Self-training and co-training in biomedical word sense disambiguation

Antonio Jimeno-Yepes
National Library of Medicine
8600 Rockville Pike
Bethesda, 20894, MD, USA
antonio.jimeno@gmail.com

Alan R. Aronson
National Library of Medicine
8600 Rockville Pike
Bethesda, 20894, MD, USA
alan@nlm.nih.gov

Abstract

Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, attempting to select the proper sense of ambiguous words. Due to the scarcity of training data, semi-supervised learning, which profits from seed annotated examples and a large set of unlabeled data, are worth researching. We present preliminary results of two semi-supervised learning algorithms on biomedical word sense disambiguation. Both methods add relevant unlabeled examples to the training set, and optimal parameters are similar for each ambiguous word.

1 Introduction

Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, attempting to select the proper sense of ambiguous words. Supervised learning achieves better performance compared to other WSD approaches (Jimeno-Yepes et al., 2011). Manual annotation requires a large level of human effort whereas there is a large quantity of unlabeled data. Our work follows (Mihalcea, 2004) but is applied to the biomedical domain; it relies on two semi-supervised learning algorithms.

We have performed experiments of semi-supervised learning for word sense disambiguation in the biomedical domain. In the following section, we present the evaluated algorithms. Then, we present preliminary results for self-training and co-training, which show a modest improvement with a common set-up of the algorithms for the evaluated ambiguous words.

2 Methods

For self-training we use the definition by (Clark et al., 2003): "a tagger that is retrained on its own labeled cache on each round". The classifier is trained on the available training data which is then used to label the unlabeled examples from which the ones with enough prediction confidence are selected and added to the training set. The process is repeated for a number of predefined iterations. Co-training (Blum and Mitchell, 1998) uses several classifiers trained on independent views of the same instances. These classifiers are then used to label the unlabeled set, and from this newly annotated data set the annotations with higher prediction probability are selected. These newly labeled examples are added to the training set and the process is repeated for a number of iterations. Both bootstrapping algorithms produce an enlarged training data set.

Co-training requires two independent views on the same data set. As first view, we use the context around the ambiguous word. As second view, we use the MEDLINE MeSH indexing available from PubMed which is obtained by human assignment of MeSH heading based on their full-text articles.

Methods are evaluated with the accuracy measure on the MSH WSD set built automatically using MeSH indexing from MEDLINE (Jimeno-Yepes et al., 2011) [1] in which senses are denoted by UMLS concept identifiers. To avoid any bias derived from

[1] Available from: http://wsd.nlm.nih.gov/collaboration.shtml

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 182–183,
Portland, Oregon, USA, June 23-24, 2011.

the indexing of the UMLS concept related to the ambiguous word, the concept has been removed from the MeSH indexing of the recovered citations.

10-fold cross validation using Naïve Bayes (NB) has been used to compare both views which achieve similar accuracy (0.9386 context text, 0.9317 MeSH indexing) while the combined view achieves even better accuracy (0.9491).

In both algorithms a set of parameters is used: the number of iterations (1-10), the size of the pool of unlabeled examples (100, 500, 1000) and the growth rate or number of unlabeled examples which are selected to be added to the training set (1, 10, 20, 50, 100).

3 Results and discussion

Results shown in Table 1 have been obtained from 21 ambiguous words which achieved lower performance in a preliminary cross-validation study. Each ambiguous word has around 2 candidate senses with 100 examples for each sense. We have split the examples for each ambiguous word into $2/3$ for training and $1/3$ for test.

The baseline is NB trained and tested using this split. Semi-supervised algorithms use this split, but the training data is enlarged with selected unlabeled examples. Self-training and the baseline use the combined views while co-training relies on two NB classifiers, each trained on one view of the training data. Even though we are willing to evaluate other classifiers, NB was selected for this exploratory work since it is fast and space efficient. Unlabeled examples are MEDLINE citations which contain the ambiguous word and MeSH heading terms. Any mention of MeSH heading related to the ambiguous word has been removed. Optimal parameters were selected, and average accuracy is shown in Table 1.

Method	Accuracy
Baseline	0.8594
Self-training	0.8763 (1.93%)
Co-training	0.8759 (1.88%)

Table 1: Accuracy for the baseline, self-training and co-training

Both semi-supervised algorithms show a modest improvement on the baseline which is a bit higher for self-training. Best results are achieved with a small number of iterations (< 5), a small growth rate (1-10) and a pool of unlabeled data over 100 instances. Noise affects the performance with a larger number of iterations, which after an initial increase, shows a steep decrease in accuracy. Small growth rate ensures a smoothed increase in accuracy. A larger growth rate adds more noise after each iteration. A larger pool of unlabeled data offers a larger set of candidate unlabeled examples to choose from at a higher computational cost.

4 Conclusions and Future work

Preliminary results show a modest improvement on the baseline classifier. This means that the semi-supervised algorithms have identified relevant disambiguated instances to be added to the training set.

We plan to evaluate the performance of these algorithms on all the ambiguous words available in the MSH WSD set. In addition, since the results have shown that performance decreases rapidly after few iterations, we would like to further explore smoothing techniques applied to bootstrapping algorithms and the effect on classifiers other than NB.

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine, administered by ORISE.

References

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

S. Clark, J.R. Curran, and M. Osborne. 2003. Bootstrapping POS taggers using unlabelled data. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 49–55. Association for Computational Linguistics.

A. Jimeno-Yepes, B.T. McInnes, and A.R. Aronson. 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation(accepted). *BMC bioinformatics*.

R. Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*.