Name: Cuong Anh Le
Supervisor: Prof. Akira Shimazu
Starting Time of Research: April 2004

Progress Research Report

Study on Word Sense Disambiguation with Feature Selection and Bootstrapping techniques

1. Research Purpose

I planned to do my study on Word Sense Disambiguation (WSD), an important problem in Natural Language Processing. Solving the problem brings advanced NLP, which contributes the reliable Internet software development.

WSD is the task of choosing a right sense of an ambiguous word given a context. It is obviously essential for language understanding applications, while also at least helpful for other applications whose aim is not understanding language such as machine translation, information retrieval, among others.

2. Proposed Approach

I focus my effort to resolve two important problems for WSD. The first is known as the problem of feature selection. Its task is to find the knowledge resources, which are useful for WSD, and to extract the best subset of the features representing this knowledge. Second is the problem of using unlabelled data to improve the performance of a WSD system. Bootstrapping, or semi-supervised learning, has become an important topic in computational linguistics. For many language-processing tasks, there is an abundance of unlabelled data, but labeled data is lacking and too expensive to create in large quantities, making bootstrapping techniques desirable. Techniques in this problem play an important role in building a real system, therefore planned to be received much our attention. A framework for building a WSD system has been sketchy designed. Using data including labeled and unlabelled data, the system first strengthens the labeled data by adding to them the predicted instances from unlabelled data and the process is repeated. Finally, obtained labeled data is used to construct a final classifier.

In my plan, a Word Sense Disambiguation system will be constructed for verifying proposed methods, also as the experimental step to build a real system for WSD.

3. Progress of this year

In this year, problems and methods in feature selection and co-training were investigated. We have focused our efforts on finding the useful features for WSD, and obtained the result that was presented in [1]. In this paper, a set of features is selected firstly by adding more rich knowledge to topic context, which is represented by ordered words in a local context and collocations, and then the features are chosen using a forward sequential selection algorithm. With obtained features the NB classifier can achieve higher accuracy in comparison with the best previously published results.

One other field also considered is combining individual classifiers to improve performance of the final decision. Our approach comes from the argument that various ways of using context in WSD can be considered as distinct representations of a polysemous word, and then all these representations are used jointly to identify the meaning of the target word. Under such a consideration, we can then apply some general framework for combining classifiers such product rule, sum rule, min rule, so on. One of our results is presented in [3].

4. Future Work

In next time, we continue the previous work and do further in new fields specially in co-training methods, particularly as follows:

- extract highly confident patterns for WSD
- combining classifiers based on Dempster-Shapfer theory
- methods in combining labeled and unlabelled data
- clustering in WSD

5. Refereed papers

[1] Le, C. A. and Shimazu A., High Word Sense Disambiguation using Naive Bayesian classifier with rich features, *The 18th Pacific Asia Conference on Language, Information and Computation* (2004), pp. 105-113

[2] Le C. A. and Shimazu A., Improving Word Sense Disambiguation accuracy using Naive Bayesian classifier with rich features, *5th International Symposium on Knowledge and Systems Sciences*, Ishikawa 10-12 November (2004), pp.87-91.

[3] Le, C. A., Huynh V-N, and Shimazu A., Combining classifiers with Multiple-Representation of Context in Word Sense Disambiguation, *The Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 05),* Hanoi 5-2005(accepted)