Abstract

Natural Language Processing (NLP) in Bengali is in its developing stage. NLP has significant overlap in the field of Computational Linguistics, and is often considered a sub field of artificial intelligence. Natural language generation systems convert information from computer data base into readable human (natural) language. In NLP, the techniques are developed which aim the computer to understand the command given in natural language and perform according to it. Different government organizations, universities and institutions like IIT Bombay[1], Central Institute of Indian Languages (CIIL), Mysore, C-DAC, Bangalore, Hyderabad University, Hyderabad etc. are working in natural language processing of Indian languages including Bengali to bring up in a common international platform of human language processing and to overcome the language barrier among different linguistics communities throughout the globe[2], Antony (2013); Antony (2011). Moreover, NLP is system through which man can interact with machine in their mother tongue. According to the researchers, Automatic Multiword Expression Detection is one of the very important and indispensible parts of a Natural Language Processing applications. Developing a Automatic Multiword Expression Detection method is a challenging and highly complicated task due to the idiomatic nature of the Multiword Expressions (MWEs) in the languages. Most of the phrase and idioms of any languages are fall under Multiword Expressions.

Multiword expressions are expressions consisting of two or more words that correspond to some conventional way of saying things Manning (2007). Multiword Expressions (MWEs) are lexical items that consist of multiple orthographic words (e.g., ad hoc, by and large, New York, 'kick the bucket' (means die), in Bengali 'উভয় সঙ্কট', 'গৃহপালিত', খাওয়া দাওয়া and মিট মিট etc. Due to idiomatic nature of MWEs they cause problems in Natural Language Processing (NLP) applications and are responsible for the shortcomings. If semantic, idiomaticity of MWEs can be handled properly, it will be useful in NLP applications areas such as summarisation, question-answering, parsing, language modeling and language generation, Information Retrieval (IR) and

[1] ***http://www.cfilt.iitb.ac.in/***

[2] ***http://www.ethnologue.com/language/mni***

Machine Translation (MT).

Morphologically, some MWEs allow some of their constituents to occur inflections freely while restricting (or preventing) the inflection of other constituents. In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation.

Syntactically, some MWEs behave like phrases while the others are phrases or words occur in fixed order while the others permit various syntactic transformations. Semantically, the compositionality of MWEs is gradual, ranging from fully compositional to fully idiomatic Bannard *et al*., (2003). In Bengali Multiword Expressions has various features namely Noun+Noun (compound noun), Noun+Verb (conjunct verb) and Verb+Verb (compound verb) etc.

A detailed review of the literature on previous related work on multiword expressions in different languages along with Bengali MWEs around different types of approaches has been carried out in chapter 2. In particular, we started with reviewing methods and approaches for extraction and detection of MWEs and Multiword terms i.e. domain-specific MWEs. Methods for MWEs were classified as rule based method, statistical method and hybrid method respectively. A summary of the reviewed literature has been presented in this chapter.

In this thesis we investigate whether the occurrence of Multiword Expressions is domain specific or not, their idiosyncratic nature, characteristics, types, word co-occurrences, features extraction, MWEs identification and their detection method. We present a general methodology for detecting Multiword Expressions (in Bengali) from a corpus.

Corpus contains text data in Bengali language. The collection of written text data is done from corpora from historical background, banking sector, News Paper[3], Phrases and Idioms from various English and Bengali Books from different sources, shopping mall web sites etc. in which some MWEs are of popular use in our real life that contain a wide variety of texts corresponding to most common language use over a given time span. We considered word token to be an occurrence of a word in the corpus. We then tagged the corpus using statistical Part of Speech(POS) tagger model focused to find word level like word counts, words collocation and concordances(like

[3]**https://www.anandabazar.com/**

index)and words fixation. Based on these MWEs are extracted and indentified using linguistic feature of the Multiwords or using statistical approaches. We developed an algorithm that proposes MWEs identification and detection in Bengali language. We used a large monolingual corpus to rank and filter these text corpus i.e. to find association measure of words occurrences.

In the present work, a tagset for Bengali language based on ILPOST (Indian Language Part of Speech Tagset) framework has been developed as a part of the present study. It has been designed fulfill the morph syntactic necessities of the language in accordance with language specific and writing conventions followed in Bengali. The Tagset consist of 14 categories of POS (based on CIIL standard), types and attributes. The statistical POS approach is used for tagging the corpus.

An overview of the system architecture of MWEs extraction and detection is presented in chapter 5. A brief introduction TnT model for MWEs extraction, data preparation, various MWEs extraction approaches are described in this chapter.

The research work has been initiated with a study on Multiword Expressions on its paradigms and various approaches, namely Linguistic Approaches, Statistical Approaches, Hybrid Approaches of MWEs extraction Weiwei (2012). Furthermore, a computational model with four categories of features namely Noun-Noun, Noun-Verb, Reduplication and Idiomatic Compound Noun with different n-grams (mostly bigrams and trigram in lesser number) has been developed for features extraction from the tagged corpus. Since Hybrid method is more accurate than other methods as it combine linguistic and statistical method to extract MWEs (Frantzi *et al*., 2000), a graphical user interface (GUI) tool named 'AMWED' has been developed by applying hybrid approach of MWEs extraction and detection, using Net Beans IDE 7.3, JDK 6 and JRE 6. The front-end of the tool has been implemented in java and its interface is connected with a text file of Bengali lexical items called "lexicon" as the back-end. The selection of textual database is for simplicity and to extend support for multiple platforms without the need of the installation of any DBMS server like MYSQL etc. by the end user are discussed in chapter 6. All the experimental results presented in this thesis were obtained by using 'AMWED'. It gives an average accuracy of 83.25% and from this result it is obvious that with an increase in the

number of features in the model and increasing the size of the corpus, the percentage of accuracy and efficiency would be further increased.

The summary and conclusions drawn from the present work together with some future works arising from the research is presented in Chapter 7.

References

Antony P.J. and Soman, K.P. (2011). *Parts Of Speech Tagging for Indian Languages: A Literature Survey.* International Journal of Computer Applications (0975 – 8887).Volume 34– No.8, November 2011.

Antony P.J. (2013). *Machine Translation Approaches and Survey for Indian Languages. Computational Linguistics and Chinese Language Processing.* Vol. 18, No. 1, March 2013, pp. 47-78.

Bannard, C., Baldwin, T., & Lascarides, A. (2003). *A statistical approach to the semantics of verb-particles.* In proceedings of the ACL 2003 workshop on Multiword expressions, (pp. 65–72)., Morristown, NJ, USA. Association for Computational Linguistics.

Frantzi, K, S. Ananiadou, and H. Mima. 2000. Automatic Recognition of Multi-word term: the C-value/NC-value Method. International Journal on Digital Libraries, 3[2]:115–130, August.

Manning, C. & Schutze, H. (1999). Foundations of Statistical Natural Language Processing, chapter 5: Collocations. MIT Press.

Weiwei Huo (2012). Automatic Multi-word Term Extraction and its Application to Web-page Summarization,page-43, Guelph, Ontario, Canada.