New Approach of Solving Semantic Ambiguity Problem of Bangla Root Words Using Universal Networking Language (UNL)

M. F Mridha
Department of CSE,
University of Asia Pacific,
Dhaka, Bangladesh
E-mail: mdfirozm@yahoo.com

Aloke Kumar Saha
Department of CSE,
University of Asia Pacific,
Dhaka, Bangladesh
E-mail: aloke71@yahoo.com

Jugal Krishna Das
Department of CSE,
Jahangirnagar University,
Savar, Dhaka, Bangladesh
E-mail:drdas64@yahoo.com

Abstract—Ambiguity problems of Bangla ROOT words are the difficult problems in the field of natural language processing of Bangla sentence. In this work we looked at it through the knowledge based perspective and language phenomena in terms of rules and dictionary features. This work is more focused on solving the problems of Bangla sentences with semantic problem, 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.

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 i) Different types of Ambiguity that is caused by Bangla Root Word ii) UNL Expression of the Bangla Root word 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.

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.

In Bangla: "সময় তীরের মত চলে"
In English: Time flies like an arrow.

This sentence has two possible interpretations. In first case, *flies* named '*time*' loves an arrow, whereas the other interpretation is '*time*' passes *like an arrow* (as fast as *an arrow*). In this case, both the meanings of a sentence are semantically valid and acceptable. Such sentences are said to be inherently ambiguous. Even human being needs a context to select the appropriate meaning. But there is other kind of ambiguity which concerns NLP. Certain sentences are interpreted only in one way by human being but multiple parses of such sentences are possible for a machine. Moreover, a machine cannot select the meaningful interpretation out of the given possible parses because of the lack of world knowledge.

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

978-1-4799-5180-2/14/$31.00 ©2014 IEEE

attachment. Depending on the site of attachment there are at least two possibilities, noun attachment or verb attachment.

For example,

In Bangla: সে বইটি টেবিলের উপরে রাখল।

In English: He kept the book on the table.

The sentence can be explained in two ways. First the prepositional phrase *'on the table'* is attached to the noun *'the book'* making a noun phrase *'the book on the table'* which is semantically incorrect. Because, if we replace the NP '*the book on the table*' in the sentence with just '*the book*' the sentence becomes incomplete. The other suggests that the prepositional phrase *'on the table'* is attached to the main verb *'kept'* of the sentence. That is, *'the book'* is not directly associated with *'the table'*. However, *'the book'* is related to *'the table'* through the verb *'kept'*. Thus *'on the table'* is attached to the verb *'kept'*

Table 1:Bangla Ambiguous Sentence of Root Word "পড়"

Root word	Meaning when used in sentence	UNL Relation	Example	
পড়	fall	agt	সে পড়ে গেল।	He has fallen down.
	read	Agt,obj	বাচ্চাটি বই পড়ে	The kid is reading book.
	go	Agt,to,pur	বাচ্চাটি পড়তে শিক্ষকের কাছে গেছে।	The kid has gone to the teacher for tuition.
	raining	rain	বৃষ্টি পড়ছে।	It is raining.
	finish	Agt,pos, gol	ছেলেটিপড়াশোনাশেষ করেছে।	The boy has finished his study.
	study	Aoj,tim	মেয়েটিপড়াশোনায়অ নেকভাল।	The girl is good at study.
	forgive	Aoj,mod, pos	ছাত্রটি ভুলের জন্য শিক্ষক এর পায়ে পড়ল।	Student teacher's legs were mistakes.
	rests	Agt,man	মা ঘুমিয়ে পড়েছে।	Mother has fallen asleep.
	broken	Obj,met	ঘড়িটি পড়ে ভেঙ্গে গেছে।	The clock has broken by falling down
	teach	agt	শিক্ষক পড়াচ্ছেন।	The teacher is teaching.

Table 2: Bangla Ambiguous Sentence of Root Word "ধর"

Root word	Meaning when used in sentence	UNL Relation	Example	
ধর	catch	Agt,obj	ছেলেটি মাছ ধরছে।	The boy is catching fish.
	caught	obj	চোরটি ধরা পড়েছে।	The thief has been caught.
	hold	Agt,pos, obj	বাচ্চাটি মাকে ধরে রেখেছে।	The kid is holding his mother.
	walk	Aoj,obj,obj	মেয়েটির চলার ধরণ ভাল না।	Walking style of the girl is not good.
	hate	Agt,pos, obj	সেতারহাতধরলনা।	He did not hold her hand.
	burn	Agt,obj	বাড়িটিতে আগুন	The house is on fire.
	took	Aoj,pos, mod,obj	সে তার নিজের পথ ধরল।	He took his own way.
	support	Agt,man, pos,obj	আমারহাতধরেইসেএ তদূরআসছে।	He reached so far taking my support.
	turn	Obj,gol,obj	সে মুসলিম ধর্ম গ্রহন করল।	He turned into Muslim.
	fitting	Aoj,plc	ঝুড়িতে আম ধরছে না।	Mangoes are not fitting in basket.

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.

We 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 মূল 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 ধানে †ekxw`b hv‡e bv|

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

প্লেন AvR hv‡e bv|

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

‡m বাড়ি hvq|

Here, root word [hv] and the immediate previous word [বাড়ি].

bZzb ক্লাটে 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
[hv]{}"spread(icl>occur, equ>distribute, obj>thing, scn>thing)" (ROOT, BANJANT)

The dictionary entry should be of all possible of dimensions for each root words from which the rules will select required combination according to the meaning of sentence.

Consider the following examples.
Arif won't invite Maria.
Arif won't come Maria.
Here, the verbs are '*invite*' and '*come*' respectively. In the first sentence the verb '*invite*' results in a grammatical sentence when followed by an noun phrase (NP) whereas in the second sentence, it seems that the verb '*come*' does not allow an noun phrase to follow it. Hence, it can be said that a verb like '*invite*' takes an noun phrase complement, whereas a verb like '*come*' does not. There does not seem to be any general way from which one can predict whether a given verb does or does not take a following noun phrase. This is not dependent on the meaning as well.

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 [□(শূন্য বিভক্তি),iya(ইয়া), ite(ইতে)] .
(d) For cholito (চলিত) the suffixes are [□(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 [□(শূন্য বিভক্তি)] then the corresponding suffix (বিভক্তি) for cholito (চলিত) is [□(শূন্য বিভক্তি)].
(b) If the suffixes (বিভক্তি) for sadhu (সাধু) language is ite (ইতে) then the corresponding suffix (বিভক্তি) for cholito (চলিত) is [□(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 [□(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) ffixes (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.

A. 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:

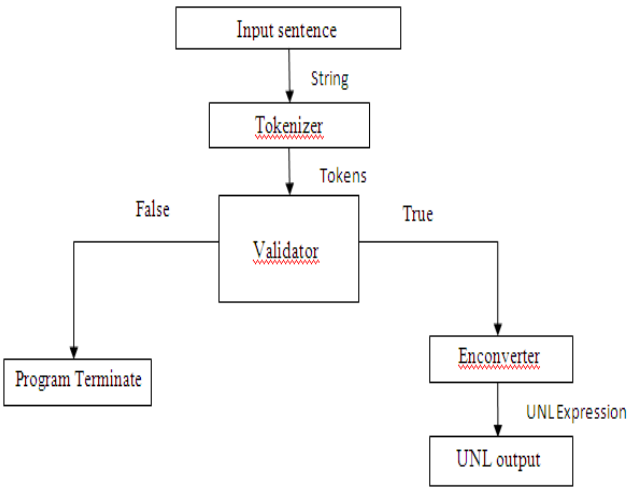

Figure 1 Proposed program Architecture

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

Example: “**আমি আম খাই**”

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

Example: “**আমি**” “**আম**” “**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 আম খাই**” – Valid

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

Example: “**আমি আম খাই**”

1. In put Bangla sentence: “**আমি আম খাই**”

[S:00]
{org:en}
I eat mango
{/org}
{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,mango(icl>edible_fruit>thing))
{/unl}
[/S]

978-1-4799-5180-2/14/$31.00 ©2014 IEEE

2. In put Bangla sentence: "সময় তীরের মত চলে"

[S:00]
{org:en}
Time flies like an arrow.
{/org}
{unl}
obj(fly(icl>occur,equ>pass,obj>thing).@entry.@present,time(icl>abstract_thing,equ>occasion))
man(fly(icl>occur,equ>pass,obj>thing).@entry.@present,like(icl>how,obj>thing))
obj(like(icl>how,obj>thing),arrow(icl>mark>thing).@indef)
{/unl}
[/S]

3. In put Bangla sentence: "সে বইটি টেবিলের উপরে রাখল"

[S:00]
{org:en}
He kept the book on the table.
{/org}
{unl}
aoj(keep(icl>be,equ>continue,obj>action,aoj>thing).@entry.@past,he(icl>person))
obj(keep(icl>be,equ>continue,obj>action,aoj>thing).@entry.@past,book(icl>publication>thing).@def)
plc(book(icl>publication>thing).@def,on(icl>how,com>surface,obj>concrete_thing,plc<uw))
obj(on(icl>how,com>surface,obj>concrete_thing,plc<uw),table(icl>place>abstract_thing,com>restaurant).@def)
{/unl}
[/S]

Some Rules of insertion of Ambiguous Root words:

1. +{V:+.@future.@sg.@male:null}{[করিবে]:null:null}
2. +{V:+.@future.@sg.@female:null}{[করিবে]:null:null}
3. +{V:+.@past.@custom.@sg.@female:null}{[করে]:null:null}
4. +{V:+.@past.@custom.@pl.@female:null}{[করেছি]:null}
5. +{V:+.@past.@custom.@pl.@male:null}{[করেছি]:null:null}
6. >{N,INANI,ACT,TCL:null:agt}{V,^VOCURR,^COO:+AGTRES,+TCL:null}
7. >{N,INANI,ACT:null:agt}{N,KAR,^VOCURR,^COO:+AGTRES:null}
8. >{N,INANI,MACH:null:agt}{V,^VOCURR,^COO:+AGTRES:null}
9. >{N,INANI,MACH:null:agt}{N,KARFSG,^VOCURR,^COO:+AGTRES:null}
10. >{N,INANI,MACH:null:agt}{N,KARFPL,^VOCURR,^COO:+AGTRES:null}
11. >{N,INANI,ACT,TCL:null:agt}{V,^VOCURR,^COO:+AGTRES,+TCL:null}
12. >{N,INANI,ACT:null:agt}{N,KAR,^VOCURR,^COO:+AGTRES:null}
13. >{N,CASE,^CAG:null:agt}{V,^AGTRES,^VOCURR,^COO:+AGTRES:null}
14. >{PRON,BEN:null:ben}{V:+BENRES:null}
15. +{V:+.@present.@sg.@male:null}{[**খাই**]:null:null}
16. +{V:+.@present.@sg.@male:null}{[**খাই**]:null:null}
17. +{V:+.@custom.@present.@sg.@female:null}{**খাই**]:null}
18. +{V:+.@future.@sg.@male:null}{[খাবে]:null:null}
19. +{V,m,sg:+.@future.@sg.@male:null}{খাইব]:null:null}
20. +{V:+.@future:null}{[খাব]:null:null}

V. CONCLUSION

Semantic Analysis method improves correct method of enconversion of UNL expression of Bangla language. A new technique has been proposed in this paper. Here 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, 2003, pp.271-276
[5] Bengali Academy (2004), Bengali-English Dictionary, Dhaka.
[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. (2000). Creating a Universal Networking Module within an Advanced NLP system. Proceedings of the 18th International Conference on Computational Linguistics, pp. 83-89.

[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 2010, page, 460-465 , ISBN: 978-1-4244-7817-0

[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'10, Paro, Bhutan, August 2010.

978-1-4799-5180-2/14/$31.00 ©2014 IEEE