Research Paper Analysis:

Some Techniques Used for Processing Bengali Corpus to Meet New Demands of Linguistics and Language Technology

-By Niladri Sekhar Dash

Guilherme D. Garcia

Ball State University

RESEARCH PAPER ANALYSIS

Bengali / Bangla is one of the seventh most widely spoken languages with around 230 million native speakers and 37 million second language speakers worldwide. Bangla is unique and diverse in its characteristics and grammatical construction, even in phonetic variations. Having this diverse linguistics features and rich morphology background, Bengali is still lacking the most important resource for language engineering due to its unavailability and enough access opportunities electronically. Bengali language corpus has hardly been developed in recent days. A few attempts have been taken, some progress have also been made in recent past but that's not well enough. Before getting into the analysis, I would like to provide a brief overview of corpus linguistics.

A corpus is defined as a collection of transcribed speech or written text compiled mainly to enhance the linguistic research and the study of corpus is corpus linguistics. It is the study of language as expressed in corpora of "real world text". Corpus is considered as the best and most useful resource to study different linguistic phenomena like morphological structure, syntactic structure, phonological variation etc. This corpus has immense importance in Natural Language Processing, so as in Artificial Intelligence. Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. However, this is now the biggest field for research.

The language which has more developed corpus, is mainly researched more and developed more. In this paper, the author discussed about some new techniques in analyzing corpus in terms of Bengali language and showed how it worked providing relevant examples and demonstrated the implications too.

RESEARCH PAPER ANALYSIS

After the development of the Bengali text corpus (Dash 2005) electronically, linguists started using it in different ways for text processing. But to process these tests, then we needed techniques. Still then no special text processing techniques for Bengali has been developed. So, the researchers had to depend on the universal test processing techniques, the same situation is still on march. This is what the author talked about throughout his paper. Today researchers use the corpus to conduct research electronically but it was not the same always because everything had to be manually done. Now these method can run automatically on corpus to obtain data and information required for designing systems for language technology, Dash (2007) stated. However these techniques helped us to open up new avenues to language research and applications which were unknown before.
The corpus processing techniques are the outcomes of close interface developed between language database and computer. Now these techniques are allowing us to analyze text corpora from newer perspectives and to shed lights on the language properties found in corpus.

This development opened more scopes for the linguists to research on Bengali language. Many of our intuition based observation about our language and its properties have been changed because those directly contradict with the new observations made after the processing of Bengali corpus.

New Corpus Processing Techniques:

Frequency refers to the number of occurrences or hits. If a word, phrase tag etc has a frequency of 10, it means it is found 10 times or it exists 10 times. It’s an absolute figure not calculated using any specific formula. The author explained the importance of frequency in text processing. He mentioned that it is important in understanding of language

RESEARCH PAPER ANALYSIS

and the developing systems. However, knowledge of frequency is also important in language teaching. The most frequently used words are used to train the second language learners as well as to test their competency level. In usage based model of language learning, it has been argued that "information about the frequency of use of various linguistic items has direct effect on language education (Johns 1991)".

The frequency list is generally demonstrated in two two different modes. One is Alphabetically and other is numerically. If you consider a large population of sample like of billions, it is likely to get the accurate frequency count from there in numerical orders. Depending on the target users, this frequency count can also be arranged in ascending and descending order. "The frequency list provides us the rudimentary ideas about the basic structure of a language to plan our future course of investigation and analysis accordingly(Dash 2007)."

Concordance is one of the most useful processing techniques which allows us to display the total list of occurrences in its own contextual environment. It helps us to get the understanding of different syntactic, semantic, lexical patterns of words as well as genre and type of a text. This is more flexible to use.

Lexical Collocation is the method that helps us to evaluate the value of consecutive occurrences of any two words in a piece of text. To what extent the actual patterns of lexical occurrence differ from the patterns that have been expected, is analyzed with this technique. For example in Bengali, kācā is an adj, means "raw", in English, found to be associated with more than 30 different words to denote equal number of collocation and sense variation. *kācā māch* "raw fish", *kācā ghar* "mud house", *kācā lok* "novice" etc. So with reference to this contexts of use of words in a piece of text, we can empirically determine which pairs of words maintain substantial collocation relation between them. This methods supports the

RESEARCH PAPER ANALYSIS

claim that “mental lexicon is made up not only with single words but also with larger multiword units- both fixed and variable”

Key-Word-In-Context, abbreviated as KWIC is mainly another form of concordance. The main difference is realized by noticing the central point of attention. In concordance, the central point is “the target word”, where in KWIC “environment” takes the place. This is more filtered version of concordance.

Local Word Grouping (LWG) is unlikely concordance and KWIC. “It is aimed at throwing lights on patterns of use of words, idioms, phrases and other language properties from different perspective”.

Lemmatization of words: Lemma is a positional attribute, the basic form of word, typically the form found in dictionaries. Lemmatization is a process of assigning a lemma to each word form in a corpus using an automatic tool called a lemmatizer.
All derived forms of a verb can be assembled under a single group to link up with lemma. For example- “make”- makes, made, making etc. This saves times of a researcher and makes the task easier. Otherwise we got lost in billions of corpora.

Parsing Sentence is a kind of annotation, operated after the grammatical annotation. It involves syntactic analysis of sentences collected in corpus in accordance with the grammar of a language. This is done completely automatically or sometimes by partial manual assistance or combining both. This exhibits the syntactic functions and relations with other constituents in the sentence. This is a very good visualization system but in case of Bengali language, the system however doesn’t yield good outputs, since normal Bangla sentences don’t adhere to the model grammar used for English or other languages, the author stated.

RESEARCH PAPER ANALYSIS

Most of the corpus processing techniques available have been designed for English, German, Dutch, Portuguese, French etc. Though they are modeled as universal techniques, but they don't work the same in any particular language. At least we can say that in term of Bengali language after this research. Since they don't yield to expected results due to certain technical problems related to Bengali orthography and text samples, all the techniques need strategic modification before introducing as Bengali text processing techniques. In essence, I'm chiming in my voice with Bosh that it has been realized that unless these are converted to an acceptable standard, blind application of these technique on the Bengali corpus will yield wrong results to tarnish the actual image of the language, which is not acceptable in any sense.

References

Dash,S.N.(2007).Some Techniques Used for Processing Bengali Corpus to Meet New Demands of Linguistics and Language Technology. SKASE *Journal of Theoretical Linguistics, 4(1).*

Dash, S.N.(2005). *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications.

Johns,T.(1991). Should you be persuaded: two samples of data-driven learning materials. *LR Journal E 4*, 1-16.