CHAPTER 6

CONCLUSION AND FUTURE WORK

The focus of this thesis is the design of machine learning based techniques for sentence-level Paraphrase discovery from large scale corpora. In order to meet this objective, various approaches for Paraphrase Recognition have been proposed and investigated. Further a Fuzzy Hierarchical Clustering method has been proposed for Paraphrase Extraction. This chapter presents the major conclusions drawn from the research and also gives directions for future work.

6.1 CONCLUSION

The recognition of semantically similar content forms an essential part of Natural Language Processing systems. The challenging nature of Paraphrase discovery has motivated both Academic and Industry majors to undertake extensive research in this domain. The design of efficient systems for Paraphrase Recognition and Extraction constitute the major research goals of this thesis. The first contribution of this research is the design of machine-learning based approaches for sentence-level Paraphrase Recognition. Four different approaches have been investigated, two of which utilize features extracted from the original input text, while the other two rely on intermediate representations. In the first approach, a Radial Basis Function Neural Network has been designed to classify the input sentence pair based on lexical, syntactic and semantic features extracted from it. The approach was found to have only a marginal improvement over a Support Vector machine based

recognizer. The investigations have indicated that the set of features used has a significant impact on the performance of the Paraphrase Recognition system. Hence a Genetic Algorithm based feature selection strategy was used to improve the performance of the SVM based PR system. Lexical features, dependency tree based features and semantic features such as noun, verb similarity were found to be best suited for the Paraphrase Recognition task. Another inference drawn from the study was that using additional knowledge in the form of a table of equivalent phrases improves the performance of the Paraphrase Recognizer.

The third approach for Paraphrase Recognition has employed UNL as an intermediate representation. The choice of UNL was motivated by the fact that it would be suitable for cross-language inputs. Features extracted by matching of UNL words and relations were used as input to an SVM Classifier for detecting paraphrases. The fourth Paraphrase Recognition system employs Predicate Argument Structures as an intermediate representation which has facilitated matching of semantic roles. The two stage system works by first segregating input sentence pairs into various categories based on Predicate Argument matching. In the second stage, different features have been extracted from each category to classify the sentence pair. Of the intermediate representation oriented PR systems, the approach using UNL has lower performance than all the other proposed approaches due to the simple feature set used. On the other hand, the system using the Predicate Argument matching approach was found to yield the best result on standard paraphrase corpora. The better performance can be attributed to matching of semantic roles followed by the usage of surface-level features.

The next major contribution of this research is the design of an efficient approach for Paraphrase extraction from large scale corpora. The method works in two stages and involves Fuzzy Hierarchical Clustering

followed by paraphrase recognition. Clustering has been carried out in two phases with grouping of sentences containing same or similar verbs followed by splitting of resultant clusters based on nouns. This technique was found to yield promising results when compared to traditional k-means and FCM clustering approaches. Additionally, the paraphrase recognition system has been applied in Student Answer Evaluation and plagiarism detection tasks. Experimental results have proved that the PR system is effective in both tasks. Finally a News Clustering system has also been designed using the Paraphrase Extraction technique and has resulted in satisfactory performance.

6.2 FUTURE WORK

A potential area for future work is the study of cognitive processes for establishing semantic similarity and the design of paraphrase recognition systems that simulate such processes. Another possible direction includes the development of wholly unsupervised, efficient techniques for paraphrase recognition / extraction. Finally the applicability of the proposed approaches in tasks such as multi-document summarization and Information Extraction can also be explored.