Analysis of Bangla Root Word for Universal Networking Language (UNL)

Aloke Kumar Saha  
Department of CSE,  
University of Asia Pacific,  
Dhaka, Bangladesh

M. F Mridha  
Department of CSE,  
University of Asia Pacific,  
Dhaka, Bangladesh

Jugal Krishna Das  
Department of CSE,  
Jahangirnagar University,  
Savar, Dhaka, Bangladesh

ABSTRACT

Root Word (RW) Analysis Approach for Universal Words (UWs) in Bangla Language is a technique which helps to find out more correct universal words more accurately and quickly than other available techniques. The main problem with universal word creation is that, as Bangla sentences varies semantically depending on the use of the same root in different sentences, it is important to get accurate universal words or universal expressions according to this variation of the sentence structure. The length of the sentence also plays an important role in determining the performance of the output. If length of the sentence is large enough and more complex, it will be problematic for the parser to make specified language which may not imply accurate meaning as required. In this paper a new method, termed as Root word Analysis Approach for converting Bangla to UNL has been investigated Universal Word construction. Root word Analysis Approach begins generation in a constructive fashion by adding some rule/criterion one after another which also shows how to use different meaning of root words using dictionary according to root words in different sentences.

Keywords

Bangla Root Word, Morphology, Bangla Suffix, Bengali Cases and their inflexions and UNL.

1. INTRODUCTION

Bangla is spoken by about 245 million people of Bangladesh and two states of India, but most of the computer based resources and technical journals are in English. Due to the language barrier, the common people face big obstacle to enjoy the optimum benefits of modern information and communication technology (ICT) as well as huge enriched English knowledge database around the globe. The Universal Networking Language (UNL), which is a formal language for symbolizing the sense of natural language sentences, is a specification for the exchange of information. The main aim of the UNL project is to overcome language barrier. Currently, the UNL includes 16 languages, which are the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish), in addition to the ten other widely spoken languages (German, Hindi, Italian, Indonesian, Japanese, Latvian, Mongol, Portuguese, Swahili and Thai). Bengali is not yet included. In the last few years, machine translation techniques have been applied to web environments. The growing amount of available multilingual information on the Internet and the Internet users has led to a justifiable interest on this area. Hundreds of millions of people of almost all levels of education, attitudes and different jobs all over the world use the Internet for different purposes [1], where English is the main language of the Internet. Because English is not understandable for most of the people, Interlingua translation programs are needed to develop. The main goal of the UNL system, which allows users to visualize websites in their native languages, is to provide a common representation for accessing Internet of multilingual websites by the majority of the people over the world. For this common representation, lexical knowledge is a critical issue in natural language processing systems, where the development of large-scale lexica with specific formats capable of being used by distinguished applications, in particular to multilingual systems, has been given special focus. Our goal is to include Bengali in this system with less effort.

In this paper we present the Root Word analysis of Bangla Sentences for UNL system. The major components of our research works touches upon i) development of grammatical attributes for Bangla root word to construct Bangla Word Dictionary and use of morphological analysis ii) UNL Expression of the Bangla attributes and iii) Bangla sentences analysis. In section 2 we describe the Bangla Sentence structure. In sections 3 and 4, we present our main works that include all the above three components.

2. BANGLA SENTENCE

STRUCTURE

2.1 Basic structure differences

Direct conversion of English sentences into Bengali sentences very often suffers from different kinds of difficulties and ambiguities. For this, conversion is done with the help of the inner significance of the sentence rather than direct word-to-word conversion. In these cases, at first, the purpose of the sentence is understood and then the sentence is interpreted according to that purpose [9][10]. It can be done by proceeding in a systematic way with the help of logical algorithm which is much complex. So direct mapping between the structure of English and Bengali sentences required for computer aided solution. But for many sentences, this mapping is not possible. As for example, we may consider the following sentence:

Please lend me your ears. -Its Bengali meaning is  “আমার শুনে শুনুন।”

In this example, the English word “lend” means “শুনে শুনুন” but this is not present in Bengali meaning rather the Bengali word “শুনুন” is present whose English synonym is “hear” which is not present in the English sentence at all.
2.2 Difference between the Forms of Verbs

In English there are four basic forms of verbs (present, past, continuous and past participle), for example: “go” verb has different forms like: go, went, going and gone. But in Bengali there are a lot of forms of a verb, which differ according to tense, person and number. So it needs to develop a system that is able to find out the appropriate form of verb.

3.2 Bangla Primary Suffixes (রূপান্তর)

We know that the core of the verb is called root and if number of suffixes are added to roots then they form verbs. When sound or sounds [8] are added with roots and form nouns or adjectives then the root words are called root verbs and the sound or sounds are added with root verbs are called Primary Suffixes. For example চল (Root verb)+বাহিরে (Primary Suffix)=চলবাহিরে (Noun) and চল (Root verb)+আসন (Primary Suffix)=চলআসন (Adjective). Some others primary suffixes are মাটি, অন্না, ক্ষুদ্র, অন্না, আ তাত্ত্বিক আ।

3.3 General Prefixes

General Prefixes are the words that are used before words to express various meanings of the same words. There are around fifty (50) prefixes used in Bengali sentences. In Shangskrit Bengali we use twenty (20) prefixes[2] say, (prefix). Some others primary suffixes are মাটি, অন্না, ক্ষুদ্র, অন্না, আ তাত্ত্বিক আ।

3.4 UNL Structure of Bengali Case

As in the traditional understanding case (in Bengali Kaarok) in Bengali denotes the relationship of the nominals with the main verb of the clause. The cases are broadly classified into 6 categories, each having a finer categorization into sub-types. The traditional case with their inflexions is shown in table.
Table 1 Bengali Cases and their inflexions

<table>
<thead>
<tr>
<th>Classical Case vs. Bengali Kaarok</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative case (Kartri Kaarok)</td>
<td>0 (null), ২(ে),৩(য), ৪(০)</td>
<td>রা(রা), প্রথা(প্রথা)</td>
</tr>
<tr>
<td>Accusative case (Karma Kaarok)</td>
<td>০(null), ২(েকে), ৩(েকে), ৪(েরে)</td>
<td>ফিকে(দিগোকে)</td>
</tr>
<tr>
<td>Instrumental case (Karan Kaarok)</td>
<td>দারা(দারা), ২(দিয়া), ৩করী (কর্তিক)</td>
<td>দারা(দারা), ২(দিয়া), ৩করী(কর্তিক)</td>
</tr>
<tr>
<td>Dative case (Sampradaan Kaarok)</td>
<td>০(null), ২(েকে), ৩(েকে), ৪(েকে), ৫(েকে)</td>
<td>ফিকে(দিগোকে)</td>
</tr>
<tr>
<td>Ablative case (Apaadana Kaarok)</td>
<td>ইরে(ইরে), ২(ইরে), ৩(ইরে)</td>
<td>ইরে(ইরে), ২(ইরে), ৩(ইরে)</td>
</tr>
<tr>
<td>Genitive case (Sambandha pad)</td>
<td>রা(রা), প্রথা(প্রথা)</td>
<td>দারা(দারা), দিয়া(দিয়া)</td>
</tr>
<tr>
<td>Case of time-place (Adhikaran Kaarak)</td>
<td>২(ে), ৩(য), ৪(০)</td>
<td>রা(রা), প্রথা(প্রথা)</td>
</tr>
</tbody>
</table>

Table summarizes the correspondence between Bengali cases and the UNL relations.

Table 2 Bengali cases versus UNL Relations

<table>
<thead>
<tr>
<th>Classical Case</th>
<th>Corresponding UNL relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kartri Kaarok</td>
<td>agt, cag, ptn, aoj, cao, and</td>
</tr>
<tr>
<td>Karma Kaarok</td>
<td>obj, ben, cob</td>
</tr>
<tr>
<td>Karan Kaarok</td>
<td>ins, met</td>
</tr>
<tr>
<td>Sampradaan Kaarok</td>
<td>ben, gol, pur, rsn</td>
</tr>
<tr>
<td>Apaadana Kaarok</td>
<td>frm, src, plf, tmf</td>
</tr>
<tr>
<td>Sambandha pad</td>
<td>mod, pos, pof</td>
</tr>
<tr>
<td>Adhikaran Kaarok</td>
<td>plc, plt, tim, tnt, to, gol, scr, opl</td>
</tr>
</tbody>
</table>

3.5 Grammar based Analysis

We can Analysis root word কর(কর), ধর(ধর), পড়(পড়), পড়া(পড়া), চল(চল), থেক(থেক), লেখ(লেখ) using these following rules that discuss above.

Table 3 Bengali root word কর (কর)

<table>
<thead>
<tr>
<th>Tense</th>
<th>Language</th>
<th>First person</th>
<th>Second person</th>
<th>Third person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Present</td>
<td>Shadhu (শাদু)</td>
<td>করে+ই+ম</td>
<td>করে+ই+ম</td>
<td>করে+ই+ম</td>
</tr>
<tr>
<td></td>
<td>Suffix-0</td>
<td>Kor+i</td>
<td>Kor+i</td>
<td>Kor+i</td>
</tr>
<tr>
<td>Cholito (চোলিত)</td>
<td>Suffix-0</td>
<td>করে+ই+ম</td>
<td>করে+ই+ম</td>
<td>করে+ই+ম</td>
</tr>
<tr>
<td></td>
<td>Suffix-0</td>
<td>Kor+i</td>
<td>Kor+i</td>
<td>Kor+i</td>
</tr>
<tr>
<td>Continuo</td>
<td>Shadhu</td>
<td>করি তেছে</td>
<td>করি তেছে</td>
<td>করি তেছে</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kor+i</td>
<td>Kor+i</td>
<td>Kor+i</td>
</tr>
</tbody>
</table>

4. BANGLE TO UNL ENCONVERSION

An En-Converter is a software that automatically or interactively en-converts natural language text into UNL. As it generates UNL from natural languages, enables peoples to make UNL documents without any knowledge about UNL. It means that users of the UNL system do not need to learn UNL. En-Converter is a language independent parser that provides synchronously a framework for morphological, syntactic and semantic analysis. It would be impossible to solve an ambiguity in morphological analysis without the use of syntactic or semantic information. Also, it would be impossible to solve an ambiguity in syntactic analysis without the use of semantic information. En-Converter generates UNL expressions from sentences (or lists of words of sentences) of a Bengali language by applying en-conversion rules. In addition to the fundamental function of en-conversion, it checks the formats of rules, and outputs the messages for any errors. It also outputs the information required for each stage of en-conversion in different levels. With these facilities, a rule developer can easily develop and improve rules by using En-Converter. En-Converter loads the en-conversion rules and the rule checker works while converting rules. Once the rules are made, they are stored automatically and can be used directly the next time without rule conversion.

1. Convert or load the rules.
2. Secondly, it inputs a string or a list of morphemes/words of a sentence of Bengali language.
3. Input a Bengali sentence.

Then, it starts to apply rules to the Node-list from the initial state. En-Converter applies en-conversion rules to the Node list. The process of rule application is to find a suitable rule and to take actions or operate on the Node-list in order to create a syntactic functionalities and UNL network using the nodes in the Analysis Windows. If a string appears in a window, the system will retrieve the Word Dictionary and apply the rule to the candidates of word entries. In this case, if a word satisfies the conditions required for the window of a rule, this word is selected and the rule application succeeds. This process will be continued until the syntactic functions and UNL network are completed and only the entry node remains in the Node-list [s7].

a. Apply the rules and retrieve the Word Dictionary Finally, it outputs the UNL network (Node-net) to the output file in the binary relation format of UNL expression.

b. Output the UNL

Output the UNL expressions with the exception of the first process of rule conversion and loading, once En-Converter starts to work, it will repeat the other processes for all input sentences.
Example: Consider the following simple English sentence “It is very hot in Dhaka today”.

The corresponding Bengali text and its word by word meaning in English is given as,

ঢাকায় আজ খুফ গরম - Dhaka-in today very hot

The node list is enclosed within “<<<<” and “>>>>”. The analysis window is enclosed within “[“ and “]].” The nodes delimited by “/” are those explored and fixed by the system. Morphological analysis is carried out in a series of steps.

1. /<</[ঢাকা]/[আজ]/“আজ খুফ গরম”/>>/
2. /<</[ঢাকা]/“আজ খুফ গরম”/>>/
3. /<</[ঢাকা]/আজ/[খুফ]/[গরম]/>>/
4. /<</[ঢাকা]/[আজ]/[গরম]/>>/
5. /<</[গরম]/>>/
6. /<</[গরম]/>>/

Step 1: The noun ঢাকা (Dhaka) of the type PLACE and case maker থেকে (in) are combined and an attribute PLC is added to the noun to indicate that plc relation can be made between the main predicate of the sentence and this noun.

Step 2: In this step, the system right shifts here because there is no combination or modification rule between noun and adverb.

Step 3: The system recognizes গরম (hot) as a predicate of the sentence. So, it generates man relation between the adverb খুফ (very) and the predicate গরম (hot). The adverb খুফ (very) is deleted afterwards.

Step 4: Then the analyzer looks ahead further right beyond the noun phrase আজ (today) to get the predicate গরম (hot). A tim relation is created between আজ (today), and গরম (hot) and finally, আজ (today) is deleted.

Step 5: By using PLACE attribute of the noun ঢাকা (Dhaka), the system generates plc relation between the noun ঢাকা and the predicate গরম.

Step 6: A right shift at this point brings the Sentence Tail (STAIL) under the LAW and thus signals the end of analysis. This right shift rules also attaches the attribute @entry to the last word left in the Node-list and thus the predicate গরম (hot) is preserved till the end.

The UNL output of the corresponding sentence is as follows:

```
unl
man (hot (icl>state),@entry,@present, very (intensifier))
tim(hot(icl>state),@entry,@present, today(icl>period))
plc (hot (icl>state),@entry,@present, Dhaka (icl>place))
[/unl]
```

5. CONCLUSION

Root word Analysis Approach targets to improve the universal word creation ability of Bangla language. A new approach proposed in this thesis. The new technique uses a constructive approach to the universal language of Bangla language to determine. The novelty of this method is that it is simple and easy approach is used to modify the properties and diversified use of words on proposals for a given word that is to be determined. This analysis functions initially tried to solve this problem for a number of proposals for example, the approach required for semantically correct rate for the true meaning of giving.

6. REFERENCES