UPV-SI: Word Sense Induction using Self Term Expansion*

David Pinto[1,2] and **Paolo Rosso**[1]
[1]Polytechnic University of Valencia
DSIC, Valencia, Spain, 46022
[2]B. Autonomous University of Puebla
FCC, Puebla, Mexico, 72570
{dpinto, prosso}@dsic.upv.es

Héctor Jiménez-Salazar
Autonomous Metropolitan University
Department of Information Technologies
Cuajimalpa, DF, Mexico, 11850
hgimenezs@gmail.com

Abstract

In this paper we are reporting the results obtained participating in the "Evaluating Word Sense Induction and Discrimination Systems" task of Semeval 2007. Our totally unsupervised system performed an automatic self-term expansion process by mean of co-ocurrence terms and, thereafter, it executed the unsupervised KStar clustering method. Two ranking tables with different evaluation measures were calculated by the task organizers, every table with two baselines and six runs submitted by different teams. We were ranked third place in both ranking tables obtaining a better performance than three different baselines, and outperforming the average score.

1 Introduction

Word Sense Disambiguation (WSD) is a particular problem of computational linguistics which consists in determining the correct sense for a given ambiguous word. It is well-known that supervised algorithms have obtained the best results in public evaluations, but their accuracy is close related with the amount of hand-tagged data available. The construction of that kind of training data is difficult for real applications. The unsupervised WSD overcomes this drawback by using clustering algorithms which do not need training data in order to determine the possible sense for a given ambiguous word.

This paper describes a simple technique for unsupervised sense induction for ambiguous words. The approach is based on a self term expansion technique which constructs a set of co-ocurrence terms and, thereafter, it uses this set to expand the target dataset. The implemented system was performed in the task "SemEval-2007 Task 2: Evaluating Word Sense Induction and Discrimination Systems"(Agirre and A., 2007). The aim of the task was to permit a comparison across sense-induction and discrimination systems. Moreover, the comparison with other supervised and knowledge-based systems may be also done, since the test corpus was borrowed from the well known "English lexical-sample" task in SemEval-2007, with the usual training + test split.

The self term expansion method consists in replacing terms of a document by a set of co-related terms. The goal is to improve natural language processing tasks such as clustering narrow-domain short texts. This process may be done by mean of different ways, often just by using a knowledge database. In information retrieval, for instance, the expansion of query terms is a very investigated topic which has shown to improve results with respect to when query expansion is not employed (Qiu and Frei, 1993; Ruge, 1992; R.Baeza-Yates and Ribeiro-Neto, 1999; Grefenstette, 1994; Rijsbergen, 1979).

The availability of Machine Readable Resources (MRR) like "Dictionaries", "Thesauri" and "Lexicons" has allowed to apply term ex-

*This work has been partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 430–433,
Prague, June 2007.

pansion to other fields of natural language processing like WSD. In (Banerjee and Pedersen, 2002) we may see the typical example of using a external knowledge database for determining the correct sense of a word given in some context. In this approach, every word close to the one we would like to determine its correct sense is expanded with its different senses by using the WordNet lexicon (Fellbaum, 1998). Then, an overlapping factor is calculated in order to determine the correct sense of the ambiguous word. Different other approaches have made use of a similar procedure. By using dictionaries, the proposals presented in (Lesk, 1986; Wilks et al., 1990; Nancy and Véronis, 1990) are the most sucessful in WSD. Yarowsky (Yarowsky, 1992) used instead thesauri for their experiments. Finally, in (Sussna, 1993; Resnik, 1995; Banerjee and Pedersen, 2002) the use of lexicons in WSD has been investigated. Although in some cases the knowledge resource seems not to be used strictly for term expansion, the aplication of co-occurrence terms is included in their algorithms. Like in information retrieval, the application of term expansion in WSD by using corelated terms has shown to improve the baseline results if we carefully select the external resource to use, with a priori knowledge of the domain and the broadness of the corpus (wide or narrow domain). Evenmore, we have to be sure that the Lexical Data Base (LDB) has been suitable constructed. Due to the last facts, we consider that the use of a self automatically constructed LDB (using the same test corpora), may be of high benefit. This assumption is based on the intrinsic properties extracted from the corpus itself. Our proposal is related somehow with the investigations presented in (Schütze, 1998) and (Purandare and Pedersen, 2004), where words are also expanded with co-ocurrence terms for word sense discrimination. The main difference consists in the use of the same corpora for constructing the co-ocurrence list.

Following we describe the self term expansion method used and, thereafter, the results obtained in the task #2 of Semeval 2007 competition.

2 The Self Term Expansion Method

In literature, co-ocurrence terms is the most common technique used for automatic construction of LDBs (Grefenstette, 1994; Frakes and Baeza-Yates, 1992). A simple approach may use n-grams, which allows to predict a word from previous words in a sample of text. The frequency of each n-gram is calculated and then filtered according to some threshold. The resulting n-grams constitutes a LDB which may be used as an "expansion dictionary" for each term.

On the other hand, an information theory-based co-ocurrence measure is discussed in (Manning and Schütze, 2003). This measure is named pointwise Mutual Information (MI), and its applications for finding collocations are analysed by determining the co-ocurrence degree among two terms. This may be done by calculating the ratio between the number of times that both terms appear together (in the same context and not necessarily in the same order) and the product of the number of times that each term ocurrs alone. Given two terms X_1 and X_2, the pointwise mutual information between X_1 and X_2 can be calculated as follows:

$$MI(X_1, X_2) = \log_2 \frac{P(X_1 X_2)}{P(X_1) \times P(X_2)}$$

The numerator could be modified in order to take into account only bigrams, as presented in (Pinto et al., 2006), where an improvement of clustering short texts in narrow domains has been obtained.

We have used the pointwise MI for obtaining a co-ocurrence list from the same target dataset. This list is then used to expand every term of the original data. Since the co-ocurrence formula captures relations between related terms, it is possible to see that the self term expansion magnifies less the noisy than the meaninful information. Therefore, the execution of the clustering algorithm in the expanded corpus should outperform the one executed over the non-expanded data.

In order to fully appreciate the self term expansion method, in Table 1 we show the co-

ocurrence list for some words related with the verb "kill" of the test corpus. Since the MI is calculated after preprocessing the corpus, we present the stemmed version of the terms.

Word	Co-ocurrence terms
soldier	kill
rape	women think shoot peopl old man kill death beat
grenad	todai live guerrilla fight explod
death	shoot run rape person peopl outsid murder life lebanon kill convict...
temblor	tuesdai peopl least kill earthquak

Table 1: An example of co-ocurrence terms

For the task #2 of Semeval 2007, a set of 100 ambiguous words (35 nouns and 65 verbs) were provided. We preprocessed this original dataset by eliminating stopwords and then applying the Porter stemmer (Porter, 1980). Thereafter, when we used the pointwise MI, we determined that the single ocurrence of each term should be at least three (see (Manning and Schütze, 2003)), whereas the maximum separation among the two terms was five. Finally, we selected the unsupervised KStar clustering method (Shin and Han, 2003) for our experiments, defining the average of similarities among all the sentences for a given ambiguous word as the stop criterion for this clustering method. The input similarity matrix for the clustering method was calculated by using the Jaccard coefficient.

3 Evaluation

The task organizers decided to use two different measures for evaluating the runs submitted to the task. The first measure is called unsupervised one, and it is based on the Fscore measure. Whereas the second measure is called supervised recall. For further information on how these measures are calculated refer to (Agirre et al., 2006a; Agirre et al., 2006b). Since these measures give conflicting information, two different evaluation results are reported in this paper.

In Table 2 we may see our ranking and the Fscore measure obtained (UPV-SI). We also show the best and worst team Fscores; as well as the total average and two baselines proposed by the task organizers. The first baseline (Baseline1) assumes that each ambiguous word has only one sense, whereas the second baseline (Baseline2) is a random assignation of senses. We are ranked as third place and our results are better scored than the other teams except for the best team score. However, given the similar values with the "Baseline1", we may assume that that team presented one cluster per ambiguous word as its result as the Baseline1 did; whereas we obtained 9.03 senses per ambiguous word in average.

Name	Rank	All	Nouns	Verbs
Baseline1	1	78.9	80.7	76.8
Best Team	2	78.7	80.8	76.3
UPV-SI	3	66.3	69.9	62.2
Average	-	63.6	66.5	60.3
Worst Team	7	56.1	65.8	45.1
Baseline2	8	37.8	38.0	37.6

Table 2: Unsupervised evaluation (Fscore performance).

In Table 3 we show our ranking and the supervised recall obtained (UPV-SI). We again show the best and worst team recalls. The total average and one baseline is also presented (the other baseline obtained the same Fscore). In this case, the baseline tags each test instance with the most frequent sense obtained in a train split. We are ranked again in third place and our score is slightly above the baseline.

Name	Rank	All	Nouns	Verbs
Best Team	1	81.6	86.8	76.2
UPV-SI	3	79.1	82.5	75.3
Average	-	79.1	82.8	75.0
Baseline	4	78.7	80.9	76.2
Worst Team	6a	78.5	81.8	74.9
Worst Team	6b	78.5	81.4	75.2

Table 3: Supervised evaluation (Recall).

The results show that the technique employed have learned, since our simple approach obtained a better performance than the baselines, especially the one that have chosen the most frequent sense as baseline.

4 Conclusions

We have reported the performance of a single approach based on self term expansion. The technique uses the pointwise mutual information for calculating a set of co-ocurrence terms which then are used to expand the original dataset. Once the expansion has been done, the unsupervised KStar clustering method was used to induce the sense for the different ocurrences of each ambiguous word. We obtained the third place in the two measures proposed in the task. We will further investigate whether an improvement may be obtained by applying term selection methods to the expanded corpus.

References

E. Agirre and Soroa A. 2007. SemEval-2007 Task 2: Evaluating Word Sense Induction and Discrimination Systems. In *SemEval-2007*. Association for Computational Linguistics.

E. Agirre, O. Lopez de Lacalle Lekuona, D. Martinez, and A. Soroa. 2006a. Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In *Textgraphs 2006 workshop, NAACL06*, pages 89–96.

E. Agirre, O. Lopez de Lacalle Lekuona, D. Martinez, and A. Soroa. 2006b. Two graph-based algorithms for state-of-the-art WSD. In *EMNLP*, pages 585–593. ACL.

S. Banerjee and T. Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *CICLing 2002 Conference*, volume 3878 of *LNCS*, pages 136–145. Springer-Verlang.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W. B. Frakes and R. A. Baeza-Yates. 1992. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic.

M. Lesk. 1986. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *ACM SIGDOC Conference*, pages 24–26. ACM Press.

D. C. Manning and H. Schütze. 2003. *Foundations of Statistical Natural Language Processing*. MIT Press. Revised version May 1999.

I. Nancy and J. Véronis. 1990. Mapping dictionaries: A spreading activation approach. In *6th Annual Conference of the Centre for the New Oxford English Dictionary*, pages 52–64.

D. Pinto, H. Jiménez-Salazar, and P. Rosso. 2006. Clustering abstracts of scientific texts using the transition point technique. In *CICLing*, volume 3878 of *LNCS*, pages 536–546. Springer-Verlang.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3).

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48, Boston, MA.

Y. Qiu and H. P. Frei. 1993. Concept based Query Expansion. In *ACM SIGIR on R&D in information retrieval*, pages 160–169. ACM Press.

R.Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern information retrieval*. New York: ACM Press; Addison-Wesley.

P. Resnik. 1995. Disambiguating Noun Groupings with Respect to WordNet Senses. In *3rd Workshop on Very Large Corpora*, pages 54–68. ACL.

C. J. Van Rijsbergen. 1979. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow.

G. Ruge. 1992. Experiments on linguistically-based term associations. *Information Processing & Management*, 28(3):317–332.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

K. Shin and S. Y. Han. 2003. Fast clustering algorithm for information organization. In *CICLing*, volume 2588 of *LNCS*, pages 619–622. Springer-Verlang.

M. Sussna. 1993. Word sense disambiguation for free-test indexing using a massive semantic network. In *2nd International Conference on Information and Knowledge Management*, pages 67–74.

Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. 1990. Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of Rogets categories trained on large corpora. In *14th Conference on Computational Linguistics*, pages 454–460. ACL.