Thesis Title: **A Study of Classifier Combination and Semi-Supervised Learning for Word Sense Disambiguation**

by Le Anh Cuong
Supervisor: Prof. AKIRA SHIMAZU

SUMMARY

Word Sense Disambiguation (WSD) involves the association of a polysemous word in a text or discourse with a particular sense among numerous potential senses of that word. In this thesis, we present a study of classifier combination and semi-supervised learning for WSD. In addition, we also work on context representation and feature selection which play important roles in obtaining high accuracy of WSD task. The thesis consists of six chapters. The first chapter presents an overview of the thesis. The second chapter first presents our survey on the approaches in previous studies, and then presents the three supervised learning algorithms (Naive Bayes, Support Vector Machines, and Maximum Entropy Models) which are used as the basic algorithms in proposed methods. The next three chapters deal with three main tasks of thesis, respectively: context representation and feature selection; classifier combination; and exploiting unlabeled data. Chapter 6 contains the summary of the thesis and future research directions.

The motivations and proposed solutions for the problems mentioned in this thesis are based on investigating and solving the limitations of related WSD studies. Experimental results were conducted on standard datasets (Senseval-2 and Senseval-3) and were compared to state-of-the-art systems in the field. The major contributions of the thesis are summarized as follows:

- The first contribution comes from the work on context representation and feature selection. The proposed method for knowledge sources determination (i.e. context representation) is based on the simultaneously use of the Forward Sequential Selection and Backward Sequential Selection algorithms. After obtaining the selected knowledge sources, we applied a filter method with using two feature measures, frequency and information-gain, to select useful individual features.

- The second contribution comes from the work on combining classifiers for WSD. In this work, two new approaches of classifier combination for WSD, which are based on Dempster-Shafer theory of evidence and OWA operators have been presented. Various combination rules were derived. The second-layer combination strategies were proposed including meta combination and meta-stacking. Various tests corresponding to various combination models were conducted and the obtained results show that meta-voting in meta combination strategies gives the best result and much improves accuracy of individual classifiers.

- The third contribution comes from the work on semi-supervised learning for WSD, that helps reduce the need of labeled training data by gaining additional information from a large amount of unlabeled data. We followed the approach in which the original labeled data is iteratively extended from unlabeled data.

Two particular bootstrapping algorithms investigated are self-training and co-training. We first identify problems which may occurring in this approach and then proposed solutions for them. As the result, a new bootstrapping algorithm with several variants are generated. With the new algorithm, unlabeled data is shown effective in improving WSD. Furthermore, a novel combination model for post semi-supervised learning has been proposed, which aims to combine advantages from classifier combination and semi-supervised learning. Experimental result of this model reaches the state-of-the-art of WSD systems.