Development of Bangla Text Generator for Translation from English

Abstract

The work presented here deals with the development of Bangla language text generator module for developing English to Bangla Machine Aided Translation (MAT) System named as AnglaBangla System. AnglaBangla is a rule based MAT system with source language being English and the target language being Bangla. Source language analyzer and target language generator are the two major building blocks of AnglaBangla System. The challenge was to adopt the AnglaBharati technology and develop the Bangla language text generator. For English language analyzer, the system uses AnglaBharati technology, which gives the Pseudo Lingua for Indian Language (PLIL) [1] as an output. This paper covers the important modules of the text generator part; viz. morphological synthesis of verbs, as verb is the most important decision making element in a sentence for generating proper translation; and the role of the English preposition in the Bangla translation. Selection of Bangla pronominal form for an English pronoun has also been discussed here.

1. Introduction

India being a multilingual country, English continues to be the link language for administration, education and business. However, as English continues to be the link language, a machine translation system caters to English as the source language (SL) and the target language (TL) being all Indian languages, is considered to be a priority.

Fully automatic general purpose high quality machine translation systems are extremely difficult to build. In spite of the difficulties, it is possible to employ the computers for Machine Translation, although it sounds paradoxical. The solution lies in separating language based analysis of texts from knowledge and inference based analysis. The former is left to the machine and latter is being taken care of by the human readers. Thus, the aspects which are difficult for the human being are handled by the machine and easier aspects are left to the human being. The aim is to minimize the effort of the human being and thereby increase his productivity; hence we refer to it as Machine Aided Translation (MAT), not machine translation.

2. AnglaBangla System

AnglaBangla, an English to Bangla Machine Translation System, is a derivative of AnglaBharati. It uses a pseudo-interlingua approach. The interlingua-based approach depends upon the theme that a suitable universal intermediate representation can be defined for the source text, which is independent of the target language. This intermediate representation is presumed to have resolved all ambiguities, and so it should be possible to generate text in any Indian language as the target form this representation [2]. AnglaBharati is a pattern directed rule based system with context free grammar like structure for analysis of English as source language. The analysis generates a 'pseudo-target', referred as PLIL (Pseudo Lingua for Indian Language) applicable to a group of Indian languages. The input to the text generator is PLIL, which is a transformed (transferred) tree structure as per Indian languages derived from the parse tree of the source language (English).

This intermediate structure can be converted to any Indian language through a language specific text generator. The basic blocks of the AnglaBharti system consists of two major blocks, one is English language analyser and another one is target language generator, which is Bangla Text generator for AnglaBangla System. The authors have developed mainly a Bangla text generator and an English-Bangla lexical database to develop AnglaBangla MAT System.

PLIL Structure:

<S>:

<adv> <verb_pattern> <toinf_pattern> <sen_type> <sub_np> <connective> <pp> <connective> <obj_np> <connective> <toinf_pattern> <verb_pattern> <adv> ,vp> .sviram

Where S is sentence, adv is adverb, toinf is to-infinity, sen_type is sentence type, sub_np is subject noun phrase, pp is prepositional phrase, obj_np is object noun phrase and sviram is sentence end.

PLIL gives the information about the subject, object and verb of the input English sentence with all possible lexical meaning. If the sentences have two objects, it can also be identified by the PLIL. It provides the information about the gender, number and person (GNP) of each noun and Tense Aspect and Modalities (TAM) of each verb. In AnglaBharati verbs are divided in five sub categories according to their aspects, namely verb_1 to verb_5, these are the simple past, present, future and other two are continuous and perfect.

The example of PLIL of an English sentence is shown below:
English Sentence: He played football.
Bangla Translation: tini phuTabala khelechhilena.
PLIL:
<aff{sub_np (he noun masculine singular third [human] [**tini:** m 8] [] [])} {obj1_np (football noun neuter singular third [**thing**] [**phuTabala:** m 3] [] []} k1{main_vp_active (play_1 verb_3 normal normal masculine singular third [**khela**] 10 [] [])}> .svram

PLIL gives the information of the type of the English sentence, in this case, which is affirmative. Subject noun phrase is *'he'* and corresponding POS is noun, and the GNP is masculine, singular and third. The first object is *'football'*, whose POS is noun and the GNP is neuter, singular and third. Main active verb is *'play'* and the POS is verb, aspect is simple past defined as verb_3, there is no auxiliary verb i.e. normal and the modality is also normal. Here play_1 occurs due to the fact that in the AnglaBharati lexicon there are two meanings for the root verb *'play'*, one is for playing football, cricket etc. and one is for playing any musical instrument. In this sentence first meaning has been considered which is playing football, so play_1 came in the PLIL. The numbers, m 8, m 3 and 10 written in the PLIL are the paradigm numbers for the nouns *'he'* and *'football'* and for the verb *'play'* respectively. This has been discussed later in the paper. The text cited in bold in the above example are the Bangla meaning of the corresponding English word.

3. Bangla Text Generator

Bangla Text Generator uses pseudo target generated by the AnglaBharati System as an input and it is used for the synthesis of the Bangla language. A number of grammatical rules of the Bangla language in the form of expectation and constraints are used to select the appropriate case markers, affixes, post-positions etc.

In the next sections of this paper, the authors have mainly concentrated on the main modules of the text generator, which are the morphological synthesis of verb, preposition-postposition disambiguation, pronominal form selection and the honorific and non-honorific identification of Bangla verbs. Before describing the text generator module one needs to discuss about the basics of the AnglaBharati lexical database structure. Because it is the fuel for the translation engine and the PLIL uses this database.

Lexical Database: It contains various details of each root word in English, like their syntactic categories, possible senses, keys to disambiguate their senses, corresponding words in target language with their all possible tags etc. All these information for a given root word are required at the time of implementing any rule in the text generator. Alternative meaning for the unresolved ambiguities are retained in the pseudo target language. Till now about 50,000 root words have been incorporated in the AnglaBangla system.

3.1 Morphological Synthesis of Verb

There are number of factors that determine the correct translation of target language. AnglaBangla system has the capability of generating sentences with different tense, moods, persons etc.; each of these requires modification of main verb of the input sentence, which requires morphological synthesis of verb. For this, verb roots are already been categorized and each category are identified and implemented by means of paradigm numbers. It is basically a table for each category providing all forms of that category, like, two verb roots *'kara'* (do) and *'chala' (walk)* have the same paradigm number because their inflected forms for different tense and person are same. So in the verb paradigm table only one form has been kept instead of two and in the lexicon the same paradigm number categorizes both. By this way paradigm tables are generated for different part-of-speech. With the help of the verb paradigm table, generation of different forms of verb in simple past, present and future tense has been handled. The other aspects of verbs like, perfect tense, continuous tense etc. have been handled by another file. The file contains a structure, which has five fields. These are i) Finiteness-this field gives information about the sentence type, like, normal, command, request, let etc. ii) Auxiliary verb iii) Main verb type iv) Phrase field and v) Suffix. Suffix field contains the suffixes, which will be concatenated with the root verb at the time of synthesis. In the Bangla verb morphological synthesizer, we have identified 34 paradigms for Bangla verbs; those can cover all the 730 number of Bangla verb [3]. Primarily total number of paradigms was 11, but after that all the finer differences of the different verbal forms has been taken care and correspondingly categorization has been increased from 11 to 34. For example in the previous 11 paradigm system, kATa(cut) and chhA.NTa(trim) two verbs were categorized by the same paradigm number because linguistically they are same. But at the time of suffix concatenation the number of character deleted from end of the verb roots are different for the two aforementioned verbs due to the presence of '.' in the chhA.NTa verb. For this reason the category number has been increased. At the time of creating the paradigm file the honorific and non-honorific forms of Bangla verbs have also been considered.

Like other languages, Bangla also have some irregular verbs, like *'yA'(go)*, 'Asa'(come) etc. For different persons and tenses inflected forms of these verbs are much different from the other root verbs. These verbs cannot be categorized with other verbs. In the Bangla text generator, the irregular verbs are handled in a different manner, not by paradigm tables. In AnglaBangla, one exceptional table has been maintained for these types of verbs, where all the forms of the particular root verbs are stored.

For command or request type of sentences, the Bangla verbal forms are completely different from other. So, verb morphological synthesizer also considers these types. For example, two English sentences, "Please **do** this job" and "I **do** this job", first one is request type of sentence and second one is affirmative type of sentence. But for both the cases same **do** verb has been used in English, but the Bangla forms are different for these two types of sentences, although the English verb is the same. For the first sentence do will be translated as *'karuna'* (honorific) or *'kara'*(non-honorific) and for the second type it should be *'karachi'*. So, this needs implementation of special rules in the text generator.

3.2 Preposition Disambiguation

Prepositions are taken care in a program considering the semantic aspects of the nouns attached to them. In AnglaBangla about 50 prepositions have been treated along with their multiple meanings. Bangla has different post-position in-place of English preposition. English prepositions are handled in Bangla by substituting suffixes attached with the nouns and /or introducing post-positional words after nouns. But the correspondence between English preposition and Bangla postposition is not one to one. In MAT, sense disambiguation of preposition is necessary when the target language has different representation for the same preposition. Let's consider some of the English sentences along with their Bangla translations:

1.English Sentence: A girl **with** beautiful eyes.
Bangla Translation: sundara **chokhera** ekaTi meYe.

2.English Sentence: A boy **with** high fever
Bangla Translation: prachaNDa jbare **AkrAnta** ekaTi chele

3.English Sentence: He wrote **with** a pen.
Bangla Translation: se pena **diYe** lekhe

4.English Sentence: Milkman mixes water **with** milk.
Bangla Translation: dudhaoYAlA **dudhera sAthe** jala meshAna

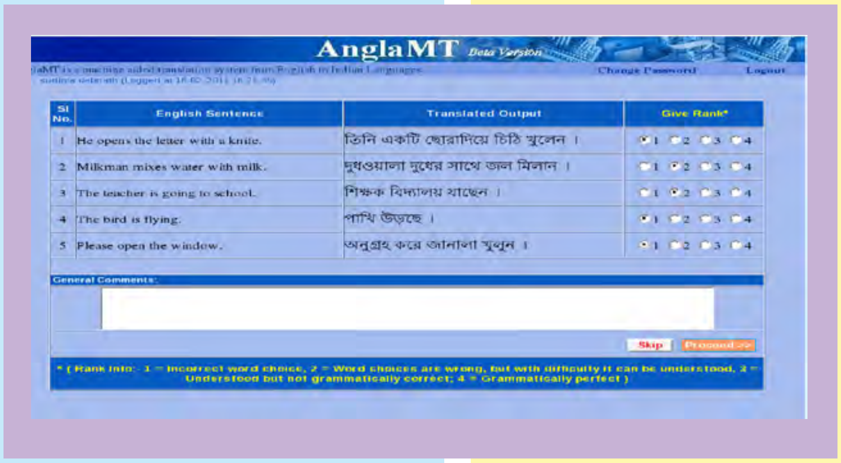
Here for one particular English preposition **'with'**, Bangla translations (depicted in bold letter) are different for different sentences according to their semantics. This module resolves the multiple meanings of the prepositions by considering semantic tags for nouns before and after the prepositions. Here for each of the above-mentioned sentences consist different types of nouns i.e. the different semantic senses of nouns attached with the preposition. Each falls in different categories. In the first two examples, 1 and 2 looks similar, but for the first one semantic of the noun attached with the pronoun is a body_part (eyes) whereas for the second case it is illness (fever). So, the translations of 'with' are different for two cases. Similarly, for the example 3, this is an instrument (pen). In the example 4, the nouns attached with the preposition **'with'** are liquids, i.e. water and milk. So for these case the translation of **'with'** is different than others. The letters cited in bold in the Bangla translations are due to the English preposition **'with'**.

3.3 Pronoun Translation

Correct selection of pronominal form for Bangla for a particular English pronoun is a big job for the text generator. For example, if the English sentence is "I have to do this", then corresponding Bangla for the English pronoun I should be *'AmAke'* instead of *'Ami'*, which is the lexical meaning of **'I'**. Here **'I'** is the subject noun phrase, so in order to get the correct translation of the particular sentence, the subject noun phrase,**'Ami'** should be changed to **'AmAke'**, which is dependent upon the auxiliary verb **'have to'** in the English sentence. This rule is also valid for the auxiliary verbs, **'has to'**, **'had to'** etc. But for the English sentence, **"I want to go there"**, the corresponding Bangla translation of **'I'** is different which is **'Ami'**.

In another case, for the English sentences, **"I have a pen"**, **"He has a pen"**, **"They had a pen"**, the corresponding Bangla translations of the pronouns I, he, they, which are the subject noun phrases, will be changed to their possessive form, i.e. **'AmAra'**, **'tAra'**, **'tAdera'** etc. So, this rule is valid when the main verb is 'have' or **'has'** or **'had'**. This rule is also valid for the main verbs **'should'**, **'should have'**, **'ought'**, **'ought to have'**, **'ought to'** etc. For example, **"He should go"**, **"We ought to go"**, **"He should have gone"** etc. In all cases, pronominal form of subject noun phrase will be translated to its possessive form in Bangla. This rule is also valid for negation, i.e. **'should not',"ought not"**etc.

Again correct form of the pronoun is also dependent on the prepositional entries. This has been handled with the help of the post-positions or the suffixes attached with the nouns in the Bangla translation. For example, if the English sentence is **"He talked to me"**, then corresponding Bangla translation is **"tini AmAra sAthe kathA balechilena"**. Here **'he'** is the subject noun phrase, **'to'** is the preposition, **'talked'** is the main verb and **'me'** is the object noun phrase. The rule for this type of sentences is, if the English preposition is **'to'** and the corresponding Bangla postposition is **'sAthe'**, then the lexical meaning of the object noun phrase will be changed to its possesive form, like 'AmAra', 'tAra' etc and the meaning of the subject noun phrase will remain unchanged i.e the lexical meaning of the particular pronoun. In the above-mentioned example, Bangla translation of me is translated as 'AmAra', which is the possesive form of 'AmAke' and the corresponding Bangla translation of he is


AnglaMT Beta Version
aMT is a machine aided translation system from English to Indian Languages
Change Password
Logout
sudipta debnath (Logged at 18-02-2011 16:21:46)
General Comments:
Skip
Proceed >>

Sl No.	English Sentence	Translated Output	Give Rank*
1	He opens the letter with a knife.	তিনি একটি ছোরাদিয়ে চিঠি খুলেন ।	● 1 ○ 2 ○ 3 ○ 4
2	Milkman mixes water with milk.	দুধওয়ালা দুধের সাথে জল মিলান ।	○ 1 ● 2 ○ 3 ○ 4
3	The teacher is going to school.	শিক্ষক বিদ্যালয় যাচ্ছেন ।	○ 1 ● 2 ○ 3 ○ 4
4	The bird is flying.	পাখি উড়ছে ।	● 1 ○ 2 ○ 3 ○ 4
5	Please open the window.	অনুগ্রহ করে জানালা খুলুন ।	● 1 ○ 2 ○ 3 ○ 4

* (Rank Info:- 1 = Incorrect word choice, 2 = Word choices are wrong, but with difficulty it can be understood, 3 = Understood but not grammatically correct; 4 = Grammatically perfect)

Fig. 1.Output of the AnglaBangla MAT System 'tini'. So here the actual lexical meaning of the object noun phrase (me) has been changed but the translation of subject noun phase (he) has been kept unchanged. Lets consider some more examples,

(a) "He is firing on me".
tini AmAra opara agnibarShaNa karachhena
(b) "He is pleased with me".
tini AmAra opara khushi.

The above-mentioned rule is also valid if the Bangla post-position is 'opara' for any cor re sponding English preposition.

4. Present Status

A complete Bangla text generator has already been developed as an integral part of An English to Bangla MAT (Machine Aided Translation) system named as AnglaBangla, which can translate all types of simple sentences and some of the complex sentences. Rules also have been incorporated for command and request type of sentences and giving correct results. Currently the system translates input sentences line by line. Many a times system also gives more than one translation for a given English sentence as shown in figure 1. User can select any one of them as a suitable translation of the particular English sentence.

Conclusion

Improved output quality of MT System can be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators, and in some cases can even produce output that can be used "as is". In general, there are number of hand crafted rules have been implemented in current text generator module. Paninian framework [4] provides a convenient way of imposing constraints. But, it may not be possible to build such rules which can completely disambiguate with the help of limited information available, and requires context for resolution.However, current systems are unable to produce output of the same quality as a human translator, particularly where the text to be translated uses colloquial language. Again system needs to implement a target language model at the output of the text generator to generate perfect one translation or decrease the number of parsed outputs for a given input sentence.

References

[1] Sinha R.M.K. and Jain Ajay, AnglaHindi: An English to Hindi Machine Translation System, MT Summit IX, New Orleans, USA, Sept.23-27, 2003.

[2] Sinha R.M.K. and others, ANGLABHARTI: A Multilingual Machine Aided Translation Project on Translations from English to Hindi, 1995 IEEE International Conf. on Systems, Man and Cybernetics, Vancouver, Canada, 1995, pp 1609-1614.

[3] Mukhopadhyay Ashoke, Samsad Banan Avidhan'(Bengali Spelling Dictionary), published in Shishu Sahitya Samsad Pvt. Ltd., February 2003.

[4] Cardona.G,"Pannini: A Survey of Research",Motilal Banarasidas, Delhi,1776.