CHAPTER 6

CONCLUSION AND FUTURE WORKS

6.1 Conclusion

Word sense disambiguation is an enabling technology. In this thesis, we focus on Thai word sense disambiguation research. Improvements in word sense disambiguation performance will lead to improvements in machine translation, Cross Language Information Retrieval, Information Retrieval.

This thesis presents the development of methodology of word sense disambiguation in Thai by using purely knowledge lean unsupervised learning techniques that do not rely on any knowledge intensive resources like sense–tagged text or dictionaries. However, we use sense-tagged corpus in the evaluation phase only to evaluate the maximum accuracy of discovered sense groups when we perform the experiments with test data.

We use co-occurrence feature as a feature vector in the context representation. We use co-occurrence within a small window as the local context with a window size ± 2 that is non-overlapping windows. We compare the thesis approach results of English word and Thai words หัว /hua4/ which is a noun and เก็บ /kep1/ which is a verb respectively with each word baseline. It shows that our approach can outperform the majority baseline system.

In this thesis, we use Singular Value Decomposition (SVD) which is the dimensionality reduction technique. SVD can capture conceptual similarities among the contexts and don't just rely on the surface forms of words in the text when performed on a large sample of text. Noise also can be removed by trimming the least important dimensions of the subspace.

The thesis approach is a novel since the work focuses on Thai language which has not received much previous attention in the Natural Language Processing literature. There were some research works such as Schütze's work (Schütze, 1998) presented his approach that can work with English word. Although our method which is different and similar with his research works, we want to support and verify that the research approach can also work well for English words. We use English word in the experiments to show that our method can give acceptable performance for English word.

6.2 Future Works

There are many issues that arose during this thesis that suggest future directions for research. These include ideas to improve existing techniques, as well as some new variations that might lead to better disambiguation. Our methodology is suitable to a broad range of problems that extends well beyond word sense disambiguation. What follows are our plans for future work.

1. We will employ different richer feature types such as collocation words or dependency word to result in richer context representation.

2. In order to test the performance of the method on naturally ambiguous words, a large number of instances have to be disambiguated by hand. As this is a time consuming task, it is convenient to generate artificially ambiguous words: pseudowords (Gaustad, 2001). A pseudoword is the union of two or more natural words. The pseudowords are composed of two pairs from different topics. For example, two or more words, e.g., *banana* and *door*, are conflated into a new type: *banana/door*. All occurrences of either word in the corpus are then replaced by the new type. The evaluation of disambiguation performance for pseudowords can be made easily since one can go back to the original text to decide whether a correct decision was made.

3. What we did in the experiments is that we use Singular Value Decomposition (SVD) to reduce the dimension of word co–occurrence matrix. The reduced matrix which is used in this thesis is 300 columns. We have not yet conducted the experiments to test the impact of results if the dimension sizes of matrix are reduced to vary numbers of columns.

4. To determine the effect of using different reduction factors such as *Independent Component Analysis* (ICA) which are similar method to SVD (Hyvärinen and Oja, 2000). ICA is a statistical formalism that takes higher-order dependencies into account. By assuming independence, ICA is capable of detecting a set of hidden vectors if only different linear mixtures of these vectors are observable. If we look at the task of word sense induction, our starting point is that we can consider the co-occurrence vector of an ambiguous word as a linear mixture of its unknown sense vectors. If corpora from different domains are available, this should give us the different linear mixtures that are required for ICA.

5. In this thesis, we use co-occurrence within a small window. The extended work can be done by using large window, topics and paragraphs. We can use co-occurrence within a maximal distance of 50 words in each direction was considered. Apart from this, we can examine on varying word side such as 2, 5, and 10 on each side in order to determine which vector reflects the best semantics.

6. In this thesis, the experimental results are based on the experimental setting of sample size of test dataset. We select arbitrary the number of test data size. If the test data size is larger size, the test data will have sentences which cover every possible sense to be tested. It will yield better result. We can conduct the experiments on variation the number of test data size to test the impact of result accuracy.

7. This thesis approach takes explicitly specified to create the number of clusters which are 20 clusters and they are from manual inspections of cluster results and are not fully automatic. The determining of optimal number of clusters which are automatically derived by the algorithm can be considered in the area of future work.

8. The predefined threshold value is defined to measure the cosine similarity between the context vector of test data and sense vector. The similarity measure will determine context vector of test data should be assigned to which sense cluster. In our experiments, we use arbitrarily a cosine threshold of 0.5 since we assume that it will not effect to any cluster assignment as we count this threshold of 0.5 is the average of cosine value. More experiments with vary threshold values should be further conducted to determine which threshold value yields the best result.