AnglaBangla: English to Bangla MT system based on Angla MT Paradigm

Abstract

This paper describes the methodology of how an English to Bangla Machine Aided Translation System, namely AnglaBangla, has been developed by adapting the AnglaBharati framework, developed at IIT, Kanpur.

1. Introduction

Prof. R.M.K Sinha, Professor, Computer Science and Engineering, IIT, Kanpur propounded that "India has 22 major regional languages written in 10 different scripts. However, English, though spoken by a miniscule 3 percent of the population, is still the de-facto link language for administration, business and control. All grass root information of land, agriculture, health and education needs to be disseminated in the respective regional languages for effective communication and understanding. Hence, translation is as important as basic and necessary infrastructure like roads, water and transportation for a country like India." Therefore, there is a tremendous need to bridge the digital divide using Machine Translation Tools."AnglaBharati" is such a blessing for humanity developed by Prof. Sinha. This directed rule based system analyzes English using context free grammar and generate Pseudo Lingua for Indian Languages (PLIL). This system was primarily developed for Hindi and its' flexibility of adaptation to other Indo-Aryan and Dravidian languages seeking to integrate other language to build a giant network of translation of Indian languages. It has successfully been integrated to Bangla, Urdu, Malayalam, Punjabi, and waiting to be integrated to other North-Eastern languages as well as some South-Indian languages. The main focus of this paper is to describe the flexible structure of the system along with its component and the adaptation procedure of Bangla. This English to Bangla machine translation system is known as "AnglaBangla".

2. AnglaBharati Framework

This framework contains the following components.

User Interface: This Interface is used to enter sentence in English and gives the translation in target language.

Preprocessor/Postprocessor: The module used by the preprocessor to retrieve various entities from the text, which can be treated as separate entities before sending the text to the Translation engine. It also gives the information about the Postprocessor, which reassigns the code value in the target language text.

Example Base: This module translates the sentence by matching it with the stored patterns.

Rule Base: This is the main analyzer module of English. It contains the rules for parsing sentences. This is the main translation engine, which identifies the structure of English sentence and transforms it into an intermediate form according to the structure of the target language.

Morphological Analyser: Morph searches a given word or group of words in the lexicon and generates the files in the form of prolog files for Rule Base.

Text Generator: Text generator converts the pseudo code (PLIL) given by the rule base of the English language analyser of the system into the target language.

3. Adaptation for AnglaBangla

For the development of AnglaBangla MAT System, CDAC, Kolkata mainly developed the two main modules of AnglaBharati, these are the Lexical database and the Bangla Text Generator module.

Lexical Resource

AnglaBharati already had a framework for building the English to target language lexicon generation module. It contains various details of each root word in English viz. 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.

Text Generator

Target language generation part is taken care of by this module. Input of the module is the PLIL, which has been generated by the AnglaBharati architecture for English as a source language. Following paragraphs will discuss about the main tasks, which had been carried out for the development of Bangla text generation module.

Handling of verb forms: Developers mainly concentrated on the basic key module of the text generator part i.e. morphological synthesis of verbs, as verb is the most important decision making element in a sentence for generating proper translation. The system has the capability of generating sentences with different tenses, moods, persons etc. each of which requires modifying main verb of the basic sentence, which needs morphological synthesis of verb. For this, verb roots have already been categorized and each category has been identified and implemented by means of paradigm numbers. It is basically a table for each category providing all forms of that category. For example, two verb roots '*kara*'(to do) and '*chala*'(to walk) have the same paradigm number as their inflected forms for different tense and person are the same. In AnglaBangla, this number of paradigms is 34 and corresponding rules are written for each paradigm in the text generator and that number has been reflected in the lexicon.

Handling of Preposition: In MAT, sense disambiguation of preposition is necessary when the target language has different representations for the same preposition of English. In AnglaBangla, prepositions are taken care of in a program considering the semantic aspects of the nouns attached to them. English prepositions are handled in Bangla using inflections and /or using post-positional words. The correspondence between English preposition and Bangla postposition is not one to one, rather, it depends on the Bangla grammar rules. Almost 50 prepositions have been handled in the AnglaBangla text generator module and many linguistic rules have been implemented to select the correct post-position for Bangla.

Handling of Pronoun: 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*'. 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.

Handling of Honorificity: Bangla Language attaches honorific marker to the final verb forms modified by the subject noun phrase. Let us consider the following sentences,

English Sentence: The teacher is going to school
Bangla Translation: shixaka bidyAlaYe *yAchchhena*
English Sentence: The student is going to school
Bangla Translation: chhAtra bidyAlaYe *yAchchhe*.

For the first sentence, the honorific form of go verb i.e. *yAchchhena* will be the Bangla translation of going, and for the second one it would be the non-honorific form of go verb i.e. *yAchchhe*. This sort of modification has been handled in the final verb form generation part which generates final translation of the verbal group, where a separate paradigm table that have the verb form generation information has been employed and it is indexed. The index calculation is based on the aspect, modality and the gender, number and person information.

Nominal agglutinization: Unlike Hindi, target form generation of noun, adjective or verb the suffixes of Bangla gets agglutinated with the source form with some modification at the end of the root word. This has been handled in the combine phase where the final form of nouns and adjectives has been generated. This is described in the Table 1.

English Sentence	Bangla Translation	Comments
This is Ram's book.	*eTA rAmera bai.*	Addition of '*era*' with proper noun *rAm*.
This book belongs to brother of Ram.	*eTA rAmera bhAiYera bai..*	Addition of '*Yera*' with noun *bhAi*.
Gita should come here.	*gItAra ekhAne AsA uchita.*	Simple addition of '*ra*' with proper noun *gItA*.

Table 1. Examples of English to Bangla Translation

Depending upon the vowel or consonant present in the last character and the penultimate character of the nominal/adjectival part the suffix has been added. For handling the final form generation of capitalized nominal within a sentence, the same logic has been deployed but it is handled in the post-processing phase, as the preprocessor modules insert different tag for this type of entries and saves the entry in a table which is being handled in the post-processing phase. Let us consider the translation of the English sentence “Netaji is the pride of India.” which is, ‘*netAjI bhAratera garba*’ where “*era*” has been added with ‘bhArat’(India) .

Handling of Adverbial Entries: Adverbial entries have been handled by the following two aspects:

1. It has been handled from the lexicon, where the adverbial entries have been kept with associated paradigm number and general meaning. If the sentence does not require special handling of the adverb sequence then its’ meaning is picked up and placed in appropriate place in the target sentence.

2. It has also been handled from the rule-base description because the following construct needs special handling

Different adverbial constructs	Examples
Adverb as subject	Now is the time.
Adverb after subject	He also eats fish.
Preposition as adverb at the sentence end	Whom do you vote for?
Adverb after auxiliary verb	He is not exactly a friend.
Adverb between verbs	I can never remember his name.
Adverb of a question sentence	Why are you smiling?

3 The constraints of handling the polemic sentences construction where the words like ‘seldom’, ‘never’, ‘rarely’ never comes at the end of a sentence has been added, for example as in the following construction “Seldom had the engine given a bad performance.”. While using adverbs of frequency in the negative form, normally the adverb is placed before the main verb. ‘Never’, ‘seldom’, ‘rarely’ and other adverbs of frequency with a negative sense are not usually used in the negative form.

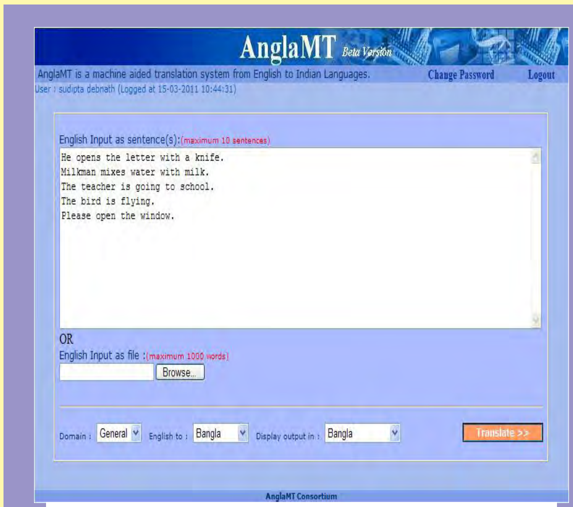
Multiword Expression (MWE)

In the process of development of AnglaBangla MAT System for a specific domain, MWE played an important role. In the area of Natural Language Processing(especially for Machine Translation) Multi Word Expression (MWE) Extraction is an important component of Machine Aided Translation System. For MWEs, structure and meaning cannot be derived from their component words, as they occur independently and whose internal structure is of no real importance to the overall analysis of the sentence. For example, *Black widow* or *Milk of Magnesia*. In the first example the actual meaning of the MWE is spider and in the second one it is the name of a medicine, which are in no way related to their component word's meaning. MWEs include nominal compounds, proverbs, idioms, phrasal verbs and collocations. MWEs, containing proper nouns, can be described with simple grammar and their internal structure is of no real importance to the overall analysis of the sentence. MWEs may occur as a whole sentence, or a part of a sentence.

Thus, table driven or template matching translation of MWEs can provide better output. So, there is a tremendous need for collection of MWEs for increasing the efficiency of Machine Aided Translation (rather domain dependent Machine Aided Translation). In order to collect the MWEs, AnglaBangla used rule-based approach. After applying rules on the parsed output of the English corpus, all the probable MWEs were collected and then these have been finalized with the help of human intervention.

4 Present Status

A complete English to Bangla MAT (Machine Aided Translation) system named as AnglaBangla has already been developed, 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. User can select any one of them as a suitable translation of the particular English sentence. User can also edit the translated output with the help of a on-screen Bangla Keyboard. User can also provide rank of the translated output.



AnglaMT is a machine aided translation system from English to Indian Languages.

User : sudipta debnath (Logged at 15-03-2011 10:44:31)

English Input as sentence(s):(maximum 10 sentences)

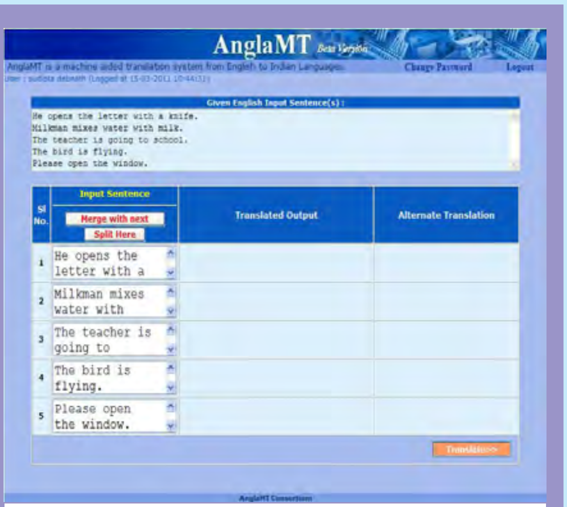
```
He opens the letter with a knife.
Milkman mixes water with milk.
The teacher is going to school.
The bird is flying.
Please open the window.
```

OR

English Input as file :(maximum 1000 words)

Domain : General English to : Bangla Display output in : Bangla

Fig 1: Input Sentence(s)



AnglaMT is a machine aided translation system from English to Indian Languages.

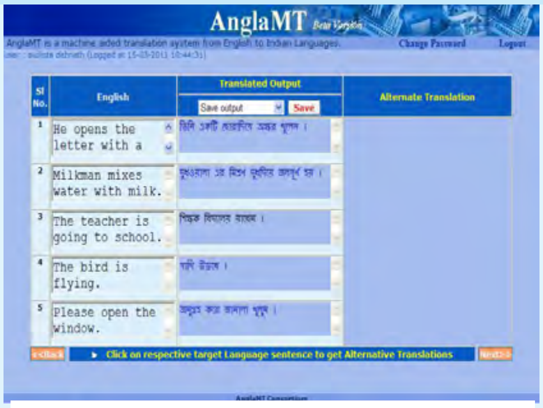
User : sudipta debnath (Logged at 15-03-2011 10:44:31)

Given English Input Sentence(s) :

```
He opens the letter with a knife.
Milkman mixes water with milk.
The teacher is going to school.
The bird is flying.
Please open the window.
```

Sl No.	Input Sentence	Translated Output	Alternate Translation
1	He opens the letter with a		
2	Milkman mixes water with		
3	The teacher is going to		
4	The bird is flying.		
5	Please open the window.		

Figure 2: Sentence Boundary Finalization

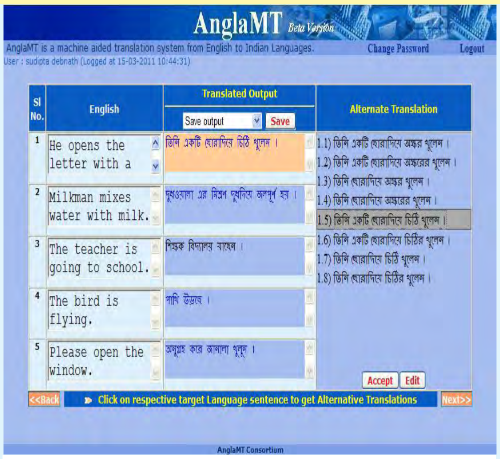


AnglaMT is a machine aided translation system from English to Indian Languages.
User : sudipta debnath (Logged at 15-03-2011 10:44:31)

Sl No.	English	Translated Output Save output ▾ Save	Alternate Translation
1	He opens the letter with a	তিনি একটি ছোরাদিয়ে অক্ষর খুলেন ।	
2	Milkman mixes water with milk.	দুধওয়ালা এর মিশ্রণ দুধদিয়ে জলপূর্ণ হয় ।	
3	The teacher is going to school.	শিক্ষক বিদ্যালয় যাছেন ।	
4	The bird is flying.	পাখি উড়ছে ।	
5	Please open the window.	অনুগ্রহ করে জানালা খুলুন ।	

Click on respective target Language sentence to get Alternative Translations

Figure 3: Translated Output(s)

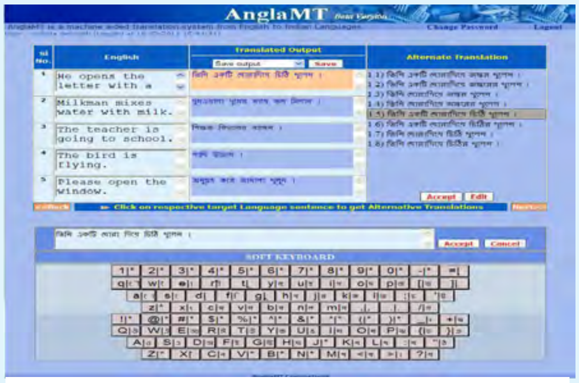


AnglaMT is a machine aided translation system from English to Indian Languages.
User : sudipta debnath (Logged at 15-03-2011 10:44:31)

Sl No.	English	Translated Output Save output ▾ Save	Alternate Translation
1	He opens the letter with a	তিনি একটি ছোরাদিয়ে চিঠি খুলেন ।	1.1) তিনি একটি ছোরাদিয়ে অক্ষর খুলেন । 1.2) তিনি একটি ছোরাদিয়ে অক্ষরের খুলেন । 1.3) তিনি ছোরাদিয়ে অক্ষর খুলেন । 1.4) তিনি ছোরাদিয়ে অক্ষরের খুলেন ।
2	Milkman mixes water with milk.	দুধওয়ালা এর মিশ্রণ দুধদিয়ে জলপূর্ণ হয় ।	1.5) তিনি একটি ছোরাদিয়ে চিঠি খুলেন ।
3	The teacher is going to school.	শিক্ষক বিদ্যালয় যাছেন ।	1.6) তিনি একটি ছোরাদিয়ে চিঠির খুলেন । 1.7) তিনি ছোরাদিয়ে চিঠি খুলেন । 1.8) তিনি ছোরাদিয়ে চিঠির খুলেন ।
4	The bird is flying.	পাখি উড়ছে ।	
5	Please open the window.	অনুগ্রহ করে জানালা খুলুন ।	

Click on respective target Language sentence to get Alternative Translations

Figure 4: Translated Output(s)



AnglaMT is a machine aided translation system from English to Indian Languages.

Sl No.	English	Translated Output Save output ▾ Save	Alternate Translation
1	He opens the letter with a	তিনি একটি ছোরাদিয়ে চিঠি খুলেন ।	1.1) তিনি একটি ছোরাদিয়ে অক্ষর খুলেন । 1.2) তিনি একটি ছোরাদিয়ে অক্ষরের খুলেন । 1.3) তিনি ছোরাদিয়ে অক্ষর খুলেন । 1.4) তিনি ছোরাদিয়ে অক্ষরের খুলেন । 1.5) তিনি একটি ছোরাদিয়ে চিঠি খুলেন ।
2	Milkman mixes water with milk.	[illegible]	1.6) তিনি একটি ছোরাদিয়ে চিঠির খুলেন । 1.7) তিনি ছোরাদিয়ে চিঠি খুলেন । 1.8) তিনি ছোরাদিয়ে চিঠির খুলেন ।
3	The teacher is going to school.	শিক্ষক বিদ্যালয় যাছেন ।	
4	The bird is flying.	পাখি উড়ছে ।	
5	Please open the window.	অনুগ্রহ করে জানালা খুলুন ।	

Click on respective target Language sentence to get Alternative Translations

তিনি একটি ছোরা দিয়ে চিঠি খুলেন ।

Figure 5: Translated Output(s) with edit

System Features:

- System accepts sentence input on-line or from file
- Supports .doc, .txt and .pdf as input file.
- User selectable single/multiple translation for a single sentence input
- User can save input and output in Unicode format
- Output supports UTF, ITRANS and Devnag riscripts
- Integrated Bangla keyboard for user editing
- User can evaluate the translation out put by selection of rank

Conclusion

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 translation or to decrease the number of parsed outputs for a given input sentence.