International Journal of Advanced Science and Technology
Vol.74 (2015), pp.49-62
http://dx.doi.org/10.14257/ijast.2015.74.06

A New Approach: Automatically Identify Naming Word from Bengali Sentence for Machine Translation

Md. Syeful Islam[1] and Dr. Jugal Krishna Das[2]
[1]*Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh, Phone: +8801916574623*
[2]*Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh, Phone: +8801712509082*
[1]*syefulislam@yahoo.com*, [2]*drdas64@yahoo.com*

Abstract

More than hundreds of millions of people of almost all levels of education and attitudes from different country communicate with each other for different purposes using various languages. Machine translation is highly demanding due to increasing the usage of web based Communication. One of the major problem of Bengali translation is identified a naming word from a sentence, which is relatively simple in English language, because such entities start with a capital letter. In Bangla we do not have concept of small or capital letters and there is huge no. of different naming entity available in Bangla. Thus we find difficulties in understanding whether a word is a naming word or not. Here we have introduced a new approach to identify naming word from a Bengali sentence for machine translation system without storing huge no. of naming entity in word dictionary. The goal is to make possible Bangla sentence conversion with minimal storing word in dictionary.

Keywords: *Machine translation, UNL, Rule based analysis, Morphological analysis, Post Converted, Naming word, Knowledge base*

1. Introduction

Today the demand of inter communication between all levels of peoples in various country is highly increased. This globalization trend evokes for a homogeneous platform so that each member of the platform can apprehend what other intimates and perpetuates the discussion in a mellifluous way. However the barriers of languages throughout the world are continuously obviating the whole world from congregating into a single domain of sharing knowledge and information. Therefore researcher works on various languages and tries to give a platform where multi lingual people can communicate through their native language. Researcher analyze the language structure and form structural grammar and rules which used to translate one language to other. From the very beginning the Indian linguist Panini proposed vyaakaran (a set of rules by which the language is analyzed) and gives the structure for Sanskrit language. After the era of Panini various linguist works on language and proposed various technique. But the most modern theory proposed by the American linguist Noam Chomsky is universal grammar which is the base of modern language translation program.

From the last few years several language-specific translation systems has been proposed. Since these systems are based on specific source and target languages, these have their own limitations. As a consequence United Nations University/Institute of Advanced Studies (UNU/IAS) were decided to develop an inter-language translation program [1]. The corollary of their continuous research leads a common form of languages known as Universal Networking Language (UNL) and introduces UNL system. UNL system is an initiative to overcome the problem of language pairs in automated

ISSN: 2005-4238 IJAST

translation. UNL is an artificial language that is based on Interlingua approach. UNL acts as an intermediate form computer semantic language whereby any text written in a particular language is converted to text of any other forms of languages [2-3]. UNL system consists of major three components: language resources, software for processing language resources (parser) and supporting tools for maintaining and operating language processing software or developing language resources.

Like other machine translation, the parser of UNL system take input sentence and start parsing based on rules and convert it into corresponding universal word from word dictionary. The challenge in detection of named is that such expressions are hard to analyze using machine translation parser because they belong to the open class of expressions, *i.e.*, there is an infinite variety and new expressions are constantly being invented. Bengali is the seventh popular language in the world, second in India and the national language of Bangladesh. So this is an important problem since search queries on word dictionary for naming word while all naming word (proper noun) cannot be exhaustively maintained in the dictionary for automatic identification.

This work aims at attacking this problem for Bangla language, especially on the human names detection from Bengali sentence without storing naming word in dictionary. In this paper we apply this generalized system into machine translation program UNL and any linguistic can apply this technique to their machine translation system.

This research paper is organized as follows: Section 2 deals with the problem domain and Section 3 provides the theoretical analysis of Machine Translation and Universal Networking Language. In Section 4 functioning of UNL En-Converter are described. Section 5 Introduce the new approach to identify naming word from a Bengali sentence for machine translation program. Section 6 results analysis demonstrate the new invented approach by applying it on UNL. Finally Section 7 draws conclusions with some remarks on future works.

2. Problem Domain

Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as Bengali) to another (such as English). To process any translation, human or automated means the meaning of a text in the original (source) language must be fully restored in the target language. While on the surface this seems straightforward, it is far more complex. Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another. This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), *etc.*, in the source and target languages, as well as familiarity with each local region.

UNL is one type of machine translation system. UNL represent sentences in the form of logical expressions, without ambiguity. The purpose of introducing UNL in communication network is to achieve accurate exchange of information between different languages. Information expressed in UNL can be converted into the user's native language with higher quality and fewer mistakes than the computer translation systems. Researchers already start works on Bengali language to include it with UNL system.

Human language like Bangla is very rich in inflections, vibhakties (suffix) and karakas, and often they are ambiguous also. That is why Bangla parsing task becomes very difficult. At the same time, it is not easy to provide necessary semantic, pragmatic and world knowledge that we humans often use while we parse and understand various Bangla sentences. Bangla consists of total eighty-nine part-of-speech tags. Bangla grammatical structure generally follows the structure: subject-object-verb (S-O-V) structure [4-5]. We also get useful parts of speech (POS) information from various inflections at morphological parsing. But the major problem is identifying the name from

sentence and convert is very difficult. In this section we try to clear the problem domain and define some point why the processing of naming word is difficult.

In terms of native speakers, Bengali is the seventh most spoken language in the world, second in India and the national language of Bangladesh. There is a huge no. of naming word existing in this language and every time new expressions are constantly being invented.

Named identification in other languages in general but Bengali in particular is difficult and challenging as:

- There is huge no. of naming word available in Bangla and it's not wise decision to store all of naming word in word dictionary. It causes slow performance.
- Unlike English and most of the European languages, Bengali lacks capitalization information.
- Bengali person names are more diverse compared to the other languages and a lot of these words can be found in the dictionary with some other specific meanings.
- Bengali is a highly inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms.
- Bengali is a relatively free order language.

In Bengali language conversion, En-Converter parse the sentence word by word and find word from dictionary and apply rules. When En-Converter doesn't find any word from dictionary then En-Converter creates a temporary entry for this word. In the maximum case the temporary entry is name word. We apply some rules to ensure that this words which is not in dictionary (temporary entry) are naming word.

The later sections we discuss about a new technique of identify the naming word from Bangla sentence and define a post converter for convert the Bangla name to universal word. And finally we apply this in UNL to demonstrate the new approach.

3. Machine Translation and Universal Networking Language

The Internet has emerged as the global information infrastructure, revolutionizing access to information, as well as the speed by which it is transmitted and received. With the technology of electronic mail, for example, people may communicate rapidly over long distances. Not all users, however, can use their own language for communication.

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another. Machining translation is the translation of text by a computer, with no human involvement.

On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus and statistical techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies. Here we discuss one of machine translation system Universal Networking language (UNL).

The Universal Networking Language (UNL) is an artificial language in the form of semantic network for computers to express and exchange every kind of information. Since the advent of computers, researchers around the world have worked towards developing a system that would overcome language barriers. While lots of different systems have been

developed by various organizations, each has their special representation of a given language. This results in incompatibilities between systems. Then, it is impossible to break language barriers in all over the world, even if we get together all the results in one system.

Against this backdrop, the concept of UNL as a common language for all computer systems was born. With the approach of UNL, the results of the past research and development can be applied to the present development, and make the infrastructure of future research and development.

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. The UNL expresses information or knowledge in the form of semantic network with hyper-node. UNL semantic network is made up of a set of binary relations, each binary relation is composed of a relation and two UWs that old the relation [6].

4. UNL En-Converter

To convert Bangla sentences into UNL form, we use En-Converter (EnCo), a universal converter system provided by the UNL project. The EnCo is a language independent parser, which provides a framework for morphological, syntactic and semantic analysis synchronously. Natural Language texts are analyzed sentence by sentence by using a knowledge rich lexicon and by interpreting the analysis rules. En-Converter generates UNL expressions from sentences (or lists of words of sentences) of a native language by applying En-conversion rules. In addition to the fundamental function of En-conversion, it checks the formats of rules, and outputs massages for any errors. It also outputs the information required for each stage of En-conversion in different levels. With these facilities, a rule developer can easily develop and improve rules by using En-Converter [7].

First, En-Converter converts rules from text format into binary format, or loads the binary format En-conversion rules. The rule checker works while converting rules. Once the binary format rules are made, they are stored automatically and can be used directly the next time without conversion. It is possible to choose to convert new text format rules or to use existing binary format rules.

➔ Convert or load the rules.

Secondly, En-Converter inputs a string or a list of morphemes/words of a sentence s native language.

➔ Input a sentence.

Then it starts a apply rules to the Node-list from the initial state (Figure 1).

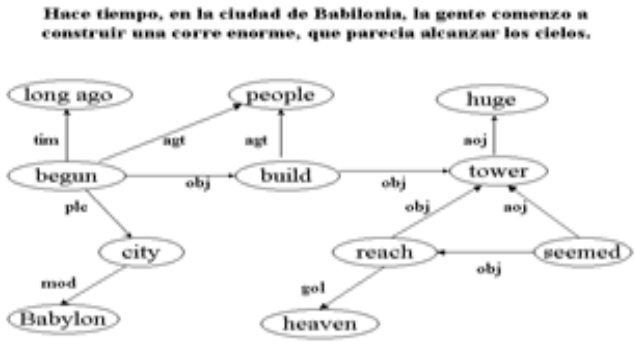

Figure 1. Sentence Information is Represented as a Hyper-Graph

En-Converter applies En-conversion rules to the Node-list its windows. The process of rule application is to find a suitable rule and to take actions or operate on the Node-list in order to create a syntactic tree and UNL network using the nodes in the Analysis windows. If a string appears in the window, the system will retrieve the word dictionary and apply the rule to the candidates of word entries. If a word satisfies the conditions required for the window of a rule, this word is selected and the rule application succeeds. This process will be continued until tree and UNL network are completed and only the entry node remains in the Node-list.

➔ Apply the rules and retrieve the Word Dictionary.

Finally the UNL network (Node-net) is outputted to the output file in the binary relation format of UNL expression.

➔ Output the UNL expressions.

With the exception of the first process of rule conversion and loading, once En-Converter starts to work, it will repeat the other processes for all input sentences. It is possible to choose which and how many sentences are to be En-converted.

4.1. Temporary Entries

Temporary entry is a kind of flag to mark an unknown word to further analysis for identifying tis word is naming word or not. In UNL there is an attribute name TEMP to flagging an unknown word. For UNL the following two cases, En-Converter creates a temporary entry by assigning the attribute "TEMP"("TEMP" is the abbreviation for "Temporary").

- Except for an Arabic numeral or a blank space, if En-Converter cannot retrieve any dictionary entry from the Word Dictionary for the rest of the character string, or
- When a rule requiring the attribute "TEMP" is applied to the rest of the character string.

The temporary entry has the following format and it's UW, shown inside the double quotation "and", is assign to be the same as its headword (HW).

[HW] {} "HW" (TEMP) <, 0, 0>;

The next chapter we proposed the technique of identifies the naming word from Bangla sentence and defines a post converter for convert the Bangla name to universal word.

5. Automatically Naming Word Identification Approach

The naming word conversion is relatively simple in English language, because such entities start with a capital letter. In Bangla we do not have concept of small or capital letters. Thus we find difficulties in understanding whether a word is a naming word or not. For example, the Bangla word "BISWAS" can be a proper noun (*i.e.*, a family name of a person) as well as an abstract noun (with the meaning of faith in English). For example, in order to understand the following Bangla sentence, we must need an intelligent parser. A parser takes the Bangla sentence as input and parses every sentence according to various rules [8-9].

Here we proposed a method for machine translation to identify naming word from any Bengali sentence with minimal storing word in word dictionary which is a combination of dictionary-based, rule-based approaches. In this approach, En-Converter identifies the naming word using rules and morphological analysis. The approaches are sequentially described here and demonstrated in Figure 2.

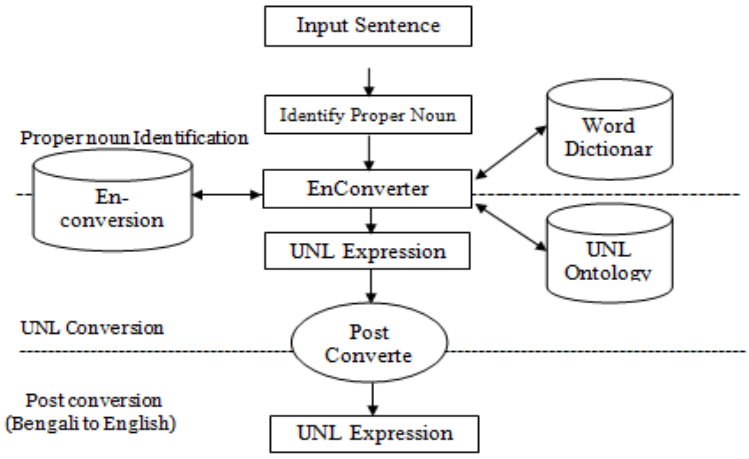

Figure 2. Naming Word Identification and Conversion in UNL

5.1. Naming Word Detection Approach

In machine translation system firstly take an input sentence and parse it word by word and search it from dictionary the relative word. If not found then mark it as a temporary word and try to recognize that the temporary word is a naming word based on defines rules. If this process is fail then morphological analysis is used. The approaches of naming word detection are sequentially described here and demonstrated in Figure 3.
Here we describe the process in three steps.

1) Dictionary based analysis for naming word detection
2) Rule-based analysis for naming word detection
3) Morphological Analysis

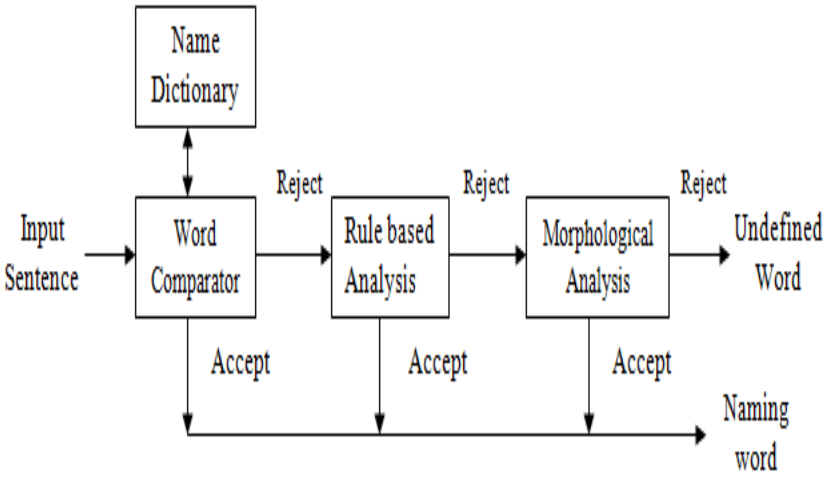

Figure 3. Naming Word Identification Technique

5.1.1. Dictionary based Analysis for Naming Word Detection: If a dictionary is maintained where we try to attach most commonly used naming word and it is known as Name dictionary. Here we describe the dictionary based naming word detection technique sequentially.

Firstly the input sentence is processed on en-converter which finds the word on word dictionary. If the word is found then it is converted into relative universal word. If not in dictionary then En-Converter creates a temporary entry for this word.

Secondly the en-converter finds the word with flag TEMP into name dictionary. If it is found then it is conceder as the word may be noun and apply rules to ensure.

Finally if the word is not in name dictionary then it sends into morphological analyzer to conform that the word is naming word (proper noun).

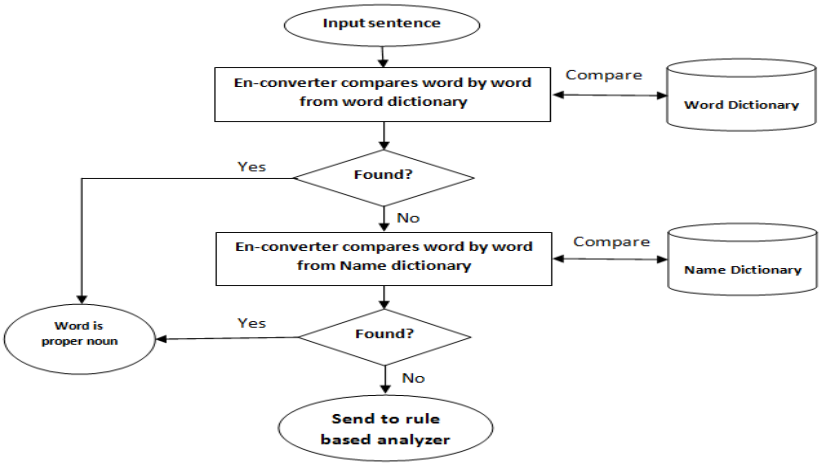

Figure 4. Flow Chart of Dictionary based Analyses

5.1.2. Rule-based Analysis for Naming Word Detection: Rule-based approaches rely on some rules, one or more of which is to be satisfied by the test word. Some words which use in Bangla sentence as a part of name. Here we take a technique to identify naming word using such word (part of name). Firstly we make dictionary entry with pof (part of name) and other relevant attribute. To identify naming word from Bangla sentence use pof word some typical rules are given below.

Rule 1- If the parser find the following word like মোঃ, মুঃ, মোসাম্মৎ, মোছাম্মৎ, আঃ, আব্দুল, আব্দুর, মিঃ, মিস, মিসেস, বেগম, বিবি, ডক্টর, ডঃ, আলী, শেখ, স্বামী, সৈয়দ, রেভারেন্ড, শ্রী, শ্রীযুক্ত etc. then the word is considered as the first word of name and set the status of the word first part of name (FPN). The word or collection of words after FPN with status TEMP is also considered as part of name.

Rule 2- If the parser find the following word (title words and mid-name words to human names) like চৌধুরী, মিয়া, মিঞা, চট্টোপাধ্যায়, মুখপাধ্যায়, খান, হোসেন, হোছাইন, রহমান, হোসাইন, ঘোষ, বোস, বসু, মিত্র, রায়, সরকার, খান, আহমেদ, রহমান, হক etc. and কুমার, চন্দ্র, রঞ্জন, শেখর, প্রসাদ, আলী, আলম *etc.*, after temporary entry word. Then last part of name (LPN) and temporary entry word along with such words are likely to constitute a multi-word name (proper noun). For example, রবি বসাক, পল্লব কুমার মল্লিক are all name.

Rule 3- If there are two or more words in a sequence that represent the characters or spell like the characters of Bangla or English, then they belong to the name. For example, বি এ (BA), এম এ (MA), এম বি বি এস (MBBS) are all name. Note that the rule will not distinguish between a proper name and common name.

Rule 4- If a substring like বাবু, দাদা, দা, সাহেব, কাকু, গঞ্জ, গ্রাম, পুর, গড়, নগর occurs at the end of the temporary word then temporary word is likely to be a name.

Rule 5- If a word which is in temporary entry ended with এ- র, রা, এর, র, র, রা, এরা, কে, দের, তে, য় then the word is likely to be a name.

Rule 6- If a word like- সরনী, রোড, স্ট্রিট, লেন, থানা, স্কুল, বিদ্যালয়, কলেজ, নদী, লেক, হ্রদ, সাগর, মহাসাগর, পাহাড়, পর্বত is found after temporary word then NW along with such

word may belong to name. For example - বিজয় সরনী, রাসেল স্ট্রিট, চিম্বুক পাহাড় all are name.

Rule 7- If the sentence containing বলেন, বললেন, বলল, শুনল, হাসল, লিখল, লিখলেন,খেলেন, দেখল after temporary word then the temporary word is likely to be a proper noun.

Rule 8- If at the end of word there are suffixes like টা, খানা, খানি, টাতে, টায়, টিকে, টাকে, টকুন, গুলা, গুলো, গুলি etc., and then word is usually not a proper noun.

5.1.3. Morphological Analysis for Naming Word Detection: When previous two steps fail to identify naming word or there is confusion about the word is naming word or not then we applying morphological analysis to sure that the unknown word is naming word (proper noun). In this case we conceder the structure of words and the position of word in the sentence and identify the word type [10].

Rule 1- Proper noun always ended with 1st, 2nd, 5th and 6th verbal inflexions (Bibhakti). So when parser fined an unknown word with 1st, 2nd, 5th and 6th bibhakti then the word may be proper noun [11-12].

Rule 2- From sentence structure if parser fined an unknown word is in the position of karti kaarak and word is ended with 1st bibhakti then it is conceder as a proper noun.

Rule 3- If the unknown word position is in the position of karma kaarak and it is indirect object and word is ended with 2nd bibhakti then it is conceder as a proper noun. But for direct object it is not a proper noun.

Rule 4- If any sentence contains more than one word ended with 1st bibhakti then the first word is with flag unknown word then must be karti kaarak and the word is proper noun.

Rule 5- In case of apaadaan kaarak, if any word in the sentence ended with 5th bibhakti and the word is unknown then it must be noun. If most of all other's noun are in word dictionary so unknown word ended with 5th bibhakti must be proper noun.

Rule 6- In case of adhikaran kaarak, if any word in the sentence ended with 7th bibhakti and the word is unknown then it may be the name of place. If the word is not in dictionary then it is conceder as proper noun (name of any pace).

In that time, when En-Converter identify an word or collection of word as a proper noun and En-Converter convert into UNL expression it keep track temporary word using custom UNL relation tpl and tpr. We use two relation "tpr" and "tpl" to identify the word which should converted. The relation "tpr" use when En-Converter finds the temporary word after pof (part of name) attribute and "tpr" use when En-Converter finds the temporary word before pof (part of name) attribute. When En-Converter found "tpr" relation then it converts the word which is after blank space. For the case of "tpl" it converts the first word. If proper noun contains only one word with attribute TEMP then using this TEMP attribute converter convert.

5.2. Function of Post Converter

In previous section we identify naming word from bangle sentence. Now we need to convert it to target language. That's why we have designed a post converter. UNL En-converter converts the sentence into corresponding intermediate UNL expression. But there is little bit problem, En-Converter convert those word which is in word dictionary. In the case of temporary word which is already identified as a proper noun or part of proper noun cannot converted and it is same as in Bengali sentence. It is difficult to other people who cannot read bangle language. So it must be converted into English for UNL

expression. Post converters do this conversion. The function of post converter demonstrated in Figure 5.

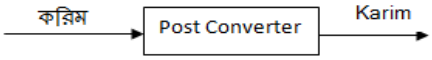

Figure 5. Function of Post Converter

5.3. Proposed En-conversion Process

Here I use simple phonetic bangle to English translation method. Simple en-converter takes the input as a temporary word and converts it into corresponding target language using phonetic approach. Here we describe the conversion of bangles naming word to English word for UNL. We use same En-Converter which is used in UNL. Firstly need to create dictionary for Bengali to English conversion. Then rules are define for converter which uses these rules for conversion. When En-Converter found "tpr" relation then it converts the word which is after blank space. For the case of "tpl" it converts the first word. The functions of post converter are sequentially described here and architecture of Post converter demonstrated in Figure 6.

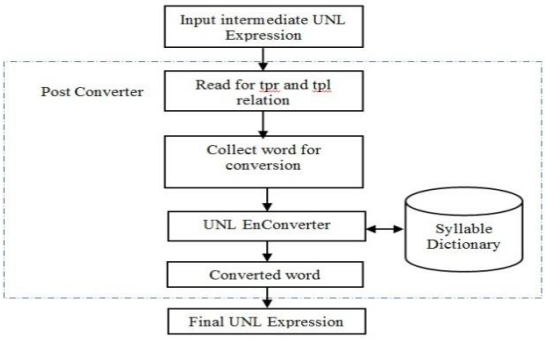

Figure 6. Architecture of Post Converter (Bengali to English)

5.3.1. Algorithm: How post converter converts Bangla word for intermediate UNL expression to English for final UNL expression. The process are describe step by step-

Step 1: In first step the UNL expression is the inputs of post converter for find the final UNL expression.

Step 2: In second step post converter read the full expression and fined relation tpr or tpl. The relation tpr and tpl are used to identify the word which should convert.

Step 3: At third step Post converter collect Bangla word which are converted within this post converter using the help of above two relations. When En-Converter found "tpr" relation then it collects the word which is after blank space and for tpl it collects the first word.

Step 4: In this steps converter convert the word applying rules and finding the corresponding English syllable or word form syllable dictionary.

Step 5: In this step we get the converted word which is placed on final UNL expression.

Step 6: This steps is the final steps which generate final UNL expression.

Here we have listed some dictionary entries for post converter. Table 1 shows the Bengali vowel and Table 2 shows the shows the Bengali consonant and the corresponding entries in dictionary. In Table 3 it shows some dictionary entries for consonant plus vowel (kar). Here we only try to present how post converter converts Bengali to English. In future we define the full phonetics for Bengali to English conversion.

Table 1. Dictionary Entries for Bengali Vowel

Bangla vowel	Dictionary entries
অ	[অ]{} "a" (TEMP) <.,0,0>
আ	[আ]{} "a" (TEMP) <.,0,0>
ই	[ই]{} "i" (TEMP) <.,0,0>
ঈ	[ঈ]{} "ei" (TEMP) <.,0,0>
উ	[উ]{} "oo" (TEMP) <.,0,0>
ঊ	[ঊ]{} "u" (TEMP) <.,0,0>
ঋ	[ঋ]{} "rri" (TEMP) <.,0,0>
এ	[এ]{} "a" (TEMP) <.,0,0>
ঐ	[ঐ]{} "oi" (TEMP) <.,0,0>
ও	[ও]{} "o" (TEMP) <.,0,0>
ঔ	[ঔ]{} "ou" (TEMP) <.,0,0>

Table 2. Dictionary Entries for Bengali Consonant

Bangla consonant	Dictionary Entries
ক	[ক]{} "k" (TEMP) <.,0,0>
খ	[খ]{} "kh" (TEMP) <.,0,0>
গ	[গ]{} "g" (TEMP) <.,0,0>
ঘ	[ঘ]{} "gh" (TEMP) <.,0,0>
ঙ	[ঙ]{} "ng" (TEMP) <.,0,0>
চ	[চ]{} "c" (TEMP) <.,0,0>
ছ	[ছ]{} "ch" (TEMP) <.,0,0>
জ	[জ]{} "j" (TEMP) <.,0,0>
ঝ	[ঝ]{} "jh" (TEMP) <.,0,0>
ঞ	[ঞ]{} "niya" (TEMP) <.,0,0>
ট	[ট]{} "t" (TEMP) <.,0,0>
ঠ	[ঠ]{} "th" (TEMP) <.,0,0>
ড	[ড]{} "d" (TEMP) <.,0,0>
ঢ	[ঢ]{} "dh" (TEMP) <.,0,0>
ণ	[ণ]{} "n" (TEMP) <.,0,0>
ত	[ত]{} "t" (TEMP) <.,0,0>
থ	[থ]{} "th" (TEMP) <.,0,0>
দ	[দ]{} "d" (TEMP) <.,0,0>
ধ	[ধ]{} "dh" (TEMP) <.,0,0>
ন	[ন]{} "n" (TEMP) <.,0,0>
প	[প]{} "p" (TEMP) <.,0,0>
ফ	[ফ]{} "f" (TEMP) <.,0,0>
ব	[ব]{} "b" (TEMP) <.,0,0>
ভ	[ভ]{} "v" (TEMP) <.,0,0>
ম	[ম]{} "mm" (TEMP) <.,0,0>
য	[য]{} "z" (TEMP) <.,0,0>
র	[র]{} "r" (TEMP) <.,0,0>
ল	[ল]{} "l" (TEMP) <.,0,0>
শ	[শ]{} "s" (TEMP) <.,0,0>

ষ	[ষ]{} "sh" (TEMP) <.,0,0>
স	[স]{} "s" (TEMP) <.,0,0>
হ	[হ]{} "h" (TEMP) <.,0,0>
ড়	[ড়]{} "r" (TEMP) <.,0,0>
ঢ়	[ঢ়]{} "rh" (TEMP) <.,0,0>
য়	[য়]{} "y" (TEMP) <.,0,0>
ৎ	[ৎ]{} "t" (TEMP) <.,0,0>

Table 3. Dictionary Entries for Bengali Consonant Plus Bengali Kar

Bangla parts of word	Dictionary entries
কা	[কা]{} "ka" (TEMP) <.,0,0>
কি	[কি]{} "ki" (TEMP) <.,0,0>
কী	[কী]{} "kei" (TEMP) <.,0,0>
কু	[কু]{} "koo" (TEMP) <.,0,0>
কূ	[কূ]{} "ku" (TEMP) <.,0,0>
কৃ	[কৃ]{} krriu" (TEMP) <.,0,0>
কে	[কে]{} "ka" (TEMP) <.,0,0>
কৈ	[কৈ]{} "koi" (TEMP) <.,0,0>
কো	[কো]{} "ko" (TEMP) <.,0,0>
কৌ	[কৌ]{} "kou" (TEMP) <.,0,0>
ক্র	[ক্র]{} "kra" (TEMP) <.,0,0>
ক্য	[ক্য]{} "kka" (TEMP) <.,0,0>

Similarly for all consonant it should need to entries in word dictionary.
Example 1-
Let's an intermediate UNL expression-

{unl}
agt(read(icl>see>do,agt>person,obj>information).@entry.@present.@progress,করিম:
TEMP :05)
{/unl}

Post converter firstly read the full sentence. When it finds the word "করিম" with attribute TEMP converter collect this word and push it into post converter. Then applying rules it convert into UNL word "karim". Converter parses "করিম" letter by letter.

ক = ক + অ -> Ka
রি = -> ri
ম -> m

That's mean "করিম" converted in "Karim"

Thus Post converter converts all naming word Bengali to English. Here we only try to present how post converter converts Bengali to English. In future we define the full for Bengali to English conversion.

6. Result Analysis

In this section we apply our proposed system in machine translation. For test case basis we choose UNL for applying this invented system. We can apply this approach at any kind of machine translation. To convert any Bangla sentence we have used the following files.

- ✓ Input file
- ✓ Rules file
- ✓ Dictionary

We have used an Encoder (EnCoL33.exe) and here I present some print screen of en-conversion.
Screen print shows the Encoder that produces Bangla to UNL expression or UNL to Bangla (Figure 7).

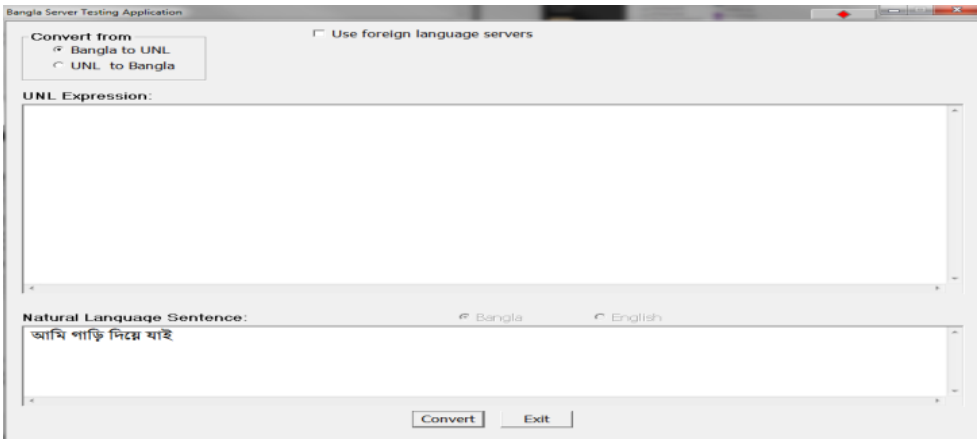

Figure 7. Encoder for En-conversion

When users click on convert button it generates corresponding UNL expression. The bellow screen shows this operation (Figure 8).

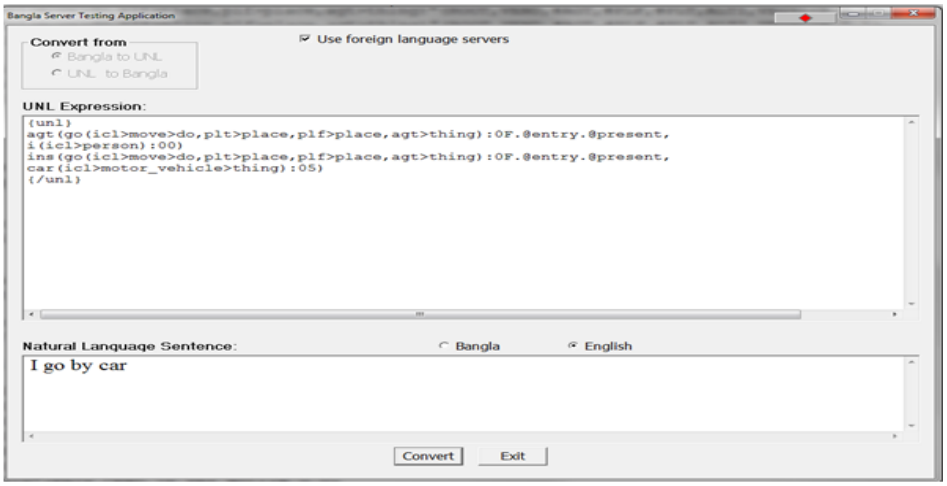

Figure 8. En-conversion

Based on the three steps pronoun detection technique we define rules for UNL system which identify the naming word (proper noun) from a Bengali sentence and create relative UNL expression from the sentence.

6.1. Example:

Let a sentence "মিসেস রহিমা লিখছিলেন" for conversion. To convert this sentence the following dictionary entries are needed for conversion the sentence.

[মিসেস] {} "Mrs." (icl>fpofname, iof>person,com>female)(N)

[লিখ] {} "write (icl>do, agt>person, obj>abstract_thing plc>thing,ins>functional_thing)" (ROOT, CEND)

[এছিলেন]{} "" (VI, CER, 3P, PST) T, CEND, ^ALT); Where, N denotes noun, ROOT for verb root, CEND for Consonant Ended Root, ^ALT for not alternative, VI is attribute for verbal inflexion.

For this sentence Mrs. Find from dictionary with attribute pof and the temporary word রহিমা y combined using relation tpr. For the word রহিমা the temporary entry as like:

[রহিমাy]{} "রহিমা" (TEMP) <.,0,0>;.[]{}

In that case when LAW found such word "মিসেস" and a blank space after it and if the next word is TEM then EnConverter takes the two words cindered as a name of female. And other parts of sentence conversion are similar.

{unl}
agt(write(icl>do,agt>person,obj>abstract_thing,plc>thing,ins>functional_thing).@entry.@past,
Mrs. রহিমা (TEMP))
{/unl}

Post Converter takes the UNL expression and converts the proper noun "রহিমা" Bengali to English "rahima".

র = র + অ -> ra
হি -> hi
মা -> ma
That's mean "রহিমা" converted in "rahima"

After conversion the final UNL expression is like as-

{unl}
agt(write(icl>do,agt>person,obj>abstract_thing,plc>thing,ins>functional_thing).@entry.@past,
misses rahima (icl>person))
{/unl}

Thus the UNL converter and Post converter can identify any naming word from Bengali sentence and convert it corresponding UNL expression. We can apply our approach at any machine translation with some modification. It reduce processing time and save memory space of databases.

7. Conclusion

Here we have defined a procedure to identified naming word (proper noun) from Bengali sentence and conversion method from bangle to UNL expression. We have also demonstrated how UNL converter identified naming word from Bengali sentence and the UNL expression conversion by taking a sentence as an example. Here we define our work as two parts, firstly identified a naming word (proper noun) from Bengali sentence and secondly applying it in machine translation and as an example we convert this naming word into UNL form. In the second parts we use a converter named as post converter which use simple phonetic method to convert Bengali to English. Any linguistic can choose this technique and apply this in there machine translation system. It save memory space and reduce the time for searching word from dictionary. We will also works on Bengali language and future plan to provide a complete faster and accurate machine translation technique for Bangla language.

References

[1] http://www.undl.org last accessed, **(2014)**, July 23.
[2] H. Uchida, M. Zhu and D. Santa, "A Gift for a Millennium", The United Nation University, Tokyo, Japan, **(2000)**.
[3] H. Uchida and M. Zhu, "The Universal Networking Language (UNL) Specification", Version 3.0, Technical Report, United Nations University, Tokyo, **(1998)**.
[4] D. C. Shuniti Kumar and B.-P. Bangala Vyakaran, Rupa and Company Prokashoni, Calcutta, **(1999)** July, pp. 170-175.

[5] D. S. Rameswar, S. Vasha Biggan and B. Vasha, "Pustok Biponi Prokashoni", **(1996)** November, pp. 358-377.
[6] R. T. Martins, L. H. M. Rino, M. D. G. V. Nunes and O. N. Oliveira, "The UNL distinctive features: interfaces from a NL-UNL enconverting task".
[7] EnConverter Specifications, version 3.3, UNL Center/ UNDL Foundation, Tokyo, Japan, **(2002)**.
[8] S. Abdel-Rahim, A. A. Libdeh, F. Sawalha and M. K. Odeh, Universal Networking Language(UNL) a Means to Bridge the Digital Divide, Computer Technology Training and Industrial Studies Center, Royal Scientific Sciety, **(2002)** March.
[9] Md. N. Y. Ali, J. K. Das, S. M. A. Al-Mamu and A. M. Nurannabi, "Morphological Analysis of Bangla Words for Universal Networking Language", Third International Conference on Digital Information Management (ICDIM **2008**), London, England. pp. 532-537.
[10] S. Dashgupta, N. Khan, D. S. H. Pavel, A. I. Sarkar and M. Khan, "Morphological Analysis of Inflecting Compound words in Bangla", International Conference on Computer, and Communication Engineering (ICCIT), Dhaka, **(2005)**, pp. 110-117.
[11] H. Azad, Bakkotottyo, Second edition, Dhaka, **(1994)**.
[12] J. Parikh, J. Khot, S. Dave and P. Bhattacharyya, "Predicate Preserving Parsing, Department of Computer Science and Engineering", Indian Institute of Technology, Bombay.

Authors

Md. Syeful Islam obtained his B.Sc. and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh in 2010 and 2011 respectively. He is now working as a Senior Software Engineer at Samsung R&D Institute Bangladesh. Previously he worked as a software consultant in the Micro-Finance solutions Department of Southtech Ltd. in Dhaka, Bangladesh. His research interests are in Natural Language processing, AI, embedded computer systems and sensor networks, distributed Computing and big data analysis.

Dr. Jugal Krishna Das obtained his M.Sc. in Computer Engineering from Donetsk Technical University, Ukraine in 1989, and Ph.D. from Glushkov Institute of Cybernetics, Kiev in 1993. He works as a professor in the department of Computer Science and Engineering, Jahangirnagar University, Bangladesh. His research interests are in Natural Language processing, distributed computing and Computer Networking.

