Algorithm for Natural Language Processing:
A Bengali Language Perspective

Md. Ahsan Arif, Md. Mobarak Hossain, Arif Tanvir

Assistant Professor, Lecturer
Dept. of Computer Science and Engineering
Asian University of Bangladesh, Dhaka-1230, Bangladesh

ABSTRACT

Presently, the use of native language through Internet and any other media is highly appreciable. For the people of Bangladesh and the West Bengal of India, it is important to read all information in Bengali through Internet or other media (like any conference, important foreign meeting or any other means of communication abroad). The Universal Networking Language (UNL) has been asserted on the communication issue in different languages now. In the UNL system, there are 16 languages successfully translated to English language but the Bengali sentences are not appropriately translated to the UNL expressions. Therefore, it is a crying need for translating Bengali sentences to the UNL expressions firstly. In this paper, I have an urge to review the various research works that has been carried out in the recent past years in different areas of Bengali Language Processing (BLP) in Bangladesh, which has pointed out the challenges and recommendations to overcome the problem ahead of successful implementation of the BLP. I have proposed some ideas of the computational analysis of some structures of Bengali sentences and conversion of Bengali sentences to the UNL expressions. I have provided here a format of a Bengali UNL Dictionary and added each Bengali word with English conversion. I also added English Grammar like tense, number, subject, object, verb, adjective, noun, pronoun form of each word picked from the database oriented dictionary. In this connection, I adopted each and every form of verb according to the tense and defined as the number, the type of sentences, person, subject, object, auxiliary verbs etc. in the dictionary for the required verb. I have also added every kind of word in Bengali (adjective, adverb, group of Bengali words of which English conversion is expressed by mono spelled word) and the phrases, which cannot be translated by any grammatical rule. The rules of Bengali and English Grammar and en-conversion rules of Bengali sentences are considered in my paper as proposed methodology. I developed an algorithm that is according to the analysis of above and also implemented this algorithm to get better output. The algorithm is applied on the number of sentences within the paragraphs. I have taken some examples of implementation of the analysis as some case studies.

Keywords: Universal Networking Language, UNL, Bengali Language Processing, Bengali Grammatical Rules, Bengali Word Dictionary, Natural Language Processing, Bengali to English Conversion.

1. INTRODUCTION

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. Moreover tens of millions of people of almost all levels of education and attitudes of different jobs all over the world use the Internet for different purposes. The Internet has to face the complexity of multilingualism. English is the main language of the Internet. But English is not understandable for most of the people. So, it is needed to develop Interlingua translation program.

The source language is transformed into a language independent representation, which can be transformed into the target language in the Interlingua program. As a consequence, United Nation University/ Institute of Advanced Studies were decided to develop an inter-language translation program. They started a research to lead a common form of languages known as Universal Networking Language (UNL) [1]. The UNL project is concerned about developing an intermediate form computer semantic language where any written text in a particular language can be converted to any other form of language. The strength of the UNL system lies in the
fact that it emphasizes to represent the semantics of a native language sentence ignoring the complexities of natural language. The en-converter converts each native language sentence to a UNL document and de-converter translates the UNL document to any native language. The UNL document is itself in English as it is known to linguistics. The development of native language specific components dictionary and analysis rules is carried out by researchers across the world. Until now, the UNL project includes 16 official languages, but bangle is not yet included.

2. FINDINGS FROM SCHOLAR’S (PREVIOUS OUTCOMES)

We studied lots of scholar’s research regarding sentence conversion from one language to another language using UNL expressions. After studying we found some findings which are shown as below:

According to this example:- ‘রাহিম বইটি পড়েছেন’ In some research paper we have seen that they break the word “পড়েছেন” as “পড়” and “ইেতেছেন”. [2] At first, how can we define the root word is “পড়”, it may be break as “পড়ি” and “ইেতেছে” because “পড়ি” also a root word at their dictionary. And they provide a structure for “ইেতেছে” at dictionary but they did not provide any ‘HW’ for “ইেতেছে”. So how can it possible to get English word for “পড়েছেন” from dictionary. If we entry the Bengali word “পড়েছেন” directly in dictionary and in this way if we entry every word like this word it is more easy to convert Bengali sentence to English according our algorithm. Because we easily identify person, tense, gender when we read this word by using our human intelligence. So we can us this human formula like as when we read “পড়েছেন”, we can easily understand that this word is in present continuous form, 3rd person singular number.

In another research paper [4] we have seen a technique to define attribute such as with the example “সে বই পড়েছে” means “He goes to School”. Here if we question the verb- বই পড়েছে i.e; go in English –সে কোথায় যায়? Means “Where does he go?” The answer is School.

In this issue, it is highly demanding to develop analysis rules to convert Bengali sentence to UNL expression. So, we are working in this topic to convert Bengali to UNL expression (to English). In this paper, we present the UNL system for Bengali. The major components of our research works

(i) To develop Bengali-UNL dictionary, this will be compactable with programming.
(ii) Develop and use of analysis rules and
(iii) Translation schema (parsing) for translation program.

<table>
<thead>
<tr>
<th>Word-(English meaning)</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>কে -?who</td>
<td>Agt</td>
</tr>
<tr>
<td>কি?- what</td>
<td>Obj</td>
</tr>
<tr>
<td>কেমন?- how</td>
<td>Aoj</td>
</tr>
<tr>
<td>কি নিয়ে? - by what</td>
<td>Man</td>
</tr>
<tr>
<td>কেমন করে?- by how</td>
<td>Man</td>
</tr>
<tr>
<td>এটার হত তারকাটা? from one to another</td>
<td>Fnt</td>
</tr>
<tr>
<td>কি হলো?- result</td>
<td>Gol</td>
</tr>
<tr>
<td>কিভাবে পুরু? - how started</td>
<td>Frm</td>
</tr>
<tr>
<td>কিভাবে শেষ? - how ended</td>
<td>Gol</td>
</tr>
<tr>
<td>কোন? - which</td>
<td>Mod</td>
</tr>
<tr>
<td>কখন?- when</td>
<td>Tim</td>
</tr>
<tr>
<td>কখন থেকে?- from when</td>
<td>Tmf</td>
</tr>
<tr>
<td>কখন পর্যন্ত?- upto when</td>
<td>Tnt</td>
</tr>
<tr>
<td>কাওয়া</td>
<td>(অবস্থা)- where (location)</td>
</tr>
<tr>
<td>কাওয়া</td>
<td>- from where</td>
</tr>
<tr>
<td>কাওয়া</td>
<td>(পরিস্থি)- where (destination)</td>
</tr>
</tbody>
</table>

In this table, it is shown that the attribute is defined by some characters, which are defined in the following way:

(i) Attribute @present
(ii) Attribute @affirmative
(iii) Attribute @past

It is true that attribute is the important part for converting sentences. Attributes are mainly for the purpose to describe the subjectivity information of sentences. And Relations and UWs are used to describe the objectivity information of sentences. Attributes modify UWs or semantic networks (scope: compound concepts) to indicate subjectivity information such as about how the speaker views these states-of-affairs and his attitudes toward them and to indicate the property of the concepts [5]. But in that paper the proposed method for finding attribute is not work for finding attribute, this method is worked for findings relation. They used a sentence as example that is “সে মুলো শায়” means “He goes to School”. We can define attribute for this sentence as – this sentence is in present tense so it’s attribute is @present and it is an affirmative sentence so the attribute is @affirmative. And attribute is defined by ‘@’ sign. Which attributes they provided in this paper,
these are not attribute, these are relation which define between two words or more. If we question the verb we can get an answer and according to answer we can define relation between verb and other words. For this sentence “সে চুলে গলে” means “He goes to School”. If we question the verb we get “গলে” which is verb and “সে চুলে” is agent (agt). And we need a method for defining verb and we need an algorithm for finding verb or other materials for converting Bengali sentences.

Gender and number are also important for identifying proper categories of nouns. In another paper we have seen a method of identifying number [6]. And we have seen some example of dictionary entry for number. As like: Number may be singular (“েছেল”, boy or “েছেলটি”, the boy, “েছেল”, book, “েছেলটি”, the book) plural (“েছেলরা”, boys “েছেলটি”, the boys, “েছেলরা”, the books etc.). So, from the word “েছেল” we get “েছেলরা”, “েছেলটি”, “েছেলটি” etc. and from the word “েছেল” we get “েছেলটি”, “েছেলটি” etc. Some dictionary entries may look like.

1. We have added every form of a verb based on tenses with affirmative and negative form. As like: if we consider a bangali verb word “খাবিরা”, the another form of this word are:

“খাই”, “খাও”, “খায়”, “খায়ি”, “খানি”, “খাবি”,
“খাবিরি”, “খাবিরি”, “খাবিরি”, “খাবিরি”,
“খাবিরি”, “খাবিরি”, “খাবিরি”, “খাবিরি”, “খাবি থাকবে” etc.

2. There have some Bengali words, when we convert these words in English it is converted by two words. As like: For the Bengali word “থাকে” the English word is “eating” and auxiliary verb is “is”. So we have added this Bengali word “থাকে” directly in dictionary and the English word for this word “is eating” we add “eating” as a main word and “is” as auxiliary verbs for this word.

3. There have some Bengali words as “ভিলে”, when we convert this word in English it is converted like as “in train”. So we have directly added this word in dictionary as “ভিলে” and the English word for this word is “train” and we added “in” as preposition in dictionary.

4. There have some Bengali words which provide an English word when it is stay as a single word in sentence but when it stays as double in sentence it provides another English word. For example: the Bengali word “মােঝ” the English word for this is “sometimes”. But the Bengali word “মােঝ” the English word for this is “middle”. So we have provided the two words in dictionary and for implementing our algorithm easily and perfectly, we have added an extra part is dictionary is “word number” and “duplicate word”. Such as when the parser find a single word like “মােঝ” it provides “middle” but when it finds a double word like “মােঝ” it provides “sometimes”.

5. We also provided phrases in dictionary. Because phrases are not implemented by any algorithm. For a Bengali phrase “গালে কিন্না বলে মাঝে কিনা থায়” for this sentence we provided the full English translation in
dictionary as “a mad man and an animal have no difference”.

6. We have defined subject type for every verb because in Bengali grammar there have many variations in subject. Such as- the Bengali word “গোলাপী” is used as subject and also used as a color name. So when we add “গোলাপী” as subject we define verb as (animal). For example: “গোলাপী খেলেছ” here “খেলেছ” is human verb, so “গোলাপী” is subject not color name. And for the example “গোলাপী রঙের জামা পরা ময়েটি খেলেছ” here for the “খেলেছ” we have added subject type for this verb word as animal. So here “গোলাপী” is not subject, it is a color.

3.1 PROPOSED DICTIONARY DESIGN FOR DATABASE SYSTEM

3.1.1 EXAMPLES OF BENGALI WORD FROM DICTIONARY

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Text</td>
<td>pk</td>
</tr>
<tr>
<td>B</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>E</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>Key_Duplicate</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>WordsNumber</td>
<td>Number</td>
<td>Not null</td>
</tr>
<tr>
<td>SentenceManner</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>KindsOfSentence</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>Tens_of_Verb</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>Persons</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>Gend</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>Number</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsNoun</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsPronoun</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsAdjective</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsAdverb</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsPreposition</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsConjunction</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsInterjection</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>IsVerbalNoun</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>RelationWithNextWord</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>RootWord</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>PrepositionIs</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>AuxiliaryVerbs</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>AttributeOfObject</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>BaseWord</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>EnglishBaseWord</td>
<td>Text</td>
<td>Not null</td>
</tr>
<tr>
<td>SubjectType</td>
<td>Text</td>
<td>Not null</td>
</tr>
</tbody>
</table>

Field Name | Data type | Constraint |
-----------|-----------|------------|

We have constructed an algorithm and implemented this algorithm according to our analysis. We worked for assertive and interrogative sentences. We have implemented the algorithm step by step as like: (we discussed for an example)

For example the sentence “বাংলাদেশ ক্রিকেট দল বিষয়ক জন্য প্রতিদিন সকালে মাত্সুয়ারা পুল নতুন কোচের সাথে ৩টা সেটিংয়াম ক্রিকেট অনুশীলন করে।”

Fig.1: Dictionary data part-1

Fig.2: Dictionary data part-2
1. At first we have checked the punctuation for identifying the type of sentence and the part of sentence. For this sentence, the punctuation is “।” so it is an assertive sentence.

2. Now we have started to convert each of the every word in English from dictionary. For this sentence, at first the parser finds English word for “বাংলাদেশ”, it finds an English word “Bangladesh” for this word. We also have implemented an algorithm for the noun or name such as “রিহম” it writes as “Rahim”. After that the next word “ফ্ল্যাক্কেট”, it finds an English word for this word is “cric ket”, then for the word “দল” it finds “team”; for “বিশ্বকাপ” it finds “the world cup”. The word “জেয়র” for this word it finds an English word and in the duplication column it has duplicate word which we added the same word that means there have some words which constructed with this word as like “জেয়র জনয”。So its then search for all other duplicate word and it finds “জেয়র জনয” and the English word for this “to win”. Following this process we also find out the English word for “বাংলাদেশের ক্রিকেট দল” and the word is “practices”. After that it finds English word for “ঢাকা” and the word is “stadium”. We added manner or attribute for every word as like if there have “জনয, কারেণ” after any word so the attribute or manner is reason, there have some words “ভাও, করে, করেত-করেত…..” if there have like these word after any word this word’s attribute is “Manner1!” and if there have the words “শেষ, মিশ, সাথে, ডিব্রে…..” after any word this word is “manner2”. There have some common words “বাংলাদেশের, দক্ষিণাত্যের…..” the attribute or manner for these words is Adverb of frequency, if there have the words “হত্তে, থেকে, চেয়ে…..”, before these words are destination1 (starting point) and after these words is destination2. If the ending part of any word is “এ, য়, ভে” these words are indicate the attribute “place”. If there have these word like “সকালে, বিকালে, ১০টায়” these words are indicated the attribute “time”. If there have “কে” in the ending part of any word it is “object1” attribute and the there have a word which does not with any other attribute and if the word place is before the verb word and this word is not subject so it is “object2” attribute. The rest words are subject.

3. We have follow a structure [7] for converting assertive sentence and the structure is

(Segment + Adverb of Frequency + Verb + Object1 + Object2 + Destination1 + Destination2 + Manner1 + Manner2 + Place + Time + Reason)

So according to attribute we can convert any Bengali sentence into English. So:-

The English conversion for our example according to the structure:

Bangladesh cricket team (subject) practices (Verb) cricket (object1) spontaneously (manner1) with new (adjective) coach (manner2) in stadium (place) in the morning (time) to win world cup (reason).

Here the output for this sentence from our Translator and also the Google Translator.
Fig.3: Proposed Translator

Bangladesh cricket team practices cricket spontaneously with new coach in Dhaka stadium every morning to win the world cup.
4. CONCLUSION AND FUTURE WORK

This research proposed the Digitally Sentence Conversion Technique from Bengali to UNL. In this paper, we have overcome the problems from previous research findings and we have tried to recover these problems through our algorithmic techniques. We have developed the design of dictionary entries for various Bangali words, which will be required in converting Bangali sentences into UNL Expression. In this regard we designed and implemented the grammatical analysis rule for conversion. We hope, our approach is very much scientific to apply in Conversion of Natural Bangali Sentences and this method will be well accepted for parsing technique [8] which will be the path of En-Converter or En-Conversion [9].

We have developed a sample dictionary with some bangali words and we have implemented an algorithm for assertive and interrogative sentences. So our future works are:

1. Developing a complete Bangali to English Dictionary which is uploaded in a Bangali Language Server.
2. Developing an algorithm for every kind of sentences.
3. Developing a good Bangali to English machine translator system which will help students, researchers and all other people and also all other sector to use English knowledge of the Internet as well as journals.

REFERENCES


[7]s@ifur's Grammar & Freehand Writing, Alauddin Bhuiyan, Tawrit Ahmed Chowdhury, Mahamodul Alom


