Kernel Methods for Word Sense Disambiguation and Acronym Expansion

Mahesh Joshi Ted Pedersen[*] Richard Maclin
Department of Computer Science
University of Minnesota
Duluth, MN, USA 55812
{joshi031,tpederse,rmaclin}@d.umn.edu

Serguei Pakhomov
Division of Biomedical Informatics
Mayo College of Medicine
Rochester, MN, USA 55905
Pakhomov.Serguei@mayo.edu

Abstract

The scarcity of manually labeled data for supervised machine learning methods presents a significant limitation on their ability to acquire knowledge. The use of kernels in Support Vector Machines (SVMs) provides an excellent mechanism to introduce prior knowledge into the SVM learners, such as by using unlabeled text or existing ontologies as additional knowledge sources. Our aim is to develop three kernels – one that makes use of knowledge derived from unlabeled text, the second using semantic knowledge from ontologies, and finally a third, additive kernel consisting of the first two kernels – and study their effect on the tasks of word sense disambiguation and automatic expansion of ambiguous acronyms.

Introduction

Word sense disambiguation (WSD) is the task of assigning the correct meaning to a polysemous word based on the context in which the word occurs. The correct sense is selected from a sense inventory (i.e., a known set of possible meanings for the polysemous word). Automatic acronym expansion is a special case of WSD where an ambiguous acronym is to be assigned the correct expansion based on its context. Here onward, we use the word *disambiguation* to refer to WSD as well as acronym expansion. Most popular approaches to these problems make use of supervised machine learning methods, which require a set of manually labeled or sense-tagged training instances of the word or acronym to be disambiguated. The amount of labeled data required to generate a robust model (which is accurate and generalizes well) using any learning algorithm is usually quite large, the lack of which imposes a significant limitation on the knowledge that a learning algorithm can acquire. This is the so-called knowledge acquisition bottleneck.

A large corpus of unlabeled text is easier to obtain as compared to labeled data. Also, semantic ontologies like WordNet (Fellbaum 1998) contain a large amount of manually crafted information and can be useful in WSD. Both these sources of knowledge can be utilized effectively for the task of disambiguation by making use of kernel methods for Support Vector Machines (SVMs). Previously, (Mihalcea & Moldovan 1999) have demonstrated that combining the information from scarce labeled data, the semantic knowledge from an ontology and context available from unlabeled text, an augmented set of features can be provided to the learner to improve its performance.

Methodology

Supervised Approaches: The supervised machine learning approach to disambiguation focuses on two aspects: (i) better learning algorithms; and (ii) better features via feature engineering for the task under consideration. A variety of learning algorithms have been proposed in the machine learning literature and many of them have been shown to perform comparably for the task of disambiguation. Among the ones shown to perform well are the naïve Bayes learner, decision trees, maximum entropy classifiers and SVMs. With regard to feature engineering, various features such as bag-of-words (sometimes restricted to a fixed window around the ambiguous word or acronym), collocations, Part-of-Speech tags, syntactic features and semantic relationships from an ontology have been explored and found useful.

Kernel Methods: The aim of kernel methods for disambiguation is to utilize resources such as unlabeled text and semantic ontologies to enrich the SVM learner with knowledge about senses and their distinction, thus reducing the reliance on labeled data to a minimum. The kernels that we have developed using unlabeled text are based on the latent semantic kernel approach by (Cristianini, Shawe-Taylor, & Lodhi 2001) and the domain kernel approach by (Gliozzo, Giuliano, & Strapparava 2005). The primary motivation for this is the availability of large corpora containing unlabeled instances of the ambiguous words and acronyms in our datasets.

Experiments and Results

Medical Terms: (Joshi, Pedersen, & Maclin 2005) have experimented with the U.S. National Library of Medicine (NLM) WSD collection, which consists of 50 terms that represent ambiguous concepts from the Unified Medical Language System (UMLS). Among the algorithms that were

[*]Dr. Pedersen has been partially supported in carrying out this research by a National Science Foundation Faculty Early CAREER Development Award (#0092784).

used, SVMs proved to be the best. The average accuracy of SVMs across all the tests run was 76.26%. Bag-of-word features with a feature selection criterion of frequency cutoff of 4 gave the best results on average. In our results mentioned below, we have used these results as the baseline for calculating accuracy improvements using text-based kernels.

We now present our preliminary results using kernels created from the unlabeled text that is available with the NLM WSD collection. We selected 11 of the 50 words from the NLM collection, which have a fair distribution of senses. The selection criteria were: (i) majority sense should not exceed 75% and (ii) the *None of the above* sense should not exceed 25%. In the best case, average accuracy across all the 11 words improved from 71.64% to 74.62%, using unigrams as features from the labeled as well as unlabeled data. There was no clear trend co-relating the accuracy obtained using text-based kernels with the amount of unlabeled data available for a given word. In some of the experiments accuracy was found to degrade for some words that have more unlabeled data, whereas accuracy was boosted by as much as 12 percentage points for the word *mosaic* which has the smallest amount of training data among all the words. In general, words that have a highly balanced sense distribution in the labeled data with only 2 senses that are almost equally dominant, showed the best improvement using text-based kernels.

Acronyms in Clinical Notes: Previous work by (Pakhomov 2002) and (Pakhomov, Pedersen, & Chute 2005) involves disambiguation of acronyms in the clinical notes database of the Mayo Clinic and automatic generation of training data from the corpus of clinical notes for the task of acronym disambiguation, the idea being that an acronym referring to some expansion and that expansion itself will tend to occur in similar contexts. To establish a baseline for our kernel experiments, we have performed a comparative study of the naïve Bayes classifier, decision trees and SVMs on the 8-acronym dataset used by (Pakhomov, Pedersen, & Chute 2005) using features specific to the clinical notes such as the gender code of the patient, section identifier of the clinical note and department code where the clinical note originated. The bag-of-word and collocation features employed in these experiments were based on a *flexible window* approach. A flexible window of size 5 uses five significant features (content words that meet other feature selection criteria) on each side of the ambiguous acronym, irrespective of their actual distance from the acronym but within the boundary of a clinical note. Unigrams and bigrams with a flexible window of 2 combined with Part-of-Speech features and clinical-notes-related features yielded the best results. SVMs with the simple linear kernel were the best performers among the algorithms compared, with an average accuracy of 93.26%.

Ongoing and Future Work

Subsequent to the preliminary results outlined above, we are actively working on tuning the features used for creating our text-based kernels. We plan to develop kernels based on knowledge that can be obtained from ontologies such as WordNet for general English and UMLS for medical terms and concepts. This approach has been successfully used for the task of text categorization in (Basili, Cammisa, & Moschitti 2005).

We then plan to explore a dynamically weighted combination kernel that is made out of the first two kernels, with higher weight being given to the kernel that represents the knowledge source more relevant to the contexts under consideration. The relevance of a knowledge source can be decided by a simple overlap of the words in contexts with those available in the knowledge source.

We are also working on the feature engineering aspect of our disambiguation tasks. The flexible window approach and the use of clinical-notes-related features represent our work in this direction so far. Further, we plan to use each individual word in all the known expansions of an acronym as a special bag-of-word feature with higher weight. We believe that this will improve performance for acronym expansion as there is a good chance that some individual word of the true expansion of an acronym occurs in its context. Finally, we would like to use the features in close vicinity of *all* the occurrences of the ambiguous word or acronym within the same context as special features having a higher weight than features at more distance.

References

Basili, R.; Cammisa, M.; and Moschitti, A. 2005. Effective use of WordNet semantics via kernel-based learning. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 1–8.

Cristianini, N.; Shawe-Taylor, J.; and Lodhi, H. 2001. Latent semantic kernels. In *Proceedings of 18th International Conference on Machine Learning (ICML'01)*, 66–73.

Fellbaum, C. 1998. Wordnet, an electronic lexical database. MIT Press.

Gliozzo, A.; Giuliano, C.; and Strapparava, C. 2005. Domain kernels for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 403–410.

Joshi, M.; Pedersen, T.; and Maclin, R. 2005. A comparative study of support vector machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI'05)*, 3449–3468.

Mihalcea, R., and Moldovan, D. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI'99)*, 461–466.

Pakhomov, S.; Pedersen, T.; and Chute, C. 2005. Abbreviation and acronym disambiguation in clinical discourse. In *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA'05)*, 589–593.

Pakhomov, S. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *40th Meeting of the Association for Computational Linguistics (ACL'02)*, 160–167.