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

M. F Mridha
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 of 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,

সে বই পড়ে । "means He reads a book"
সে হাটতে গিয়ে পড়ে গেল ।means "he fell during the walking"

here same word "পড়ে" has two different meaning.

TABLE I. BANGLA AMBIGUOUS SENTENCE OF ROOT WORD "চল্"

Root word	Meaning when used in sentence	Relation	Example	
[চল্]	Leave	Agt,obj,gol,pur	সে কথা শেষ না করেই চলে গেল।	He has left the place not to finish the word.
	Continuity	Pos,obj,man	তার ব্যবসা ভালই চলছে।	His business is being run well.
	Go	Rec,plt,obj	চল গ্রামে যাই।	Let's go to village.
	Obey	Obj,mod	পিতামাতার	Obey the

		কথা মেনে চলা উচিত।	word of the parents.
Invalid	Qua,obj,plc	সব নিয়ম এখানে চলে না।	All kinds of rules will not be implemented here.
Tradition	Obj,mod	সেসবের চল এখন আর নেই।	The tradition of the said things are not being continued.
Style	Pos,mod,aoj	তার চাল– চলন ঠিক সুবিধার না।	His living style is not satisfactory.
Expiry	man	টাকাটা আর চলবে না।	Taka has been expired.
Variable	obj	চলকের মান পরিবর্তিত হতে পারে।	The value of the driver may be changed.
Recent	Obj,mod,tim	চলতি মাসে দাম আরও বাড়তে পারে।	The value may rise in this current month.

TABLE II. BANGLA AMBIGUOUS SENTENCE OF ROOT WORD "কহ"

Root word	Meaning when used in sentence	Relation	Example	
[কহ]	Say	Agt,nam	কহেন কবি কালিদাস।	The poet Kalidas has said.
	Option	Mod,aoj,obj	কহিবার কোন উপায় থাকল না।	There is no option to say.
	Repeat	Aoj,agt	কহিয়া কহিয়া আমি ক্লান্ত।	I have been tired by saying.
	Invite	plc	কহ কানে কানে।	Say in the ear.
	Cannot say	obj	কহিতে না পারি।	It cannot be said.
	speakable	Man,qua, aoj, pur	কহতব্য কত কিছুইতো ছিল।	There was so many things to speak.
	speakable	Aoj,rsn	কথাটি কহনযোগ্য নয়।	The word is not for speak.
	cannot say	obj	কথাটি কহা হইলো না।	The word has not been told.
	Said	Obj,agt,pos, scn,met	কথাটি কহিয়া মনে শান্তি পেলাম।	I have got the peace in my mind by saying

			the word.
Flower	Aoj,pos,mod	কহলার আমাদের জাতীয় ফুল।	Lily is our national flower.

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 bangle 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:

ÒmK‡jB GK ¶z‡zi gv_v gywo‡q‡Q †`LwQ, cix¶vq mevB †dj K‡i‡Q|Ó

Here, Ò¶z‡zi 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 ÒGK 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 †ekxw`b hv‡e bv|

here, root word [hv] and the immediate previous word [‡ekxw`b].

‡Uªb AvR hv‡e bv|

here, root word [hv] and the immediate previous word [AvR].

‡m ‾‹z‡j hvq|

here, root word [hv] and the immediate previous word [‾‹z‡j].

bZzb evwo‡Z K‡e hv‡eb?

here, root word [hv] and the immediate previous word [K‡e].

It needs to know semantically the use of root word in each sentence same root carry different meaning. For this it needs to analyze the words in the sentences before and after the main root/verb. The parser need to know which particular dictionary entry has to retrieve to make the universal word.

If the meaning of the root word is: [hv]- 'go'; - then the dictionary entry will be [hv]{}"go(icl>do)"(ROOT, BANJANT)

when it is of [hv]- spread than.

Rules for solving ambiguity of Bangla Root word[13][14][17].

Rule sets 1:

a) Some Suffixes (বিভক্তি) are used immediately after the root (ধাতু) for sadhu (সাধু) & cholito (চলিত) both languages.

b) Some Suffixes (বিভক্তি) are changed according to Person (পুরুষ) for sadhu (সাধু) & cholito (চলিত) both languages.

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

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

e) For Person (পুরুষ) the suffixes are [lam(লাম) , le(লে), l(ল), ch(ছ), che(ছে), chi(ছি), bo(ব), be(বে)] .

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

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

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

c) If the suffix (বিভক্তি) for sadhu (সাধু) language is iya (ইয়া) then the corresponding suffix for cholito (চলিত) is e [(এ)].

Rule set 3: Person (পুরুষ) (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 [ɸ(শূন্য বিভক্তি)] or O(ও) and e [(এ)] or y(য়) respectively.

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

c) If the suffixes (বিভক্তি) for 1st person is lam(লাম) or chilam (ছিলাম) then the corresponding suffix (বিভক্তি) for 2nd person and 3rd person is l(ল) or chil(ছিল) and le(লে) or chile(ছিলে) respectively.

d) If the suffixes (বিভক্তি) for 1st person is bo(ব) 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 ay(আয়) respectively.

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

Shadhu Suffix

["”]{} " " (PROT,KBIVOKTI,SHADHU,INDIFINIT)<B,0,0>
[ই]{} " " (PROT,KBIVOKTI,SHADHU,INDIFINIT)<B,0,0>
[ইতে]{}"“”
(PROT,KBIVOKTI,SHADHU,CONTINUOUS)<B,0,0>
[ইয়া]{}" " (PROT,KBIVOKTI,SHADHU,PERFECT)<B,0,0>

Cholito Suffix

["”]{}" " (PROT,KBIVOKTI,CHOLITO,INDIFINIT)<B,0,0>
[" ”]{}" " (PROT,KBIVOKTI,CHOLITO,CONTINUOUS)<B,0,0>
[এ]{}" " (PROT,KBIVOKTI,CHOLITO,PERFECT)<B,0,0>

Person Suffix

[ই]{}" " (PROT,KBIVOKTI ,1P)<B,0,0>
[0]{}" " (PROT,KBIVOKTI, 2P)<B,0,0>
[ও]{}" " (PROT,KBIVOKTI, 2P)<B,0,0>
[এ]{}" " (PROT,KBIVOKTI, 3P)<B,0,0>
[য়]{}" " (PROT,KBIVOKTI, 3P)<B,0,0>
[ছি]{} "(PROT,KBIVOKTI,1P)<B,0,0>
[ছ]{}" "(PROT,KBIVOKTI,2P)<B,0,0>
[ছে]{}" "(PROT,KBIVOKTI,3P)<B,0,0>
[লাম]{}" "(PROT,KBIVOKTI,1P)<B,0,0>
[লা]{}" "(PROT,KBIVOKTI,2P)<B,0,0>

[লে]{}" "(PROT,KBIVOKTI,3P)<B,0,0>
[ব]{}" "(PROT,KBIVOKTI,1P)<B,0,0>
[বে]{}" "(PROT,KBIVOKTI,2P)<B,0,0>
[বে]{}" "(PROT,KBIVOKTI,3P)<B,0,0>

IV.   OUR PROPOSED METHOD

There have 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 UNL expression uses three dimensions 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.

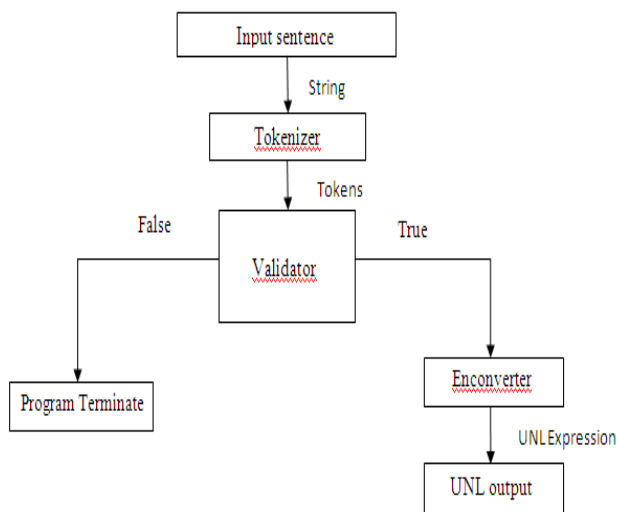Flow Chart of our proposed program Architecture is shown below:

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 খাই"

Tokenizer: In here the input sentence "String" is devide into tokens.

Example: "আমি"   "ভাত"   "খাই"
Token   Token   Token

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 তুমি খাই" - Invalid
"আমি ভাত খাই" – Valid

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

Example: "আমি ভাত খাই"

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

[1] H. Uchida, M. Zhu. The Universal Networking Language (UNL) Specification Version 3.0 Edition 3 ,Technical Report, UNU, (2005/6-UNDL Foundation, International Environment House, Tokyo, 2004)

[2] H. Uchida, M. Zhu, "The Universal Networking Language (UNL) Specification Version 3.0", Technical Report, United Nations University, Tokyo, 1998

[3] S. Abdul-Rahim, A.A. Libdeh, F. Sawalha, M. K. Odeh, "Universal Networking Language(UNL) a Means to Bridge the Digital Divide", Computer Technology Training and Indistrial Studies Center, Royal Scientific Sciety, March 2002.

[4] M. M. Asaduzzaman, M. M. Ali, "Morphological Analysis of Bengali Words for Automatic Machine Translation", International Conference on Computer, and Information Technology (ICCIT), Dhaka, pp.271-276, 2003.

[5] Bengali Academy Bengali-English Dictionary, Dhaka (2004).

[6] Enconverter Specifications, version 3.3, UNL Center/ UNDL Foundation, Tokyo, Japan 2002.

[7] Enconverter Specification Version 3.3, (UNU Centre, Tokyo 150-8304, Japan 2002)

[8] DeConverter Specification, Version 2.7, (UNL Center, UNDL Foundation, Tokyo 150-8304, Japan 2002)

[9] D.M. Shahidullah. Bengali Baykaron, (Ahmed Mahmudul Haque of Mowla Brothers prokashani, Dhaka 2003)

[10] Zakir Hossain, Shahid Al Noor, Muhammad Firoz Mridha Some Proposed Standard Models for Bengali Dictionary Entries of Bengali Morphemes for Universal Networking Language. IJCSNS International Journal of Computer Science and Network Security, V OL.12 No.11, November 2012 .

[11] Bouguslavsky, I., Frid, N. and Iomdin, L. Creating a Universal Networking Module within an Advanced NLP system. Proceedings of the 18th International Conference on Computational Linguistics, pp. 83-89. (2000).

[12] Aloke Kumar Saha, Muhammad F. Mridha, Manoj Banik, and Jugal Krishna Das. Specification of UNL Deconverter for Bengali Language. International Journal of Scientific & Engineering Research, Volume 3, Issue 9, September-2012 ISSN 2229-5518.

[13] Muhammad Firoz Mridha, Md. Zakir Hossain, Shahid Al Noor, "Development of Morphological Rules for Bangla Words for Universal Networking Language" IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.10, October 2010.

[14] Muhammad Firoz Mridha, Kamruddin Md. Nur, Manoj Banik and Mohammad Nurul Huda, "Structure of Dictionary Entries of Bangla Morphemes for Morphological Rule Generation for Universal Networking Language". International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) 2011.

[15] Muhammad Firoz Mridha, Kamruddin Md. Nur, Manoj Banik and Mohammad Nurul Huda, "Generation of Attributes for Bangla Words for Universal Networking Language(UNL)". International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) 2011.

[16] Md. Sadequr Rahman, Sangita Rani Poddar, Muhammad Firoz Mridha, Mohammad Nurul Huda, "Open Morphological Machine Translation: Bangla to English". NWESP, page, 460-465 , ISBN: 978-1-4244-7817-0, 2010.

[17] Muhammad Firoz Mridha, Md. Nawab Yousuf Ali, Manoj Banik3, Mohammad Nurul Huda,Chowdhury Mofizur Rahman, Jugal Krishna Das, "Conversion of Bangla Sentences to Universal Networking Languages, " SKIMA'l0, Paro, Bhutan, August 2010.