Development of Morphological Rules for Bangla Words for Universal Networking Language

Muhammad Firoz Mridha, Md. Zakir Hossain, Shahid Al Noor
Department of Computer Science
Stamford University Bangladesh
Dhaka, Bangladesh

Abstract
This paper describes a method for the development of Bangla Enconversion within the framework of the Universal Networking Language (UNL).We also discuss some issues and problems related to the UNL representation that affect the quality of generation. Additionally, the lingware engineering is introduced as a technique to enhance the quality and increase the development efficiency. In this paper, we propose a pioneer work that analyzes the Bangla words morphologically from which we obtain Roots and Krit Prottoy (Primary suffixes), and develops some rules for Bangla Root, Krya Bivokti (Verbal suffixes) and Primary Suffix for the UNL.

Index Term
Morphology, Bangla Roots, Primary Suffix, Morphological rules, Universal Networking Language.

I. INTRODUCTION

Bangla (widely used as Bengali) is spoken by about 245 million people of Bangladesh and two states of India, but most of the computer based resources and technical journals are in English. Due to the language barrier, the common people face big obstacle to enjoy the optimum benefits of modern information and communication technology (ICT) as well as huge enriched English knowledge database around the globe. 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. Currently, the UNL includes 16 languages, which are the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish), in addition to the ten other widely spoken languages (German, Hindi, Italian, Indonesian, Japanese, Latvian, Mongol, Portuguese, Swahili and Thai). In the last few years, machine translation techniques have been applied to web environments. The growing amount of available multilingual information on the Internet and the Internet users has led to a justifiable interest on this area. Hundreds of millions of people of almost all levels of education, attitudes and different jobs all over the world use the Internet for different purposes [1], where English is the main language of the Internet. Because English is not understandable for most of the people, Interlingua translation programs are needed to develop. The main goal of the UNL system, which allows users to visualize websites in their native languages, is to provide a common representation for accessing Internet of multilingual websites by the majority of the people over the world. For this common representation, lexical knowledge is a critical issue in natural language processing systems, where the development of large-scale lexica with specific formats capable of being used by distinguished applications, in particular to multilingual systems, has been given special focus. Our goal is to include Bangla in this system with less effort.

In this paper, we present a UNL system for Bangla which comprises: i) development of grammatical attributes for Bangla root and Primary suffixes to construct *Bangla Word Dictionary* and use of morphological analysis, ii) UNL Expression of the Bangla attributes and *iii) development of rules* .

The organization of the paper is as follows: we describe the UNL system in Section II. In Sections III and IV, we present our main works that include all the above three components. Finally, Section V draws conclusions with some remarks on future works.

II. UNIVERSAL NETWORKING LANGUAGE

UNL, which has been developed to convey linguistic expressions of natural languages for machine translation, is an artificial language that allows the processing of information across linguistic barriers [10]. Such information is expressed in an unambiguous way through a semantic network, which composed of nodes and arcs that represent concepts and relations between them, respectively.

UNL contains three main elements:
•***Universal Words:*** Nodes that represent word meaning.
•***Relation Labels:*** Tags that represent the relationship between Universal Words (UWs).
•***Attribute Labels:*** Additional information about the UWs.

Manuscript received October 5, 2010
Manuscript revised October 20, 2010

These elements are combined in order to establish a hierarchical Knowledge Base system [10] that defines unambiguously the semantics of UWs. The UNL Development Set provides tools named EnConverter and DeConverter, which enable the semi-automatic conversion of natural language into UNL and vice-versa. The EnConverter [11], which allows morphological and syntactical ambiguities resolution, translates natural language sentences into UNL expressions and implements a language independent parser that provides a framework for morphological, syntactic and semantic analysis synchronously. On the other hand, the DeConverter [3, 12] is a language independent generator that converts UNL expressions into natural language sentences only.

A. Universal Words

A UW represented by a hypergraph is not only a unit of the UNL syntactically and semantically for expressing a concept, but also a basic element for constructing a UNL expression of a sentence or a compound concept. It is noted that UW is an element of the vocabulary of UNL system,

B. Relational Labels

The relation [1] between UWs, which has different labels according to the different roles, is a binary, where the relation label is represented as strings of three characters or less. There are many factors to be considered in choosing an inventory of relations of 46 types in UNL. For example, agt is an agent, which is defined as a thing that initiates an action. The syntax of agt in UNL is: agt(do, thing) and agt(action, thing). On the other hand, obj is an object with some attributes, where the syntax of obj in UNL is: obj(be, thing), obj(do, thing), obj(occur, thing), etc.

C. Attributes

The attributes of UWs, which are used to describe subjectivity of sentences, represent the grammatical properties of the words and show what is said from the speaker's point of view: how the speaker views what he is said. This includes phenomena technically [4, 5] called speech, acts, propositional attitudes, truth values, etc. Conceptual relations and UWs are used to describe objectivity of sentences, where attributes of UWs enrich this description with more information about how the speaker views these state of affairs and his attitudes toward them.

III. MORPHOLOGY OF BANGLA WORDS

Morphology focuses on patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. Thus morphological analysis is found to be consented on analysis and generation of word forms, and deals with the internal structure of words and how words can be formed. It may be mentioned that morphology plays an important [2, 8] role in the applications such as spell checking, electronic dictionary interfacing and information retrieving systems. In these applications, it is important that words, which are only morphological variants of each other, are identified and treated similarly. In natural language processing (NLP) and machine translation (MT) systems we need to identify words in texts in order to determine their syntactic and semantic properties [7]. A Bangla morpheme, besides the root word, is supposed to be represented in the Bangla-UNL dictionary using the following UNL format [10].

[HW] "UW" (ATTRIBUTE1, ATTRIBUTE2, …) <FLG, FRE, PRI>
HW← Head Word (Bangla Word)
UW← Universal Word
ATTRIBUTE← Attribute of the HW
FLG← Language Flag
FRE← Frequency of Head Word
PRI← Priority of Head Word

The attributes describe the nature of the head word for classifying it as a grammatical, semantic or morphological feature. So, we will be especially concerned about representation of morphemes using various attributes.

A. *Bangla Roots*

Bangla Language contains a lot of verbs in which the core part is called root. In another way if we split the verbs we get two parts *Root* and *Suffix*. For example 'K‡i' (pronounce as "kore", Ki+G), which means something to do, is a verb and has two parts: (i) 'Ki' (pronounce as "koro") is root and (ii) 'G' (pronounce as "e") is verbal suffix (Krya Bivokti in Bengali language). Some other Bangla root verbs are Pj& (pronounce as "cholo", meaning is "walk"), co& (pronounce as poro, meaning is "read"), and bvP (pronounce as "nach", meaning is "dance").

B. Verbal Suffix (Krya Bivokti)

The suffix, which is used after root to form new verb word, is called Verbal Suffix (Krya Bivokti) [13]**.** In Bangla, B (pronounce as "i"), B‡ZwQ (pronounce as "itechi"), B‡ZwQ‡jb (pronounce as "itechilen"), and †e (pronounce as "be") are all verbal suffixes.

C. Bangla Primary Suffixes

The sound [8], which is added with root, form noun or adjective then the root word is called root verb, and the corresponding sound is called Primary Suffix (Prottoy). For example, Pjb (pronounce as "cholon", meaning is "mobility of an object") is divided into two parts: Pj&& (pronounce as "cholo", meaning is "walk") and Ab& (pronounce as

"on"). Here, Pj&& and Ab& are root verb and primary suffix, respectively.

D. Morphological Analysis of Bangla verbs

Morphological analysis is applied to identify the actual meaning of the word by identifying suffix or morpheme of that word. Every 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. For example, if we consider 'Ki&'(pronounce as "koro", meaning is "do") as a root word, then addition of B (pronounce as "i") after the word 'Ki&' derives a new verb 'Kwi'(pronounce as "kori", meaning is "something is done by first person"). Thus, we get the grammatical attributes of the main word by morphological analysis. Derivational morphology is simple; a word rarely uses the derivational rule in more than two or three steps. The first step forms nouns or adjectives from verb roots, while the next steps form new nouns and adjectives [5]. We have examined derivational morphology for UNL Bangla dictionary too. For developing word dictionary and rules we divided Bangla roots in 11 groups according to primary suffixes (Prottoy) attachment.

1. আ গ্রুপ(AA GROUP)
2. ই গ্রুপ(EI GROUP)
3. আও গ্রুপ(AOW GROUP)
4. আনো গ্রুপ(ANOW GROUP)
5. অন্ত গ্রুপ(ANTO GROUP)
6. অন গ্রুপ(OAN GROUP)
7. তি গ্রুপ(TI GROUP)
8. ইয়ে গ্রুপ(YEA GROUP)
9. ওয়া গ্রুপ(WA GROUP)
10. ও গ্রুপ(OA GROUP)
11. ঊয়া গ্রুপ(UAW GROUP)

IV. RULES FOR UNL TEXT GENERATION

A. Analysis Rules

An analysis rule describes rule application conditions, a method to rewrite the attribute of node that satisfies the application condition, and construction methods of syntax tree. While applying rules, the EnConverter analyzes morphemes, syntax and semantics. Finally, it generates a syntax tree and a network.

The description format of the analysis rules is as follows [11]:

<TYPE>
["("<PRED>")]…"{"[<COND1>]":" [<ACTION1>] ":" [<RELATION1>] ":"
[<ROLE1>] "}"
"{" [<COND2>] ":" [<ACTION2>] ":" [<RELATION2>] ":"
[<ROLE2>] "}"
["("<SUF>")"]…"P("<PRIORITY>");"

***Symbol Explanation*:**
"" represents terminal symbol,
[ ] represents zero or more times,
{} and () designates an analysis windows in the node list.

Description of Condition:
<PRE> Describes condition of nodes on the left side of the left of analysis window.
<SUF> Describes condition of nodes on the right side of the right of analysis window.
<COND1> Describes condition of the node in the Left Analysis Window (LAW).
<COND2> Describes condition of the node in the Right Analysis Window (RAW).

Description of Action:
<ACTION1> Describe the rewriting of grammatical attribute in the LAW.
<ACTION2> Describe the rewriting of grammatical attribute in the RAW.

Direction of Semantic Relation:
It describes the semantic relation between the left node (LN) and the right node (RN).
<RELATION1>Describe the semantic relation of the RAW to LAW.
<RELATION2>Describe the semantic relation of the LAW to RAW.
<PRIORITY> Describes priority of the rules. Code 0-255 is used to specify the priority.

B.Types of the Analysis Rules

This part explains the action and functions of the rule types that can lie described with <TYPE> in analysis rules.

Left Composition " + | +:+ | +: c | +:*"

The RN is combined to LN to make one composition node. The syntax tree and the attribute having left node are inherited. When the RN attributes is inherited, "@" is put in the action column of the LN, the original two nodes are deleted from the node list. The composition node is inserted into the node list. After applying the rules, the composition node takes a position in the RAW.

Right Composition " - \| -:+ \| -:c \| -:* "

The LN is combined to RN to make one composition node. The composition node is inserted into the node list. After applying the rules, the composition node takes a position in the LAW.

Left Modification "<"

When the RN modifies LN, the RN is deleted from node list and the LN remains only. The node, which the <RELATION> is described, is the to-node and the other node is from-node.

Right Modification ">"

When the LN modifies RN, the LN is deleted from node list and the RN remains only.

Left Shift "L"

Shift the analysis window to the left.

Right Shift "R"

Shift the analysis window to the right.

Attribute Changing Rule ":"

This rule adds or deletes attributes from a particular node.

B. Morphological rule Generation for Bangla Words

Bangla is a semantic language, and its basic characteristic is the rich morphology in which most of its words are derived from roots. Inflections and derivations are generated by changing vowels and insertion of consonants. Bangla sentences are characterized by a strong tendency for agreement between its constituents: between verb and noun, noun and objective, in matters of numbers, gender, definitiveness, case, person, etc. These properties are expressed by a comprehensive system of affixation. To satisfy these grammatical properties, the generation rules are expected to be complex for handling the processing of generating grammatically correct Bangla sentences from UNL expression and structure. The linguistic attributes of roots, which have been used in the dictionary, are basically: SORANT, BANJANT and CASE MARKER. Finally, the variations in the written forms of Bangla are also handled by making entry for each of these forms in the dictionary. A database system has been developed for the classification and features adding for each entry in the dictionary. In Fig.1, the system gets the UW and tries to get the equivalent Bangla word from a Bangla - UNL dictionary. The selected Bangla word is then classified to Noun or Verb or Particle. The relation mapping is implemented in the enconversion rule.

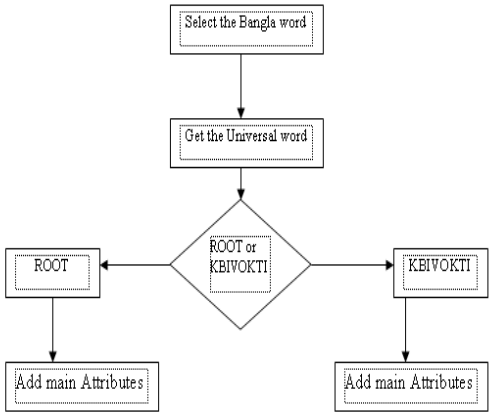

Fig.ure-1: Adding entries for ROOT and KBIVKOKTI (Verbal Suffix)

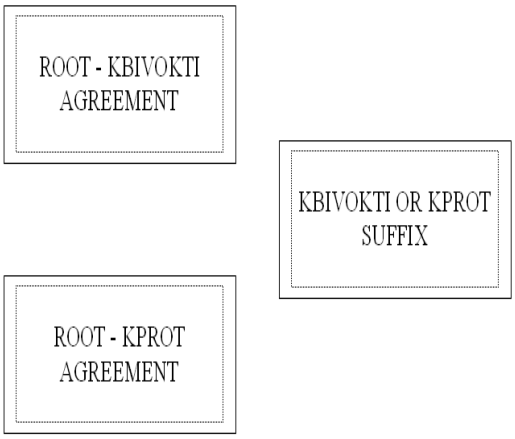

Fig. 2: Bangla morphological rules agreement

In our system, we handle this rich and complicated morphology by implementing a modular approach for coding the rules, which is shown in Fig. 2. Our implemented process of morphology generation starts by choosing the right stem, which is fixed up to accept prefixes or suffixes depending on its position and role in the sentence.

C. Morphological Rules for Bangla ROOT and KRYA BIVOKTI

From the analysis of Bangla Root and Krya Bivokti, one can readily find that they agree right composition rule.

Right Composition Rule: (For Bangla root and Krya Bivokti)

-: C {ROOT:::} { KBIVOKTI : +VERB,-KBIVOKTI::}

For example, Bangla word: পড়িয়াছি (pronounce as "poriyachi", meaning is "something have been read by first person") is divided into two morphological parts, which are

পড় (pronounce as "poro", meaning is "read") and ইয়াছি (pronounce as "yachi") where পড় is Bangla ROOT and ইয়াছি is KRYABIVOKTI (Verbal Suffix).

[পড়] {} "read(icl>do)"(ROOT, BANJANT) <B,0,0>
[ইয়াছি] {} "" (1P, PresentPerfect) <B, 0, 0>

Some proposed rules for morphological analysis of Bangla ROOTs and KRITPROTTOY:

Rule 1: (For WA(ওয়া) group and the root ended with vowel)

-:C{ROOT,SORANT,WA:::}{KPROT,SORANT,WA:-KPROT,- SORANT,-WA, +VNOUN`::}

Example: From the dictionary entry,
[চা] {} "want(icl>do)"(ROOT,SORANT,WA)<B,0,0>
[ওয়া] {} "" (PROT, KPROT, WA, VNOUN) <B,0,0>

Using this rule, the root "চা" (when it is in the LAW) is added with suffix "ওয়া" (when it is in the RAW) to form a verbal noun "চাওয়া" (pronounce as chaowa). It describes: if there is a vowel ended root of group WA(ওয়া) is in LAW and a vowel ended suffix in group WA(ওয়া) is in RAW, then two head words will be added to make "চাওয়া" (pronounce as chaowa). This rule also describes that all the attributes of the node of RAW(attributes for "ওয়া" (pronounce as "wa") are added with the attributes of new word and the following attributes KPROT, SORANT and WA are deleted and attribute VNOUN is added that denotes that new word "চাওয়া" (pronounce as chaowa) is a verbal noun.

Similarly, we can write the following rules for other groups as follows:

Rule 2: (For AA(আ) group and the root ended with consonant)
-:C{ROOT,BANJANT,AA:::}{KPROT,BANJANT,AA:-KPROTOY,- BANJANT,- AA,+MNOUN::}

Rule 3: (For EI(ই) group and the root ended with vowel)
-:C{ROOT, SORANT, EI:::}{KPROT, SORANT,EI: -KPROT,-SORANT,-EI,+MNOUN::}

Rule 4: (For EI(ই) group and the root ended with consonant)
-:C{ROOT, BANJANT,EI:::}{KPROT, BANJANT,EI: -KPROT,- BANJANT,-EI,+MNOUN::}

Rule 5: (For AOW(আও) group and the root ended with vowel)
-:C{ROOT,SORANT,AOW:::}{KPROT,SORANT,AOW:-KPROT,-SORANT,-AOW,+ MNOUN::}

Rule 6:(For AOW (আও) group and the root ended with consonant)
-:C{ROOT,BANJANT,AOW:::}{KPROT,BANJANT,AOW:-PROTOY,-BANJANT,-AOW, +MNOUN::}

Rule 7: (For ANOW(আনো) group and the root ended with vowel)
-:C{ROOT,SORANT,ANOW:::}{KPROT, SORANT, ANOW:-KPROT,-SORANT,-ANOW, +MNOUN::}

Rule 8: (For ANOW (আনো) group and the root ended with consonant) -:C{ROOT,BANJANT,ANOW:::}{KPROT,BANJANT, ANOW: -KPROT, -BANJANT,-ANOW,+MNOUN::}

Rule 9: (For ANTO(অন্ত) group and the root ended with consonant)
-:C{ROOT,BANJANT,ANTO:::}{KPROT, BANJANT,ANTO:-KPROT,- BANJANT,-ANTO, +MADJ::}

Rule 10: (For OAN(অন্) group and the root ended with consonant)
-:C{ROOT, BANJANT,OAN:::} {KPROT, BANJANT,OAN: -KPROT, -BANJANT,-OAN, +MNOUN::}

Rule 11: (For OAN (অন্) group, the root ended with vowel)
-:C{ROOT, SORANT,OAN:::} {KPROT, BANJANT,OAN: -KPROT, - BANJANT,-OAN ::}

Rule 12: (For TI(তি) group and the root ended with consonant only)
-:C{ROOT, BANJANT,TI} {KPROT, BANJANT,TI: -KPROT,- BANJANT,-TI,+MADJ::}

Rule 13: (For YEA(ইয়ে) group and the root ended with vowel only)
-:C{ROOT, SORANT,YEA} {KPROT,SORANT,YEA: -KPROT, -SORANT,-YEA, +MADJ::}

Rule 14: (For YEA(ইয়ে) group and the root ended with consonant only)

-:C{ROOT, BANJANT,YEA:::} {KPROT, BANJANT,YEA: -KPROT, - BANJANT, - YEA ,+MADJ::}

Rule 15: (For OO(ও) group and the root ended with vowel only)
-:C {ROOT, SORANT,OO:::}{KPROT,SORANT,OO: - KPROT, -SORANT,-OO, + MADJ ::}

Rule 16: (For OO(ও) group and the root ended with consonant only)
-:C{ROOT,BANJANT,OO:::}{KPROT,BANJANT,OO: - KPROTOY, - BANJANT,-OO+ MADJ ::}

Rule 17: (For UAW(ঊয়া) group and the root ended with consonant)
-:C {ROOT, BANJANT, UWA:::} {KPROT, BANJANT, UWA: - KPROTOY,- BANJANT ,-UWA, +MADJ ::}

V. CONCLUSION

In this paper we discussed the format and working mechanism of different types of rules that have been generated by us. These mainly comprise the morphology rules for Bangla verb. The dynamically addition or deletion of various attributes were also discussed when two different Bangla Words get combined.

We have tried to follow some systematic way in using the available EnCo tool for compensating the lacking of high level programming constructs and modularity features. This study is limited to fewer number of Bangla words. The authors would like to build a Bangla language server that will contain a complete Bangla Word Dictionary and full set of rules.

REFERENCE

[1] S. Abdel-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.

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

[3] Serrasset Gilles, Boitel Christian, (1999) UNL-French Deconversion as Transfer & Generation from an Interlingua with Possible Quality Enhancement through Offline Human Interaction. *Machine Translation Summit-VII*, Singapore.

[4] M. E. H. Choudhury, M. N.Y. Ali, M.Z.H. Sarkar, R. Ahsan, "Bridging Bangla to Universal Networking Language- A Human Language Neutral Meta- Language", International Conference on Computer and Information Technology (ICCIT), Dhaka, 2005,pp.104- 109

[5] M.E.H. Choudhury, M.N.Y. Ali, "Framework for synthesis of Universal Networking Language", East West University Journal, Vol. 1, No. 2, 2008, pp. 28-43

[6] M.N.Y. Ali, J.K. Das, S.M. Abdullah Al Mamun, M. E.H. Choudhury, "Specific Features of a Converter of Web Documents from Bengali to Universal Networking Language", International Conference on Computer and Communication Engineering 2008(ICCCE'08), Kuala Lumpur, Malaysia.pp. 726-731

[7] S. Dashgupta, N. Khan, D.S.H. Pavel, A.I. Sarkar, M. Khan, "Morphological Analysis of Inflecting Compound words in Bangla", International Conference on Computer, and Communication Engineering (ICCIT), Dhaka, 2005, pp. 110-117

[8] M.N.Y. Ali, J.K. Das, S.M. Abdullah Al Mamun, A. M. Nurannabi,"Morpholoical Analysis of Bangla worfs for Universal Networking Language",icdim,08.

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

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

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

[12] Deconverter Specifications, version 2.7, UNL Center/ UNDL Foundation, Tokyo, Japan, 2002

[13] D. S. Rameswar, "Shadharan Vasha Biggan and Bangla Vasha", Pustok Biponi Prokashoni, November 1996, pp.358-377

Muhammad Firoz Mridha is now a full time faculty in Stamford University Bangladesh. He obtained MSc in Computer Science and Engineering from United International University (UIU) in 2010 and BSc in Computer Science Engineering 2006 from Khulna University of Engineering and Technology. His research interest includes Natural Language Processing, Artificial Intelligence and Software Engineering.

Md. Zakir Hossain is now a full time faculty in Stamford University Bangladesh. He obtained MSc and BSc in Computer Science and Engineering from Jahangirnagar University in 2010 and 2006 respectively. His research interest includes Natural Language Processing, Image Processing, Artificial Intelligence and Software Engineering.

Shahid Al Noor is a full time faculty of Stamford University Bangladesh in the department of Computer Science. He was also a former faculty member of Prime University, Bangladesh in the department of Electronics and Telecommunication Engineering. He Obtained both of his MSc and BSc in Information and Communication Engineering from University of Rajshahi, Bangladesh in 2006 and 2005 respectively. His research interest includes Distributed Networks, Wireless Sensor Networks and Computer Vision and Natural Language Processing.