Developing a Bangla to English Machine Translation System using Parts of Speech Tagging

Md. Shahnur Azad Chowdhury
tipu_iiuc@yahoo.com
Assistant Professor, International Islamic University Chittagong, Chawk Bazar Chittagong

Abstract

Machine translation is always a challenging job. This paper reviews an efficient implementation of Machine Translation (MT) System from Bangla to English. Normally there are three stages for machine translation: 1) Tagging 2) Transfer and 3) Generation. In the propped system the source text is analyzed using a set of grammatical rules, transferred and synthesized with the direct help of some dictionaries (i.e. lexicons). In this system, a Word Corresponding Lexicon in addition to the Root Lexicon and the Suffix Lexicon and a typical set of Bangla to English transfer rules are used.

Keywords
Lexicon, tag vector, morphology, machine translation.

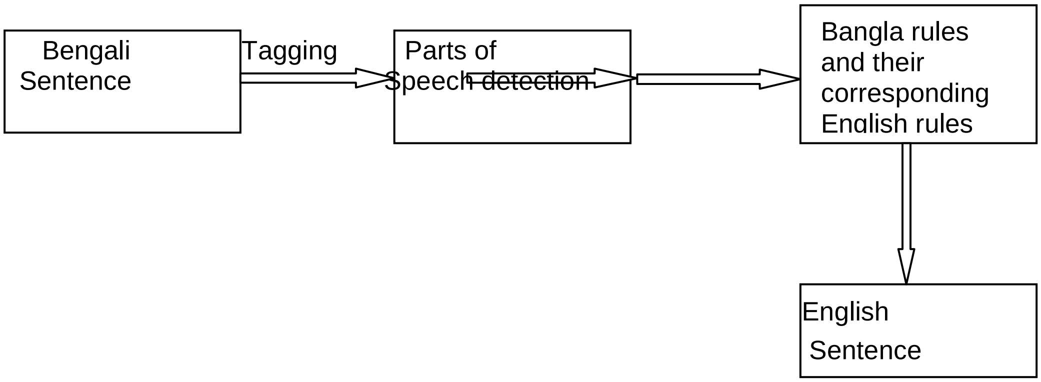
1. Introduction: Machine translation means translation of natural language from one to another. A significant part of the development of any machine translation (MT) system is the creation of lexical resources that the system will use. Dictionaries are of critical importance in MT.A well defined Bengali word dictionary with necessary suffixes to be added or dropped is incorporated in the system I proposed. The very crucial issue is to find out the Parts of speech and also the relevant aspects like Number, Person, Mode, Tense and Emotion of any word. If the aspects mentioned and the dictionary entry can be correctly identified then the translation can also be possible almost correct in all aspect. The third important issue is to map the Bengali to English sentence structure rules for each type of sentence and Tense. For accomplishing this task there is defined a sophisticated Bengali-English grammatical rule set in the proposed system.

2. Literature Review

It is really difficult to build up a complete MT system for natural Languages.MT includes natural Language understanding and generation. The proposed system represent a new solution for building an MT system of English to Bengali translation, by modifying the rule based transfer approach of MT system.(**Md. Golam Robiul Alam,Md. Monirul Islam,Nowrin Islam,2010**).In this paper we will dicuss about feature based morphological parsing for Bangla which gives us parts of spech and other morphological features in addition to the morpheme division[2][8].At first we give an idea of normal morphological parsing and in the end we shed light on the comperison between the approaches.(**Sajib Das Gupta and Dr Mumit Khan,2004)**.We used an approch where the English Parse tree,which is generated via CYK parsing algorithm,is changed into manother form of English Parse tre,which in turn can be easily transferred into Bengali Parse tree.(**Sajib Dasgupta,Abu Wasif, Sharmin Azam,2004**).Our main objective is to convert simple and short Bangla Text into English text.Sometimes position of words in a sentance doesnot worth much although it expresses the same meaning as it did previously.So it is difficult to pick all the variations that may be possible in Bangla.For this we concentrate on sinple and short sentence where the meaning of the sentence is clear and easily transferable.In addition,full automation of good quality translation is a virtually impossible and human intervention either before or after computer process (known as per editing and post editing respectively) would be essential.(**Shehab Raihan,Muhammad Masroor Ali,2004**).Since dictionary development consumes a large portion of the total resources allocated for the development of an MT system,it requires significant consideration .However ,these seems to be little literatiure available about efficient procedures and practices for developing dictionaries.And so far very little work has been done in MT in the field of automatic translation,Parsing and System Analysis.In this paper we have discussed the issues relevant to the development of MT dictionary for Bangla.(**Mortaza Ali and Muhammad Masroor Ali, 2002)** The Bangla interface helps those people who do not know english well,However they want to get themselves involved in the field of database technology.Through this interface they will be able to design a databese ,view and update data on database and also maintain the consistancy of the database.(**Md. Kamrul Hasan,Md. Abdul Alim,Md. Wahedul Islam, 2002**).In this paper we will also show some method to resolve some pronouns to summerize the text so that coherence and important information is conserved.For our experiment we used unicode based document corpus and also the programs are written to handle unicode based text.(**Bhasa:a Corpus-Based Information Retrieval and Summariser for Bengali Text.,7th ICCIT 2004,BRAC University,26-28 December 2004**).

3. The System Architecture

The block diagram given in the figure represents the glance of system architecture. In the system design, three major steps are analysis, does an analysis on Bangla input sentence using Tag Vector and some other well-defined Bangla grammars, transferring step, translates underlying representation of Bangla words into underlying representation of English words, and the final step is the synthesis step, involves the representation of English language using a set of English grammar rules.


Bengali Sentence
Tagging
Parts of Speech detection
Bangla rules and their corresponding English rules
English Sentence

4.1 Morphological Analysis

In this section we have pointed out some of the information about words that may be included in a Bangla MT dictionary. [10]

Grammatical properties

Any Bangla words fall in the one of the five categories noun, pronoun, adjective, verb and indeclinable.

4.1.1.1 Nouns

Bangla nouns may be concrete or abstract. Concrete nouns can be classified as proper noun ((করিম, তাজমহল), common noun (সুন্দর ফুল), material noun (সানা, দুধ), and collective noun (সভা, দল).Collective noun can be classified as proper collective noun (`য়া, সৌন্দয)) and verbal noun (শয়ন, গমন).Additional information about noun that can be included in the dictionary is its number, gender and case. Number of noun can be classified as singular (ছেলে) and plural (ছেলেরা).Gender of nouns can be classified as masculine (বাবা), feminine (মা) and common and neuter(শিশু).Case of a noun may be nominative (ছেলে) and locative (বাগানে).

4.1.1.2 Pronouns

There are eight different types of pronouns in Bangla. They are: (i) Personal(আমি, তিনি),(ii)Interrogative(আমি,কারা),(iii)Relative(যারা,যাদের),(iv)Demanostrative(এটা,সেটা)),(v)Indefinite(কউ, কান)),(vi)Reflexive(নিজে, স্বয়ং),vii) Distributive(প্রত্যেক, যেকোন),(viii)Collective(সকল, সব)).Additional features of pronouns are person, number and singular and plural pronouns when asking a question(কে, কারা),.Unlike English Bangla distinguishes between masculine and feminine pronoun. There is however a neuter pronoun meaning “it”.

4.1.1.3 Adjective

Adjectives fall into four subcategories :proper adjective(বাংলাদেশী কাপড়),),Adjectives of quality (সুন্দর ফুল),adjective of quantity(অধিক, দ্বিগুন),pronominal adjective(য কোন লোক).In bangle adverbs fall in adjectives in Bangla, adjective that modifies a verb (আসলেই হবে). ,adjective that modifies a verb, Adjectives that modifies another adjectives(খুব ভাল লোক),),adjective that modifies an indeclinable, adjective that modifies a sentence(খুব ভাল লোক),

4.1.1.4 Verb:

Verb is most important word category it can be finite(আমি পড়ি) or nonfinite(পড়তে যাব).)..Verb can be classified as intransitive(ছলেরা খেলছে),),transitive(বল খেলছে),),ditransitive(বাবা আমাকে কলম দিয়েছেন),causative(মা শিশুকে চাঁদ দেখান),compound(ঘটনাটি শুনে রাখ) and complex (করলাম).)verb.Additinal information about verbs may be its mode: indicative(আমি বই পড়ি),imperative(মন দিয়ে পড়),subjunctive(পড়লে পাস করবে), optative (তার মঙ্গল হোক).Other features of verb that can be put in an MT dictionary are its tense and person.

4.1.1.5 Indeclinable:

Indeclinable are of four kinds: conjunction,Interjection((আহ্!),post position(জন্যে, কাছে), and reasoning(শো শো).).Bangla conjunction subcategories into: cumulative(ও,এবং),adversative(অথচ, বরং),disjunctive(কিংবা, অথবা).

Bangla Morphology

An important distinction between the development of paper based dictionaries and MT dictionaries is the morphological component. A morphological component must be added to the system to save time, space and effort. So, during the development of an MT dictionary one should try to describe all regular inflections, derivations and compounding in general rules, with additional explicit rules for irregular inflection, derivation and compounding. [10]

class	person	present	present Continuous	Present perfect	Future	Past	Past Continuous	past Perfect	Past Habitual
Class 1 (পড়)	First	ই (পড়ি)	ছি (পড়ছি)	এছি (পড় ছি)	বা (পড়বাা)	লাম (পড়লাম)	ছিলাম (পড়চিলাম)	এছিলাম (প ড়ছিলাম)	তাম (পড়তাম)
	Second (Familiar)	া প ড়া	ছা (পড় ছা)	এ ছা (প ড় ছা)	ব (পড়বা)	ল (পড় ল)	ছি ল (পড়ছি ল)	এছি ল (প ড়ছিল)	ত (কর ত)
	Second (polite)	এন প ড়ন	ছন (পড়ছন)	এ ছন (প ড় ছন)	বন (পড় বন)	লন (পড় লন)	ছি লন (পড়ছি লন)	এছি লন (প ড়ছি লন)	তন (ক রছি লন)
	Third	এ প ড়	ছ (পড় ছ)	এ ছ (প ড় ছ)	ব (পড় ব)	লা (পড় লা)	ছি লা (পড়ছি লা)	এছি লা (প ড়ছি লা)	তা (পড় তা)
Class 2 (রাখ)	First	ি (রাখি)	ছি (রাখছি)	এছি (র খছি)	বা (রাখ বা)	লাম (রাখলাম)	ছিলাম (রাখছিলাম)	এছিলাম (র খছিলাম)	তা (রাখতাম)
	Second (Familiar)	া (রা খা)	ছ (রাখ ছা)	এ ছা (র খ ছা)	ব (রাখ ব)	ল (রাখ ল)	ছি ল (রাখছি ল)	এছি ল (র খছি ল)	ত (রাখ ত)
	Second (polite)	এন (রা খন)	ছন (রাখ ছন)	এ ছন (র খ ছন)	বন (রাখ বন)	লন (রাখ লন)	ছি লন (রাখছি লন)	এছি লন (র খছি লন)	তন (রাখ তন)

Table 1: Conjugation of Verbs

5. Structure of Tag Vector

For tagging any word with its various aspects we have used a sixteen-bit tag vector. Where parts of speech (POS.). Person, Mode, Tense number and emotion are put in different length. Three bits are kept for parts of speech. In POS. there are noun, pronoun, adjective, verb and preposition. Noun is divided into proper noun and dictionaries word whereas adjective is divided into proper adjective and modal adjective. Verb is divided into finite verb and infinite verb and infinite verb whereas infinite verb is divided into gerund and participle. For person identification, it is divided into first, second and third person. The mode in Bangla is mainly of two-type প্রাণীবাচক (*Pranibachok-living*) and অপ্রাণীবাচক *(Opranibachok-nonliving*). The প্রাণীবাচক *(Pranibachok-living)* may be general, honor, disgrace and deictic. The অপ্রাণীবাচক *(Opranibachok-nonliving)* are disgrace and general. Tense are mainly three types: present, past and future where present tense may be divided into present indefinite, present continuous and present perfect. Past tense can be divided into past indefinite, past continuous and past perfect. The future tense is only one type, which is future indefinite. In Bangla number is two types: Singular and Plural. We reserved three bits in tag vector to represent emotional state of sentence. So the tag vector is defined by sixteen bits data [4]

6. Grammatical Rules and Actual Mapping

We can investigate how the comparative grammar relates a representation for Bangla sentence to the corresponding representations for English sentence. The comparative grammar has bilingual dictionary rules. In the simplest case, these may just relate source lexical items to target lexical items. The comparative grammar also contains some structural rules, which relate other parts and nodes of the two functional structures to each other. [1]

The following table shows some Bangla rules and their corresponding English rules. [2]

Bangla Rule	English Rule
S=NP+NP+VP	S=NP+NP+VP
S=NP+PP+NP+PRIN	S=NP+PRIN+NP+PP
S=NP+PP+NP+AP	S=NP+AP+NP+PP
S=NP+OBJ1+PP+OBJ2+AP	S=NP+AP+OBJ1+OBJ2+PP
SNP+OBJ1+PP+OBJ2+PRIN	S=NP+PRIN+OBJ1+OBJ2+PP

Table 2: Bangla and English grammatical Rules

In our proposed system, every word carries its own aspects including Parts of Speech, Verb, Person, and Number etc.

So, after analyzing any word depending upon the aspects and sentence rules and mapping the corresponding English words, Bangla sentence can be converted into English.

7. Algorithm:

```
Step1: Start traversing the sentence given in Bengali.
Step 2: Match the first word with the words in the root lexicon.
Step 3: Translate the Bengali word into corresponding English word from the words stored in the dictionary.
Step 4: Find out the attributes of the word found by analyzing the 16 bit Tag vector.
Step 5: If the word is recognized as subject then
 Repeat step 2 to 3 for the all the words of the sentence
Step 6: after getting all the words translated, add grammatical suffixes as s, es, ed, t etc with the verb which is necessary.
Step 7: Rearrange the word sequence according to Bengali to English grammatical rules.
```

Example: For example we may consider the Bangla sentence সে ভাত খায় .*(Se Vat Khay)* . The sentence is traversed from left word by word. It considers the first word সে . The word সে is first analyzed. It is of third person; singular number then searching the corresponding word lexicon the English word for this Bengali word সে (Se) is found He. ভাত (Vat) is third person, singular number and খায় (Khay) is the verb whose tense classification, depending upon the suffix is found a tense in present indefinite form. So, first of all, the corresponding English words are matched then grammatical suffixes in English i.e. s/es/ing/ed or auxiliaries like am/is/are/have/has/had etc are attached if necessary. Finally the system looks at the corresponding grammar rules for conversion which is found as NP+NP+VP=NP+VP+NP. Placing the corresponding English words and adding necessary English suffixes with nouns and verbs find the English sentence as He eats rice for the Bangla sentence সে ভাত খায় .*(Se Vat Khay)*

8. **Conclusion**

Although Bangla is our mother tongue, there is hardly any work on complex machine translation. I tried in this paper to introduce a new approach towards the research of MT from Bangla to English for the simple Bengali sentences. This is an efficient system for simple sentences. The grammar and examples cited here are simple ones also. But this work may be a starting for future development of an efficient MT engine from Bangla to English.

References

[1] Md. Abdullah Al Mamun,Mohammed Iftekhar Ahmed,Mohammed Alauddin Bhuian,Mohammed Riaz Selim and Zafar Iqbal,"An Implementation of Machine Translation between Bangla and English",(ICCIT)2002,NSU,Dhaka,Bangladesh,PP290-293.

[2] Sajib Dasgupta, Abu Wasif, Shamim Azam "An Optimal way of Machine Translation from English to Bengali",7th International Conference on Computer and Information Technology,(ICCIT)2004,Brac University Dhaka Bangladesh,PP648-653.

[3] Md. Shahnur Azad Chowdhury, Nahid Mohammed Minhaz Uddin, Mohammed Imram, Mohammed Mahadi Hasan and Md. Emdadul Hoque,"Patrs of Speech Tagging of Bangla Sentence", 7th International Conference on Computer and Information Technology, ICCIT (2004), BRAC University, Dhaka, Bangladesh, PP632-637.

[4] Md. Hanif Siddiui, A.K Mohammed Shohel Rana, Abdullah Al Mahmud and Taufique Sayed,"Parts of Speech Tagging Using Morphological Analysis in Bangla". 6th International Conference on Computer and Information Technology, ICCIT (2003), BRAC University, Dhaka, Bangladesh, PP374-379.

[5] A New Approach to Develop a Bangla to English Machine Translation System Md. Golam Rabiul Alam,Md. Monirul Islam and Nowrin Islam, Computer Scinece and Engineering Research Journal Voil.06(2009-2010),CUET,PP19-25.

[6] Feature Unification for Morphological Parsing in Bangla, Sajib das Gupta and Dr Mumit Khan,7th ICCIT,2004,26-28,December,BRAC University,Dhaka

[7] Translation From English To Bangla, Judith Francisca, Md. Mamun Mia, Dr. S. M. Monzurur Rahman, Indian Journal of Computer Science and Engineering (IJCSE)

[8] Bangladesh to English Translation by Rule-Based Aproach,Shehab Raihan,Muhammad Masroor Ali,ICCIT 2004,26-28 December 2004

[9] Bangladesh to English Translation by Rule-Based Aproach,Shehab Raihan,Muhammad Masroor Ali,ICCIT 2004,26-28 December 2004

[10] Development of MT Dictionary for Bangla Language,(Mortaza Ali and Muhammad Masroor Ali,5th ICCIT 2002,East West University,27-28 December 2002)

[11]Development of a Bangla SQL interface for DBMS, Md. Kamrul Hasan, Md. Abdul Alim,Md. Wahedul Islam, 5th ICCIT 2002,East West University,27-28 December 2002)

[12] Bhasaa Corpus-Based Information Retrieval and Summaries for Bengali Text, 7th ICCIT 2004,BRAC University,26-28 December 2004