English to Bangla Phrase–Based Machine Translation System with the Help of Rules

A.O.M Asaduzzaman, Md. Nazrul Islam

Abstract— **In this paper we present a machine a machine translation system to translate an English sentence into a Bangla sentence of equivalent meaning. We have considered the translation of simple, declarative English sentences to Bangla for Present, Past and Future tenses. To do this work we employed Rule-based decision making principle along with phrasal decomposition of the sentences. A bilingual dictionary has been developed to provide the morphological properties and contextual information of the English words with their corresponding Bangla meaning. The proposed MT system starts with parts of speech tagging of the English sentences. Then the system applies a phrasal decomposition method to obtain individual phrases out of the sentences. To make the translation process easier a preprocessing of the phrases before translation has also been considered.**

Index Terms— **Machine Translation, Rule-based system, parts of speech tagging, bilingual dictionary, phrasal decomposition**

I. INTRODUCTION

Bangla (or Bengali) is one of the richest and largely spoken languages in the world. With nearly 230 million people speak in Bangla as their native language. It is ranked sixth based on the number of speakers [1]. However, only a very small number of tools and resources are available in Bangla.

Machine translation (MT) system generates a translation of natural language from one to another. The overall system involves analysis of the input sentence in the source language to discover its grammatical structure and transfer it to target language structure. The component words in the source language are identified with their morphological characteristics and then translated to obtain the meaning in the target language. A significant part of the development of any machine translation system is the creation of lexical resources that the system will use.

Natural Language Processing (NLP) is quite difficult task. There were many researches in the field of language translation but there is no fully successful language translation machine so far. Since it is a Human Language Technology (HLT), there are lots of varieties and lots of opportunities for research. There are varieties types of sentences both in English and Bangla language. In this paper we have considered only simple sentences. The system starts with the identification of parts of speech and assigns a notation corresponding to its type (i.e. noun, pronoun, adjective etc.). Along with the identification of parts of speech it also splits the sentences into various phrases. The phrases are then preprocessed for translation make easier. With the help of bilingual dictionary and contextual information, the system finally generates equivalent translation.

Manuscript received Dec 22, 2014

A.O.M Asaduzzaman, Assistant Professor, Dept. of CSE, Islamic University, Kushtia, Bangladesh

Md. Nazrul Islam, Assistant Professor, Dept. of CSE, Islamic University, Kushtia, Bangladesh

II. LITERATURE REVIEW

Machine translation (MT) has a long history of ambitious goals and unfulfilled promises. It is really difficult to build up a complete MT system for natural languages. Although being the widely used language, Bangla language still lacks significant research in the area of natural language processing, especially in MT.

Dasgupta [2] proposed an approach for English to Bangla MT that uses syntactic transfer of English sentences to Bangla aiming at optimal time complexity. They used the CYK algorithm to parse English sentences in Chomskey Normal Form (CNF). Then they used transfer rules and a bilingual dictionary to convert English parse trees to Bangla parse trees. Output translation was done with morphological generation. A hybrid scheme was shown in [3]. An algorithm for language translation using Artificial Intelligence (AI) was proposed in [4]. They considered only the present indefinite and present continuous forms of English sentences. Machine Translation of news headlines were described in [5]. They proposed an Example-Based Machine translation system. Naskar and Bandyopadhyay [6] showed a technique of handling prepositions in English to Bangla machine translation system. A semi-supervised approach was proposed in [7] for Bangla to English phrase based MT.

The quality and capability of an MT system largely depends on the size and quality of the bi-lingual dictionary. A bi-lingual dictionary was developed in [8]. The MT dictionary should contain a large collection of source language words with their meaning in the target language, and the morphological properties of the source language words. The selection of meaning of the source language words from the MT dictionary and to incorporate the meaning in the target language sentence requires intelligent algorithms. English to Bangla machine translation system employing heuristic or artificial intelligence in algorithms for word recognition and translation was proposed in [9]. Mapping rules had also been defined and used in the development of MT system [10] to obtain the equivalent Bangla grammatical structure of the structure of an input English sentence.

Different researchers employed different techniques, methods, and algorithm for Machine Translation. All of their efforts and outcomes have fueled our current research in MT especially in English to Bangla translation. To enhance and

contribute to the research in MT, the proposed system represented a new hybrid rule-based and phrase-based approach for English to Bangla machine translation.

III. PROPOSED ARCHITECTURE

The overall architecture of the proposed system is shown in Fig-1. The system is organized mainly in three phases some of which has several steps to be considered. In the first phase the system dissects the input paragraph into various sentences. Each sentence then steps into various processes to obtain the desired translation. In the last phase, the task is to assembly or post-processing of all the sentences into paragraph.

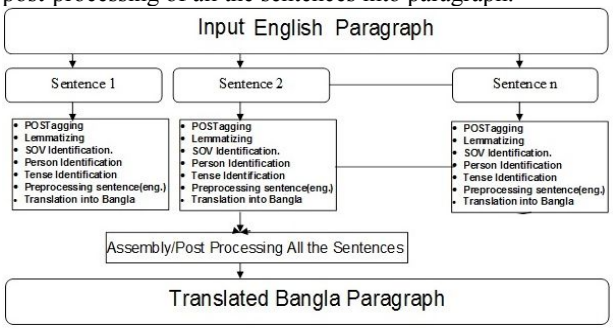

Fig. 1 : Architecture of English to Bangla Machine Translator

3.1 Sentence Translation

After dissecting the paragraph into sentences each sentence is ready for translation. This phase is composed of seven correlated sub phases. These are i) Parts of Speech (POS) Tagging, ii) Lemmatizing, iii) Subject Object Verb (SOV) Identification, iv) Person Identification, v) Tense Identification, vi) Preprocessing English sentence, and vii) Translation.

3.1.1 Parts of Speech Tagging:

Parts of Speech tagging is the most crucial area of translation. Error free translation is much dependent on a perfect POS tagging. Tagging is the process of marking up the words in a text corresponding to a particular part of speech, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. To tag a sentence with Parts of Speech various notations are used, which are collectively known as Tagset. There are various Tagset available; among them we used Penn Treebank Notation approached by University of Pennsylvania. There are 36 Tags in Penn Treebank notation. Some examples of Penn Treebank Tagset are shown in Table-1.

Table-1

Tag	Description
CC	Coordinating Conjunction e.g. *and, but, or.*. CD Cardinal Number
DT	Determiner
EX	Existential *there*
NN	Noun, singular or mass
NNP	Proper Noun, singular
RB	Adverb
RBR	Adverb, comparative
VB	Verb
--	----------

We used here Brill Tagger [11] method for Parts of Speech tagging. The method is an "error-driven transformation-based tagger". It is Error-Driven in the sense that it recourse to supervised learning. Transformation-based in the sense that a tag is assigned to each word and changed using a set of predefined rules. If the word is known, it first assigns the most frequent tag, or if the word is unknown, it naively assigns the tag "noun" to it. Applying over and over these rules, changing the incorrect tags, a quite high accuracy is achieved. The method is approximately 96-97% correct for Parts of Speech tagging.

3.1.2 Lemmatization and Stemming

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. This refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*. For example: am, is, are →be; car, cars, car's, cars' →car etc.

Stemming usually chops off the end of words and often includes the removal of derivational affixes. We have performed Stemming using the most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective is Porter's algorithm and Lemmatization in our own rule based approach. The main objective of stemming and lemmatization is to transform a word into its base form, to recognize it with a lexicon.

The plural number related words end with –ies, es, s, ves etc. Examples are: *pony →ponies*, *knife →knives*. With stemming the last ies, es, s, ves etc. are chopped off from the word. For example, for *knives*, after stemming it will be *'kni'*. Now it is the tern of lemmatization to make it *"knife"*. As our system is rule based we applied rules in lemmatization.

3.1.3. Subject Object Verb (SOV) Identification

To Identify SOV in a sentence its mandatory to detect the verb of the sentence at first. In our consideration Verb is the delimiter between Subject and Object, and Verb is the catalyst to identify Subject and Object.

3.1.3.1. Verb Identification

To identify Verb of a sentence, the sentence is put in an Array, and then checks the words for the verb TAG or for the Lemma 'be', 'have', 'shall' or will. The system continues checking until the last word and return *Verb Phrase* (VP).

3.1.3.2. Subject and Object Identification

Subject Identification:

Case 1: If there is no verb in the sentence, then the total sentence is considered as Subject. Case 2: If the first word is a Verb, the sentence is considered without Subject.

Case 3: If the verb identification is done in the middle of the sentence, then the section before the Verb is considered as Subject, i.e. SP (Subject Phrase).

Object Identification:

Case 1: If there is no Verb identified, then there would be no object.

Case 2: If the verb pointer is set to at the end of the sentence, then there would be no Object.

Case 3: If the verb pointer is set at the middle of the sentence, the section after the verb pointer is considered as Object, i.e. OP (Object Phrase).

3.1.4. Person Identification and Tense Identification

Person and Tense are identified according to the grammatical rules of the English sentence. Person identification is done through searching in Subject Phrase (SP). In English, Tense is determined from the Auxiliary Verb i.e. be-verb or have-verb, on the contrary there is no auxiliary verb in Bangla, tense is identified from the main verb. In Bangla.Tense is determined for making the Bangla main verb from the English sentence. With the tense information the main verb of Bangla is created with the modification of the root verb (main verb) of English sentence. From the Verb Phrase (VP) the Tense is identified. A function takes the VP as argument and returns the identified tense.

3.1.5. Preprocessing English Sentence

In this logical phase, the English sentence is prepared for translation into Bangla. Here the phrasal partitioning of SP, OP are done.

3.1.5.1. Subject and Object Preprocessing

The formation of English sentence is Subject+Verb+Object (SVO) and in Bangla it is (SOV). There are various noun/pronoun modifiers like prepositions, articles in English grammar. Generally the SP and OP contain the same kind of phrases, i.e. Noun Phrase (NP), Prepositional Phrase (PP), Adjective Phrase (AdjP), Adverbial phrase (AP), Gerund etc. To make the translation process easier, the SP and OP is preprocessed. Preprocessing contains various steps, which can be described as with the SP:

Step-1: Split into phrases

At this preprocessing step of SP or OP the translator traverses every word. According to the grammatical rules it sub-divides the SP or OP into various phrases depending on the various modifiers such as Prepositions, Adjectives, Adverbs, Gerund etc. present in SP or OP.

Step-2: Process modifiers of the step-1 Phrases

In this step, translator traverses every phrase and finds if there is any noun or pronoun modifier, if there is a modifier found it sets the modifier next to noun. However, NP contains nouns or pronouns that may appear with the determiner 'the'. We eliminate the determiner 'the' from NP for translation convenience. Again, Determiner Phrase (DTP) starts with a determiner i.e. 'a', 'an' or 'the'. If it starts with the determiner 'the' then 'the' is eliminated from DTP, otherwise no changes will be made. Preprocessing of other phrases is dealt in the similar fashion that the modifiers are transferred next to the noun. The Gerund, Adverbial Phrase, or other words excluding the above fact are processed as a single word Phrase. Again, if there is a pronoun next to 'to' we have eliminated the 'to', if there is any phrase that starts with 'the' we eliminated 'the' from the phrase. Also, if the Verb is only 'have', 'has' or 'had' it changes the subject noun or pronoun into its possessive form, but no changes if it is in OP, if it is pronoun then it changes like he-> his, I -> my and if it is noun then it adds a 's(apostrophe s) at the end of the noun. As for example, the sentence ***I have a Cow (I = আমি)***, will be preprocessed as ***my** have a cow **(my =আমার** একটি গরু **আছে)***

Step 3: Reverse step-2 Phrases

This is the last step of this Subject and Object preprocessing logical phase. The phrases are positioned one after another and are processed for the last time. In this step, the step-2 phrases are reversed from last to first, with one exception, if there is any pronoun (I,we,his,him etc.) in Subject Phrase (SP), pronoun phrase is set at the first place then other phrases are reversed sequentially last-to-first with their corresponding position. Fig.-2 illustrates the process with an example sentence: "The man in the jungle walking with his white horse".

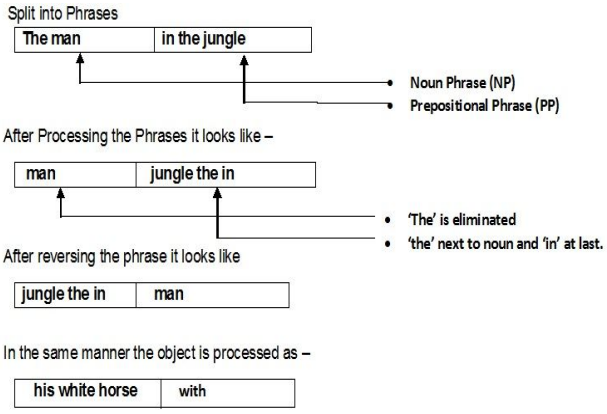

Fig.-2: Subject and Object preprocessing steps

3.1.5.2. Verb Preprocessing

In this step no changes are made in the Verbal Phrase (VP) during translation. The conditions are checked and needed eliminations are made. We preprocessed the sentences to bring flexibility in the translation phase. In this preprocessed form the translation process will be easier.

3.1.6. Translation

3.1.6.1. Subject and Object Translation

Subject and Object are translated separately, but their methods for translation are closely related. The translation of subject is done with the translation of various preprocessed phrases forming the Subject Phrase. Object translation is done with the preprocessed phrases forming the Object Phrase. In our work, we have maintained an English to Bangla Dictionary database to find the Bangla word for each lemma Word. It is to be noted that the preprocessed phrases contain Noun, Pronoun or Adjective. In translating the preprocessed phrases various morphological analyses is done and the translation proceeds accordingly. In general, the modifer modify the Noun/Pronoun according to grammatical rule which we have included into the module. However, there are other cases to consider. While translating the preprocessed noun phrases, there could be two situations: The determiner 'the' may or may not be present in the phrase. If there is 'the' next to Noun, the translator adds 'টি' at the end of the Bangla word for Noun. For example, man the (original: the man) → মানুষটি. Other considerations are: if Noun is plural, add 'গুলো' at the end of word, example: tables → টেবিলগুলো, if the plural word ends with ('s) , add 'র' at the end of word, example:

tables's → টেবিলগুলোর. Example translations based on the modifiers are shown below:

For the Noun Phrases (with determiner 'the') if the modifier is

for → add 'র জন্য' at the end, example, man the for (*original: for the man*) মানুষটির জন্য

in → add 'তে' at the end, example, man the in (*original: in the man)* → মানুষটিতে

by → add 'র দ্বারা' at the end, example, man the by → মানুষটির দ্বারা

For the Noun Phrases (Without determiner 'the') if the modifier is

for →If there is a Kar(কার – া, ি ,ী, ু , ূ, ে etc.) at the end → add 'র জন্য' at the end, ex: শেখা→শেখার জন্য

If there is no Kar and no Sorborno(স্বরবর্ণ) at the end → add 'ে+র+<স্পেস>+জন্য' ex: ফল→ফলের জন্য

at → If there is aa-kar(আ-কার) at the end→ add 'য়' at the end; ex: ঢাকা→ঢাকায়

If there is a Kar(ি ,ী, ু , ূ, ে etc.) except aa-kar at the end → add 'তে' at last, ex: বাড়ি→বাড়িতে

If there is no Kar and no Sorborno (স্বরবর্ণ) at the end → add 'ে' ex: ফল→ফলে

If there is Sorborno at end → add 'য়+ে' ex: সোনাই→ সোনাইয়ে

For the Pronoun phrase if the modifier is

for → add '<space> জন্য' at the end, example: for you → your for → তোমার জন্য

with → add '<space> মধ্যে' at the end, example: with you → your with → তোমার সাথে

With these morphological analysis, SP and OP are translated with a slight modification in OP, where the meaning of you is তোমাকে, whereas in SP the meaning of you is তুমি/আপনি. In fact, in Object phrase the meaning of pronoun is preserved by rule.

3.1.6.2. Verb Translation [12]

While translating verbs, our Translator leverages the power of the fact that Bengali derives its inheritance from *Sanskrit*. In Bengali, final forms of verb can be *generated* by knowing the base form of the verb (root verb), the person, the tense forms and whether verb is *active* or *passive*. For this moment we ignored the passive form translation. Construction of final verb forms in Bengali can be viewed as a method of multiplication between a *scalar* quantity and a *matrix*. In this analogy, all final verb forms can be viewed as the elements of resultant matrix from the product between scalar root verb (base form) and a *universal* matrix containing all verb modifier suffixes as its elements. The suffix matrix is universal in a sense that all verb forms for all tense can be generated from this matrix. Their product rules can be defined using the rules of joining words, the juncture rules which are known as *Sondhi* (সন্ধি).

We now discuss how our Translator system implements this abstract notion of verb form generation. Let us consider an example sentence, "I am reading" which has *"am reading"* as VP. It has two verbs: "am" with lemma "be" and "reading" with lemma "read". Here verb is acting actively. While constructing Bengali verb Translator ignores auxiliary verbs such as "am" here. To construct the final verb form in Bengali, Translator first performs a dictionary lookup for the lemma "read" whose base form in Bengali is "পড়". Then it looks up for the element in the suffix table corresponding to *Present continuous* tense, *person* and *active* verb. The corresponding suffix element in Bengali is "ছি". Other elements in the suffix table for such situations are given in the Table 1.

Table 1: Verb modifier suffix table for first person when verb is active

	Infinitive	Continuous	Perfect	Perfect Continuous
Present	"ি"	"ছি"	"েছি"	**"ছি"**
Past	"েছিলাম"	"ছিলাম"	"েছিলাম"	**"ছিলাম"**
Future	**"ব"**	**"তে থাকব"**	**"ব"**	**"তে থাকব"**

Having known the base form of the verb and relevant suffix, Translator joins them following the juncture rules. In particular, for this example final Bengali verb "পড়ছি" is generated as

পড় × ি ⇨পড়ি.
পড় × ছি ⇨পড়ছি.

For this example, the rule of joining is rather trivial. However, for many situations the joining rules can be quite involved. Generations of final verbs for all tense forms with subject being first person and verb being active, are illustrated below.

	ি	ছি	েছি	ছি		পড়ি	পড়ছি	পড়েছি	পড়ছি
প	"েছিলাম	ছিলাম	"েছিলাম	ছিলাম	⇨	পড়েছিলাম	পড়ছিলাম	পড়েছিলাম	পড়ছিলাম
ড়	ব	তে থাকব	ব	তে থাকব		পড়ব	পড়তে থাকব	পড়ব	পড়তে থাকব

Above verb form generating equation can be abstractly written as
Root verb form × Universal suffix matrix ⇨ Final verb forms matrix

Where a part of "Universal suffix matrix", relevant for *first person*, is illustrated in the Table 1. This aptly demonstrates the structural advantages for machine synthesis of Bengali sentences. These underlying structures owes to the fact that Bengali derives its origin from *Sanskrit*.
Universal suffix matrix for second person and third person also given bellow –

Table 2: Verb modifier suffix table for second person when verb is active

	Infinitive	Continuous	Perfect	Perfect Continuous
Present	"ো"	"ছো"	"েছো"	**"ছো"**
Past	"েছিলে"	"ছিলে"	"েছিলে"	**"ছিলে"**
Future	**"বে"**	**"তে থাকবে"**	**"বে"**	**"তে থাকবে"**

Table 3: Verb modifier suffix table for third person when verb is active

	Infinitive	Continuous	Perfect	Perfect Continuous
Present	"ে"	"ছে"	"েছে"	**"ছে"**
Past	"েছিলো"	"ছিলো"	"েছিলো"	**"ছিলো"**
Future	**"বে"**	**"তে থাকবে"**	**"বে"**	**"তে থাকবে"**

CONSTRUCTION OF FINAL BENGALI SENTENCE

Having constructed Bengali subject, object and verb separately, our translator joins them together to form the final Bengali sentence in the S-O-V order. For the example sentence, "আমি" (I), the subject (S), "একটি বই" (a book), the object (O) and "পড়ছি" (am reading), the verb (V) are joined together to form the final Bengali sentence "আমি একটি বই পড়ছি।".

ASSEMBLY OR POST-PROCESSING ALL THE SENTENCES TO PARAGRAPH

This is the final step towards the translation of input English paragraph. After translating all English sentences, the system assembled them one after another to produce the final output in Bangla as shown in Fig.-1. The punctuation symbol (only '.' was considered) acted as the sentence delimiter.

CONCLUSION

In this work, we tried to make a basic platform for translator that makes translation of English sentences to Bangla easier. The proposed system deals with only simple affirmative sentences. The other types of sentences are left for future work. The quality and effectiveness of the system depends largely on how strong the rules are and how large the English to Bangla dictionary is maintained. Our system works fine with grammatically correct English sentences and the words found in the dictionary. Again, same word can have different meaning in different sentences. However, the proposed system considered only a single meaning. In future work, a semantic analysis part can be added to the system to identify semantic error and correctly choose the word meaning among different available meanings. Increasing the effort in other areas like updating the dictionary, finding the rules for complex and compound sentences, introducing a machine learning system etc. can make the translator for general purpose use.

REFERENCES

[1] http://en.wikipedia.org/wiki/List_of_Languages_by_total_speakers
[2] Sajib Dasgupta, Abu Wasif and Sharmin Azam, An Optimal Way Towards Machine Translation from English to Bangla, Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT), 2004.
[3] Saha Goutam Kumar, " The EB-ANUBAD translator: A hybrid scheme, " in Journal of Zhehiang University of Science, pp 1047-1050, 2005.
[4] S. Ahmed, M.O. Rahman, S.R. Pir, M.A. Mottalib, and Md. S. Islam, " A New Approach towards the Development of English to Bangla Machine Translation System", in International Conference on Computer and Information technology (ICCIT), Jahangirnagar University, Dhaka, Bangladesh, 2003.
[5] Diganta Saha and Sivaji Bandyopadhyay, " A semantic-based English-Bengali EBMT system for translating news headlines", Proceedings of MT Summit X second workshop on example-based machine translation.
[6] Sudip Kumar Naskar and Sivaji Bandypadhyay, " Handling s of of prepositions in English to Bengali Machine Translation, Proceedings of the EACL workshop, 2006.
[7] Maxim Roy, " A Semi-Supervised approach to Bengali-English Phrase-Based Statistical Machine Translation, Proceedings of the 22nd Canadian Conference on Artificial Intelligence.
[8] Mortuza Ali and Muhamm Masroor Ali, " Development of Machine Translation Dictionaries for Bangla Language", in the Proceedings of the International conference on Computer and information Technology (ICCIT), Dhaka, Bangladesh, pp.267-271, 2002.
[9] S.A. Rahman, K.S. Mahmud, B. Roy and K.M.A. Hasan, " EnGlish to Bengali Translation using a New Natural Language Processing algorithm", in the Proceedings of the International conference on Computer and information Technology (ICCIT), Dhaka, Bangladesh, pp. 294-298, 2003.
[10] Sabbir Ahmed, Md. Obaidur Rahman, Saifur Rahman Pir, M.A. Mottalib and Md. Saiful Islam, " A new Approach Towards the Development of English to Bangla Machine Transaltion System", in the Proceedings of the International conference on Computer and information Technology (ICCIT), Dhaka, Bangladesh, pp. 360-364, 2003.
[11] http://en.wikipedia.org/wiki/Brill Tagger
[12] "A brief introduction to ANUBADOK (The Bengali Machine Translator)", by Golam Mortuza Hossain.