UBC-UMB: Combining unsupervised and supervised systems for all-words WSD

David Martinez,Timothy Baldwin
LT Group, CSSE
University of Melbourne
Victoria 3010 Australia
{davidm,tim}@csse.unimelb.edu.au

Eneko Agirre, Oier Lopez de Lacalle
IXA NLP Group
Univ. of the Basque Country
Donostia, Basque Country
{e.agirre,jibloleo}@ehu.es

Abstract

This paper describes the joint submission of two systems to the all-words WSD subtask of SemEval-2007 task 17. The main goal of this work was to build a competitive unsupervised system by combining heterogeneous algorithms. As a secondary goal, we explored the integration of unsupervised predictions into a supervised system by different means.

1 Introduction

This paper describes the joint submission of two systems to the all-words WSD subtask of SemEval-2007 task 17. The systems were developed by the University of the Basque Country (UBC), and the University of Melbourne (UMB). The main goal of this work was to build a competitive unsupervised system by combining heterogeneous algorithms. As a secondary goal, we explored the integration of this method into a supervised system by different means. Thus, this paper describes both the unsupervised system (UBC-UMB-1), and the combined supervised system (UBC-UMB-2) submitted to the all-words task.

Our motivation in building unsupervised systems comes from the difficulty of creating hand-tagged data for all words and all languages, which is colloquially known as the knowledge acquisition bottleneck. There have also been promising results in recent work on the combination of unsupervised approaches that suggest the gap with respect to supervised systems is narrowing (Brody et al., 2006).

The remainder of the paper is organized as follows. First we describe the disambiguation algorithms in Section 2. Next, the development experiments are presented in Section 3, and our final submissions and results in Section 4. Finally, we summarize our conclusions in Section 5.

2 Algorithms

In this section, we will describe the standalone algorithms (three unsupervised and one supervised) and the combination schemes we explored. The unsupervised methods are based on different intuitions for disambiguation (topical features, local context, and WordNet relations), which is a desirable characteristic for combining algorithms.

2.1 Topic Signatures (TS)

Topic signatures (Agirre and de Lacalle, 2004) are lists of words related to a particular sense. They can be built from a variety of sources, and be used directly to perform WSD. Cuadros and Rigau (2006) present a detailed evaluation of topic signatures built from a variety of knowledge sources. In this work we built those coming from the following:

- the relations in the Multilingual Central Repository (TS-MCR)
- the relations in the Extended WordNet (TS-XWN)

In order to apply this resource for WSD, we simply measured the word-overlap between the target context and each of the senses of the target word. The sense with highest overlap is chosen as the correct sense.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 350–353,
Prague, June 2007.

2.2 Relatives in Context (RIC)

This is an unsupervised method presented in Martinez et al. (2006). This algorithm makes use of the WordNet relatives of the target word for disambiguation. The process is carried out in these steps: (i) obtain a set of close relatives from WordNet for each sense (the relatives can be polysemous); (ii) for each test instance define all possible word sequences that include the target word; (iii) for each word sequence, substitute the target word with each relative and query a web search engine; (iv) rank queries according to the following factors: length of the query, distance of the relative to the target word, and number of hits; and (v) select the sense associated with the highest ranked query.

The intuition behind this system is that we can find related words that can be substituted for the target word in a given context, which are indicative of its sense. The close relatives that can form more common phrases from the target context determine the target sense.

2.3 Relative Number (RNB)

This heuristic has been motivated as a way of identifying rare senses of a word. An important disadvantage of unsupervised systems is that rare senses can be over-represented in the models, while supervised systems are able to discard them because they have access to token-level word sense distributions.

This simple algorithm relies on the number of close relatives found in WordNet for each sense of the word. The senses are ranked according to the number of synonyms, direct hypernyms, and direct hyponyms they have in WordNet. The highest ranked sense is taken to be the most important for the target word, and all occurrences of the target word are tagged with that sense.

2.4 k-Nearest Neighbours (kNN)

As our supervised system, we relied on kNN. This is a memory-based learning method where the neighbours are the k most similar contexts, represented by feature vectors ($\vec{c_i}$) of the test vector ($\vec{f}$). The similarity among instances is measured by the cosine of their vectors. The test instance is labeled with the sense that obtains the maximum sum of the weighted votes of the k most similar contexts. Each vote is weighted depending on its (neighbour) position in the ordered rank, with the closest being first. Equation 1 formalizes kNN, where C_i corresponds to the sense label of the i-th closest neighbour.

$$\arg\max_{S_j} = \sum_{i=1}^{k} \begin{cases} \frac{1}{i} & \text{if } C_i = S_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The UBC group used a combination of kNN classifiers trained over a large set of features, and enhanced this method using Singular Value Decomposition (SVD) for their supervised submission (UBC-ALM) to the lexical-sample and all-words subtasks (Agirre and Lopez de Lacalle, 2007). However, we only used the basic implementation in this work, due to time constraints.

2.5 Combination of systems

We explored two approaches to combine the standalone systems. The first consisted simply of adding up the normalized weights that each system would give to each sense. We tested this voting approach both for the unsupervised and supervised settings.

The second method could only be applied in combination with the supervised kNN system. The idea was to include the unsupervised predictions as weighted features for the supervised system. We refer to this method as "stacking", and it has been previously used to integrate heterogeneous knowledge sources for WSD (Stevenson and Wilks, 2001).

3 Development experiments

We tested the single algorithms and their combination over both Semcor and the training distribution of the SemEval-2007 lexical-sample subtask of task 17 (S07LS for short). The goal of these experiments was to obtain an estimate of the expected performance, and submit the most promising configuration. We present first the tests on the unsupervised setting, and then the supervised setting. It is important to note that the hand-tagged corpora was not used to fine-tune the parameters of the unsupervised algorithms.

3.1 Unsupervised systems

For the first evaluation of our unsupervised systems, we relied on Semcor, and tagged 43,063 instances of the 329 word types occurring in SemEval-2007

System	Recall
RNB	30.6
TS-MCR	**57.5**
TS-XWN	47.0
TS-MCR & TS-XWN	57.3
RBN & TS-MCR & TS-XWN	53.6

Table 1: Evaluation of standalone and combined unsupervised systems over 43,063 instances from Semcor

System	Recall
TS-MCR	60.1
TS-XWN	54.3
TS-MCR & TS-XWN	61.1
TS-MCR & TS-XWN & RIC*	**61.2**

Table 2: Evaluation of standalone and combined unsupervised systems over 8,518 instances from S07LS training

all-words. Due to time constraints, we were not able to test the RIC algorithm on this dataset. The results are shown in Table 1. We can see that the RNB heuristic performs poorly, and that the best configuration consists of applying the single TS-MCR algorithm. From this experiment, we decided to remove the RNB heuristic and focus on the topic signatures and RIC.

We also used S07LS for extra experiments in the unsupervised setting. From the training part of the S07LS dataset, we extracted 8,518 instances of words also occurring in SemEval-2007 all-words. As S07LS used senses from OntoNotes, we relied on the mapping provided by the task organisers to link them to WordNet senses. We left RNB out of this experiment due to its low performance in Semcor, and regarding RIC, we only evaluated a sample of 68 instances. Results are shown in Table 2. The best scores are achieved when combining both sets of topic signatures. The few cases that have been disambiguated with RIC improve the overall performance slightly.

3.2 Combined system

We could not rely on Semcor in the supervised setting (we used it for training), and therefore tried to use as much data as possible from the training component of S07LS, wherein all the instances available (22,281) were disambiguated. We tested first

System	Recall
kNN	**87.4**
kNN & TS-MCR	86.8
kNN & TS-XWN	86.4
kNN & TS-MCR & TS-XWN	86.0

Table 3: Evaluation of voting supervised systems in 22,281 instances from S07LS training

System	Recall
kNN	71.7
kNN & TS-MCR & TS-XWN	**71.8**

Table 4: Evaluation of "stacking" the unsupervised systems on kNN over 8,518 instances from S07LS training

the voting combination by adding the normalized weights from the output of each system. Due to time constraints we only evaluated the combination of kNN with TS-MCR and TS-XWN. Results are shown in Table 3, where we can see that combining the unsupervised systems with voting hurts the performance of the kNN method.

Finally, we applied the second combination approach, consisting of including the predictions of the unsupervised systems as features for kNN ("stacking"). We performed this experiment on the training part of S07LS, but only for the 8,518 instances of the words occurring on the all-words dataset. The results of this experiment are given in Table 4. We observed a slight improvement in this case.

4 Final systems

For our final submissions, we chose the combination "TS-MCR & TS-XWN & RIC" for the unsupervised system (UBC-UMB-1), and the combination "kNN & TS-MCR & TS-XWN" via "stacking" for our supervised system (UBC-UMB-2). The results of all the systems are given in Table 5.

We can see that our unsupervised system ranked 10th. Unfortunately, we do not know at the time of writing which other systems are unsupervised, and therefore are unable to compare to other unsupervised systems.

Our "stacking" supervised system performs slightly lower than the kNN supervised systems by UBC-ALM (which ranks 7th), showing that our system was not able to profit from information from

System	Precision	Recall
1.	0.537	0.537
2.	0.527	0.527
3.	0.524	0.524
4.	0.522	0.486
5.	0.518	0.518
6.	0.514	0.514
7.	0.493	0.492
8. UBC-UMB-2	**0.485**	**0.484**
9.	0.420	0.420
10. UBC-UMB-1	**0.362**	**0.362**
11.	0.355	0.355
12.	0.337	0.337
13.	0.298	0.298
14.	0.120	0.118

Table 5: Official results for all systems in task #17 of SemEval-2007. Our systems are shown in bold. UBC-UMB-1 stands for TS-MCR & TS-XWN & RIC, and UBC-UMB-2 for kNN & TS-MCR & TS-XWN.

System	Precision	Recall
TS-MCR	36.7	36.5
TS-XWN	33.1	32.9
RIC	30.6	30.4
TS-MCR & TS-XWN	37.5	37.3
TS-MCR & TS-XWN & RIC	36.2	36.2

Table 6: Our unsupervised systems in the SemEval-2007 all words test data

the unsupervised systems. However, we cannot attribute the decrease only to the unsupervised features, as the kNN implementations were different (UBC-ALM relied on SVD).

After the gold-standard data was released, we were able to test the contribution of each of the unsupervised systems in the ensemble, as well as two additional combinations. The results are given in Table 6. We can see that TS-MCR is the best performing method, confirming our development experiments (cf. Tables 1 and 2). In contrast, including RIC decreased the performance by 0.7 percent points, and had we used only TS-MCR and TS-XWN our results would have been better.

5 Conclusions

In this submission we combined heterogeneous unsupervised algorithms to obtain competitive performance without relying on training data. However, due to time constraints, we were only able to submit a preliminary system, and some of the unsupervised methods were not properly developed and tested.

For future work we plan to properly test these methods, and deploy other unsupervised algorithms. We also plan to explore more sophisticated combination strategies, using meta-learning to try to predict which features of each word make a certain WSD system succeed (or fail).

Acknowledgements

The first and second authors were supported by Australian Research Council grant no. DP0663879. We want to thank German Rigau from the University of the Basque Country for kindly providing access to the MCR.

References

Eneko Agirre and Oier Lopez de Lacalle. 2004. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC)*, pages 1123–6, Lisbon, Portugal.

Eneko Agirre and Oier Lopez de Lacalle. 2007. UBC-ALM: Lexical-Sample and All-Words tasks. In *Proceedings of SemEval-2007 (forthcoming)*, Prague, Czech Republic.

Samuel Brody, Roberto Navigli, and Mirella Lapata. 2006. Ensemble methods for unsupervised WSD. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 97–104, Sydney, Australia.

Montse Cuadros and German Rigau. 2006. Quality assessment of large scale knowledge resources. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP-06)*, pages 534–41, Sydney, Australia.

David Martinez, Eneko Agirre, and Xinglong Wang. 2006. Word relatives in context for word sense disambiguation. In *Proceedings of the 2006 Australasian Language Technology Workshop*, pages 42–50, Sydney, Australia.

Mark Stevenson and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–49.