Generation of Bangla Text from Universal Networking Language Expression

Md. Nawab Yousuf Ali[1], Shaikh Muhammad Allayear[1], M. Ameer Ali[1], and Golam Sorwar[2]

[1] Department of Computer Science and Engineering, East West University, Bangladesh
[2] Southern Cross University, Australia
nawab,allayear@ewubd.edu, ameer7302002@yahoo.com, golam.sorwar@scu.edu.au

Abstract. This paper presents a work on generating Bangla sentences from an interlingua representation called Universal Networking Language (UNL). UNL represents knowledge in the form of semantic network like hyper-graphs which contains disambiguated words, binary semantic relations, and speech act like attributes associated with the words, assisted by the semantically rich lexicon and a set of analysis and generation rules. We have developed a set of generation rules for converting UNL expression to Bangla sentences. Our experiment shows that these rules successfully generate correct Bangla sentences from UNL expressions.

Keywords: Interlingua, Universal Networking Language, Universal Word, Morphology, DeConverter, Generation Rules.

1 Introduction

A significant part of the development of any machine translation (MT) system is the creation of lexical resources that the system will use. Dictionaries are of critical importance in MT. They are the largest components of an MT system in terms of the amount of information they hold. Generation of natural language from a machine processable, precise knowledge representation has to grapple with the problem of redundancy and impreciseness inherent in any natural language. In the UNL system [1], natural language analysis is carried out by two tools: EnConverter (EnCo) [2] and DeConverte (DeCo) [3]. Both tools are associated with a word dictionary of the native language and a set of language specific analysis and generation rules. EnCo converts a natural language text into corresponding UNL expressions whereas DeCo converts UNL expressions to a variety of natural languages. We have worked on various types of simple sentences. For brevity, in this paper we present the steps to construct a simple affirmative sentence. In our earlier work [4], [5-6], we have shown how to convert Bangla sentences into UNL expressions. In this paper we present a set of generation rules to convert UNL expressions into Bangla sentences. By using these generation rules along with a supporting word dictionary Bangla . Analysis of Hindi

V.V. Das, E. Ariwa, and S.B. Rahayu (Eds.): SPIT 2011, LNICST 62, pp. 157–163, 2012.

grammar for parts of speech tagger has been performed by Chakrabarti and Bhattacharyya [7], Chakrabarti*et al.* [8] and Singh *et al.* [9] and generation of Hindi from UNL has been analyzed by Dwivedi [10]. Hindi grammar has been analyzed to create UNL based MT system for Hindi language. Hindi generation rules for Hindi EnConverter have been created by analyzing Hindi grammar by Giri [11], Dave *et al.* [12], and Bhattacharyya *et al* [13]. The analysis of Tamil morphology for the development of Tamil EnConverter for EnConversion of Tamil to UNL has been performed by Dhanabalan*et al* [14]. Similar kinds of works have been done in many other countries such as French, Spanish [15], Chinese, English, Russian, and German [16]. For Bangla language processing, research has been done for morphological analysis of Bangla words [17], parsing methodology for Bangla sentences [18] and dictionary development of Bangla words[19]. Suffix, prefix and inflexions are discussed in [20]. No previous attempt has been made to convert Bangla texts into UNL expressions and UNL expressions to Bangla texts.

Section 2 explains the different stages in the generation process and outlines the format of dictionary entries. After proposing the generation rules and illustrating the experimental results in section 3, we summarize the paper with some concluding remarks in section 4.

2 Stages in the Generation Process

The generation process consists of three main stages morphological generation of lexical words, function words insertion, and syntax planning. In morphological analysis, Bangla nouns inflect for number and case, and can be described as having major categories of the forms based on the oppositions direct-oblique and singular-plural. They can be categorized into masculine and feminine gender in terms of their agreement with adjectives and verbs. In UNL, plural nouns are represented using the attribute *@pl,* and singular ones remain unspecified (absence of *@pl*refers to a singular noun). Gender and vowel endings are stored in the UNL-Bangla dictionary. The morphological rules based on word paradigms generate a noun form using all this information, *viz.*, lexical, relational, and UNL attributes.Bangla verbs inflect based on vowel ended and consonant ended roots [6]. Inflections are marked either on the main verb or on its auxiliaries that appear as free morphemes.

Roles of Root and Verbal Inflexion in the Formation and Meaning of a Verb. A root contains the core meaning, which relates with the action or state of the verb, whereas verbal inflexion (VI) defines the formation of the verb and reflects person, tense (in case of finite verb) and other properties [7].

Variations of Roots. For development of the lexicon for UNL compatible Bangla Word Dictionary and rules for morphological and semantic analyses, Vowel Ended and Consonant Ended roots have been divided into several groups [6]. It has been observed that some of the roots change their forms when they combine with some specific VIs to make verbs. All the variations of a root appear in the lexicon at different entities, though they all contain same UW but in case of grammatical

attributes we use ALT (for first alternative), ALT1(for second alternative) and ALT2 (third alternative) etc. and rest of the attributes will be the same for all variations.

UNL encodes case information by using relation labels assigned as per the properties of the connected nodes. Consider, for example, the translation of sentence, 'আমি কলম দিয়ে চিঠি লিখছি', pronounce as "Ami kolom die chitthi likhchhi" means *I am writing a letter with a pen*. Here, the case marke 'দিয়ে' *(diye)* and 'ছি' *(chhi)* are inserted to derive the relation 'কলম', *kolom* (pen) and 'লিখছি', *likhchhi* (writing) have with the verb 'write'. Given a node along with all its lexical attributes from the UNL Bangla dictionary, an appropriate case marker is inserted. Similarly, other function words like- conjunctions, disjunctions, particles, *etc.*, are also inserted to represent clausal information.

Dictionary Format. Each entry of the Word Dictionary is composed of three kinds of elements: the **Headword (HW),** the **Universal Word (UW)** and the **Grammatical Attributes** [4], [16].

Data Format: **[HW]{ID}"UW"(Attribute1, Attribute2,...)<FLG, FRE, PRI>**

According to the dictionary format some examples are as follows:

[আমি]{}"i(icl>person)"(PRON,HPRON, P1,SG,SUBJ)<B,1,1>
[ভাত]{}"rice (icl>food)"(N)<B,0,0>

The entries in parentheses are morpho-syntactic and semantic attributes of Bangla words which control various generation decisions choosing special case makers.

3 Conversion of UNL Expression to Bangla Texts

A set of generation rules is to be used to generate native language sentences from UNL expressions. The DeConverter finds the most suitable rule to create a native language sentence. A set of native language sentences from UNL expressions will finally be generated after applying all the necessary rules. Among the various types of generation rules described in [3], A*ttribute changing* (**:**) rule plays an important role to insert words and morphemes from node-nets of UNL expressions into node-list for making the sentences. Attribute changing rule is used to rewrite the attributes of the nodes in both left and right Generation Windows. If any rewriting action occurs in the node of the Left Generation Window, the position of the Generation Window moves to the left so that the Right Generation Window will always be placed on the rewritten node. It is also used for insertion node. In that case, either node in the Generation Window must be indicated as being an inserted node. If the left node is an inserted node, the Right Generation Window will move so that it is placed on it after the rule application.

3.1 Proposed Rules

In this section, formats of some generation rules have been proposed that are to be used for converting UNL expressions to Bangla sentences. We define format of rules

to insert subjective pronouns for agent (agt) relation of both alternative and not alternative roots. We also define rules to insert subjective pronouns for thing with attributes (aoj) relation of both alternative and not alternative roots. Format of rules is also defined to insert verbal inflexions at the end of roots for first, second and third persons. Finally, format of rules to insert nouns before roots and to insert articles for singular and plural are defined. For example, the format of rules to insert subjective pronouns of alternative roots for agent relation is defined as follows,:"HPRON,(x)P,SUBJ, [^] @respect, | [^]@contempt, | [^]HON, [^] NGL :: agt:"{ROOT,VEND,[^]@present|@progress|@complete,VEG(y),ALT,#AGT,^(x)p:(x)p::}P10;Here, the grammatical attributes 'HPRON' for human pronoun, '(x)P' indicates person and the value of 'x' denotes first, second or third person., 'SUBJ' for subject of a sentence, @respect for respected person, @contempt for neglected person, 'agt' for agent relation, 'ROOT' for verb root, 'VEND' for vowel ended root, @present for present tense, @progress for continuous tense, @complete for perfect tense, VEG(y) for vowel ended group, where 'y' denotes group number, 'ALT' for alternative root '#AGT' indicates that the corresponding root involves with agent relation. and 'p' is the temporary attribute for person to prevent recursive operations. Similarly, the format of rules to insert noun before roots is defined as follows,:"N,[^]@pl,^SUBJ:SUBJ:agt:"{ROOT,VEND,#AGT,^3p,[^]sg|pl:3p,sg|pl::}P10; Here, N denotes noun, 'sg' for singular and 'pl' for plural numbers. Within the limited scope of this paper we avoided presenting all the format of rules. Interested readers are referred to [21], for a detailed description of format of all the rules.

3.2 Experimental Results

This section describes the conversion procedures and the experimental results of the UNL expressions into Bangla sentence. The UNL expressions of the sentence, *John has eaten a mango with spoon* is shown in Table 1 by using Russian and English language server [22]. In the UNL expressions, *agt*(agent), *obj* (object) and *ins* (instrument) are the **semantic relations**. The relaters eat(icl>consume>do,agt>living_thing,obj>concrete_thing), John(icl>name,iof>person,com>male)), mango(icl>edible_fruit>thing)) and spoon(icl>cutlery>thing) are the **Universal Words (UWs)** [1]. These are language words with restrictions mentioned in parentheses for the purpose of denoting a unique sense. *icl* stands for *inclusion* and *iof* stands for *instance of*. Attribute @entry typically attached to the main predicate. We have used a DeConverter [23] tool for our experiment. The tool takes as its input a UNL expression file (Table 1), a set of generation rules (Table 3) and a dictionary file (Table 2) and generates sentence of the target language.

Table 1. UNL expressions of the sentence 'John has eaten mango with a spoon'

agt(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).@entry.@present,john(icl>name,iof>person,com>male))
obj(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).@entry.@present,spoon(icl>cutlery>thing).@indef)
ins(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).@entry.@present,mango(icl>edible_fruit>thing))

Table 2. Dictionary entries for respective Bangla sentence

[জন]{}"John (iof>person)"(N, NPRO, 3P, SG,SUBJ)<B,1, 1>
[চামচ] {}"spoon(icl>cutlery>thing)"(N,#INS,CEND)
[আম] {} "mango(icl>edible_fruit>thing)"(N,NCOM, #OBJ,CEND)<B,0,0>
[খা]{}"eat(icl>consume>do,agt>living_thing,obj>concrete_thing)"(ROOT,VEND,#AGT, #OBJ, VEG1)<B,0,2>
[খাচ্ছে] {}"VI" (VI,VEND,3P,PRS,CMPL,CHL)
[দিয়ে] {}"INF"(INF, 3RD, CEND)

Table 3. Generation rules for converting UNL expression to Bangla sentence

Rule 1: (Noun insertion)
:"N, SUBJ,^@pl,::agt:"{ROOT,VEND,#AGT,^3p,^sg:3p,sg::}P10;
Rule 2: (Right shift)- R{:::}{SUBJ:::}
Rule 3: (Blank insertion)- :{SUBJ,^blk:blk::}"[],BLK:::"P10;
Rule 4: (Right shift)- R{:::}{SUBJ:::}
Rule 5: (Right shift)- R {SUBJ:::}{:::}
Rule 6: (Noun insertion)
:"N,^ins:ins:ins:"{ROOT,VEND,#INS:::}P9;
Rule 7: (Right shift)- R{:::}{N,INS:::}
Rule 8:(Insertion rule of case maker)
:{N,INS:::}"[[INF]],INF,CEND:::"P10;
Rule 9: (Right shift)- R{:::}{N,INS:::}
Rule 10: (Right shift)- R{N,INS:::}{:::}
Rule 11:(Blank insertion)- :{INF,^blk:blk::}"[],BLK:::"P10;
Rule 12:(Blank insertion)- :{N,INS,^blk:blk::}"[],BLK:::"P10;
Rule 13:(Right shift)- R{:::}{N,INS:::}
Rule 14:(Right shift)- R{N,INS:::}{:::}
Rule 15:(Right shift)- R{:::}{INF:::}
Rule 16:(Right shift)- R{INF:::}{:::}
Rule 17: (Noun insertion)
:"N,^obj:obj:obj:"{ROOT,VEND,^#AGT,#OBJ:::}P9;
Rule 18: (Right shift)- R{:::}{N,#OBJ:::}
Rule 19:(Blank insertion)- :{N,#OBJ,^blk:blk::}"[ ],BLK:::"P10;
Rule 20: (Right shift)- R{:::}{N,#OBJ:::}
Rule 21:(Right shift)- R{N,#OBJ:::}{:::}
Rule 22:(Right shift)- R{:::}{ROOT,VEND:::}
Rule 23: (Verbal inflexion insertion)
:{ROOT,VEND,3p,#AGT,@present,@complete,^@progress,^kbiv:kbiv::}
"[[KBIV]],KBIV,VEND,3P,PRS,CMPL,^PRGR"P10;
Rule 24: (Right shift)- R(V:::}{:::}

These generation rules will be applied to the nodes in the node-list for operation on them and/or inserting nodes from the Node-net into the Node-list.

Rule 1 describes when root "খা" (khe) is in the RGW (Right Generation Window) the noun "জন" (John) is to be inserted in the RGW. Rule 2 is applied to shift the windows of DeConverter to right. The blank insertion rule (rule 3) is applied to insert a blank space between noun and root. After applying right shift rules (rule 4, 5), rule 6 is to be used to insert noun "চামচ", *chamoch* (spoon) in the RGW. If the noun "চামচ" is in the RGW, the windows will be shifted to right by applying rule 7. The case maker insertion rule (rule 8) is to be applied to insert 'দিয়ে' (diye) on the right side of the RGW. Two right shift rules 9 and 10 are to be applied to shift the windows two steps right followed by two blank insertion rules 11and 12 to insert blank spaces between case maker and root, and noun "চামচ", and case maker respectively. Subsequently, noun 'আম", *aam* (mango) is to be inserted into the node-list (rule 17) after applying right shift rules 13, 14, 15 and 16. To make a blank space between noun, 'আম", and root, rule 19 is to be applied after using right shift rule 18. Finally, right shift rules 20, 21 and 22 are to be applied followed by a verbal inflexion insertion rule (rule 23). The right shift rule 24 completes the sentence generation processes of DeConverter. After completing the deconversion procedures, DeCo generates the following Bangla sentence, জন চামচ দিয়ে আম খেয়েছে.

We have experimented different types of simple sentences by varying subjects, persons as well as tenses. Our experiment showed that Bangla sentences are generated correctly by the proposed generation rules.

4 Conclusions

We have proposed a set of generation rules to generate Bangla sentences from UNL expressions. The paper also focused on the dictionary formats of Bangla words and case makers considering grammatical and semantic attributes using standard dictionary format of UNL. We have analyzed various types of simple Bangla sentences. By using the generation rules we successfully translated correct Bangla text from UNL expressions. It is now possible to generate any simple Bangla sentence from UNL expressions. Our long term plan is to develop a mechanism which will allow us to translate any language into corresponding Bangla texts through UNL expressions. This paper focused only on simple sentences. Currently, we are experimenting on both compound and complex sentences and respective generation rules. Our generation rules are defined by following standard formats so that generation rules of other languages can be benefited from our formats. Completion of the generation rules for all types of sentences will be a major step towards developing a generic Bangla language translator.

References

1. Uchida, H., Zhu, M., Senta, T.C.D.: Universal Networking Language, UNDL Foundation, International environment house, 2005/6, Geneva, Switzerland
2. EnConverter Specification, Version 3.3, UNL Center/UNDL Foundation, Tokyo 150-8304, Japan (2002)

3. DeConverter Specification, Version 2.7, UNL Center, UNDL Foundation, Tokyo 150-8304, Japan (2002)
4. Ali, M.N.Y., Das, J.K., Abdullah Al Mamun, S.M., Nurannabi, A.M.: Morphological Analysis of Bangla words for Universal Networking Language. In: ICDIM 2008, London, England, pp. 532–537 (2008)
5. Ali, M.N.Y., Sarker, M.Z.H., Das, J.K.: Analysis and Generation of Bengali Case Structure Constructs for Universal Networking Language. IJCA International Journal of Computer Applications, 34–41 (March 2011)
6. Ali, M.N.Y., Sarker, M.Z.H., Farooque, G.A., Das, J.K.: Conversion of Bangla Sentence into Universal Networking Language Expression. IJCSI 8(2) (March 2011)
7. Chakrabarti, D., Bhattacharyya, P.: Syntactic Alternation of Hindi Verbs with Reference to Morphological Paradigm. In: Language Engineering Conference, Hyderabad, India (December 2002)
8. Chakrabarti, D., Sarma, V., Bhattacharyya, P.: Hindi Verb Knowledge Base and Noun Incorporation in Hindi. In: Third Global WorldNet Conference, Jeju Island, Korea (January 2006)
9. Singh, S., Gupta, K., Shrivastava, M., Bhattacharyya, P.: Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi. In: COLING/ACL, Sydney, Australia (July 2006)
10. Vijay, D.: Generation of Hindi from Universal Networking Language. IIT Bombay M Tech Thesis
11. Giri, L.: Semantic Net Like Knowledge Structure Generation from Natural Languages. IIT Bombay B Tech Dissertation (2000)
12. Deve, S., Bhattacharyya, P.: Knowledge Extraction from Hindi Text. JIETE 18(4) (2001)
13. Deve, S., Parikh, J., Bhattacharyya, P.: Interlingua Based English Hindi Machine Translation and Language Divergence. Journal of Machine Translation (JMT) 16(4), 251–304 (2001)
14. Dhanabalan, T., Saravanan, K., Geetha, T.V.: Tamil to UNL EnConverter. ICUKL, Goa, India (2002)
15. Gilles, S., Christian, B.: UNL-French Deconversion as Transfer & Generation from an Interlingua with Possible Quality Enhancement through Offline Human Interaction. Machine Translation Summit-VII, Singapore (1999)
16. Ali, M.N.Y., Das, J.K., Abdullah Al Mamun, S.M., Choudhury, M.E.H.: Specific Features of a Converter of Web Documents from Bengali to Universal Networking Language. In: ICCCE 2008, Kuala Lumpur, Malaysia, pp. 726–731 (2008)
17. Asaduzzaman, M.M., Ali, M.M.: Morphological analysis of Bangla Words for Automatic Machine Translation. In: International Conference on Computer and Information Technology, Dhaka, Bangladesh, pp. 271–276 (December 2003)
18. Asaduzzaman, M.M., Ali, M.M.: A Knowledge Based Approach to Bangla-English Machine Translation for Simple Assertive Sentences. International Journal of Translation 15, 77–97 (2003)
19. Islam, M.S.: Research on Bangla Language Processing in Bangladesh: Progress and Challenges. In: 8th ILDC, Dhaka, Bangladesh (June 2009)
20. Khairunnahar, K.: Morphological Analysis of Bangla Prefix. The Dhaka University Journal of Linguistic 1(2), 157–168 (2008)
21. http://www.ewubd.edu/~nawab
22. http://www.unl.ru/deco (last access: July 20, 2011)
23. http://www.undl.org/ (last access: July 20, 2011)