Development of Bangla Word Dictionary through Universal Networking Language Structure


The mission of the UNL project is to allow people across nations to access information in the Internet in their own languages. A good number of researchers in computational linguistic all over the world have already started developing UNL system for their respective native language. But the UNL system is not developed for Bangla. Researchers have been working on this issue. But so far no immense attempt has been made to develop Bangla word dictionary through universal networking language structure. In this paper, we particularly emphasize on development of Bangla Words along with their grammatical attributes with the help of previous works done by English-UNL-Dictionary and morphological analysis in the framework of UNL with goal to produce a Bangla Word Dictionary for UNL which can be used for Enconverter and Deconverter (to convert Bangla natural language sentences to UNL expressions and vice versa.

Keywords: Universal Networking Language (UNL), Universal Words (UW), Bangla-UNL Dictionary, UNL Knowledge Base (UKB)

1. Introduction

Universal Networking Language is a computer language that allows computers to process information and knowledge across the language barrier. The Universal Networking Language is an artificial language specifically designed for use as a bridge language. The objective of the Universal Networking Language Program (UNLP), as initiated by the United Nations and devised by the Universal Networking Digital Language (UNDL) Foundation, is to enable all people to generate information and have access to cultural knowledge in their native languages.

The UNL system was launched in 1996 by the Institute of Advanced Studies of United Nations University in Tokyo, Japan. The goal is to eliminate the massive task of translation between two languages and reduce language to language translation to a one time conversion to UNL. Once information written in one language is “enconverted” into UNL it will be able to be shared by anyone in the world.

To become a member of UNL society, the researchers of every native language develop an enconversion program that contains word dictionaries; grammatical rules etc. to convert native language texts into UNL texts and also have to develop a deconversion program to convert UNL texts to the respective native languages.

* Khaled Bin Yousuf, Department of Computer Science & Engineering, East West University, Bangladesh. Email: khaled_ewu@ovi.com
** Md. Nawab Yousuf Ali, Department of Computer Science & Engineering, East West University, Bangladesh. Email: nawab@ewubd.edu
Bangla can also be the member of UNL society, because it is the 6th most widely spoken language with more than 230 million speakers, most of whom live in Bangladesh and the Indian state of West Bengal. Therefore, it is essential to take into account a morphological analysis of Bangla Words and also develop the Bangla word dictionary for the UNL system to include Bangla as a member of UNL. It would allow a vast number of people to access and share through the Internet a vast storehouse of knowledge.

The UNL system allows people to communicate with peoples of different languages in their mother tongue. The UNL system basically consists of language servers, UNL editors and UNL viewers. The structure of UNL shown in Figure 1. A conversion system from native languages into UNL is called “enconverter”, and one that deconverts from UNL into native languages is called “deconverter”.

A framework has been developed for converting Bangla texts to UNL expression and vice versa. But for the limitations of words in word dictionary and analysis rules, it is difficult for the researcher to develop this project efficiently. In this paper we develop template of universal words of Bangla Head Word for word dictionary through Universal Networking Language Structure provided by the UNDL foundation.

The organization of this paper is as follows: Section 2 we highlights the literature review of the paper, Section 3 describes the structure of the Universal Networking Language. Format of Bangla word dictionary with examples has been explained in section 4 while section 5 explains the dictionary development of Bangla words. Conversation of a Bangla sentence to UNL expression has been shown in section 6. And finally Section 7 draws conclusions with some remarks on future works.

2. Literature Review
For development Bangla word dictionary firstly Universal Networking Language have been analyzed where we focused on UNL Expressions, UNL relations, Attributes, Universal words and UNL knowledge base. All these are key factors for preparing the word dictionary. Morphology and Punjabi word format have been analyzed by Joshi; Gill has developed a rule based part of speech tagger for Punjabi. Analysis of Hindi grammar for parts of speech tagger has been performed by Chakrabarti and Bhattacharyya. Generation of Hindi from UNL has been analyzed by Dwivedi. We have also gone through some other research papers that describe the template of dictionary entries of Bangla roots and primary suffix and Bangla to English dictionary to prepare the Bangla word dictionary by UNL structure which is provided by UNDL foundation.

3. UNL Structure

UNL system is composed of five elements. They are Universal Words, Attributes, Relations, UNL Expression and Knowledge Base.

A. Universal Words

Universal words are UNL words that carry knowledge or concepts. There are two types of UWs: permanent and temporary. Permanent UWs represent concepts of common use and are included in the UW dictionary. Temporary UWs may represent new concepts, too specific or not translatable so that they are not included in the dictionary. For example: Dhaka(iof>capital), Bangladesh(iof>country).

B. Attributes

Attributes are annotations used to represent grammatical categories, mood, aspect, etc. Every attribute starts with “@” symbol. For example: @Present: I go to school, @Past: It was raining yesterday, @Future: I will do the work.

C. Relations

UNL relation is a semantic network that makes relation between two words in a sentence. There are currently 46 relations in the UNL, and they define the syntax of UNL. They include argumentative (agent, object, goal), circumstantial (purpose, time, place), logic (conjunction, and disjunction) relations, etc. For example: I eat rice. Here agt(agent) relation is made between I and eat. & obj(object) relation is made between rice & eat.

D. UNL Expression

The UNL expresses information or knowledge in the form of semantic network. UNL semantic network is made up of a set of binary relations where each binary relation is composed of a relation and two UWs that hold the relation. A binary relation of UNL is expressed in the following format: <relation> (<uw1>, <uw2>)

In <relation>, one of the relations defined in the UNL specifications is described. In <uw1> and <uw2> the two UWs that hold the relation given at <relation> are described.

E. Knowledge Base
UNL Knowledge Base (UNLKB) defines every possible relation between concepts. The possible relations are defined based on a hierarchy of UWs (UW System). The UW System is built up by inclusive relations between concepts according to property inference mechanism of concepts. The architecture of the UW System allows introducing and defining any concept no matter how particular or specific it is.

Such UNLKB is a semantic network comprising every directed binary relation between UWs. It plays two roles: 1) defines the semantics (concepts) of UWs, and 2) provides linguistic knowledge of concepts. Concepts of UWs and linguistic knowledge of the concepts are defined by possible relations each concept can have with others. Such UNLKB not only provides linguistic knowledge in the form that computer can understand but also provides the semantic background of UNL expressions, that is the UNLKB ensures the meanings of UNL expression.

From the figure 2 we can represent how a sentence is translated to machine language through UNL structure easily. For that we use the following English sentence “Bangladesh is beautiful?!”

In the example above, “Bangladesh(iof>country)” and “beautiful(icl>attractive)”, which represent individual concepts, are UWs; “aoj” (= attribute of an object) is a directed binary semantic relation linking the two UWs; and “@interrogative”, “@present”, “@exclamation” and “@entry” are attributes modifying UWs. and icl and iof are relations where “ifo” indicates is an instance of and “icl” indicates is a kind of.

4. Template of Bangla Word Dictionary

The Word Dictionary is a collection of the word dictionary entries. Each entry of the Word Dictionary is composed of three kinds of elements: the Headword (HW), the Universal Word (UW) and the Grammatical Attributes. The complete set of UWs composes the UNL dictionary. The UNL dictionary is complemented with local bilingual dictionaries (here we use Bangla) , connecting UWs with head word from natural language . Local dictionaries are formed by pairs of form:

<Headword, UW>. Being Headword any word from a given natural language and UW the corresponding representation of its sense in UNL.
The following is a pair linking a Bangla Headword with its UW:

[ঢাকা] {} “Dhaka(iof>capital>thing)” (N, NPRO, VEND, CAPT, DIVS, SASI, ASIA, DIST ,CITY, #PLC, #PLF, #PLT)

Each Dictionary entry has the following format of any native language word [10]. Data Format:

[HW] {ID} “UW” (Attribute1, Attribute2,…) <FLG, FRE, PRI>

Here,

HW ← Head Word (Bangla word)  
ID ← Identification of Head Word (omitable)  
UW ← Universal Word  
ATTRIBUTE ← Attribute of the HW  
FLG ← Language Flag (we use B for Bangla)  
FRE ← Frequency of Head Word  
PRI ← Priority of Head Word

Format of an element of Bangla-UNL Dictionary would be:

Figure 3: Format of Word Dictionary

We are concerned how to make Bangla Word Dictionary for UNL. In UNL Knowledge Base (UKB) made by the UNL center of UNDL Foundation (Last updated version) on November 18, 2012, there are 96280 formats of Universal Words (UWs). We can find the UWs for each of the Bangla HW by searching the UNL KB to develop Bangla Word Dictionary for UNL. As per our observation this is not the suitable way to find out the UWs for the Bangla HW.

Firstly, it is a lengthy process to build Bangla Word Dictionary for UNL by searching the appropriate UWs from a huge number of words formats in UNL KB. Secondly, a word may have two or more meanings. Such types of words are represented with
various concepts in UNL KB. So, which one to choose out of two or more meanings for a Head Word is a tough job and we cannot get out proper/accurate words for the corresponding Bangla HWs.

We have found a new way (easiest and shorten) of searching based on existing works of other languages especially for English. Firstly, we can take some manually translated texts from Bangla to English in different forms and then convert them into UNL expressions (using English-UNL EnConverter to UNL expressions).

For example, **Assertive sentence**: আমি বাংলাদেশ হইতে পাকিস্তান যাচ্ছি (Aami Bangladesh hoite Pakistan jachhi) in English “I am going to Pakistan from Bangladesh.”

**Interrogative sentence**: কি আমি বাংলাদেশ হইতে পাকিস্তান যাচ্ছি (Ami ki Bangladesh hoite Pakistan jachhi) in English “Am I going to Pakistan from Bangladesh?”

**Negative sentence**: আমি বাংলাদেশ হইতে পাকিস্তান যাচ্ছি না (Aami Bangladesh hoite Pakistan jachhi na) in English “I am not going to Pakistan from Bangladesh.”

If we convert the first sentence by the English-UNL Converter we get the following UNL expressions shown in Table 1

<table>
<thead>
<tr>
<th>Table 1: UNL expression of the sentence “I am going to Pakistan from Bangladesh”</th>
</tr>
</thead>
<tbody>
<tr>
<td>agt(go(icl&gt;move&gt;do,plt&gt;place,plf&gt;place,agt&gt;thing).@entry.@present.@progress,i(icl&gt;person))</td>
</tr>
<tr>
<td>to(go(icl&gt;move&gt;do,plt&gt;place,plf&gt;place,agt&gt;thing).@entry.@present.@progress,Bangladesh)</td>
</tr>
<tr>
<td>plf(go(icl&gt;move&gt;do,plt&gt;place,plf&gt;place,agt&gt;thing).@entry.@present.@progress,Pakistan)</td>
</tr>
</tbody>
</table>

The same way, if we convert the two other sentences above, we get the same concepts of the words “I”, “am”, “going”, “to”, “Pakistan”, “from”, and “Bangladesh” respectively. As we know that Dictionary Entries are made using HW (Head Word), UW (Universal Word) and GA (Grammatical Attributes) so that, the Bangla Words “আমি” (aami) “বাংলাদেশ” (Bangladesh), “হইতে” (hoite), “পাকিস্তান” (Pakistan) and “যাচ্ছি” (jachhi) can be represented as. [আমি]“i(icl>person)” [বাংলাদেশ] “iof>place>thing” [পাকিস্তান] “iof>capital>thing” [হইতে] “go (icl>move>do)” and [যাচ্ছি] “.”

Similarly, by manually translating the different types of simple Bangla sentences (with variety of words) to English sentences and then English sentences to UNL expressions, we can get the appropriate concepts of thousand of Bangla Words to build the Bangla Word Dictionary for UNL.
Khaled, Ali & Shaheed

Secondly, we can take texts from some reliable translated sources (from Bangla to English) from Bangla Academy Scientific literatures. Then we can convert them into UNL expressions as above sentences and again can get the constraint lists of thousands of words for dictionary entries.

During formation of Bangla Word Dictionary for UNL we have resolved many ambiguities. Say, many Bangla Words have two or more English meanings. Similarly, many English Words also have two or more Bangla meanings. These concepts are not enough for representing the words for the dictionary entries. We have developed templates or sentence conversions.

5. Dictionary Development of Bangla Words

Based on process of developing Universal words and grammatical attributes [6,7], the template of the dictionary is:

\[ [\text{HW}] {\text{"Universal Word"}}(N, \text{NPRO/ NCOM/ NCON/ NABS, VEND/ CEND/ DIST/ CAPT}/ \text{CITY/ CONT, #PLC, #PLF, #PLT}) \]

We have formats of word dictionary for Bangla Head word as follows which include capitals, countries, and continents of the world.

\[ [\text{Dhaka}(iof>capital>thing)"(N, \text{NPRO, VEND, CAPT, DIVS, SASI, ASIA, DIST ,CITY, #PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Canberra}(iof>capital>thing)"(N, \text{NPRO, VEND, CAPT, OCEA ,CITY, #PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Colombo}(iof>city>thing)"(N, \text{NPRO, VEND, CAPT, DIVS, SASI, ASIA, DIST ,CITY, #PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Bangladesh}(iof>country>thing)"(N, \text{NPRO, CEND, ASIA,#PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Pakistan}(iof>country>thing)"(N, \text{NPRO, CEND, ASIA,#PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Asia}(iof>continent>thing)"(N, \text{NPRO, VEND, #PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Australia}(iof>continent>thing)"(N, \text{NPRO, VEND, #PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{Europe}(iof>continent>thing)"(N, \text{NPRO, VEND, #PLC, #PLF, #PLT}) <B,0,0> \]

\[ [\text{italic}(icl>adj,equ>slanted)"(N,LOGY, NCOM, VEND)<B,0,0> \]

\[ [\text{lion}(icl>big_cat>thing)"(N,NCOM, #OBJ,ANIM,CEND)<B,0,0> \]

Where attributes N stands for Noun, NPRO means proper noun, NCOM means common noun, VEND for vowel ended, CEND for consonant ended, and CAPT means capital respectively. Also Here

Grammatical Part: N, NPRO, NCOM
Morphological Part: VEND, CEND
6. Conversion of a Bangla Sentence into UNL Expression

A set of relevant dictionary entries and rules are to be used to convert a native language sentence into UNL expressions. Here we convert a Bangla sentence with some relevant dictionary entries to UNL representation as follows:

For example, if we convert the following sentence says,

**Sentence:** আমি বাংলাদেশ থেকে পাকিস্তান চ্ছি

**Pronounce as** “Aami Bangladesh hoite Pakistan jachhi.”

**Meaning:** “I am going to Pakistan from Bangladesh.”

To convert this sentence the following dictionary entries are needed.

- [আমি] { “ i(id>person) ” (PRON,NPRO, SG,1P,SUB)
- [বাংলাদেশ] { ”Bangladesh (iof>place>thing)” (N, NPRO, CEND, DIVS, SAS,CITY, #PLC, #PLF)
- [হইতে] { ” (ABY, ANUS#FRM)
- [পাকিস্তান] { ”Pakistan (iof>capital>thing)” (N, PROP, VEND, CAPT, DIVS, SAS, ASIA, DIST ,CITY, #PLC, #PLF, #PLT)
- [যা] { ”go (icl>move>do)” (VR,VE,DEF,VEG3)
- [চ্ছি] { ” (VI,P1,PRT,PRG, DEF,CH)

The above sentence is converted into the following UNL expression by the English-UNL converter shown in table 2.

<table>
<thead>
<tr>
<th>Table 2: UNL Expression of the Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>{unl}</td>
</tr>
<tr>
<td>agt(go(icl&gt;move&gt;do,plt&gt;place,plf&gt;place,agt&gt;thing),@entry.@present.@progress,i(icl&gt;person))</td>
</tr>
<tr>
<td>to(go(icl&gt;move&gt;do,plt&gt;place,plf&gt;place,agt&gt;thing),@entry.@present.@progress,Bangladesh)</td>
</tr>
<tr>
<td>plf(go(icl&gt;move&gt;do,plt&gt;place,plf&gt;place,agt&gt;thing),@entry.@present.@progress,Pakistan)</td>
</tr>
<tr>
<td>}</td>
</tr>
</tbody>
</table>

7. Conclusions and Future Works

We have proposed dictionary formats of Bangla head words for UNL based Bangla word dictionary considering the grammatical and semantic attributes using standard dictionary format of UNL.

We have successfully converted some simple Bangla sentences using the dictionary developed by our team .It shown that our dictionary entries are working properly on conversation.

We have not yet developed all kinds of format for different Bangla words and case makers. Our future plan is to make more templates for other head words including their prefixes, suffixes and inflexions and templates for morphological and semantic
rules based on the format provided by the UNL center of the UNDL foundation which will be helpful for the further research work on UNL based Bangla Machine Translation.

References


*UNL Word Dictionary*, viewed 20 April 2013,

*UNL Ontology-UW System*, viewed 20 April 2013,


Khaled, Ali & Shaheed


