CHAPTER 8

Conclusions

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

8.1 Summary of Works

The present work investigated the effectiveness of various POS tagging algorithms and computational linguistics approaches from a natural language processing perspective. The present study began with the obtainment of a general overview of the part of speech tagging and its different paradigms and approaches. A study has been done on some prominent tagging approaches like HMM, CRF model, Rule based approach on their applications in part of speech tagging. It is learned that statistical techniques were more successful than rule-based methods, but there are many constraints to adopt the statistical methods for Manipuri POS tagging. A large tagged corpus is required to develop a statistical POS tagger with high accuracy rate. In contrast, the computational works in Manipuri is still in the infant stage and Manipuri has no enough tagged corpus to implement the same in statistical method. Moreover, the rule based tagger has many advantages over statistical tagger, including a vast reduction in stored information, the perspicuity of a small set of meaningful rules, ease of finding and implementing improvements to the tagger. Based on the handcrafted linguistics rules and affix stripping algorithms a rule-based Manipuri Part of Speech Tagger is implemented. All these works are presented in the various chapters as follows:

Chapter 2 presents brief review of the prior work in part of speech tagging however our focus has been made on part of speech taggers of Indian languages.

In chapter 3 an overview of part of speech tagging and its different paradigms and approaches are presented. Some applications of part of speech tagging in the field of computational linguistics are also presented.

Chapter 4 represents a brief description of Manipuri including the people who speak the language and geographical location. It also presents the typological features of Manipuri in its linguistics perspectives.

In chapter 5 the morphosyntactic categories in Manipuri are described elaborately. It also presents the discussion of tagsets of various languages and development Manipuri Tagset based on ILPOST framework with a little customization to meet the morphosyntactic requirements of the language.

Chapter 6 begins with general definition of computational morphology. It then presents a discussion on roots and affixes of Manipuri and three major word formation processes viz., Affixation, Compounding and Derivation. This chapter also discusses different algorithms of affix stripping techniques and proposed a new affix stripping algorithm. Further, the chapter presents some experimental results.

Chapter 7 presents an overview of rule-based part of speech tagging. It then presents the proposed algorithm of rule based part of speech tagger of Manipuri and features of the graphical user interface POS tagger tool which is developed by applying Manipuri linguistics rules. Further, the chapter presents some experimental results.

8.2 Summary of Contributions

The main contributions of this thesis include development of a tagset for Manipuri based on ILPOST framework and it has been customized for Manipuri to meet the morphosyntactic requirements of the language, development of a morpheme segmenter by using an affix stripping algorithm and development of rule based Manipuri POS tagging tool "POSTIM" to aid researchers working in the area of computational linguistics to tag Manipuri lexical items with proper morphosyntactic categories and attain high accuracy level. Other important contributions of the thesis include the studies and analysis of various POS tagging algorithms and computational linguistics approaches. Finally, it includes studies and analysis of Linguistics rules of Manipuri.

8.3 Future Directions

In this thesis a rule-based Manipuri Part of Speech Tagger is implemented as part of the larger goal of computational analysis of Manipuri language. The tagged output of the tagger can be used as corpus in analysis of Manipuri language by applying many statistical methods like unigram, bigram and HMM model etc. This tagger gives a good accuracy rate of tagging Manipuri lexical items excluding MWE and named entity.

The future work would be to design a tagging model by hybridization of rule-based and statistical method to attain better accuracy rate using the Bureau of Indian Standards (BIS) POS tagset that has been prepared for the Indian Languages by the POS Tag Standardization Committee of Department of Information Technology (DIT), New Delhi, India. The design should be enabled to detect MWE and named entity recognition in tagging Manipuri text.