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 *conditional maximum entropy 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. Using the domain enrichment of WordNet, a 7% 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 [3], researchers showed the latest contributions to WSD. Some supervised systems competed in the English lexical sample task. The Johns Hopkins University system combines, by means of a voting-based classifier, several WSD subsystems based on different methods: decision lists [4], cosine-based vector models, and two Bayesian classifiers. The Southern Methodist University

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

system is an instance-based learning method but also uses word-word relation patterns obtained from WordNet and Semcor. LazyBoosting [5] is based on the AdaBoost.MH algorithm.

[6] 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 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.

At SENSEVAL-2, the Standford University implements several combinations of simple classifiers, and one of them makes use of conditional ME models. In [7] ME is used to perform semantic classification on machine translation tasks, 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. 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. An enrichment of WordNet is proposed using subject field codes [8]. At SENSEVAL-2, this enrichment were used by ITC-irst systems [9]. Another proposal using IPTC[1] subject codes can be seen in [10].

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 [11].

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 [11].

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.

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

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

The implementation of the ME-based WSD system was done in C++ and 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 [12].

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"* and *word(x,-1)="big"* could be used. 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 type of 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. For example, $word(x, -1) \in \{$ *"best"*, *"big"*$\}$. 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. All nouns from the DSO sense-tagged English corpus [13] have been selected and evaluated. This corpus is structured in files containing tagged examples of several nouns and verbs. Tags correspond to senses in WordNet 1.5 [14]. In order to make use of WordNet Domains, tagged senses were mapped to WordNet 1.6 [15][3]. This corpus has been parsed using MiniPar [16].

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

The set of features defined for the training of the system is described in figure 1. The majority of them depend on nearest words (for example, s or L types comprise all possible features defined by the words occurring at positions w_{-3}, w_{-2}, w_{-1}, w_{+1}, w_{+2}, w_{+3} related to the ambiguous word). Features are

[3] http://www.lsi.upc.es/ nlp/tools/mapping.html

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.

Table 1 shows the best results obtained for a sub-set of nouns 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 compare best values of WDD (the left half of the table) and WSD (the right half) for each word.

Table 1. Example of best-feature-selection for WDD and WSD

	Doms	Ex	Features	Accur	MFS	Sens	Features	Accur	MFS
action,N	4	1049	sprdm	59.35	46.75	5	0sprdmK10	52.69	46.75
activity,N	2	786	0sBCprdmK10	86.95	85.65	3	0sprdm	71.31	68.75
art,N	2	393	Most frequent	97.51	97.51	4	0sprdm	65.19	47.95
body,N	2	390	0LSsBCprdm	86.27	77.91	4	0LSsBCprdm	68.59	60.51
book,N	3	615	0sbcprdmK10	84.35	80.60	4	0sprdmK10	70.07	64.97
business,N	6	1483	0sbcprdmK10	64.97	50.30	7	0sBCprdmK10	64.15	50.30
case,N	3	1419	0sbcprdmK10	74.62	66.76	9	0sbcprdmK10	56.82	32.53
center,N	3	546	0LSsBCprdm	80.90	58.33	6	0sbcprdmK10	72.36	58.33
church,N	2	367	0sprdm	70.45	67.11	3	0sprdmK10	67.08	62.05
condition,N	2	624	0sbcprdmK10	87.88	84.59	3	0LSsBCprdm	83.38	79.63
course,N	4	337	0sBCprdmK10	78.85	49.36	5	0sBCprdmK10	72.14	42.32
interest,N	5	1476	0sprdmK10	71.79	45.86	6	0sprdmK10	70.87	45.86
line,N	14	1320	0LSsBCprdm	65.26	42.52	22	0sprdmK10	56.02	22.73
work,N	3	1419	0LSBCprdm	80.63	71.71	6	0sprdmK10	54.58	32.83

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, *Doms* (for "domains") and *Sens* (for "senses") columns show the number of classes effectively learned, *Features* the feature selection with the best result, *Ex* (for "examples") the number contexts, and *Accur* (for "accuracy") the number of correctly classified contexts divided by the total number of contexts. Column *MFS* is the accuracy obtained by most-frequent-sense classification.

The data summarized in table 1 reveal that all types of features, relaxed and non-relaxed ones, are useful. Nouns are better classified than verbs. Moreover, each word has its own best-feature-selection. If such strategy of selection is assumed, better values of accuracy are expected than applying the same types of features to all words. Obviously, these results are unreliable because train and test data overlap, but a toy test using SENSEVAL-2 data point to the usefulness of such information.

Table 2 shows the evaluation results for all nouns in DSO using several different sets of features. The first consequence of using domains instead synsets is the reduction of the number of classes (from an average of 4.8 senses to 3.5 domains per noun), 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.

Table 2. Word domain and word sense disambiguation results

Features	WDD	WSD	Diff
Most frequent	68.7	58.7	+9.98
LB	73.5	64.6	+8.94
SP	74.8	66.6	+8.20
0LB	75.4	67.1	+8.34
0SP	75.7	67.8	+7.96
sp	77.2	69.5	+7.70
0sp	77.7	70.2	+7.52
sprdm	78.1	70.6	+7.48
0sprdm	78.4	71.0	+7.37
0LSsBCprdm	78.6	71.0	+7.58
0sprdmk10	78.7	71.4	+7.26
0sBCprdmk10	78.7	71.4	+7.27
0sbcprdmk10	78.7	71.4	+7.33

As a direct consequence of the polysemy reduction, an average gain in accuracy of 7% had been achieved. The DSO corpus has 121 nouns and 938 senses: 100 nouns reduce its polysemy to 629 subject field codes. Examining the last selections, there is not a great influence of adding more features to the training . In fact, using a 10-fold cross-validation paired Student's t-test [17] with a confidence value $t_{9,0.975} = 1.833$, BC features has no effect on accuracy (but they are worse than bc). The best feature selection ($0sbcprdmk10$), fails the significance test with the other four first ranked selections (with $sprdm$ succeeds) except for $0sBCprdmk10$ which is worse than it.

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.

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 [8] 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, all nouns were selected from the DSO corpus. A gain of a 7% 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. In: Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh (2001) 79–86
2. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
3. Preiss, J., Yarowsky, D., eds.: Proceedings of SENSEVAL-2. In Preiss, J., Yarowsky, D., eds.: Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France, ACL-SIGLEX (2001)
4. Yarowsky, D.: Hierarchical decision lists for word sense disambiguation. Computers and the Humanities **34** (2000) 179–186
5. 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)
6. Pedersen, T.: A baseline methodology for word sense disambiguation. [18] 126–135
7. 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
8. 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)
9. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A.: Using Domain Information for Word Sense Disambiguation. [3] 111–114
10. Montoyo, A., Palomar, M., Rigau, G.: WordNet Enrichment with Classification Systems. In Preiss, J., Yarowsky, D., eds.: Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburgh, PA, USA (2001)
11. Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD thesis, University of Pennsylvania (1998)
12. Suárez, A., Palomar, M.: Feature selection analysis for maximum entropy-based wsd. [18] 146–155
13. 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)
14. 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)

15. Daude, J., Padro, L., Rigau, G.: Mapping wordnets using structural information. In: Proceedings of the 38th Anual Meeting of the Association for Computational Linguistics (ACL'2000), Hong Kong (2000)
16. 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)
17. Dietterich, T.G.: Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation **10** (1998) 1895–1923
18. Gelbukh, A.F., ed.: Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings. In Gelbukh, A.F., ed.: CICLing. Volume 2276 of Lecture Notes in Computer Science., Springer (2002)