Tejashree Khot and Sanjana Agarwal

ASSIGNMENT 6 - MACHINE LEARNING IN COMPUTATIONAL LINGUISTICS

Literature Review

There are several papers that talk about the supervised, semi-supervised and unsupervised machine learning approaches applied to Word Sense Disambiguation. It uses the SVM binary classifier to differentiate between samples into either true or false category. SVM is first adapted to multicast classification and then converted into binary classification problem. In the semi-supervised approach they’ve used algorithms like Yarowsky and Bilingual bootstrapping methods. However, in conclusion, supervised methods like decision trees, naive bayes and particularly SVMs give better accuracy and run in less amount of time(eg. Lokesh Nandanwar, Kalyani Mamulkar, 2013). According to Yoong Keok Lee and Hwee Tou Ng, 2002, experiments were conducted on different datasets (SENSEVAL-1 and SENSEVAL-2) using various types of features for multiple classifiers. They found that combining all types of features, ie. Part-of-Speech (POS) of neighboring words, single words in the surrounding context, local collocations and syntactic relations gives best results. Also, it is found that among SVM, Adaboost, Naive Bayes and Decision Trees, SVM achieve better accuracy. Moreover, local collocations features contribute the most for SVM. As said by Yu Niu et al. (2005), supervised sense disambiguation requires a lot of manually sense-tagged data. In this paper, the use label propagation algorithm based on global consistency assumption like bootstrapping. It not only achieves better performance than SVM and bootstrapping methods when less labeled data is available but also doesn’t require external resources. Moreover, an entropy based method to automatically identify a distance measure that can boost the performance of LP algorithm on a given dataset.

In a paper presented by Roberto Navigli and Paola Velardi (2005), structural or syntactic pattern recognition has proven to be effective in other fields and is hence applied for WSD. For these objects, a representation based on a “flat” vector of features causes a loss of information that negatively impacts on classification performances. Word senses clearly fall under the category of objects that are better described through a set of structured features which is done through knowledge of grammar. It returns each sense choice with weight representing the confidence of the system in its output. Therefore, it could be tuned for high precision (possibly low recall), an asset that we consider more realistic for practical WSD applications. Currently, their system is tuned noun disambiguation. Extending it is to other parts of speech is simple.

Some have tried to approach the problems of semi-supervised methods (eg. Cuong Le et al. (2006)). In this paper they’ve showed two problems of semi-supervised learning for self-training algorithm. Thus, they’ve built a new bootstrapping algorithm with several variants based on the problems. Like, to determine a subset of new labeled examples at each extension of labeled data (the first problem), they used Naive Bayes classifier, and SVM to decrease error rate of new labeled examples. And to determine how to generate the final classifier when the process of extending labeled data is completed (the second problem), they used two strategies of classifier combination including median and max rules to utilize both advantages of the last classifier (built based on the extended labeled data) and the initial supervised classifier. Thus, Naive Bayes gave an improved accuracy over SVM for about 1.9% and the proposed solutions were effective for improving semi-supervised learning.

Furthermore, Xuri Tang et al. (2010), proposes a semi-supervised approach for WSD in Word-Class based selectional preferences. There are four types of formalization models employed: Word model, Class model, Class-only model, Word-class model. The approach exploits syntagmatic semantic redundancy (among all possible sense collocations for a word collocation, the most appropriate is the one in which senses exhibit the most redundant information between each other) and paradigmatic semantic redundancy (among all possible sense collocations for a word collocation, the most appropriate is the one which is also implicitly or explicitly expressed by other synonymous, metonymic or metaphorical word collocations) in the semantic system and uses association computation and minimum description length for the task of WSD. The experiments show that the approach proposed is fairly encouraging in disambiguation of polysemous predicates, especially under semi-supervised conditions when a small portion of data is annotated.

Innovative approaches like feature clustering with semi-supervised algorithms was also done in the past. Zheng-Yu Niu et al. (2005) presents one such approach. They proposed a semisupervised feature clustering algorithm. It can deal with both seen and unseen features in feature clustering process. For the problem, they’ve considered a total of n ambiguous words. The first l words are the labeled words and the remaining words are unlabeled. Thus, with the help of a semi-supervised algorithm they’ve predicted the sense of the ambiguous word by the use of the label information and similarity information among the words. The experimental results on SENSEVAL-3 data showed that feature clustering aggressively reduced the dimensionality of feature space while still maintaining state of the art sense disambiguation accuracy. Furthermore, when combined with a semi-supervised WSD algorithm, semi-supervised feature clustering outperformed supervised feature clustering and other dimensionality reduction techniques. Additional experiments on sampled SENSEVAL-3 data indicated that the semi-supervised feature clustering method is robust to the noise in small labeled data, which achieved better performance than supervised feature clustering.

Composite kernel method and improvements in the same have been remarkably shown by two such papers. Guiliano et al. (2009) define the composite kernel to combine and extend the individual kernels representing syntagmatic and domain aspects of sense distinction. This method achieves state-of-the-art performance while requiring less labeled training data compared to its contemporary systems. It can be further improved by using syntactic information produced by a parser which could be implemented by adding a tree kernel. With a little advancement of using Kernel PCA method, Wiefeng Su et al. (2004) proposed a new semi-supervised model that takes advantage of unlabeled data, along with a composite model that combines both the supervised and semi-supervised models. Their baseline WSD model is a supervised learning model that makes use of Kernel Principal Component Analysis (KPCA). They’ve not applied the semi-supervised KPCA model indiscriminately rather have made a small assumption about the prediction confidence of the semi-supervised and the supervised KPCA method (which is called the composite semi-supervised KPCA model). The composite semi-supervised KPCA model improves on the high-performance supervised KPCA model, for both coarse-grained and fined-grained sense distinctions. In the composite model, the supervised KPCA model predicts senses with high confidence for more than 94% of the test instances. Thus, the composite semi-supervised KPCA model exploits unlabeled data to improve upon the accuracy of the supervised KPCA model.

Finally, Dinu, Georgiana, and Sandra Kübler (2007) wrote a paper on Romanian WSD and experiments with features of memory-based classifier. It states that the feature set needs to be tightly controlled and achieved optimal results with on average seven features per word. This occurs because MBL methods are affected by irrelevant or redundant features. However, this could be caused by idiosyncrasies in the Romanian data set or limited size of the training data.