature selection analysis for maximum entropy-based wsd. [20] 146–155
15. Ng, H.T., Lee, H.B.: Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In Joshi, A., Palmer, M., eds.: Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, San Francisco, Morgan Kaufmann Publishers (1996)
16. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Five Papers on WordNet. Special Issue of the International journal of lexicography **3** (1993)
17. Lin, D.: 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 (1998)
18. Tapanainen, P., Järvinen, T.: A non-projective dependency parser. In: Proceedings of the Fifth Conference on Applied Natural Language Processing. (1997) 64–71
19. ACL, ed.: Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations. In ACL, ed.: Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburgh, PA, USA (2001)
20. Gelbukh, A., ed.: Proceedings of 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002). In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, Mexico City, Springer-Verlag (2002)