

Developing a Bangla to English Machine Translation System Using Parts Of Speech Tagging: A Review

Md. Shahnur Azad Chowdhury

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 proposed 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 is introduced in addition to the Root Lexicon and the Suffix Lexicon and a typical set of Bangla to English transfer rules are used. An algorithm is developed to traverse the Bengali sentence from left to right and correspond the English words according to the meaning and context.

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 aspects. 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. In an MT lexicon bears a lot of importance. As searching the lexicon is a compulsory task so by optimizing this task the system can be developed in an efficient way. I tried to do this job in this paper.

2. Literature Review

Md. Golam Robiul Alam et al, (2010) concluded that it is really difficult to build up a complete MT system for natural Languages. MT includes natural Language understanding and generation. The proposed system represents a new solution for

Shahnur

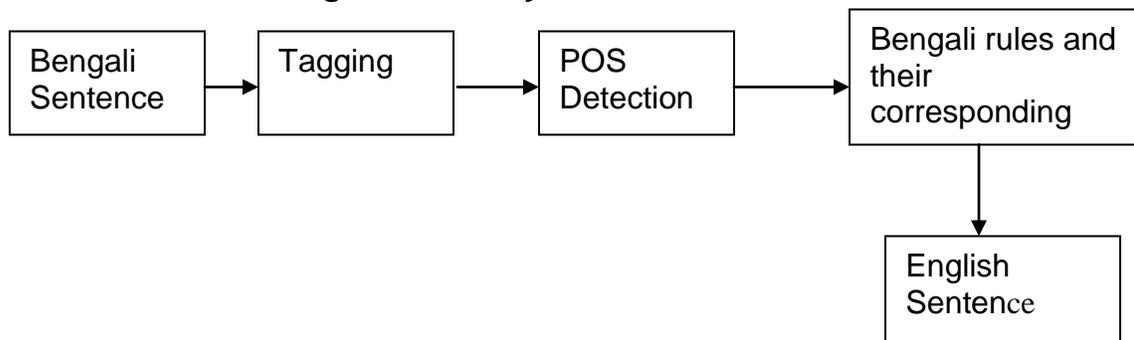
building an MT system of English to Bengali translation, by modifying the rule based Transfer approach of MT system.

Gupta and Khan, (2004) discussed about feature based morphological parsing for Bangla which gives us Parts of Speech and other morphological features in addition to the morpheme division. Dasgupta et al, (2004) used an approach where the English Parse tree, which is generated via CYK parsing algorithm, is changed into another form of English Parse tree, which in turn can be easily transferred into Bengali Parse tree. Shehab Raihan et al, (2004) converted simple and short Bangla Text into English text. Sometimes position of words in a sentence doesn't 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 they concentrated on simple 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 pre editing and post editing respectively would be essential. Ali and Ali, (2002) discussed the issues relevant to the development of MT dictionary for Bangla. Hasan et al, (2002) developed a Bangla interface which 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 database, view and update data on database and also maintain the consistency of the database.

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. The system was tested by developing the software and it worked for the simple sentences.

Figure 1: The System Architecture



4. 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 Nouns

Bangla nouns may be concrete or abstract. Concrete nouns can be classified as proper noun (করিম, আজমহল), Common Noun (নন্দর ফুল), Material Noun (সাদা, দুধ), and Collective Noun (নন্দ দল). Collective Noun can be classified as Proper Collective Noun (স্বা, সৌন্দর্য) and Verbal Noun (শয়ন, গমল). Additional information about noun that can be included in the dictionary are its Number, Gender and Case. Number of noun can be classified as Singular (ছেলে) and Plural (ছেলের). 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).

4.2 Pronouns

There are eight different types of Pronouns in Bangla. They are: (i) Personal (আমি, তুমি), (ii) Interrogative (আমি, কারা), (iii) Relative (যারা, যাদের), (iv) Demonstrative (এটা, সেটা), (v) Indefinite (কউ, কান), (vi) Reflexive (নিকে, স্বয়ং), (vii) Distributive (থেকে, যেখানে), (viii) Collective (সকল, সব). Additional features of pronouns are Person, Number - 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".

4.3 Adjective

Adjectives fall into four subcategories: Proper Adjective (বঙ্গদেশী কাপড়), Adjectives of quality (নন্দর ফুল), Adjective of quantity (অধিক, বিদ্যে), Pronominal adjective (যে কোন লোক). In bangla adverbs fall in adjectives, adjective that modifies a verb (আসলেই হবে), Adjectives that modifies another adjectives (খুব ভাল লোক), Adjective that modifies an indeclinable, Adjective that modifies a sentence (খুব ভাল লোক),

4.4 Verb:

Verb is most important word category it can be finite (আমি পড়ি) or nonfinite (পড়তে যাব). Verb can be classified as Intransitive (ছেলেরা খেলাছে), Transitive (বল খেলাছে), Causative (মা শিশুকে চাঁদ দেখান), Compound (ঘটনাটি শুনে রাখ) and Complex (করলাম) verb. Additional 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.

Shahnur

Table 1: Conjugation of Verbs

Class	person	present	present Continuous	Present perfect	Future	Past	Past Continuous	past Perfect	Past Habitual
Class 1 (co)	First	ই (পড়ি)	ছি (পড়ছি)	এছি (পড়ছি)	বা (পড়বো)	লাম (পড়লাম)	ছিলাম (পড়ছিলাম)	এছিলাম (পড়ছিলাম)	তাম (পড়তাম)
	Second (Familiar)	ত পড়া	ছা (পড়ছা)	এ ছা (পড়ছা)	ব (পড়বো)	ল (পড়ল)	ছিল (পড়ছিল)	এছিল (পড়ছিল)	ত (করত)
	Second (polite)	এন পড়ন	ছন (পড়ছন)	এ ছন (পড়ছন)	বন (পড়বন)	লন (পড়লন)	ছিলন (পড়ছিলন)	এছিলন (পড়ছিলন)	তন। (করতন)
	Third	এ (পড়)	ছ (পড়ছ)	এ ছ (পড়ছ)	ব (পড়ব)	লা (পড়লা)	ছিল (পড়ছিল)	এছিল (পড়ছিল)	ত (পড়ত)
Class 2 (রাখ)	First	ি (রাখি)	ছি (রাখছি)	এছি (রাখছি)	বা (রাখবো)	লাম (রাখলাম)	ছিলাম (রাখছিলাম)	এছিলাম (রাখছিলাম)	ত (রাখতাম)
	Second (Familiar)	Class 2 (রাখ)	ছ (রাখছ)	এ ছা (রাখছা)	ব (রাখবো)	ল (রাখল)	ছিল (রাখছিল)	এছিল (রাখছিল)	ত (রাখত)
	Second (polite)	এন (রাখন)	ছন (রাখছন)	এ ছন (রাখছন)	বন (রাখবন)	লন (রাখলন)	ছিলন (রাখছিলন)	এছিলন (রাখছিলন)	তন (রাখতন)

4.5 Indeclinable:

Indeclinable are of four kinds: Conjunction, Interjection (আহ!), Post Position (জন্মে, কাছে), and Reasoning (শেষে). Bangla conjunction subcategories into: Cumulative (ও,এক), 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]

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

Shahnur

Adjective. Verb is divided into Finite 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]

Table 2: Bangla and English grammatical Rules

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

In my 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 from left.

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.

Shahnur

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. The language translation architecture that is represented here is not developed before. According to this architecture the algorithm has been constructed. 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. The task that I have done in this paper can be extended more. A lot research is possible in this field. I have tried to keep variation among the English sentences that we have translated into Bangla sentences. But we have not completed all the variety of sentences. Since it is Natural Language Processing (NLP) the number of variation is almost unlimited. So there is unlimited opportunity to upgrade the current research.

References

- Alam, M.G.R, Islam, M. M., and Islam, N. (2009-2010), 'A New Approach to Develop a Bangla to English Machine Translation System', *Computer Science and Engineering Research Journal*, 06, 19-25.
- Ali, M., and Ali, M. M. (2002), 'Development of MT Dictionary for Bangla Language', *Proceedings of 5th International Conference on Computer and Information Technology*, East West University, 27-28 December, PP-169-172
- Chowdhury, M. S. A., Uddin, N. M. M., Imram, M., Hassan, M. M., and Hoque., M. E. (2004), 'Parts Of Speech Tagging of Bangla Sentence', *Proceedings of 7th International Conference on Computer and Information Technology*, BRAC University, Dhaka, Bangladesh, PP632-637.
- Dasgupta, S., Wasif, A., and Azam, S. (2004), 'An Optimal Way Of Machine Translation From English To Bengali', *Proceedings Of 7th International Conference On Computer and Information Technology*, Brac University, Dhaka Bangladesh, PP 648-653.
- Francisca, J., Mia, M. M., and Rahman, S. M. M., 'Translation from English to Bangla', *Indian Journal of Computer Science and Engineering (IJCSE)*. Gupta, S. D., and Khan, D. M. (2004), 'Feature Unification for Morphological Parsing in Bangla', *Proceedings of 7th International Conference on Computer and Information Technology*, 26-28 December, BRAC University, Dhaka, Bangladesh, PP345-349.
- Hasan, M. K., Alim, M. A., and Md. Islam, W. (2002), 'Development of a Bangla SQL interface for DBMS', *Proceedings of 5th International Conference on Computer and Information Technology*, East West University, 27-28 December, Dhaka, Bangladesh.
- Mamun, M. A. A., Ahmed, M. I., Bhuiyan, M. A., Selim, M. R. and Iqbal, Z. (2002), 'An Implementation Of Machine Translation Between Bangla And English', *Proceedings Of International Conference On Computer and Information Technology*, NSU, Dhaka, Bangladesh, PP290-293.
- Raihan, S., Ali, and M. M. (2004), 'Bangla to English Translation by Rule-Based Approach', *Proceedings Of 7th International Conference On Computer and Information Technology*, 26-28 December, PP-123-129.
- Raihan, S., and Ali, M. M. (2004), 'Bangla to English Translation by Rule-Based Approach', *Proceedings of 7th International Conference on Computer and Information Technology*, 26-28 December, PP 256-262.
- Siddiui, M. H., Rana, A. K. M. S., Mahmud, A. A. and Sayed, T. (2003), 'Parts Of Speech Tagging Using Morphological Analysis In Bangla', *6th International Conference on Computer and Information Technology*, BRAC University, Dhaka, Bangladesh, PP374-379.