BANGLA TO ENGLISH TEXT CONVERSION USING OPENNLP TOOLS

Sk. Borhan Uddin[1], Dr. Md. Fokhray Hossain[2] and Kamanashis Biswas[3]

[1]Bangladesh Internet Press Limited, Dhaka
[2]Department of CSE, Daffodil International University
[3]Department of CSE, Daffodil International University

Email: *oneof.rebel@gmail.com, drfokhray@daffodilvarsity.edu.bd, ananda@daffodilvarsity.edu.bd*

Abstract- *Natural Language Processing is one of the most difficult areas in artificial intelligence. Because, completely different grammatical rules (on which the languages are based on) make the task more complicated. The same problems are found in conversion of Bangla text to English text. Bangla grammar maintains so many rules which make the task harder. Although, a number of researches are done in different area such as Bangla keyboard layout design, English to Bangla translator etc., very few researches are done to translate Bangla text to English. In this paper, we developed a system to do the conversion using OpenNLP tool that performs both statistical and rule based conversion. We have found that using this conversion method it is possible to translate about 40% of text from Bangla to English correctly.*

Keywords- *Machine Translation, OpenNLP Library, POS Tagging, Statistical Context Free Description*

1. Introduction

Since 1976 EC (European Commission) uses the MT (Machine Translator) to convert text from one language to another language. This broad usage spreads its importance widely and the translation technique is also developed for regular uses. Now-a-days, Google translator [1] is one of the pioneer applications supporting a number of languages to translate from one to another. Although, it has been successfully implemented for many languages but for Bangla language it is still in developing phase. The others translators e.g. Yahoo Babel Fish [2], SDL Free Translation, Systran Language Translation [3] etc. support multi language translation like Danish, English, Chinese, Italia, Japanese, French, Greek, Korean etc. but not Bangla. Hence, we developed the system which performs both statistical and rule based conversion to translate Bangla text to English.

2. Major Tools Used in Translation

2.1OpenNLP Library

OpenNLP is an organizational center for open source projects related to natural language processing [4]. Its primary role is to encourage and facilitate the collaboration of researchers and developers on such projects. OpenNLP is a java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and co reference. This tool can be integrated with other software to assist in the processing of text.

2.2 Statistical context-free description of compound structures

The statistical context-free description of compound structures is used to analyze compound entries. The linguistic description is a context free grammar, with associated linguistic probability. For example, to analyze the Bangla sentence "আমি ভাত খেয়েছি ।" as an [((NP (PRP) (NN) (NN))] noun phrase, the system uses the following rules:

noun-phrase (NP)	=>	.99 noun
pronoun (PRP)	=>	.20 pronoun
noun + noun	=>	.80 noun

2.3 Set of Rules

A set of rules defined, which are used to establish a sentence, according to the grammar rules. This

Date of submission : 01.12. 2011 Date of acceptance : 19. 07. 2012

operation is held in primary level, when we insert a simple sentence it just match some predefined common rules patterns. If the rule is absent then generate a pattern (rule) on basis of specific tense.

2.4 POS Tagging

POS tagging is the process of assigning a parts of speech for each word in a sentence. Here we have used Penn Treebank, the linguistic corpus developed by the University of Pennsylvania. The POS tagger returns array according to tokens.

3. Conversion Process

In this section, five major functions of the system are described. These are: (1) Bangla Grammar Detection (2) POS Tagging (3) Bangla Parse Tree Generation (4) Bangla Parse Tree to English Parse Tree Matching and (5) Bangla to English Text Translation.

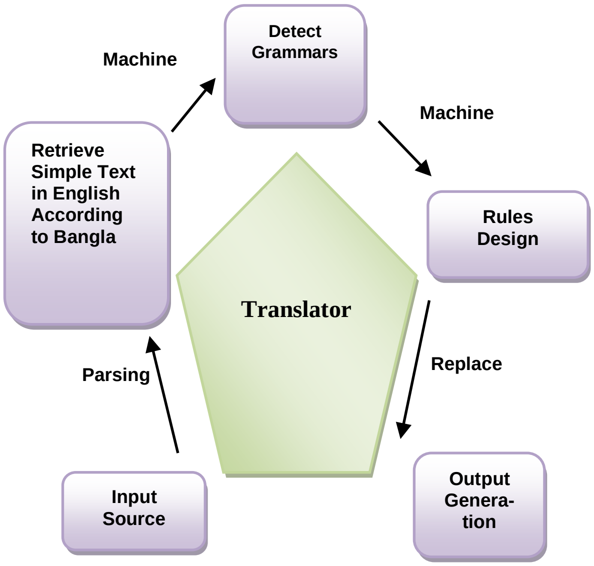


Fig. 1: Bangla to English Conversion Process

3.1 Bangla Grammar Detection

In this step, the first goal is to understand the grammar of input sentence using its components (like person, verb, objects etc). From the Bangla grammar sentence making rules, it is known that verb always sits at the end while person remains at the beginning in simple Bangla sentences. Then, to detect the tense, we just have to consider the Bangla verb. For example "আমি ভাত খেয়েছি ।" In this sentence verb is "খেয়েছি". We want to find out the tense. We have a pre-defined database where we put all Bangla verb keywords to detect tense as shown in the table in next column.

Table 1: Tense Mapping for Bangla sentence

Tense_Mapping		
ID	**BangSuffix**	**TensCode**
1.3	চ্ছে	12
1.1	য়, ই	11
1.2	চ্ছি	12
1.4	য়াছি	13
1.5	য়াছে	13
2.1	ছিল	21
2.2	ছিলাম	21
2.3	ছি	21
2.4	চ্ছেছিল	22

BangSuffix determines the keywords of the verb, and tense code determines the tense. BangSuffix 'য়' indicates tense code 11 which means Bangla Present Indefinite Tense. Now we get the tense (Bangla) from the table. In "আমি ভাত খেয়েছি ।" sentence "আমি" is determining the person. There is a table for person and number detection that helps to determine the Bangla person. At first "আমি" is converted into English i.e. 'I' then search in the table to detect the person and number. This table is also use to define pronoun. It helps to optimize the coding and searching data from database.

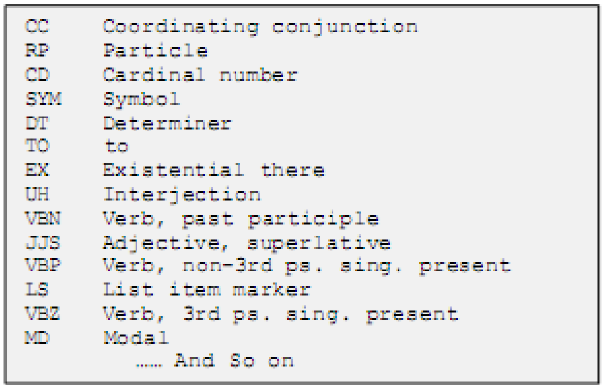
Table 2: Person detection table

Pronoun			
ID	**Pronoun**	**S_P**	**Person**
1	I	S	1
2	We	S	1
3	You	S/P	2

Pronoun			
ID	**Pronoun**	**S_P**	**Person**
4	He	S	3
5	She	S	3
6	They	P	3

3.2 POS Tagging

POS tagging is the process of assigning a parts of speech for each word in a sentence. Here we have used Penn Treebank, the linguistic corpus developed by the University of Pennsylvania [5]. The POS tagger returns array of tags and tokens. A sentence is splinted into tokens. Here we have used OpenNLP for POS tagging. Following figure shows the tag set used by OpenNLP [4].



```
CC    Coordinating conjunction
RP    Particle
CD    Cardinal number
SYM   Symbol
DT    Determiner
TO    to
EX    Existential there
UH    Interjection
VBN   Verb, past participle
JJS   Adjective, superlative
VBP   Verb, non-3rd ps. sing. present
LS    List item marker
VBZ   Verb, 3rd ps. sing. present
MD    Modal
        …… And So on
```

Fig. 2: Tag set used by OpenNLP [4]

3.3 Bangla Parse Tree Generation

This step consists of chunking and parsing. The chunkier returns phrase name based on pos tagging. Producing a full parse tree is a task that builds on the NLP algorithms, which goes in grouping the chunked phrases into a tree diagram that illustrates the structure of the sentence. The parsing algorithm is implemented by the OpenNLP library [6]. Based on chunk string parse tree is generated. The following figure shows the Bangle parse tree.

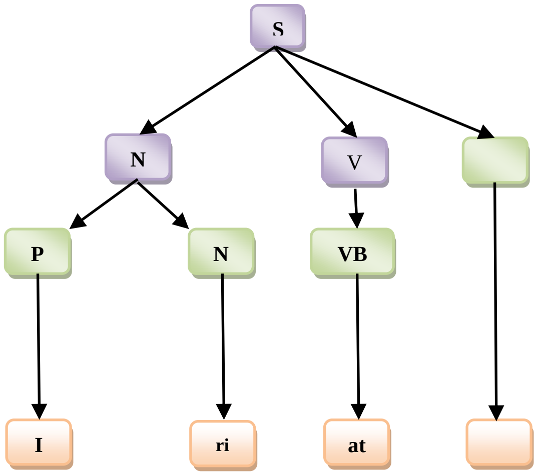


Fig. 3: Parse tree of Bangla sentence

Bangla POS tag is still in research level, so we represent Bangla parse tree in English. It represents the real Bangla sentence structure.

3.4 Bangla Parse Tree to English Parse Tree Matching

Most of the Machine Translation uses Chomsky Normal Form (CNF) [7] for defining grammar. Here, we have used a sorting method for defining English grammar. Let’s see the structure of Bangla sentence and English sentence according to CNF.

Bangla grammar in CNF is:

1. S = NP + NP + PP + VP
2. S = NP + PP + NP + AP

And, corresponding English grammar in CNF is:

1. S = NP + VP + PP + NP
2. S = NP + AP+ NP+PP

Here,
NP = DET + NOUN + PRONOUN
DET = “a”|”an”|”The”
VP = AUXILIRAY + PRINCIPLE

From the above example, for each Bangla rule there must be an English rule for generating Bangla parse tree. In CFG we have to define grammar for different sentences. But here we have used a sorting method for generating English parse tree. It helps to develop statistical knowledge of the system for further conversion. There are also some pre-defined rules that will be worked when a common pattern is matched between Bangla and English. Consider an example,

“আমি ভাত খেয়েছি ”

S = NP (PRP আমি) + NP (NN ভাত) +VP (VBPP খেয়েছি)

From the English sentence making rule, we get 'Subject + Verb + Object' which corresponding Bangla sentence rule is 'Subject + Object + Verb'. If we synchronize the Bangla rules with English rules we get,

Corresponding Bangla Sentence Syntax = I (PRP) + rice (NN) + have eaten (VBA+VBPP).

English Sentence :

$$\underbrace{\text{I (PRP)}}_{\text{Subject}} + \underbrace{\text{(have(VBA) + eaten (VBPP))}}_{\text{Verb (Auxiliary + Principle)}} + \underbrace{\text{rice(NN).}}_{\text{Object}}$$

A sorting mechanism is implemented to rearrange the word according to the rules, and generate the English parse tree.

3.5 Bangla to English Text Translation

After English grammar detection, we pass all the data (words) to a particular module known as sentence construction module as parameter. This sentence making process is done after completion of some steps. The steps are described here.

Step 1: To complete a sentence, we detect Parts-of-Speech from sentence at first. Here, we have to do POS tag again. For example, the sentence that we got at previous step is as follows.

Corresponding Bangla Sentence Syntax - I rice have eaten.

We have detected tense that is Present Perfect and person that is 1st person singular number.

English sentence with POS tag - I (PRP) rice (NN) have(VBA) eaten (VBPP).

Step 2: Rearrange the words according to the priority of POS tag. We get the priority value of Parts of Speech from the following table.

Table 3: Priority value of POS tag

POS Sort		
ID	**POSTAG**	**sortIndex**
1	DT	1
2	PRP	0
3	NNP	1.5
4	VBA	2
5	VBP	2
6	VBPP	3
7	NN	5

Example: "আমি ভাত খেয়েছি "

I (PRP) + rice (NN) + have (VBA) + eaten (VBPP)

From the priority Table we get the values of Parts-of-Speech:

PRP = 0; NN = 5; VBA = 2; VBPP = 3;

We get English Sentence (ES) = 0, 5 , 2 , 3 .
Now, we can form the sentence by rearranging these values in ascending order,
ES = 0, 2, 3, 5;
ES = PRP + VBA + VBPP + NN;
ES = I + have + eaten + rice.
Finally, we get- I have eaten rice.

4. Example Illustrating Sequential Steps

In this section, we have shown the conversion steps of another simple sentence according to rule based analysis:
At first, assume a simple sentence- " আমি ভাত খাই। "

Step 1: Split the sentence

আমি + ভাত + খাই

Step 2: Detect the tense from verb

আমি + ভাত + খাই ⟵ Bangla verb

Output: From database and defined rules we get "খাই" represents the present indefinite tense (Grammar Detection).

Step 3: Detect the person, compare the data from database

Bangla Person ⟶ আমি + ভাত + খাই

Indicating the 1st Person singular number …(Detect Person)

Step 4: Insert the corresponding English word from Database

I + rice + eat ← Fetch words from database

Output: I rice eat….. (Data fetch)

Step 5: POS tag the sentence to get the Bangla parse tree

PRP/I, NN/rice, VBP/eat ← Parse the whole sentence

Output: (S(NP((PRP/I)(NN/rice))(VP(VBP/eat))) (Sentence Parsing)
Step 6: Search the sentence pattern from database.

BS pattern: PRP + NN +VBP corresponding ES pattern: PRP + VBP + NN

Output: (S (NP (PRP/I)) (VP (VBP/eat) (NP (NN/rice))) ... English Parse tree Generation

Step 7: Get the sentence from parse tree.

PRP/ I, VBP/ eat, NN/ rice.

I eat rice. ………… Final output

5. Algorithm

It is almost impossible to develop an algorithm which is highly efficient in translating one language to another. Specially, the languages with huge number of grammatical rules lead many problems such as if we want highly accurate result then the conversion will need more time. Hence, the target is to achieve near most translation from which we can understand the sentence. The algorithm that we developed also provides the conversion which is not hundred percent accurate for complex sentence but the meaning can be understood. The algorithm is described below.

```
Step1W:  Get page or Input from any source.
Step2W: Split the Bangla sentence into word and find out the tense, person etc.
Step3W: Send text to Machine Translator as parameter.
===============MT=================
Step 1:  Receive input as text.
Step2: Determine the Bangla tense from the Bangla verb (scaning the last bangla word) get the person from first Bangla word.
Step 3: Find all the English word according to the Bangla words.
Step4: For each sentences do
        a. Tokenizing sentences
        b. Parts of speech tagging.
        c. Divide in chunk and generate parse tree.
        d. Find out the appropriate rules
        e. Rearrange the words according to the rules.
Step 5: Print English sentence.
=============END OF MT=============
Step5W: Replace Bangla content with english content.
Step6W: Show the target language.
```

6. Conclusion

A number of critical issues always make natural language processing tasks more complex. There are a number of exceptions that violate the normal rules of grammar. And this is really tough to keep track of all those situations. Hence, the efficiency in translating languages with complex grammatical rules is not too high. We have implemented the system and found that from a chunk of text it can translate 40% of sentences correctly in average. Our observations have found that for simple sentences the system can easily response with correct answer (e.g. ”আপনি কোথা হতে আসছেন?”, Where are you coming from?) but for complex sentences, (e.g. আমরা গতকাল বই ক্রয় করতে গিয়েছিলাম, কিন্তু আমরা দেরি করেছিলাম| We went buy book yesterday but we were late.) it requires more time and in some cases it cannot do it properly.

6.1 Future Work

The main challenge in Bangla to English text conversion is grammatical rules. If we can make a complete format for all rules and exceptions then the task will be simpler. Efficient AI techniques, indexing and searching mechanisms will improve the total system that may result in more accurate output.

References

[1] Google Translator, www. translate.google.com, accessed on September 12, 2011.

[2] Yahoo Babel Fish, www. babelfish.yahoo.com, accessed on September 12, 2011.

[3] Jin Yang and Elke D. Lange, Systran on Altavista A User Study On Real-Time Machine Translation On The Internet, Lecture Notes in Computer Science, 1998,

Volume 1529/1998, 275-285, DOI: 10.1007/3-540-49478-2_25

[4] OpenNLP, *www.maxent.**source**forge.net*, accessed on September 12, 2011.

[5] Penn Treebank Project, www.cis.upenn.edu */~treebank/* , accessed on September 12, 2011.

[6] *The OpenNLP*, www.opennlp.sourceforge.net */projects.html*, accessed on September 12, 2011.

[7] Chomsky Normal Form, www. *en.wikipedia.org /wiki/Chomsky_normal_form*, accessed on September 12, 2011.

Sk. Borhan Uddin has completed his graduation from Daffodil International University, Dhaka, Bangladesh in Computer Science and Engineering. His major area of research interest includes natural language processing, robotics and neural networks. At present, he is working as software engineer in Bangladesh Internet Press Limited.

Dr. Md. Fokhray Hossain obtained his B.Sc. (Honors) and M.Sc. in Physics from ***Jahangirnagar University*** in 1991 and subsequently appointed as a research fellow at ***Dhaka University*** in 1993. Dr. Hossain obtained his Ph.D. from ***University of Glamorgan***, UK back 1998 through Overseas Development Administration Shared Scholarship Scheme (ODASSS). At present, he is working as registrar at Daffodil International University.

Kamanashis Biswas has post graduated in security engineering from BTH, Sweden. His major area of research interest includes artificial intelligence, algorithm, and computer and network security issues. At present, he is working as Assistant Professor at Daffodil International University, Dhaka, Bangladesh