IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 1 (Sep-Oct. 2012), PP 14-19
www.iosrjournals.org

Outlining Bangla Word Dictionary for Universal Networking Language

Mohammad Zakir Hossain Sarker[1], Md. Nawab Yousuf Ali[2],
Jugal Krishna Das[3]
[1](Jahangirnagar University , Bangladesh)
[2](East West University, Bangladesh)
[3](Jahangirnagar University, Bangladesh)

Abstract: *Universal Networking Language (UNL) is a computer language that enables computers to process information and knowledge across the language barriers. It is an artificial language that replicates the functions of natural languages in human communication. 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. For this common representation, lexical knowledge is a critical issue in natural language processing systems. We have been working to include Bangla in the UNL system and in this paper we have discussed about the Bangla Word Dictionary that we have designed to include in the system.*

Keywords: *Universal Word, Head Word, Grammatical Attributes, Universal Networking Language*

I. Introduction

Although, there is an immense proliferation of information through Internet, it is not accessible to vast multitude of people across nations as most of the resources are in English. To overcome this problem, United Nations launched Universal Networking Language project [1] under the auspices of United Nations University, Tokyo. The project team, after reviewing all such previous attempts, developed universal networking language (UNL), a language neutral specification, and universal parser specification [2] which is considered as a milestone for overcoming the language barrier for web publication. 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. For example, Bangla corpora, once converted to UNL, can be translated to any other language given UNL system built for that language. The strength of the UNL system lies in the fact that it emphasizes to represent the semantics of a native language sentence ignoring the complexities of natural languages. The en-converter converts each native language sentence to a UNL document and de-converter translates the UNL document to any native language. The UNL document is itself in English as it is known to linguistics. The development of the native language specific components - *dictionary* and *analysis rules*- is carried out by researchers across the world. The UNL project currently includes 16 official languages such as Arabic, Chinese, English, French [3], Russian, Hindi [4] but very little effort has been made so far to convert Bangla language to UNL expressions. We have been working on this topic from the last 3 years. To convert Bangla sentences into UNL expression we needed to go through Bangla verb, verb root, consonant ended root, vowel ended root, verbal inflections, tense, case structure, persons, etc. from [5], [6], [7], [8]. Then we have studied to gather knowledge about dictionary and how it could be used in case of UNL system [1], [2], [3]. Finally after having all the knowledge we have worked on outlining Bangla dictionary to be used in Bangla to UNL conversion and vice versa.

The organization of this paper is as follows: In Section 2 we describe the Research Methodology, Section 3 has the detail about UNL, Section 4 describes our work – how we outline Bangla Dictionary to use in converting Bangla sentences into UNL expression. Finally, Section 5 draws conclusions with some remarks on future works.

II. Literature Review

For converting Bangla sentence to UNL expressions firstly, we have gone through Universal Networking Language (UNL) [1, 2, 3, 9, 10] 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 natural language sentence (here Bangla sentence) into UNL expressions. Secondly, we have rigorously gone through the Bangla grammar [4, 5, 6, 7, 8], Verb and roots (Vowel ended and Consonant Ended), Morphological Analysis, Primary suffixes [11, 12, 13, 14, 15], construction of Bangla sentence based on semantic structure. Using above references we extort ideas about

Bangla grammar for morphological and semantic analysis in order to prepare Bangla word dictionary (for verb root, verbal inflections, etc)in the format of UNL provided by the UNL center of the UNDL Foundation.

1. About UNL

1.1. What is UNL?

The UNL consists of Universal words (UWs), Relations, Attributes, and UNL Knowledge Base. The Universal words constitute the vocabulary of the UNL, Relations and attribute constitutes the syntax of the UNL and UNL Knowledge Base constitutes the semantics of the UNL.

1.2. Why the UNL is necessary?

A computer in future needs a capability to make knowledge processing. Knowledge processing means a computer takes over thought and judgment of humans using knowledge of humans. It is necessary to make a processing based on contents. Computers need to have knowledge for knowledge processing. It is necessary for computers to have a language to have knowledge like human. It is also necessary to have a language to process contents like human. The UNL is a language for computers to do so. The UNL can express knowledge like a natural language. The UNL can express contents like a natural language.

1.3. What is different from others?

Systems which can deal with knowledge and contents have already been developed. But, their representation of knowledge or contents is different from each other. Moreover, their representations are language dependent. Namely, concept primitives used to represent knowledge are language dependent.

Knowledge or contents of a system cannot be used in other systems.

The situation is same as machine translation. For example, if we put all the result of research and development of machine translation, we cannot realize multilingual machine translation systems which can break language barriers.

1.4. Advantage of common language for computers

The UNL greatly reduces development cost of developing knowledge or contents necessary to make knowledge processing by sharing knowledge and contents. Furthermore, if every knowledge necessary for doing something by software is described in a language for computers such as the UNL, software only need to interpret instructions written in the language to perform it functions. And those instructions could be shared by other software. Then we can accumulate such knowledge for computer like a library for humans.

1.5. How the UNL express information?

The UNL represents information, i.e. meaning, sentence by sentence. Sentence information is represented as a hyper-graph having Universal Words (UWs) as nodes and relations as arcs. This hyper-graph is also represented a set of directed binary relations, each between two of the UWs present in the sentence.

The UNL expresses information classifying objectivity and subjectivity. Objectivity is expressed using UWs and relations. Subjectivity is expressed using attributes by attaching them to UWs.

A UNL document, then, will be a long list of relations between concepts.

2. Outlining a Bangla Word Dictionary for UNL

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**. A headword is a notation/surface of a word of a natural language that composing the input sentence and it is to be used as a trigger for obtaining equivalent UWs from the Word Dictionary in EnConversion. An UW expresses the meaning of the word and is to be used in creating UNL networks (UNL expressions) of output. Grammatical Attributes are the information on how the word behaves in a sentence and they are to be used in enconversion rules. Each Dictionary entry has the following format of any native language word [1].

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

FRE ← Frequency of Head Word

PRI ← Priority of Head Word
Format of an element of Bangla-UNL Dictionary is shown in Fig 1

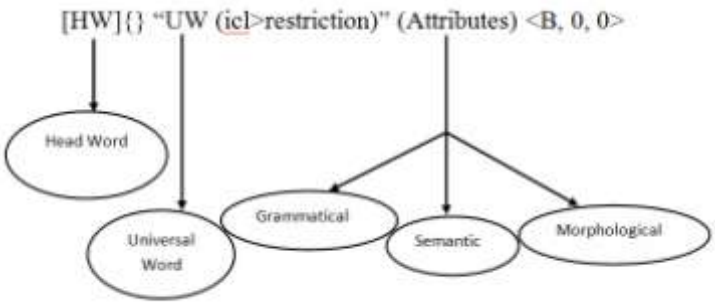

Figure 1: Format of a Bangla Word Dictionary

Now we are concerned how to make Bangla Word Dictionary for UNL. In UNL Knowledge Base (KB) made by the UNL center of UNDL Foundation (Last updated version) in 2004, there are 21862 formats of Universal Words (UWs) [1]. 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 perception this is not the suitable way to find out the UWs for the Bangla HW.

Firstly, it is a long 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 hard job and we can't get out suitable/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) [2].
For example,

Assertive sentence: Avwg fvZ LvB‡ZwQ | (aami bhat khaitechhi) in English " I am eating rice."
Interrogative sentence: Avwg wK fvZ LvB? (aami bhat khai), in English "Do I eat rice?"
Negative sentence: Avwg fvZ LvB bv | (aami bhat khai na) in English "I do not eat rice."

If we convert the first sentence by the English-UNL Converter [2] we get the following UNL expressions shown in Table 1

Table 1: UNL expression of the sentence "I am eating rice"

agt(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).@entry.@present.@progress,i(icl>person))
obj(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).@entry.@present.@progress,rice(icl>cereal>thing))

The same way, if we convert the two other sentences above, we get the same concepts of the words "I", "eat" and "rice" 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 "Avwg" (aami) "Lv" (kha) and "fvZ" (bhat) can be represented as.
[Avwg] {}" i(icl>person)" , [Lv] {}"eat(icl>consume>do)" and [fvZ] {}"rice(icl>cereal>thing)"

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.

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. For example, we use "‡m"(she) in Bangla, but in English it has two meanings "he" and "she". Again, we use "rice" in English, but in Bangla it has three meanings "fvZ"(bhat) or "PvDj"(chaul) or "avb" (dhan). "avb"(dhan) means paddy in English, when it is in the field. To resolve these ambiguities we can represent them in the dictionary as follows.
[‡m(cyi"l)] {} "he(icl>person)"
[‡m(gwnjv)] {} "she(icl>person)"

[fvZ] {} "rice(icl>cereal>thing)"
[PvDj] {} "rice(icl>grain>thing)"
[avb] {} "rice(icl>grain>thing)"
[avb] {} "rice(icl>paddy>thing)"
[`qvjy] {} "kind(icl>sympathy>thing)"
[cÖKvi] {} "kind(icl>category>thing)"

These concepts are not enough for representing the words for the dictionary entries. We have developed templates for assigning the grammatical attributes for the words, roots and their inflexions that are useful for developing EnConversion and DeConversion rules for sentence conversions.

2.1. Development of Grammatical Attributes

Representing Universal Words (UWs) for each of the Bangla Head Word we need to develop grammatical attributes that describe how the words behave in a sentence. They play very important roles for writing Enconversion and Deconversion rules because a rule uses GA in morphological and syntactic analysis, to connect or analyze one morpheme with another to build a meaningful (complete) word and to examine or define the position of a word in a sentence. When we analyze the HWs for representing them in the word dictionary as UWs, we find all the possible specifications of the HWs as attributes named grammatical attributes, so that they can be used in the dictionary for making rules (EnCo and DeCo). For example, if we consider "cvwLÓ (pakhi) meaning bird as a head word, then we can use attributes N (as it is noun), ANI (as bird is an animal), SG for singular number and CONCRETE (as it a concrete thing which is touchable).
So, this word can be represented in the dictionary as follows:
[cvwL] {} "bird(icl>animal>animate thing)"(N, ANI, SG, CONCRETE)

Head Universal Word Grammatical Attributes

Similarly, we can represent the words avb (paddy), bvP& (dance) in the Word Dictionary as follows:
[avb] {} "paddy(icl>plant>thing)"(N, PLANT, CONCRETE)
[bvP&]{} "dance(icl>do)"(ROOT, BANJANT)
Some proposed grammatical attributes for developing Word Dictionary of Bangla words and morphemes, analysis and generation rules for encoversion and deconversion are shown in Table 2.

Table 2: Proposed grammatical attributes

Grammatical Attributes	Descriptions of attributes	Examples (Here we use Bangla/English words)
ADJ	adjective	fvj (good), my›`i (beautiful) etc.
ALT	alternative root	গি (gi), যে (je) etc.
ABY	indeclinable	এবং (and), জন্য (for) etc.
BOCH	articles	টি (ti), টা (ta), গুলা (gula), গুলি (guli) etc.
BIV	normal inflexions	অন্ত (onto), অই (oi) etc.
7TH	seventh Bivokti (Inflexion)	এ(e), য়(oy), তে(te) etc.
5TH	fifth Bivokti (Inflexions)	হইতে (hoite), থেকে (theke) etc.
3RD	third Bivokti (Inflexions)	দ্বারা (dara), দিয়ে (die) etc.
2ND	second Bivokti (Inflexions)	কে (ke) etc.
CEND	verb roots or nouns that are ended with	co& (read), ai& (catch), লন্ডন (London)
CEG	verb roots of consonant ended groups	co& (read), লিখ্ (write), পারিস (Paris) etc.
CMPL	verbal inflexions that can combine with roots	য়েছি (echhi), য়েছিলাম (echhilam) etc.
CHL	inflexions that are used for cholti language	তাম (tam), লাম (lam) etc.
CONCRETE	solid thing	জমি (land), ঘর (house), etc.
FUT	verbal inflexions that are combined with roots	বে (be) etc.
FEM	female person	সে (মহিলা), she (female) etc.
HON	respected pronouns	আপনি (you), তিনি (he) etc.
HPRON	human pronoun	আমি (ami), সে (she) etc.
IMPR	verbal inflexions that can combibe with roots	ও (o), ন (n) etc.
KPROT	the suffixes that are used after roots to create	BK (ik), Ab (on) etc.
KBIV	verbal inflexions	B (i), B‡ZwQ (itechhi), †e (be) etc.

MNOUN	the suffixes that are added with roots to make	Av etc.
MADJ	the suffixes that are added with roots to make	Aš— etc.
MAL	male person	সে (পুরুষ), he (male) etc.
N	any noun	Kjg (pen), Avg(mango) etc.
NPRO	proper noun	দুলাল (Dulal),-name of a person, পদ্মা (Padma),-name of a river etc.
NCOM	common noun	gvbyl (Man), গরু (Cow), MvQ (Tree), gvQ (Fish) etc.
NMAT	material noun	Rj (water), evZvm (air), AvKvk (sky), ‡jvnv (iron) etc.
NABS	abstract noun	myL (happiness), `ytL (sadness) etc.
NP	noun phrase	কলম দিয়ে (by pen) etc.
NUM	number	৫ (5), ৭(7), ৯(9) etc.
NANI	not amimate	বই (book), কলম (pen) etc.
NGL	neglected pronouns	তুই (you), তোরা (you [pl]) etc.
PL	plural number	আমরা (amra), তাহারা (tahara) etc.
PRON	pronoun	Avwg (I), Avgiv (we) etc
PSTEM	pronoun stem	আমা (ama), তোমা (toma) etc.
PROT	all suffixes	Av (a), Ab (on), AvB (ai) etc.
1P	first person pronouns	Avwg (I), Avgiv (we) etc.
2P	second person pronouns	Zzwg (you) etc.
3P	third person pronoun	‡m cyi“l (He), ‡m gwnjv (She) etc.
PRS	the suffixes that are added with roots to create	B (i), etc.
PRS	verbal inflexions that are combined with roots	B (i), B‡ZwQ (itechhi) etc
PST	verbal inflexions that are combined with roots	লাম (lam) etc.
PRGR	verbal inflexion that are combined with roots	ইতেছি (itechi), ইতেছিলাম (itechhilam),
PSUFF	primary suffixes	অন্ত (onto), আ(a) etc.
ROOT	verb root	Pv (want), hv(go), co& (read), ai&
SHD	inflexions that are used for shadhu language	B (i), B‡ZwQ (itechhi) etc.
SG	singular number	আমি (ami), তুমি (tumi) etc.
UROOT	consonant	দুল্ (dul), খুল (khul) etc.
VEND	verb roots or nouns that are ended with vowel	Pv (want), hv(go) etc.
VEG	verb roots of vowel ended groups	Pv (want), hv(go) etc.
VERB	verb	খাই (i), খাইতেছি (khaitechi)

III. Conclusion

In this paper we have discussed about outlining the Bangla Dictionary to use in the Universal Networking Language. We have also shown the grammatical attributes, that describe how the words behave in a sentence, to represent Universal Words (UWs) for each of the Bangla Head Word. Now we have to develop the dictionary entries for various Bangla words, which will be required in converting Bangla sentences into UNL expression.

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

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

[4] Bhattacharyya, (2001) Multilingual Information Processing Using Universal Networking Language, in Indo UK Workshop on Language Engineering for South Asian Languages LESAL, Mumbai, India

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

[6] D. C. Shuniti Kumar. Bhasha-Prakash Bangala Vyakaran, (Rupa and Company Prokashoni, Calcutta, July 1999, pp.170-175)

[7] Humayun Azad. Bakkotottyo - Second edition, (Bangla Academy Publishers, Dhaka, 1994)

[8] D. S. Rameswar, Shadharan Vasha Biggan and Bangla Vasha, (Pustok Biponi Prokashoni, November 1996, pp.358-377)
[9] http://www.undl.org last accessed on 29 July 2012
[10] DeConverter Specification, Version 2.7, (UNL Center, UNDL Foundation, Tokyo 150-8304, Japan 2002)
[11] 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
[12] M.N.Y. Ali, J.K. Das, S.M. Abdullah Al Mamun, A. M. Nurannabi. "Morphological Analysis of Bangla words for Universal Networking Language", International Conference on Digital Information Management, icdim, 2008, London, England, pp. 532-537
[13] M.N.Y.Ali, A. M. Nurannabi, G. F. Ahmed, J.K. Das. "Conversion of Bangla Sentence for Universal Networking Language", International Conference on Computer and Information Technology (ICCIT), Dhaka, 2010 pp. 108-113
[14] Md. Ershadul H. Choudhury, Md. Nawab Yousuf Ali, Mohammad Zakir Hossain Sarker, Ahsan Razib. "Bridging Bangla to Universal Networking Language- A Human Language Neutral Meta-Language", In proceedings International Conference on Computer, and Information Technology (ICCIT), Dhaka, 2005 pp. 104-109
[15] Md. Nawab Yousuf Ali, Mohammad Zakir Hossain Sarker , Ghulam Farooque Ahmed , Jugal Krishna Das."Conversion of Bangla Sentence into Universal Networking Language Expression", International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011, pp. 64-73