CHAPTER 1

Introduction

The thesis contributes to the subject area of part of speech tagging, a preliminary and important component of computational linguistics or natural language processing. It is focused on the study and analysis of rule based part of speech tagging in Manipuri from a computational linguistics perspective.

Part of speech tagging is the process of assigning morphosyntactic categories of each lexical item including punctuation marks in a given text according to the context. It is an indispensable part of natural language processing. It plays a vital role in developing of any serious applications in processing all the natural languages in the world. Part of speech tagging is an initial stage of linguistics text analysis like sub-category acquisition, information retrieval, machine translation, text to speech synthesis etc [79].

Computational linguistics is an interdisciplinary field dealing with the rule based or the statistical modeling of natural language from a computational perspective. Human languages are generally known to as natural languages; the science of studying natural languages falls under the area of linguistics and its implementation using computational means is regarded as computational linguistics. Computational linguistics has theoretical and applied components, where theoretical computational linguistics takes up issues in theoretical linguistics and cognitive science, and applied computational linguistics focuses on the practical technique for the realization of linguistic theory to facilitate real-world applications.

There is a very limited research works in the field of computational linguistics in Manipuri, an agglutinating Tibeto-Burman language of North-East India. It is very difficult to classify the lexical categories of Manipuri, because most of the root words found in the language is bound and the affixes i.e. prefixes and suffixes are the determining factor of the class of the words in the language. Due to the typical features of the language, it is required to analyze the language from its linguistics perspective in developing an effective part of speech tagger of Manipuri.

1.1 Motivation

Many works related to POS tagging has been done in languages like English, Japanese, Chinese, German and Arabic etc. Several POS tagger of these languages are developed by using different algorithms. For instance, English language has developed POS tagger using rule based, statistical method, neural network and transformational based method etc [47]. In the year 1992 Eric Brill has been developed a rule based POS tagger with the accuracy rate of 95-99% [10]. POS tagging of some languages like Turkish [60], Czech [34] has been attempted using a combination of hand-crafted rules and statistical learning.

Similarly, Indian languages like Hindi, Bengali, Punjabi, Tamil, Telegu and Malayalam languages have many POS taggers but the effort is still in its infant stage. There is a few works related to POS tagging in Manipuri and other Tibeto-Burman languages in the Indian sub-continent.

Apart from being necessary for further language analysis, Manipuri part of speech tagging is of interest due to a number of language processing applications like speech synthesis and recognition, information retrieval and

extraction, partial parsing, word sense disambiguation and machine translation etc. Part-of-speech gives key information about the word and its neighbours which can be useful in a language model for different speech and natural language processing applications. Development of POS tagger for Manipuri will also influence several pipelined modules of natural language understanding system of the language. It can be learned from the existing POS tagging techniques that development of an effective POS tagger with a high accuracy rate requires either developing a comprehensive set of handcrafted linguistics rules or a large amount of annotated text. We have the following observations.

- Rule based POS taggers use hand written linguistics rules to assign tags to unknown or ambiguous words. Although, the rule based system allows the design of an extremely accurate system, it is expensive and difficult to develop a rule based POS tagger.

- Stochastic POS taggers use a large amount of annotated data or tagged corpus for the development of a POS tagger in shorter time.

- However, no tagged corpus in Manipuri is available for the development of a stochastic POS tagger.

Therefore, there is a fundamental requirement to develop an automatic part of speech tagger in Manipuri for the overall development of the language in the field of Computational Linguistics. With this motivation, we identify objectives of this thesis.

1.2 Objectives

The primary objective of the thesis can be summarized as follows:

1) To study the paradigms and different approaches to part of speech tagging.
2) To study some computational linguistics approaches and their applications in part of speech tagging.
3) To study and analyse the linguistic features of the proposed language.
4) To develop a tagset for Manipuri based on ILPOST framework with a little customization to meet the morphosyntactic requirements of the language in accordance with language specific and writing conventions followed in Manipuri.
5) To develop a morpheme segmenter, i.e., Manipuri, being an agglutinative language it is required to develop a morpheme segmenter for segmenting the lexical items into its constituent morphemes.
6) To develop a rule-based part of speech tagger for Manipuri. As Manipuri has no tagged corpus the generated tagged output of the rule based tagger can be used as the tagged corpus in other statistical methods. Tagging part of speech to each lexical item manually is time consuming and tedious task.

1.3 Methodology

1) A general overview of part of speech tagging paradigms and approaches was first obtained.
2) Study of some computational techniques of part of speech tagging and their applications in the field of natural language processing.

3) Study of linguistic features of Manipuri and collection of text data manually from newspapers, journals, novels, short-stories, dramas, text-books, dictionaries, word-books, etc., through library works.
4) A tagset for Manipuri based on ILPOST framework has developed.
5) A morpheme segmenter has developed.
6) A rule-based Part of Speech tagger for Manipuri has developed to generate the tagged output with high accuracy level.
7) The results are evaluated by using the principles of formal grammar and linguistic rules of Manipuri.

1.4 Main Contributions

The main contributions of the thesis are:

1. Studies and analysis of various POS tagging algorithms and computational linguistics approaches.
2. Studies and analysis of linguistics rules of Manipuri.
3. 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.
4. Development of a morpheme segmenter by using an affix stripping algorithm.
5. Development of rule based POS tagger for Manipuri.
6. Development of GUI 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.

1.5 Thesis outline

Chapter 2: Review of literature. This chapter presents brief review of the prior works in part of speech tagging and part of speech taggers in Indian languages.

Chapter 3: Part of Speech Tagging. This chapter presents an overview of part of speech tagging, its different paradigms and standard approaches. It also presents some applications of part of speech tagging in the field of computational linguistics.

Chapter 4: A brief overview on Manipuri. This chapter presents a brief description of Manipuri including the people who speak the language and geographical location. It also presents the typological features of Manipuri language.

Chapter 5: Annotation Guidelines in Manipuri lexical items. This chapter describes the morphosyntactic categories in Manipuri. It also presents the discussion of tagsets of various languages and development of Manipuri Tagset based on ILPOST framework with a little customization to meet the morphosyntactic requirements of the language.

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

Chapter 7: Rule based Part of Speech Tagging in Manipuri. This chapter 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. Furthermore, the chapter presents some experimental results.

Chapter 8: Conclusion. Finally, this chapter presents the conclusion. Summary of the works and contributions are outlined along with a discussion on scope for future research work.