Solving Semantic Problem of Phrases in NLP Using Universal Networking Language (UNL)

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

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

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

Abstract—This paper largely deals with the Semantic problem and generation of semantic relations, which are difficult problems in the field of natural language processing. In this work we looked at it through the knowledge based perspective and language phenomena in terms of rules and dictionary features. The work is more focused on solving the problems of Bangla sentences, thereby improving the accuracy of analysis process. The problem needs, particularly, a knowledge intensive solution. We have used insights from linguistics, towards solving this problem. Also, the usefulness of automatic extraction of features for words in the dictionary becomes evident through the work.

Keywords—Syntax and Semantic analysis; Bangla Root Word; Morphology and UNL; NLP

I. INTRODUCTION

Syntax as part of grammar is a description of how words grouped and connected to each other in a sentence. There is a good definition for syntax for programming languages: “… syntax usually entails the transformation of a linear sequence of tokens (a token is key to an individual word or punctuation mark in a natural language) into a hierarchical syntax tree”. Later we will see that the same definition also can be used for NL. Main problems on this level are: part of speech tagging (POS tagging), chunking or detecting syntactic categories (verb, noun phrases) and sentence assembling (constructing syntax tree).

Semantics and its understanding as a study of meaning covers most complex tasks like: finding synonyms, word sense disambiguation, constructing question-answering systems, translating from one NL to another, populating base of knowledge. Basically one needs to complete morphological and syntactical analysis before trying to solve any semantic problem.

In this paper we present the Root Word analysis of Bangla Sentences for UNL system. The major components of our research works touches upon

1) Different types of Ambiguity that is caused by Bangla Root Word
2) UNL Expression of the Bangla Root word and
3) 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

II. PROBLEM DEFINITION

A. Structural Ambiguity

A word, phrase, sentence or other communication is called ambiguous if it can be interpreted in more than one way. If the ambiguity is because of a multiple meanings of a word, it is called lexical ambiguity. One type of ambiguity, called structural ambiguity, arises due to more than one possible structure for the sentence.

ছেলেটি ভাল গান গায়।

It has two English translation one is “The boy sings well” or “The boy sings good song” in the first sentence it means he has a good singing voice and in the second sentence it means that the meaning of his song is good.

B. Attachment ambiguity

This is a specific type of structural ambiguity in which a clause or a phrase has more than one possible association in the tree structure of the sentence of which it is a part. If the ambiguity is about the attachment of a clause then it is called clause attachment and if it is about attachment of prepositional phrase it is called prepositional phrase attachment. Depending on the site of attachment there are at least two possibilities, noun attachment or verb attachment. For example,

ে বই পড়ে। “means He reads a book”
ে হাটে গিয়ে পড়ে গেল। “means he fell during the walking”

Here same word “পড়ে” has two different meaning.

<table>
<thead>
<tr>
<th>Root word</th>
<th>Meaning when used in sentence</th>
<th>Relation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>[চল]</td>
<td>Leave</td>
<td>Agt, obj, gol, p ur</td>
<td>সে কথা পের না করেই চলে গেল।</td>
</tr>
<tr>
<td>Continuity</td>
<td>Pos, obj, man</td>
<td>তার বাবা ভালই গেছে।</td>
<td>His business is being run well.</td>
</tr>
<tr>
<td>Go</td>
<td>Rec, p, obj</td>
<td>চাল গ্রামে যাই।</td>
<td>Let’s go to village.</td>
</tr>
<tr>
<td>Obey</td>
<td>Obj, mod</td>
<td>পিতামাতার</td>
<td>Obey the</td>
</tr>
</tbody>
</table>

www.ijacsa.thesai.org
Invalid Qua,obj,plc সব সিদ্ধি প্রাপ্ত হোলে না। All kinds of rules will not be implemented here.

Tradition Obj,mod সেকারের চন্দল এখন আর নেই। The tradition of the said things are not being continued.

Style Pos,mod,aoj তাহি চলন ঠিক সূচিত না। His living style is not satisfactory.

Expiry man চিকিৎসা আর চাহে না। Taka has been expired.

Variable obj চৌকির মান পরিবর্তিত হয়ে গেলু। The value of the driver may be changed.

Recent Obj,mod,tim চাঠির মাসে দাম আরও বাড়তে পারে। The value may rise in this current month.

<table>
<thead>
<tr>
<th>Root word</th>
<th>Meaning when used in sentence</th>
<th>Relation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>[কহ্য]</td>
<td>Say</td>
<td>Agt,nam</td>
<td>কহ্যে কবি কাৰ্লিনদা। The poet Kalidas has said.</td>
</tr>
<tr>
<td></td>
<td>Option</td>
<td>Mod,aoj,obj</td>
<td>কবিকর কান উপরে ধাকল না। There is no option to say.</td>
</tr>
<tr>
<td></td>
<td>Repeat</td>
<td>Aoj,agt</td>
<td>কাহীয়া কাহীয়া আমি কহাই। I have been tired by saying.</td>
</tr>
<tr>
<td></td>
<td>Invite</td>
<td>plc</td>
<td>কথার কানে। Say in the ear.</td>
</tr>
<tr>
<td></td>
<td>Cannot say</td>
<td>obj</td>
<td>কহিতে না পারি। It cannot be said.</td>
</tr>
<tr>
<td></td>
<td>Speakable</td>
<td>Man,qua,aoj,pur</td>
<td>কথক্যা কথ কিসুয়া ছিল। There was so many things to speak.</td>
</tr>
<tr>
<td></td>
<td>Speakable</td>
<td>Aoj,rsn</td>
<td>কথাটি কথখোস্ণ না। The word is not for speak.</td>
</tr>
<tr>
<td></td>
<td>Cannot say</td>
<td>obj</td>
<td>কথাটি কথা হইলে না। The word has not been told.</td>
</tr>
<tr>
<td></td>
<td>Said</td>
<td>Obj,agt,pol,scn,met</td>
<td>কথাটি কথায় মনে পশ্চিম কলম। I have got the peace in my mind by saying</td>
</tr>
</tbody>
</table>

C. Why Semantic analysis of Root word need?

Verb is the main part of any sentence for any native language. Any Sentence can complete without subject or object. But without verb no sentence is complete. So verb analysis is need to converting from Bangla to UNL. And verb is the combination of root word and suffixes. And root word is titled as entry node when converting any native language to UNL. And not only verb but also other word is derived from a root word that may have the different transformations. This happens because different morphemes are added with it as suffixes. Therefore, the meaning of the word varies for its different transformations. We developed the following rules to this problem.

III. AMBIGUITY OF BANGLA WORD IN SENTENCES

It is necessary to make universal words in the context of bangla sentence and their usage. Converting to the English sentence and then make the universal word for Bangla language will not be semantically correct. In that case enconversion and deconversion will not be correct also.

If we check for some sentences like below:

ÔmK‡jB GK ¶z‡i gv_v gywo‡q‡Q †`LwQ, cix¶vq mevB †dj K‡i‡Q|Ô Here, Ô‡iz‡i gv_v gywo‡q‡QÓ – this clause means to have similar in nature. And in it is a type of Bagdhara. We have to make UW directly from Bangla sentence and it bear the correct meaning than. Both the enconversion and deconversion process will satisfy it.

On the other hand there should have some dictionary entries for special Bangla words which are called ÔGK K_vq cÖKvkÓ| This means to express a group of words in a single word shortly.

w`‡b †h GKevi Avnvi K‡i- GKvnvix| †Q‡jwU w`‡b GKevi Avnvi K‡i| †Q‡jwU GKvnvix|

Both the sentences are same in meaning. So, the Enconversion as well as Deconversion system should know this.

G Mv‡Qi wkKo Mfx‡i hvq| here, root word [hv] ‘go’and the immediate previous word [Mfx‡i]. It needs to make a relation / dependency checking among these two words as meaning of root depends on its previous word. So, there should have a technique of matching which retrieves appropriate word from the dictionary.

G Pv‡j  tekxw`b hv‡te bv| here, root word [hv] and the immediate previous word [tekxw`b].

†U‡b AvR hv‡te bv|
Rule sets 1:

- Some Suffixes (বিভক্তি) are used immediately after the root (শাহু) for sadhu (সাধু) & cholito (চলিত) both languages.
- Some Suffixes (বিভক্তি) are changed according to Person (পূর্ব) for sadhu (সাধু) & cholito (চলিত) both languages.

  c) For sadhu (সাধু) the suffixes are [শূন্য বিভক্তি, iya(ইয়া), ite(ইটে)].

  d) For cholito (চলিত) the suffixes are [শূন্য বিভক্তি] and e (ে).

  e) For Person (পূর্ব) the suffixes are [শূন্য], le(লে), li(লি), ch(চ), che(চে), chi(চি), bo(বো), be(বে)].

Rule sets 2: sadhu (সাধু) and cholito (চলিত) language for different Tense (কাল)

- If the suffixes (বিভক্তি) for sadhu (সাধু) language is [শূন্য বিভক্তি] then the corresponding suffix (বিভক্তি) for cholito (চলিত) is [শূন্য বিভক্তি].

- If the suffixes (বিভক্তি) for sadhu (সাধু) language is ite (ইটে) then the corresponding suffix (বিভক্তি) for cholito (চলিত) is [শূন্য বিভক্তি].

- If the suffix (বিভক্তি) for sadhu (সাধু) language is iya (ইয়া) then the corresponding suffix for cholito (চলিত) is e (ে).

Rule set 3: Person (পূর্ব) (1st, 2nd and 3rd) (singular and plural)

a) If the suffixes (বিভক্তি) for 1st person is i(ই) then the corresponding suffix (বিভক্তি) for 2nd person and 3rd person is [শূন্য বিভক্তি] or O(ও) and e (ে) respectively.

b) If the suffixes (বিভক্তি) for 1st person is chi(চি) then the corresponding suffix (বিভক্তি) for 2nd person and 3rd person is ch(ছ) and che(চে) respectively.

c) If the suffixes (বিভক্তি) for 1st person is lam(লাম) or chilam (চিলাম) then the corresponding suffix (বিভক্তি) for 2nd person and 3rd person is l(ল) or chi(চি) and le(লে) or chile(চিলে) respectively.

d) If the suffixes (বিভক্তি) for 1st person is bo(বো) then the corresponding suffix (বিভক্তি) for 2nd person and 3rd person is be(বে) and be(বে) respectively.

e) If the suffixes (বিভক্তি) for 1st person is ai(আই) then the corresponding suffix (বিভক্তি) for 2nd person and 3rd person is ao(আও) and ai(আই) respectively.

B. Dictionary Entry[12][15][7]

Shadhu Suffix

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>B(ব)</td>
<td>B(ব)</td>
<td>B(ব)</td>
</tr>
</tbody>
</table>

Cholito Suffix

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>B(ব)</td>
<td>B(ব)</td>
<td>B(ব)</td>
</tr>
<tr>
<td>B(ব)</td>
<td>B(ব)</td>
<td>B(ব)</td>
</tr>
</tbody>
</table>

Person Suffix

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>B(ব)</td>
<td>B(ব)</td>
<td>B(ব)</td>
</tr>
</tbody>
</table>
IV. OUR PROPOSED METHOD

There are two reasons to get better performance than other methods. The first reason is the effect of the approach in case of Bangla language. For example, the existing way to find the enconverted or deconverted output which are i) Dictionary Entry Look-up, ii) Rules of morphological analysis and iii) Semantic Analysis. Since component nodes are created by using these steps, they may be less accurate sue which may not expresses semantically correct output as there are different language constraints. As a result, the converted expression may be grammatically correct one but not be meaningfully correct.

The second reason is the determination of the number of component nodes path for constructing desired output. Although the nature of input is meaningfully different as seen in Table-1 and Table 2, existing approach uses the grammatical attributes to select component nodes for all problems. However, Attribute Analysis Approach uses different options to get actual path in the component networks for different input sentence based on the nature of input.

Flow Chart of our proposed program Architecture is shown below:

![Flow Chart](image)

Fig. 1. Proposed program Architecture

Here, **Input sentence**: The Bengali sentence which will be converted to UNL expression. This sentence is given as string.

Example: “আমি ভাত থাই”

**Tokenizer**: In here the input sentence “String” is divide into tokens.

**Validator**: It check that is the is the tokens are arranged in right order or check is there any grammatical mistakes in the given sentence or “String” or “Tokens”.

Example: “আমি ভাত থাই” - Invalid

“আমি ভাত থাই” – Valid

**En-Converter**: It convert the given sentence or “String” or “Tokens” in UNL expression.

Example: “আমি ভাত থাই”

UNL Expression:

\{unl\} 

\agt(eat(icl>consume>do,agt>living_thing,obj>concrete_thing,ins>thing),@entry, @present,i(icl>person))

obj(eat(icl>consume>do,agt>living_thing,obj>concrete_thing,ins>thing),@entry, @present,rice(icl>grain>thing))

\{/unl\}

V. CONCLUSION

Semantic Analysis Approach improves correct method of enconversion of UNL expression of Bangla language. A new technique has been proposed in this thesis work. The new technique used a constructive approach to determine the universal words of Bangla language. The novelty of this method is that, it used straightforward and simple technique to determine the ambiguity of Bangla word as well as the diversified usage of words in sentences for a given Bangla sentence. Semantic Analysis Approach first tried to solve the given problem by some example sentences, than it finds out required approaches to get semantically valid equivalence to get actual meaning of the sentence.

Semantic Analysis Approach explores a new era in universal word construction, i.e., determining number of paths by analyzing the dictionary entries; which leads to creates good options to find appropriate meaning of the input sentence for proper enconversion and deconversion process.

REFERENCES


