Semi-supervised Word Sense Disambiguation using the Web as Corpus

Rafael Guzmán-Cabrera[1,2], Paolo Rosso[2], Manuel Montes-y-Gómez[3], Luis Villaseñor-Pineda[3], David Pinto-Avendaño[4]

[1] FIMEE, Universidad de Guanajuato, Mexico
guzmanc@salamanca.ugto.mx
[2] NLE Lab, DSIC, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es
[3] LabTL, Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
{mmontesg, villasen}@inaoep.mx
[4] FCC, Benemérita Universidad Autónoma de Puebla, Mexico
dpinto@cs.buap.mx

Abstract. As any other classification task, Word Sense Disambiguation requires a large number of training examples. These examples, which are easily obtained for most of the tasks, are particularly difficult to obtain for this case. Based on this fact, in this paper we investigate the possibility of using a Web-based approach for determining the correct sense of an ambiguous word based only in its surrounding context. In particular, we propose a semi-supervised method that is specially suited to work with just a few training examples. The method considers the automatic extraction of unlabeled examples from the Web and their iterative integration into the training data set. The experimental results, obtained over a subset of ten nouns from the SemEval lexical sample task, are encouraging. They showed that it is possible to improve the baseline accuracy of classifiers such as Naïve Bayes and SVM using some unlabeled examples extracted from the Web.

1 Introduction

It is well known that, in all languages, some words may have several different meanings or senses. For example, in English, the word “bank” can either mean a financial institution or a sloping raised land. Related to this language phenomenon, the task of *Word Sense Disambiguation* (WSD) considers the assignment of the correct sense to such ambiguous words based on their surrounding context [6].

There are two main kinds of methods to carry out the task of WSD. On the one hand, the knowledge-based methods, which disambiguate words by comparing their context against information from a predefined lexical resource such as Wordnet [1, 3]. On the other hand, *corpus-based methods*, which achieve the sense disambiguation by applying rules that were automatically learned from a sense tagged corpus [14]. Recent reports [8] indicate that corpus-based methods tend to be more precise than knowledge-based ones. Nevertheless, due to the lack of large sense tagged cor-

pora (as well as to the difficulty of manually creating them), the use of these kind of methods is still very limited.

In order to tackle the above mentioned problem, many researches have recently been working on *semi-supervised learning methods* [2, 4], which consider the usage of large amount of unlabeled data together with a few labeled examples. In particular, the idea of learning classifiers from a combination of labeled and unlabeled data has been successfully applied in WSD [9, 10, 13, 15, 16].

In line with these current works, we have proposed a new semi-supervised method for general text classification tasks [5]. This method differs from previous approaches in two main issues. First, it does not require a predefined set of unlabelled training examples, instead it considers their automatic extraction from the Web. Second, it applies a self-training approach that selects instances not only considering their labelling confidence by a base classifier, but also their correspondence with a web-based labelling[1]. This method has been applied with success in thematic and non-thematic text classification tasks, indicating that it is possible to automatically extract discriminative information from the Web.

In this paper, we move forward to investigate the application of the proposed *web-based self-training method* in the task of WSD. This task confronts our method with new challenges since (i) ambiguous words tend to have several "slightly" different meanings, and (ii) their classification typically rely only on a very small context. This way, the task of WSD can be considered as a narrow-domain and short-text classification problem.

The rest of the paper is organized as follows. Section 2 describes our web-based self-training approach. Section 3 presents the evaluation results of the method in a subset of ten words from the last SemEval English lexical sample exercise. Finally, Section 4 depicts our conclusions.

2 Our Semi-supervised Classification Method

Figure 1 shows the general scheme of our semi-supervised text classification method. It consists of two main processes. The first one deals with the corpora acquisition from the Web, whereas the second focuses on the self-training learning approach. The following sections describe in detail these two processes.

It is important to notice that this method can be directly applied to the task of WSD since it is, in essence, a text classification problem, where word senses correspond to classes and word contexts represent the documents.

2.1 Corpora Acquisition

This process considers the automatic extraction of unlabeled examples from the Web. In order to do this, it first constructs a number of *queries* by combining the most significant words for each sense of a polysemous word; then, using these que-

[1] Given that each unlabeled example is downloaded from the Web using a set of automatically defined class queries, each of them has a default category or web-based label.

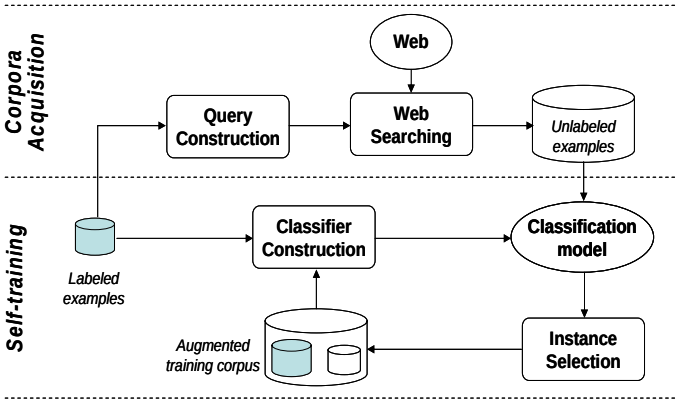

Figure 1. General overview of our text classification method

ries, it looks at the Web for some additional training examples related to the given senses.

At this point, it is important to comment that even though the idea of using the Web as corpus, it may not initially sound intuitive; there are already a number of successful efforts concerning different natural language tasks [7]. In particular, in [17], the authors proposed a method for mining the Web to improve text classification by creating a background text set. Our method is similar to this approach in the sense that it also mines the Web for additional information (extra-unlabeled examples). Nevertheless, as we will describe below, our method applies finer procedures to construct the set of queries related to each sense and to combine the downloaded information.

Query Construction.

To construct the set of queries for searching the Web, it is necessary to previously determine the set of relevant words from each sense in the training corpus. The criterion used for this purpose is based on a combination of two characteristics of the given words: on the one hand, their frequency of occurrence, and on the other hand, their information gain. Explicitly, we consider that a word w_i is relevant for a sense S if:

1. The frequency of occurrence of w_i in S is greater than the average occurrence of all words (happening more than once) in that sense. That is:

$$f_{w_i}^S > \frac{1}{|S'|} \sum_{\forall w \in S'} f_w^S \text{, where } S' = \left\{ w \in S \middle| f_w^S > 1 \right\}$$

2. The information gain of w_i in the given training set is positive ($IG_{w_i} > 0$). The idea of this condition is to select those words that help reducing the uncertainty of the value of the sense from the given set of examples.

Having obtained the set of relevant words per each sense it is possible to construct their corresponding set of queries. We decided to construct queries of three words[2]. This way, we created as many queries per sense as all three-word combinations of its relevant words. We measure the significance of a query $q = \{w_1, w_2, w_3\}$ to the sense S as indicated below:

$$\Gamma_S(q) = \sum_{i=1}^{3} f_{w_i}^{S} \times IG_{w_i}$$

Because the selection of relevant words relies on a criterion based on their frequency of occurrence and their information gain, the number of queries per sense is not the same even though they include the same number of training examples. In addition, an increment in the number of examples does not necessarily represent a growth in the number of built queries.

Web Searching.

The next action is using the defined queries to extract from the Web a set of additional unlabeled text examples from the Web. Based on the observation that most significant queries tend to retrieve the most relevant Web pages, our method for searching the Web determines the number of downloaded examples per query in a direct proportion to its Γ-value. Therefore, given a set of M queries $\{q_1,\ldots, q_M\}$ for sense S, and considering that we want to download a total of N additional examples per sense, the number of examples to be extracted by a query q_i is determined as follows:

$$\Psi_S(q_i) = \frac{N}{\sum_{k=1}^{M} \Gamma_S(q_k)} \times \Gamma_S(q_i)$$

It is important to notice that, because each downloaded example corresponds exactly to one particular query; it is possible to consider that these examples belong to a particular sense (the same sense of the query that was used to retrieved them). This information, which we previously mentioned as Web-based labeling, represents a kind of prior category for the unlabeled examples, and thus it can be of great help in improving the performance of the semi-supervised learning approach.

2.2 Semi-supervised learning

The objective of this second process is to increase the classification accuracy by gradually enlarging the originally small training set with the unlabeled examples downloaded from the Web. In particular, we designed this process based on the *self-training approach* described in [12]. In this approach, a classifier is initially trained using the small amount of labeled data; then, this classifier is used to classify the unlabeled data, and the most confident examples -in conjunction with their predicted

[2] Queries formed by more than three words tend to produce very few results; on the other hand, queries of one or two words are very general and, consequently, tend to retrieve a lot of irrelevant results.

label- are added to the training set; finally, the classifier is re-trained and the procedure is repeated.

In our case, as we previously explained, the selection of the most confident examples not only considers their labeling confidence by a base classifier, but also their correspondence with the Web-based labeling. Following, we detail our new self-training algorithm:

1. Build a weak classifier (C_l) using a specified learning method (l) and the training set available (T).
2. Classify the unlabeled Web examples (E) using the constructed classifier (C_l). In other words, estimate the sense for all downloaded examples.
3. Select the best m examples per sense ($E_m \subseteq E$; in this case E_m represent the union of the best m examples from all senses) based on the following two conditions:
   a) The estimated sense of the example corresponds to the sense of the query used to download it. In some way, this filter works as an ensemble of two classifiers: C_l and the Web (expressed by the set of queries).
   b) The example has one of the m-highest confidence predictions for the given sense.
4. Combine the selected examples with the original training set ($T \leftarrow T \cup E_m$) in order to form a new training collection. At the same time, eliminate these examples from the set of downloaded instances ($E \leftarrow E - E_m$).
5. Iterate σ times over steps 1 to 4 or repeat until $E_m = \emptyset$. In this case σ is a user specified threshold.
6. Construct the final classifier using the enriched training set.

3 Experimental Evaluation

3.1 Evaluation Data Set

The evaluation of the method was carried out on a subset of the lexical sample task from the SemEval forum[3]. In particular, we consider only *nine nouns* which have training instances for all their senses. Table 1 shows some numbers about these nouns. It is interesting to notice that there is an important imbalance problem for some nouns indicated by the standard deviation value. For instance, for the first sense of "bill" there are 685 training instances, whereas for the second there are only 54, producing an average standard deviation of 446.18.

3.2 Evaluation Measure and Baseline Results

The effectiveness of the method was measured by the *classification accuracy*, which indicates the percentage of instances of a polysemous word that were correctly classified from the entire test set.

[3] http://nlp.cs.swarthmore.edu/semeval/tasks/task17/description.shtml

Table 1. Data set statistics

Noun	Number of senses	Training instances	Test instances	Standard deviation (*of training instances per sense*)
Source	5	151	35	20.64
Bill	2	739	114	446.18
President	3	872	176	401.24
Management	2	277	44	40.30
Condition	2	130	33	59.40
Policy	2	329	39	129.4
Rate	2	1003	145	490.02
Drug	2	205	46	28.99
State	3	609	70	263.03

Table 2 shows the baseline results for two different classifiers, namely, Naïve Bayes and SVM. In all cases, we determined the context of the words using a window of five words to the left and five words to the right. In all cases, we also removed all punctuation marks and numerical symbols, as well as all stopwords.

Table 2. Baseline results using Naïve Bayes and SVM

Noun	Classification accuracy	
	Naïve Bayes	SVM
Source	77.14	74.29
Bill	92.08	95.05
President	89.20	89.20
State	78.57	78.57
Management	77.27	85.82
Condition	66.66	72.72
Policy	74.36	87.18
Rate	86.90	87.59
Drug	78.26	71.74

As it can be seen, there is a relationship between the number of training instances and their degree of imbalance (refer to Table 1) with the baseline accuracy (refer to Table 2). Therefore, this result evidences the need for increasing the size of the training sets by incorporating new unlabeled examples.

3.3 Results of the Method

This section describes the application of the proposed semi-supervised method to the task of WSD. The method, as depicted in Section 2, includes two main processes: the corpora acquisition from the Web and the self-training learning approach. Following, we detail some data from both of them.

The central task for corpora acquisition is the automatic construction of a set of queries that expresses the relevant content of each sense. For this experiment we considered the ten words with the greatest weight. Then, using these queries, we collected from the Web a set of 1,000 additional examples per sense for each polysemous word. Table 3 shows some example queries corresponding to the two different senses of the word “drug”.

Table 3. Example queries for the two senses of the word "drug"

Sense	Queries
Drug-1	drug new used drug said company drug sales companies
Drug-2	drug trafficking charges drug charges major drug major use

Regarding the learning phase, it is important to point out that there is not a clear criterion to determine the parameters *m* and σ of our self-training method. For this experiment, we determined the number of unlabeled examples that must be incorporated into the training set at each iteration based on the following condition: the added information –expressed in number of words– must be proportionally small with respect to the original training data. This last condition is very important because of the small size of word contexts. In particular, we decided to incorporate five examples per sense at each iteration. However, it is necessary to perform further experiments in order to determine the best value of *m* for this task.

Table 4 shows the results of this experiment. They indicate that our method slightly outperformed all baseline results especially when using the Naïve Bayes classifier. These results confirm our intuition that in scenarios having very few training instances it is better to include a small group of unlabeled examples that considerably augments the dissimilarities among senses than to include a lot of doubtable-quality information.

Table 4. Results of our method for the first three iterations (using Bayes and SVM)

	Bayes				SVM			
Noun	**Baseline Result**	**It. 1**	**It. 2**	**It. 3**	**Baseline Result**	**It. 1**	**It. 2**	**It. 3**
Source	77.1	80.0	80.0	80.0	74.3	77.1	80.0	68.6
Bill	92.0	92.0	92.1	91.1	95.1	95.1	95.1	93.1
President	89.2	87.5	88.1	88.1	89.2	89.8	89.8	87.5
State	78.5	80.0	78.6	80.0	78.6	78.6	78.6	78.6
Managment	77.2	79.5	79.5	79.5	85.8	81.8	81.8	81.8
Condition	66.6	66.6	66.7	63.6	72.7	72.7	75.8	75.8
Policy	74.3	76.9	76.9	74.4	87.2	87.2	87.2	74.8
Rate	86.9	86.9	89.0	89.0	87.6	87.6	86.9	74.4
Drug	78.2	80.4	80.4	-	71.7	71.7	69.6	86.9

3.4 Discussion of Results

In order to have a deep understanding of achieved results, we carried out a statistical analysis of the used corpus. The purpose of this analysis was to explain the complementary performance of the Naïve Bayes and SVM classifiers. It is necessary to remark that Naïve Bayes is a probabilistic classifier that apply the Bayes theorem under the assumption (naïvely) that exist independence on the features of the items to be classified. From this viewpoint, we suggest to use a statistical measure that takes into account the relationship among the words that made up each text (the features

used in this experiment). In particular, we applied a measure called SLMB[4] (supervised language modeling based measure) [11]. This measure uses a set of language models (based on bigrams and trigrams) to compute the entropy among the different meanings of each ambiguous word. Formally, given a corpus D (of one ambiguous word), with a gold standard consisting of k classes (or meanings) $C = \{C_1, C_2,\ldots,C_k\}$, the SLMB measure is defined as follows:

$$SMLB(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(Perplexity(C_i \mid \overline{C_i^*}) - \mu(Perplexity(C)))^2}$$

$$\mu(Perplexity(C)) = \frac{\sum_{i=1}^{k} Perplexity(C_i \mid \overline{C_i^*})}{k}$$

In these formulas, $\overline{C_i^*}$ indicates the language model obtained by using all the classes except C_i, and $Perplexity(C_i \mid \overline{C_i^*})$ denotes the perplexity of the class C_i language model with respect to the $\overline{C_i^*}$ language model. The latter formula calculates the mean of the perplexity among the different ambiguous word meanings.

Table 5 shows the results obtained by the SLMB measure and the perplexity mean for all word corpora. In all cases, we evaluate the original and the enriched corpus. On the one hand, we may observe that words *state*, *management*, *policy* and *rate* have not changed significantly their language model from the original to the enriched version of the corpus. Therefore, there were not significant changes over the dependency relationships among the words (features), which leads to obtain a similar behavior of the Naïve Bayes classifier with both corpora (original and enriched). On the other hand, we may see that the ambiguous words *president*, *source*, *condition* and *drug* have obtained important changes on the values obtained with the SLMB and perplexity mean, which means that their language models have been modified sufficiently avoiding to preserve the same or similar results with both corpora (original and enriched). However, the SVM classifier may have been benefited from this last fact. We consider that their support vectors have been enriched, which could helped the SVM classifier to have obtained better results than the Naïve Bayes did on these last ambiguous words.

[4] http://nlp.dsic.upv.es:8080/watermaker

Table 5. SLMB and perplexity mean results over both, the original and enriched corpus

Noun	SLMB			Perplexity Mean		
	Original value	Final value	Change (%)	Original value	Final value	Change (%)
Bill	23.8	30.5	28.1	140.2	166.5	18.7
State	24.4	29.0	19.0	106.3	143.5	34.9
Management	5.7	5.1	10.5	114.3	122.4	7.1
Policy	41.4	38.2	7.7	116.1	135.9	17.1
Rate	21.6	22.8	5.5	124.7	136.8	9.7
President	61.2	164.9	169.4	150.8	264.4	75.3
Source	52.3	81.7	56.2	68.8	145.4	111.1
Condition	25.8	31.5	22.4	76.8	111.5	45.1
Drug	5.8	0.4	93.4	81.9	87.4	6.8

4 Conclusions

This paper describes a novel web-based self-training method for text classification. This method differs from other semi-supervised classification approaches in that: (i) it is specially suited to work with very few training examples, and (ii) it considers the automatic extraction of additional training knowledge from the Web.

The described method was already evaluated on two different classification tasks (classification of news reports and contemporary poem authorship attribution, respectively), obtaining good results in both cases. In this paper, we went a step forward and investigated the possibility of applying this method in the task of WSD, which can be considered a narrow-domain and short-text classification problem.

The results obtained in a subset of ten nouns from the SemEval lexical sample task were not as successful as those achieved in previous tasks. Nevertheless, they evidence that unlabeled data may improve performance of potentially any corpus-based WSD system.

Acknowledgments

This work was done under partial support of CONACYT-Mexico, PROMEP-Mexico (UGTO-121), and MCyT-Spain (TIN2006-15265-C06-04).

References

1. Aguirre E., Rigau G.: A Proposal for Word Sense Disambiguation using Conceptual Distance. In: Proc. of the Int. Conf. on Recent Advances in NLP. RANLP'95. (1995).
2. Blum A. and Mitchell T., Combining labeled and unlabeled data with co-training. In Proc. COLT, pages 92-100. (1998).
3. Buscaldi, D. and Rosso P. A conceptual density-based approach for the disambiguation of toponyms. In: International Journal of Geographical Information Science, 22 (3): 143-153. (2008)
4. Goldman S. and Zhou Y., Enhancing supervised learning with unlabeled data. In Proc. ICML, pages 327-334, (2000).
5. Guzmán-Cabrera R., Montes-y-Gómez M., Rosso P., Villaseñor-Pineda L., Using the Web as Corpus for Self-training Text Categorization. Journal of Information Retrieval. Springer Netherlands, (2009). ISSN 1386-4564. Forthcomming.

6. Ide N. and Veronis J., Introduction to the special Issue on word sense disambiguation: the state of the art, Computational Linguistics. Special Issue on word sense Disambiguation, 24 (1), 1-40. (1998).
7. Kilgarriff A. and Greffenstette G., Introduction to the Special Issue on Web as Corpus, Computational Linguistics, 29, (3), 1-15. (2003).
8. Lee Y. K. and Ng H. T., An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In Proc. EMNLP, pages 41-48. (2002).
9. Mihalcea R., Co-training and Self-training for Word Sense Disambigua-tion. In Proc. CoNLL, pages 33-40. (2004).
10. Pham T. P., Ng H. T., and Lee W. S., Word Sense Disambiguation with Semi-Supervised Learning. In Proc. AAAI, pages 1093-1098. (2005).
11. Pinto D. On Clustering and Evaluation of Narrow Domain Short-Text Corpora. PhD thesis, Universidad Politécnica de Valencia, Spain, (2008).
12. Solorio T., Using unlabeled data to improve classifier accuracy. M.Sc. thesis, Computer Science Department, INAOE, Mexico. (2002).
13. Su W., Carpuat M., and Wu D., Semi-Supervised Training of a Kernel PCA-Based Model for Word Sense Disambiguation. In Proc. COLING, pages 1298-1304. (2004).
14. Tratz S., Sanfilippo A., Gregory M., Chappell A., Posse C. and Paul W., PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation, Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 264–267, 2007.
15. Yarowsky D., Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proc. ACL, pages 189-196. (1995).
16. Yu N. Z., Hong J. D., and Lim T. C., Word Sense Disambiguation Using Label Propagation Based Semi-supervised Learning Method. In Proc. ACL, pages 395-402. (2005).
17. Zelikovitz S., and Kogan M., Using Web Searches on Important Words to Create Background Sets for LSI Classification, 19th Int. FLAIRS Conf., Melbourne Beach, Florida. (2006).