DESIGN AND IMPLEMENTATION OF AN EFFICIENT ENCONVERTER FOR BANGLA LANGUAGE

M. F. Mridha[1], Aloke Kumar Saha[1], Md. Akhtaruzzaman Adnan[1], Molla Rashied Hussein[1] and Jugal Krishna Das[2]
[1]Department of Computer Science and Enginerring, University of Asia Pacific, Dhaka, Banglaesh
[2]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Banglaesh
E-Mail: mdfirozm@yahoo.com

ABSTRACT

In this paper, a distinctive approach of Machine Translation (MT) from Bangla language to Universal Networking Language (UNL) is proffered. This approach corroborates analyzing Bangla sentences more precisely. The analysis churns out a semantic net like structure expressed by means of UNL. The UNL system comprises two major components, namely, EnConverter (used for converting the text from a Native language to UNL) and DeConverter (used for converting the text from UNL to a Native language). This paper discusses the framework for designing EnConverter for Bangla language with a particular attention on generating UNL attributes and relations from Bangla Sentence input. The structural constitution of Bangla EnConverter, algorithm for understanding the Bangla sentence input and resolution of UNL relations and attributes are also conferred in this paper. The paper highlights the EnConversion analyzing rules for the EnConverter and indicates its usage in generating UNL expressions. This paper also covers the results of implementing Bangla EnConverter and compares these with the system available in a Language Server located in Russia.

Keywords: en-converter, machine translation, knowledge base, natural language parsing, universal networking language.

INTRODUCTION

According to a story narrated in the "Book of Genesis of the Tanakh" (Hebrew Bible), everyone on Earth used to speak the same language. People there learned to make bricks and build a city with a skyscraping tower. Purpose of that skyscraper is to stay in a single building and not to be scattered over the world. Eventually, they developed diverse languages over several eras and got themselves scattered over the world, as they failed to comprehend each other's language as well as motives.

Natural Language Processing (NLP) has a potential to unite the Universe again, as per the aforementioned story, but not by the same language, rather by constructing common platform for all existing languages. UNL has been used by researchers as an Interlingua approach for NLP. The World Wide Web (WWW) today has to face the complexity of dealing with multilingualism. People speak different languages and the number of natural languages along with their dialects is estimated to be close to 4000. The Universal Networking Language [1,2,3] has been introduced as a digital meta language for describing, summarizing, refining, storing and disseminating information in a machine independent and human-language-neutral form. A good number of societies over the world are lagging behind in this age of Information Technology just because of the language barrier. There is a great need to translate digital contents which include but not limited to Websites, Blogs, Online News Portals, E-books, E-Journals, E-mails into the native language for overcoming that language barrier. This paper focuses on one such technology and includes the work carried out in this direction for Bangla Language.

UNL System has EnConverter and DeConverter as two important components. The EnConverter converts source language sentences into UNL expressions [4]. The DeConverter converts UNL expressions to target language sentences. With the development of EnConverter for Bangla language, the Bangla text is converted to UNL expressions. It has a potential to translate Bangla language text to any language, if that language has its own DeConverter, which can convert UNL expressions generated by Bangla EnConverter into that destined language. This will certainly help to develop a multilingual machine translation system for Bangla Language.

The organization of this paper is as follow: In Section 2, we describe the related works, Section 3 has the short description about UNL format for representation information, Section 4 describes design of Bangla EnConverter. Finally, Section 5 draws conclusions with some remarks on future works.

RELATED WORKS

In order to design a multilingual machine translation system for Bangla Language, interlingua approach is the best match as it requires only *n* interlingua transfer modules for *n* languages [5]. A transfer module of each language requires only two components: one for converting from source language to Interlingua and other for converting Interlingua to the target language. We have used Universal Networking Language (UNL) as the Interlingua for this task, as the UNL representation has the right level of expressive power and granularity. UNL has 46 semantic relations and 86 attributes to express the semantic content of a sentence [5,6]. UNL has been developed and is managed by the Universal Networking Digital Language (UNDL) foundation, an independent NGO founded in 2001 and based in Geneva, Switzerland, the extension of an initial project launched by the Institute of Advanced Studies of the United Nations University**,** Tokyo, Japan in 1996 [7,8]. For converting Bangla sentence to UNL expressions firstly, we have gone through Universal Networking Language (UNL) [10,11,12,13] where we have learnt about UNL expression, Relations, Attributes, Universal Words, UNL Knowledge Base, Knowledge Representation in UNL, Logical Expression in UNL, UNL systems and specifications of EnConverter. All these are key factors for preparing Bangla word dictionary, enconversion and deconversion rules in order to convert a Bangla language sentence to UNL expressions. Secondly, we have rigorously gone through the Bangla grammar [13,9], Morphological Analysis [14,15,16,17], construction of Bangla sentence [9] based on semantic structure. Using above references we extract ideas about Bangla grammar for morphological and semantic analysis in order to prepare Bangla word dictionary [18,19], morphological rules and enconversion rules in the format of UNL provided by the UNL center of the UNDL Foundation.

UNL format for representation of information

We presume a UNL representation consists of UNL relations, UNL attributes and Universal Words (UWs). UWs are represented by their English equivalents. These words are listed in the Universal Word Lexicon of UNL knowledge base [1]. Relations are the building blocks of UNL sentences. The relations between the words are drawn from a set of predefined relations [3]. The attribute labels are attached with universal words to provide additional information like tense, number *etc.* For example, “সে স্কুলে যায়” in English “se school e jai” can be represented into UNL expression as:

{unl}
agt(go(icl>move>do,plt>place,agt>thing):0B.@entry.@present,he(icl>person):00)
plt(go(icl>move>do,plt>place,agt>thing):0B.@entry.@present,school(icl>building>thing,equ>educational_institute):03)
{/unl}

We can here note that *agt* is the UNL relation which indicates “a thing which initiates an action”; *obj* is another UNL relation which indicates “a thing in focus which is directly affected by an event”; @entry and @present are UNL attributes which indicate the main verb and tense information; and @sg is UNL attribute which indicates the number information.

Proposed algorithm descriptions

The Bangla EnConverter processes the given input sentence from left to right. It uses two types of windows [11], namely, analysis window and condition window in the processing. The currently focused analysis windows are circumscribed by condition windows as shown in Figure-1. Here, ‘A’ indicates an analysis window, ‘C’ indicates a condition window, and ‘ni’ indicates an analysis node.

Bangla EnConverter architecture

The architecture of Bangla EnConverter can be divided into six phases. It consists of the tasks of processing of input Bangla sentence by Bangla parser, creation of linked list of nodes on the basis of output of parser, extraction of UWs and generation of UNL expression for the input sentence. The phases in proposed Bangla EnConverter are tokenize, linked list creation, Universal Word lookup, Case marker lookup, Unknown word handling and UNL creation phase.

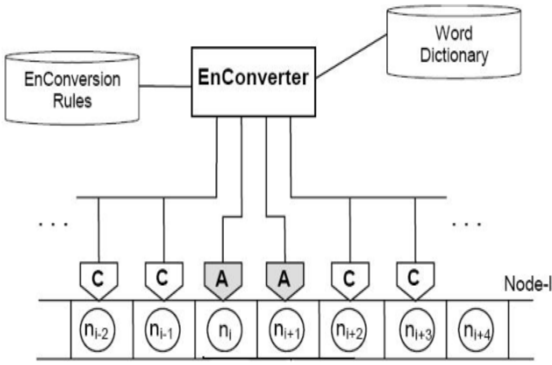

Figure-1. A schematic of EnConverter.

Tokenize phase

Bangla EnConverter uses Bangla parser for tokenize the input sentence. For parsing an input Bangla sentence to produce the intermediate outputs of tokenizer, morph analyzer, part-of-speech tagger, person, number and Bivokti computation.

Linked list creation phase

In this phase, Bangla EnConverter constructs a linked list of nodes. This linked list is constructed on the basis of information generated by the Bangla parser, Bangla- UW dictionary and root word-modifier table. Each root word of the token and verb modifiers of the main verb act as the candidates for the node. For each root word, the words obtained by combining it with root word of next consecutive tokens are searched in Bangla-UW dictionary and root word-modifier table, so that the largest token can be formed on the basis of root words stored in the Bangla- UW dictionary.

After receiving the If a token formed by the concatenation of consecutive root words is found as a single entry in Bangla-UW dictionary or in root word-modifier table, then that group of words is considered as a single token and stored as a node in the linked list, otherwise, each root word of the token is considered as a single token and stored as a node in the linked list. A node in the linked list has Bangla root word attribute, Universal Word attribute, Part-of- Speech (POS) information attribute, and a list of lexical and semantic attributes.

Universal word lookup phase

In this phase, Bangla-UW dictionary is used for mapping of Bangla root word of each node to Universal

Words and to retrieve its lexical semantic information. Exact UW is extracted from the dictionary on the basis of node's Bangla root word attribute and its grammatical category. Since, Bangla-UW dictionary may contain more than one entry for a given Bangla word, the searching process retrieves the UW that matches with the node's Bangla word and its grammatical category. For example, Bangla word, খেল khel 'play' has two entries in Bangla-UW dictionary, one as a noun and other as a verb. It selects only that entry which matches with the grammatical category of the node given by the Bangla parser. If a node is marked as unknown in the first phase of parsing, then node's Bangla word attribute is searched in dictionary with its grammatical category as 'null'. In case of multiple entries of that word, the system returns the UW of first entry and thus the unknown word becomes known during this phase. After extracting the UW, the node's UW attribute is updated and linked list of lexical and semantic attributes is extended to append the UW dictionary attributes with the attributes generated by the parser.

Case marker lookup phase

If Bangla root word attribute of a node is not found in the Bangla-UW dictionary, then it may be a case marker or function word of the language having no corresponding UW. In such a case, node's Bangla word attribute is searched in the case marker lookup file. If a word is found then the information about the case marker is added in the linked list of lexical and semantic attributes of the node and its UW is set to 'null' (because a case marker has no corresponding UW). This information plays an important role in resolving UNL relations in UNL generation phase.

Unknown word handling phase

If an unknown word is resolved in Universal Word lookup phase, then corresponding node is updated with its UW and dictionary attributes. Otherwise, these are resolved in the Case marker lookup phase. If some words still remain unknown, these words are processed in Unknown word handling phase. In this phase, system searches an unknown word in unknown word handling file. It contains only those Bangla words that are derived from some root words because all other unknown words are resolved by UW lookup phase or Case marker lookup phase.

For example, in case of unknown Bangla word, যাবে jabe 'will go' having root word যা ja 'go' has বে bae as a modifier. This modifier contains tense, number and gender information about the sentence. It plays an important role in the generation of UNL attributes. Thus, a new node is inserted in the linked list for this modifier as Bangla root word attribute and its UW attribute, POS attribute and linked list of lexical semantic attributes are all set to 'null'. As such, in case of unknown word যাবে jabe 'will go', node's Bangla word attribute is set to যা ja 'go' and a new node is inserted into the linked list with its Bangla word attribute as বে bae. If a node is updated by unknown word handling phase, it is again processed in Universal Word lookup phase for getting its UW otherwise the token remains to be unknown word.

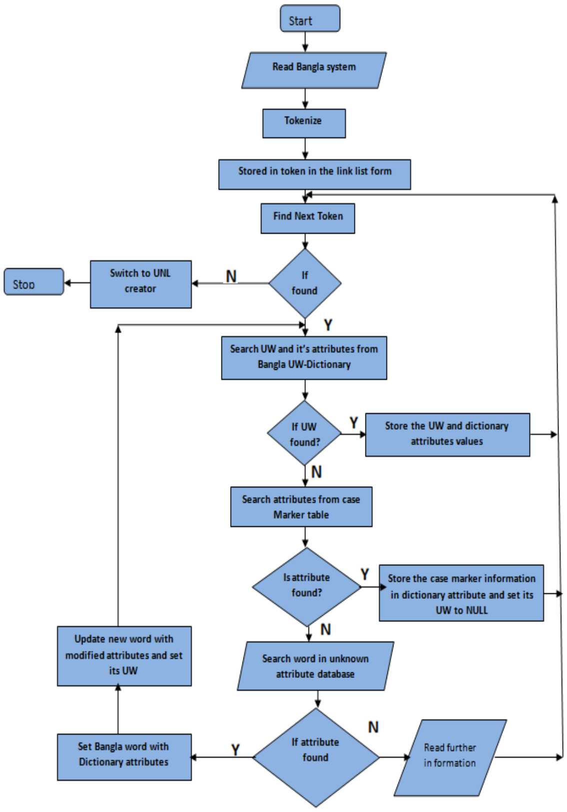
Start
Read Bangla system
Tokenize
Stored in token in the link list form
Find Next Token
If found
N
Switch to UNL creator
Stop
Y
Search UW and it's attributes from Bangla UW-Dictionary
If UW found?
Y
Store the UW and dictionary attributes values
N
Search attributes from case Marker table
Is attribute found?
Y
Store the case marker information in dictionary attribute and set its UW to NULL
N
Search word in unknown attribute database
If attribute found
N
Read further in formation
Y
Set Bangla word with Dictionary attributes
Update new word with modified attributes and set its UW

Figure-2. Flowchart of Bangla EnConverter.

Algorithm for UNL relation resolution

Bangla EnConverter System invokes the following algorithm for UNL relation resolution and generation of attributes.

i) Process each node of linked list by considering the first node as left analysis window and next node as right analysis window.

ii) Search the required rule from EnConverter analysis rules. This depends upon the dictionary attributes of left and right analysis windows.

iii) Modify the linked list to resolve the UNL relations and generate UNL attributes according to the fired rule. If no rule is fired, then go to step (v).

iv) Consider first node of modified linked list as left analysis window and next node as right analysis window. Go to step

(v) With new analysis windows. If the modified linked list contain only single node, then consider that node as "entry node" and stop further processing. It means that all the nodes are successfully processed by the system.

vi) If no rule is fired in step (ii), then shift the window to right. This effectively means that right analysis node will become the left analysis window and next node will become the right analysis window. Go to step (ii) with new analysis windows.

Enconversion of a Bangla sentence to UNL

“ivRv Zvnvi cyÎ‡K gyKzU w`‡e” pronounce as ‘*Raja tahar putroke mukut dibe*’ means “The king will give crown to his son”

Dictionary Entry:
[ivRv]{} “king (icl>sovereign>thing,ant>queen)” (N,3P)
[Zvnv]{} “he(icl>person)”(PRON,HPRON,3P)
[i]{} “” (INF, INF6TH,OBJ, POS)
[cyÎ]{} “son(icl>male_offspring>thing,ant>daughter)” (N)
[‡K]{} “” (INF, INF2ND, NOM, OBJ, BEN)
[gyKzU]{} “crown(icl>jewelled_headdress>thing)”(N)
[w`]{}“give(icl>do,equ>hand_over,agt>thing,obj>thing,ben>person)” (ROOT,VEND,VEG1,#AGT,#OBJ)
[‡e]{} “” (VI,VER,3P,FUT)

Morphological Rules:

First, second and third morphological analyses to be held between “w`” (di) & “‡e” (be), “cyÎ” (putro) & “‡K” (ke) and “Zvnv” (taha) & “i” (r) to complete the meaning of the words “w`‡e” (dibe), “cyÎ‡K” (putroke) and “Zvnvi” (tahar) respectively using the following morphological rules:

Rule1:+{ROOT,VEND,^ALT,^VERB:+VERB,-ROOT,+@::} {VI,VEND:@future::}P10;
Rule2: +{N:@::}{INF, 2NDINF, NOM:::}P10;
Rule3:
+{PRON,HPRON:@::}{INF,6THINF,NOM:::}P10;

Semantic Rules:

First semantic relation is object (obj) relation which is made between “gyKzU” (crown) “w`‡e” (will give) using the following rule,
>{N::obj:}{VERB,#OBJ:+&@future::}P10;

Second semantic relation is made between “cyÎ‡K” (to son), which is beneficiary and “w`‡e” (will give) is dative case. Rule for dative case to perform semantic analysis is:
>{N,BEN::ben:}{VERB:+&@future::}

Third semantic relation is possessive (pos) relation to be held between “Zvnvi” (his) and “cyÎ‡K” (to son) using the following rule:
>{PRON,HPRON,OBJ::pos:)}{N:::}

Forth semantic relation is agent (agt) relation to be held between “ivR v” (king) and “w`‡e” (give) using the following rule:
>{N,SUBJ::agt:)}{VERB:+&@future,&@entry::}

Experimental result and testing system

We have tested our system on several Bangla sentences. It has been seen that the system successfully handles the resolution of UNL relations and generation of attributes for these sentences. The system has been tested with the help of English sentences available at Russian UNL language server. We have manually translated the given English sentences at Russian language server into equivalent Bangla sentences and then inputted those equivalent Bangla sentences to the designed Bangla-UNL EnConverter system. We have compared the UNL expressions generated by our system with the UNL expressions generated by Russian UNL language server. This comparative analysis is given in Table-1 for five sentences. When more sentences are tested and rules will be added then accuracy will be increased.

Table-1. A comparative analysis of UNL expressions generated by Bangla EnConverter and Russian UNL language server.

S. No.	Input Bangla sentence	Relations resolved	Rules fired	UNL expressions generated by the Bangla EnConverter	UNL expressions generated by the Russian UNL

1	আপনি কখন জাবেন?	Agt, tim	R {SHEAD:::} {HPRON,SUBJ:::} P1;
			DR{HPRON,SUBJ,^blk:blk::} {BLK:::} P10;
			R {SHEAD:::} {HPRON,SUBJ:::} P1;
			R {HPRON,SUBJ:::} {QPRON:::} P1;
			DR {QPRON,^blk:blk::} {BLK:::} P10;
			R {HPRON,SUBJ:::} {QPRON:::} P1;
			R {QPRON:::} {ROOT,VEND:::} P1;
			+{ROOT,VEND,^VERB:+VERB,-ROOT,+@::} {KBIV,FUT:::} P8;
			:{:::} {VERB,KBIV:-KBIV,-VEND,+3P::} P10;
			R {QPRON:::} {VERB:::} P1;
			+ {VERB:::} {QMARK:::} P8;
			> {QPRON::tim:} {VERB,#TIM:::} P8;
			> {HPRON,SUBJ::agt:} {VERB,#AGT:::} P8;
			R{SHEAD:::}{VERB,^&@entry,^&@future,^&@interrogative:+&@entry,+&@future,+&@interrogative::
				agt:01(leave(icl>refrain>do,obj>thing,agt>thing):0C.@entry.@future.@interrogative,you(icl>person):00)
				tim:01(leave(icl>refrain>do,obj>thing,agt>thing):0C.@entry.@future.@interrogative,when(icl>how>time):05)
					tim(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@imperative.@interrogative,will(icl>legal_document>thing,pos>person))
					mod(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@imperative.@interrogative,u-initial)

2	ট্রেন আজ যাবে না।	agt tim	R {SHEAD:::} {N:::} P1;
			DR {N,^blk:blk::} {BLK:::} P10;
			R {SHEAD:::} {N:::} P1;
			R {N:::} {N:::} P1;
			DR {N,^blk:blk::} {BLK:::} P10;
			R {N:::} {N:::} P1;
			R {TODAY,N:::} {DEPART,ROOT,^VERB:::} P1;
			+{DEPART,ROOT,VEND,^ALT,^VERB:+VERB,-ROOT,+@::} {KBIV,FUT:::} P8;
			: {:::} {VERB,KBIV:-KBIV,-VEND::} P10;
			> {TODAY,N::tim:} {VERB,#TIM:::} P8;
			>{TRAIN,N:&@def:agt:} {VERB,#AGT:::} P8;
			R {SHEAD:::} {VERB:::} P1;
			DR {VERB,^blk:blk::} {BLK:::} P10;
			R {SHEAD:::} {VERB:::} P1;
			DR{VERB,^&@entry,^&@future,^&@not:+&@entry,+&@future,+&@not::} {NOT:::} P10;
				agt(depart(icl>exit>do,equ>go,plt>thing,plf>thing,agt>thing):0A.@entry.@future.@not,train(icl>public_transport>thing):00.@def)
				tim(depart(icl>exit>do,equ>go,plt>thing,plf>thing,agt>thing):0A.@entry.@future.@not,today(icl>how>time,equ>nowadays):06)
					agt(depart(icl>exit>do,equ>go,plt>thing,plf>thing,agt>thing).@entry.@not.@future,train(icl>public_transport>thing))
					tim(depart(icl>exit>do,equ>go,plt>thing,plf>thing,agt>thing).@entry.@not.@future,today(icl>how,equ>nowadays))

3	সময় চলে যাচ্ছে।	obj	R {SHEAD:::} {N:::} P1;
			DR {N,^blk:blk::} {BLK:::} P10;
			R {SHEAD:::} {N:::} P1;
			R{TIME,N:::} PASS,ROOT,CEND,,^VERB:::} P1;
			+{PASS,ROOT,CEND,^VERB:+VERB,-ROOT,+@::} {KBIV,PRS,PRG,3P:::} P8;
			:{:::}{PASS,VERB,KBIV,^&@present,^&@progress:-KBIV,-CEND,&@present,&@progress::} P10;
			> {OBJ::obj:} {VERB,#OBJ:::} P8;
			R {SHEAD:::} {VERB,^&@entry:+&@entry::} P1;
			R {VERB:::} {STAIL:::} P1;
				obj(pass_by(icl>travel>occur,equ>travel_by,cob>thing,obj>thing):07.@entry.@present.@progress,time(icl>abstract_thing,equ>occasion):00)	mod(pass(icl>accomplishment>thing,equ>base_on_balls).@entry,time(icl>abstract_thing,equ>occasion))
					mod(pass(icl>accomplishment>thing,equ>base_on_balls).@entry,away(icl>adj,equ>away))

CONCLUSIONS

In this paper, the structural constitution of Bangla EnConverter has been proposed and implemented. Bangla EnConverter uses the EnConversion analysis rules for UNL relation by resolution and generation of attributes derived from the input of Bangla ambiguous sentences. At this juncture, we have developed approximately five hundred EnConversion analytical rules. They were necessary for the development of proliferating Bangla EnConverter. This EnConverter has been tested and examined thoroughly for its performance in a public domain hosted by the Russian Language Server. The test results were encouraging as the system output in analogous with the Russian Language Server. At present, the Bangla EnConverter can process nothing further than simple sentence input. We are working on extending the scopes of the EnConverter to include clausal, interrogative and long sentences. Moreover, the effectual implementation of the ambiguity problem module in the proposed Bangla EnConverter is also being worked on till date.

REFERENCES

[1] Uchida H. and Zhu M. 1993. Interlingua for Multilingual Machine Translation CENTER OF THE INTERNATIONAL COOPERATION. MT Summit IV. pp. 157–169. , Kobe, Japan.

[2] Dey K. and Bhattacharyya P. 2005. Universal Networking Language based analysis and generation of Bengali case structure constructs. Research on Computing Science. Vol. 12, pp. 215–229.

[3] Uchida H. and Zhu M. 2001. The universal networking language beyond machine translation. International Symposium on Language in Cyberspace. pp. 1–15. , Seoul, Republic of Korea.

[4] Kumar D.C.S. 1999. Bhasha-Prakash Bangala Vyakaran. Rupa and Company Prokashoni, Calcutta.

[5] Hong M. and Streiter O. 1999. Overcoming the language barriers in the Web: The UNL-Approach. In Multilingual Corpora : encoding, structuring, analysis. 11th Annual Meeting of the German Society for Computational Linguistics and Language Technologiesing, Germany.

[6] Universal Networking Language (UNL) Specifications Version 2005, http://www.undl.org/unlsys/unl/unl2005/.

[7] Uchida H., Zhu M. and Senta T. Della. 1999. A gift for a Millennium. , Tokyo, Japan.

[8] Dave S., Parikh J. and Bhattacharyya P. 2001. Interlingua-based English-Hindi Machine Translation and Language Divergence. Machine Translation. Vol. 16, pp. 251–304.

[9] Ali N.Y., Das J.K., Al-Mamun S.M.A. and Nurannabi A.M. 2008. Morphological analysis of bangla words for universal networking language. 3rd International Conference on Digital Information Management, ICDIM 2008. pp. 532–537.

[10] Dhanabalan T., Saravanan K. and Geetha T.V. 2002. Tamil to UNL EnConverter. International Conference onUniversal Knowledge and Language. , Goa, India.

[11] Uchida H. 1987. ATLAS: Fujitsu Machine Translation System. Machine Translation Summit. , Hakone, Japan.

[12] Jain M. and Damani O.P. 2009. English to UNL (Interlingua) Enconversion. Second Conference on Language and Technology, (CLT). , Lahore, Pakistan.

[13] EnConverter Specification Version 3.3, Tokyo, Japan (2002).

[14] H. A. Bakkotottyo, Dhaka, Bangladesh (1994).

[15] Mridha M.F., Huda M.N., Rahman C.M. and Das J.K. 2010. Development of morphological rules for Bangia root, verbal suffix and primary suffix for universal networking language. ICECE 2010, pp. 570–573, Dhaka, Bangladesh.

[16] Mridha M.F., Saha A.K. and Das J.K. 2014. New Approach of Solving Semantic Ambiguity Problem of Bangla Root Words Using Universal Networking Language (UNL). International Conference on Informatics, Electronics & Vision (ICIEV). pp. 1 – 6. IEEE, Dhaka, Bangladesh.

[17] Saha A.K., Mridha M.F. and Das J.K. 2014. Analysis of Bangla Root Word for Universal Networking Language (UNL). International Journal of Computer Applications. Vol. 89, pp. 8–12.

[18] Saha A.K., Mridha M.F., Akhtar S. and Das J.K. 2013. Attribute Analysis for Bangla Words for Universal Networking Language (UNL). International Journal of Advanced Computer Science and Applications (IJACSA). Vol. 4, pp. 158–163.

[19] Mridha M.F., Rahman M.S., Huda M.N. and Rahman C.M. 2010. Structure of dictionary entries of Bangla morphemes for morphological rule generation for universal networking language. 2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM. pp. 454–459, Krackow, Germany.

[20] Mridha M.F., Banik M., Ali M.N.Y., Mohammad Huda N., Rahman C.M. and Das J.K. 2010. Formation of Bangla Word Dictionary Compatible with UNL Structure. 4th International Conference on Software, Knowledge and Information Management and Applications. , Paro, Bhutan.