CHAPTER 1

INTRODUCTION

The thesis contributes to the subject area of Automated Multiword Expressions Detection in Bengali, a preliminary and important component of computational linguistics or natural language processing. One of the important issues in both natural language understanding and generation is the appropriate processing of Multiword Expressions (MWEs). MWEs create huge problem to the precise language processing due to their idiosyncratic nature and diversity in lexical, syntactical and semantic properties. The semantics of a MWE cannot be expressed after combining the semantics of its constituents. Therefore, there is a great need for extracting MWEs especially for resource constraint languages like Bengali. The present hybrid approach contributes to locate the number of MWEs present in the corpus. In this experiment, we apply the features extraction of Bengali bigram MWEs, though it can be extended to any types of MWEs.

Natural language processing has emerged as an interdisciplinary area in the field of Computational Linguistics. It consist of computational linguistics and artificial Intelligence (AI).The Natural language processing uses some tools like data structures, algorithms, formal models for representing knowledge, models for reasoning processes, etc in the field of Artificial Intelligence. The aim of Natural language processing is to specify a comprehensive language comprehensive and production theory to such a level that a man is able to write a computer program which can understand and produce Natural language. The typical sub-areas of the field of NLP include the specification of Parsing algorithms and the study of their computational properties, acquiring knowledge representation formalisms that can help semantic analysis of the sentences, and model creating for the reasoning processes that can count for the analysis of context affecting the interpretations of sentences (Kumar, 2011).

Natural Language Processing of Indian languages is in its developing stage. Developing an Automated Multiword Expression Detection method will be a very important step for Bengali language as it will help in language processing like information extraction and retrieval, machine translation and Lexicography etc. Due to the typical typological features of the MWEs in different languages, it is required to analyze the language from its linguistics perspective in developing an effective MWEs detection method.

This thesis is about Automated Multiword Expressions Detection and their uses in Natural Language Processing (NLP) applications. Building computer systems capable of dealing with MWEs is a hard and open problem, due to the complex and pervasive nature of these constructions in language. This chapter is a general introduction to our research topic. We motivate and illustrate the importance of MWEs through many examples in English and Bengali languages. Then, we discuss the motivation and scope of the computational framework for Automated MWEs detection in this thesis.

Multiword expressions are those expressions in which two or more than two words form a new meaning by losing their individual meaning of the words (Manning and Schutze 1999). For examples, ‘kick the bucket’ (means die), in Bengali ‘উভয় সঙ্কট’, ‘গৃহপালিত’ are examples of MWEs in which individual words lost their meaning after composition. Due to the idiomatic nature of construction of the MWEs and their high frequency of occurrence in all sorts of text, they create many problems in NLP applications and are responsible for their difficulties in language processing.

Recognizing MWEs, with their degree of their idiomaticity, would be useful in applications like information retrieval, question-answering, summarisation, parsing, language modeling and language generation that require some degree of semantic processing. In this thesis, we investigate the issues of extracting and detecting Bengali MWEs, depending on the various features of MWEs in Bengali Language.

Taking MWEs into account is an important aspect of the output of NLP systems. An MT system needs training to be aware of idiomatic expression like ‘মূষল ধারে বৃষ্টি হইতেছে’, in English ‘it is raining cats and dogs’ to avoid different translations meaning. If MWEs are not translated as a unit, it can generate translation errors using computer system. If MWEs are not handled properly it may generate ungrammatical and unnatural output and create translations errors of idioms and collocations. In some cases, it may cause to confusion and misunderstanding (Churck *et al.*, 1990).

Incorrect MWE processing can lead to errors in syntactic analysis (parsing). Information about MWEs in the sentences can be obtained by looking them up in a lexicon (Seretan, 2013) or by modeling them as features (Seretan, 2011), and can help in disambiguating the parse trees. On the other hand, if MWEs are not identified properly then it will probably reduce the parser's accuracy. In MT and in parsing and other NLP applications, when the words composing MWEs are treated as separate units, the system will probably produce erroneous output.

Finally, if we ignore MWEs and do not treat them properly, important information will be lost. MWEs make sure naturalness to a system's output and they are frequently used in languages, and are hard in texts to be processed.

1.1 Motivation

Looking at languages from a computational perspective, Automated Multiword Expressions detection is one of the hard and open problems, due to the complex nature of MWE constructions in language. It is one of the most important pursuits of NLP in all languages for MWE processing for a given application. Developing Automated Multiword Expression detection with optimum level of accuracy would be significant contributions because it would lead to its use in applications like MWEs acquisition, Linguistic processing of MWEs, Lexicography and Machine translation etc.

Detecting MWEs can be both manual as well as automatic. Manual extracting and detecting, though more accurate but it is a time-consuming, long and continuous process. Hence, the automatic detection of MWEs is essential to speed up the computational processing with less chance of errors and inconsistencies. But in case of MWEs, research works are at a nascent stage and hence the output is not up to expectation due to their idiosyncrasies.

Various automatic MWEs tools have been developed for languages like English, Germany, and Portuguese etc. worldwide and for some Indian languages like Hindi, Tamil, Telugu and Odiya with limited features in their extraction using linguistic rules and stochastic models. Different kinds of MWEs have certain advantages as well as disadvantages due to language barrier. Automated Multiword

Expressions Detection is a challenging task for Indian languages which are highly inflectional and morphologically rich. Hence, the development of Automated Multiword Expressions Detection tool is a challenging task.

Development of Automated Multiword Expressions Detection method will also influence several pipelined modules of natural language understanding system of the language. It can be learnt from the existing Automated MWEs extraction in other language like English with limited features that development of an effective Automated MWEs Detection method with a high accuracy rate entails either developing a comprehensive set of linguistics rules or incorporating extra features in a large amount of annotated text. Proper identification and use of MWEs have proven to be a pain in the neck for NLP, due to lack of adequate available resources such as manually annotated corpora in various languages.

We motivate for the following observations

- To know what are Multiword Expressions in different languages especially in Bengali. Depending on the linguistic features of Bengali, categories of MWEs are studied in details for their acquisition.
- To know the problems of Multiword Expressions in Bengali to make system aware of handling MWEs in Bengali language.
- To use Bengali Multiword Expressions in Natural Language Processing.

Multiword Expressions are commonly used in everyday life. A native speaker does not follow it all the time. MWEs occur frequently in many languages from written to oral, from general to domain specific such as 'good morning', 'never mind' and 'bye bye'.

Researchers in theoretical and computational linguistics evaluated the recurrence of MWEs in a more systematic way. Linguistics studies provide many examples and figures proving how frequently MWEs occur in text collections across different languages and domains (Biber *et al.*, 1999). It is often assumed that a native speakers' lexicon contains as many MWEs as simple words (Jackendoff, 1997). Thus, any computational system dealing with human language must take MWEs into account. The following list presents some NLP tasks and applications. If we don't handle MWEs properly, it will generate ungrammatical and unnatural output (Ramich, 2015).

1. **Morphological and syntactic analysis:** To work on morphological analysis of the word structures of a language, it requires substantial amount of grammatical, linguistic and vocabulary insights of the language in question. So in order to implement the morphological model of a language, understanding linguistic insights of the word structure of a language is an essential task from computational point of view. Syntactic analysis is the analysis of a sentence or other string of words into its constituents, to show their syntactic relation to each other. It contains semantic and other related information. MWEs help in syntactic disambiguation. Incorporation of MWEs features in POS taggers and parsers is a great issue for many NLP applications.

2. **Information Retrieval (IR):** An information retrieval is a process of retrieving requested query that are given into the system as information needs. For example, we search information into the search engine .The search engine gives related information by providing several relevance with the requested search. IR model unable to differentiate idiomatic interpretations of an expression. For example , if we provide search as bank the IR model will gives several information related to financial institution, river bank etc. If the MWE river bank is Index as lexical and semantic units the information retrieval process will improve.

3. **Machine translation (MT):** MT is a system of translating text, articles from one language to another language. In case of MWEs, word to word translation is difficult and required language model to represent them as a unit. Statistical MT help to represent bilingual word sequences as MWEs that contain most of the phrases and Idioms .Thus, MWEs remain a challenging task for automatic translation independently.

4. **Computational semantics:** It deals with how to automate the process of constructing and semantic representations in NLP. Thus, it plays a vital role in

computational linguistics. Most of the NLP applications are required for semantic interpretation. For examples, word sense disambiguation, automatic summarisation, text mining etc. These applications enhanced both semantic compositionality and non-compositionality of MWEs.

5. **Optical character recognition (OCR):** Optical character recognition is an area of computer science that involves reading text from paper and translating the images into a form that the computer can manipulate (like ASCII codes). An OCR system allows us to take a book or a magazine article, supply it directly into an electronic computer file, and then edit the file using a word processor. Also a part of MWEs that can be perform using n-gram language model. Thus MWEs can help to enhanced OCR technology by reducing the length of n-gram language model.

1.2 Objectives

The primary objectives of the thesis can be summarized in following research directions

1. To investigate the Characteristics of Multiword Expressions
2. To detect Multiword Expressions automatically from texts
3. To evaluate the usefulness of a given method of MWE acquisition.
4. To study POS tagging in Bengali for Multiword Expressions identification.
5. To study and analyze the linguistics features (rules) of the proposed language i.e., Bengali for MWEs.
6. To propose and combine generic and portable techniques for automatic MWE acquisition from corpora.
7. To evaluate these techniques by measuring their usefulness in NLP applications.

1.3 Methodology

1) A general overview of Multiword Expressions paradigms and approaches was first obtained.

2) Study of computational techniques in Multiword Expressions Extraction and their applications in the field of Natural Language Processing.

3) Study of linguistic features of Bengali and collection of text data manually from newspapers, bank web sites, journals, novels, short-stories, Bengali literature text book etc., through library works and internet surfing.

4) A tagset for Bengali based on ILPOST, BIS framework has been developed.

5) Freely available Sanchy tool has been used for Bengali Part of Speech Tagging

6) A Hybrid approach of Automated Multiword Expressions detection for Bengali has been developed to generate the MWEs extraction and detection with high accuracy level.

7) The results are evaluated by using the principles of linguistic rules and Features of Bengali MWEs.

1.4 Main Contributions

The main contributions of the thesis are:

1. Studies and analysis of various algorithms for Multiword Expressions and computational linguistics approaches for their acquisition.
2. Studies and analysis of Linguistic rules of Bengali Language.
3. Development of a corpus for Bengali based on ILPOST and BIS framework for Bengali tagset and it has been customized for Bengali to meet the morph syntactic requirements of the language.
4. Development of Bengali annoted corpus.

5. Development of Automated Multiword Expressions Detection tool for Bengali language.

1.5 Thesis outline

Chapter 2: Review of literature

This chapter presents brief review of the prior work in Automated Multiword Expression Detection in English and Indian languages especially issues in Bengali Language.

Chapter 3: Multiword Expressions

This chapter presents an overview on Multiword Expressions, its different paradigms and standard approaches. It describes necessary and sufficient condition for MWEs, type of MWEs, characteristics of MWEs, and classification of MWEs. It also deals with some applications of Multiword Expressions in the field of computational linguistics.

Chapter 4: Part of Speech Tagging in Bengali

This chapter presents an overview of Part of Speech tagging in Bengali, its different paradigms and standard approaches. It also describes important guidelines and framework for developing Part of Speech Tagset in Bengali.

Tagset development is a prerequisite for natural language processing of any language. It is also a pre-processing step in Multiword extraction and detection process for which proper tagging of the corpus is required. Thus, we can say prior to the development of a tagged corpus features extraction of MWEs can be enhanced.

Chapter 5: Methodological framework for Multiword Expressions acquisition

This chapter describes methodological framework for developing Automated Multiword Expressions Detection with system architecture. It describes various modules that can be chained together in several ways.

It presents corpus preparation considering tokenization, candidate selection, filtering based on N-gram Model that is whether a word is bigram or trigram in the corpus. Thus, we filter MWEs in bigram and trigram based on the combination of NC-NC, NP-VP, NP-VA and NC-NC-NC, NC-NC-VM etc. respectively. It also describes frequency of occurrences of MWEs using statistical Co-occurrence tests.

Chapter 6: Multiword Expressions: Features Extraction and Detection

This chapter presents an overview of features extraction from the corpus based on the given features in TnT Model. If the features are matched, system gives higher accuracy of system performance which is based on Precision (P), Recall(R) and F-measure (F). Based on the available features, we can build a new classifier for a given language. After the extraction, MWEs are detected in which word form, lemma, or word sense are categories they are and list of MWEs are shown separately. This chapter also describes rule based approach that we use for detection method.

Chapter 7: Conclusion

This final chapter summarizes the research work that is carried out in this thesis. It also discusses contributions of the research work outlined. Moreover, it presents a handful of open issues related with previous chapters and proposes several directions for future work.

1.6 Chapter Summary

In this chapter, an introduction of the thesis, motivation to do the work, objectives, methodology, contributions in this topic and discuss thesis outline is briefly. Moreover, it presents a handful of open issues related with previous chapters and proposes several directions for future work.