CHAPTER 7

CONCLUSIONS AND FUTURE WORK

This chapter concludes the thesis by summarizing the works, findings and contributions of the thesis. It also presents some directions of future research work.

7.1 Summary of Contributions

In this thesis, we surveyed the area of MWEs extraction and identification in several languages and proposed several new association measures based on the *n*-grams probabilities. Our experiment on the corpus data showed that the system with bigram MWEs yield significantly better result compared to single method when used for MWEs extraction and detection. The main reason is that word-normalization or word alignment helps to shrink the word gaps between *n*-grams of different lengths, making a long *n*-gram more comparable with short *n*-grams for the word-association values. As a result, we select more meaningful MWEs and at the same time reduce the unwanted ones. In addition, hybrid approach also work well to bring some significant improvement in our experiment. However, due to the limited Multiword Expressions related resources, we have applied four feature criteria for optimizing the parameter values in the corpus. More feature selection criteria will be our future work to explore more efficient way of tri-gram and four gram words using *n*-gram probabilities. In addition, we eliminated stop word at the two ends of multiword expressions, which leads to a huge boost in the performance for all experimental results of our implemented Automated Multiword Expressions detection technique.

We further applied Hybrid approach for Multiword Expressions extraction and identification for summarizing a text file containing Multiword Expressions either from a corpus or from any web page. Such a process is necessary since a corpus may not be well structured and words may be scattered. Using features of Multiword, we are able to produce abstraction of MWEs identification from a corpus. In our experiment, we found that Multiword Expressions are suitable information units that can not only capture meaningful contents but also preserve MWEs concepts in a separate form.

In this work, new method has been developed for Multiword extraction and detection. The development of these methods for automatic identification of multiword from a corpus or from a text file contributes in an efficient way to the area of Natural Language Processing.

Each of the proposed method includes the Part-of-Speech tagging throughout the corpus, words selection, words weighting in terms of bigram, trigram etc. For each experiment, the configurations of the proposed method is presented and corresponding results are given. Also, the analysis of the experimental results and the comparison between different experiments have been covered.

We have reached most of the expected contributions as proposed initially, particularly the following essential objectives have been achieved:

1. Development of Bengali annoted corpus.
2. Development of Automated Multiword Expressions detection system in Bengali.

As described in chapter 3 of this thesis, most important parts of the text can be automatically detected using Multiword descriptions. In particular, we use frequencies of *n*-gram in order to find multiword expressions.

The comparison result as shown in Table 6.3, we used Hybrid method we used to find the result of evaluation for the Multiword Expressions identification. At the same time we discussed the worst MWEs detection if features are not matched in the model.

7.2 Future Work for Automated Multiword Expressions Detection

We indicate some possible ways and ideas to extend this work which are as follows

1. The proposed method was tested for Bengali corpus, which can be tested with other language corpus to affirm the preliminary conclusions.
2. Extending Parallel corpora for Multiword Expression detections.
3. Testing more words as bigram, trigram for multiple detection.
4. Extracting Multiword Expressions in order to apply them in other NLP applications.

5. To test our system using bilingual dataset by collecting bilingual corpus.
6. To use IPA symbol of Bengali to enhanced the research in Bengali MWEs