2010-01-01

MACHINE AND WEB TRANSLATOR FOR ENGLISH TO BANGLA USING NATURAL LANGUAGE PROCESS

Haque, Mohammad Hasibul
Daffodil International University

http://hdl.handle.net/20.500.11948/509

Downloaded from http://dspace.library.daffodilvarsity.edu.bd, Copyright Daffodil International University Library
Abstract: The modern web contents are mostly written in English and developing a system with the facility of translating web pages from English to Bangla that can aid the massive number of people of Bangladesh. It is very important to introduce Natural Language Processing (NLP) and is required to developing a solution of web translator. It is a technique that deals with understanding natural languages and natural language generation. It is really a challenging job to building a Web Translator with 100% efficiency and our proposed Web Translator basically uses Machine Translator as its mother concern. This paper represents an optimal way for English to Bangla machine and the Web translation & translation methods are used by translator. Naturally there are three stages for MT but here we propose a translation system which includes 4 stages, such as, POS tagging, Generating parse tree, Transfer English parse tree to Bengali parse tree and Translate English to Bangla and apply AI. An innovation initiative has scope of being upgraded in future and hopefully this work will assist to develop more improved English to Bangla Web Translator.

Keywords: Machine Translator, Web Translator, POS Tagging, Parsing, HTML Parsing, Verb Mapping

1 Introduction

Machine Translation has a mechanism of translating languages from one to another and it is one of the important areas of Natural Language Processing (NLP). And Web Translation is a procedure of translating contents of web pages from one language to another.

Our objective is to present a transfer methodology for both of the English sentences to corresponding Bangla sentences and English web pages to corresponding Bangla web pages and that is presenting both the Machine and Web Translation procedure. Here we propose a methodology for Machine Translation which has 4 stages. Here we have used Peen Treebank algorithm [14] which has minimum parsing steps and MIL HTML [1] parser for parsing web pages.

Yahoo Babel Fish, Google Free Online Language Translator and some others highly developed solutions provide support for multi language translation like Danish, English, Chinese, Italia, Japanese, French, Greek, Korean and so on. But we were shocked to find that there is no existing system that provides the facility for Bangla translation. However some Bangladeshi and Indian developers develop English to Bangla translation methodology using Natural Language Processing [5, 6, 7, 9, 11, 12, 13] but most of them deals only with Machine Translation. From the best of our knowledge we are one of the pioneer of developing web based English to Bangla translator.

Difference between English and Bengali sentences structures:

Direct interpretation isn’t possible because of structural difference between English and Bengali. There are many situations where direct interpretation causes problem. In this case we used Artificial Intelligence (AI). This type of interpretation is possible for human beings but quite difficult for computer. Consider some examples.

I have to go. “আমার যেতে হবে।” Here direct interpretation isn’t possible. Let’s see another example He is too weak to walk. “হে এক পুরুষ যা দাঁড়িয়ে পাড় না।” here direct mapping or interpretation isn’t possible.

2 Proposed Methodology/Architecture

We have divided the whole system into two parts:

1. Machine Translator
2. Web Translator

Fig. 2.1 shows the block diagram of overall architecture for machine and web translation.
2.1 Machine Translator

Machine Translator translates source language to target language using Machine Translation (MT). This architecture has 4 steps. (1) POS tagging (2) Generating parse tree (3) Transfer English parse tree to Bengali parse tree (4) Translate English to Bangla using AI.

2.1.1 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 [14]. The POS tagger returns array of tags and tokens. A sentence is splinted into tokens. Here we have used OPEN NLP for POS tagging.

Fig. 1 Proposed architecture (step by step)

Following Fig. shows the tag set used by OPEN NLP [2, 3].

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>SYM</td>
<td>Symbol</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>EX</td>
<td>Existential there</td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
</tr>
<tr>
<td>VBN</td>
<td>Verb, past participle</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective, superlative</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, non-3rd ps.sing.present</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker (and so on)</td>
</tr>
</tbody>
</table>

Fig. 2 Tag set used by open NLP [2,3]

2.1.2 Generating Parse Tree (Parsing)

It is one of the complex steps and consists of chunking & 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 full parsing algorithm is implemented by the OpenNLP library [2]. Based on chunk string parse tree is generated. Fig. 2 and Fig. 3 shows parse tree for English and Bangla sentence.

Open NLP can detect following phrases and based on these parse tree is generated.

Consider following example.

**English Sentence:**
He gave me a book.

**After Chunking:**
[NP He/PRP] [VP gave/VBD] [NP me/PRP] [NP a/DT book/NN].

**After Parsing:**
(TOP (S ((NP (PRP He)) (VP (VBD gave)) (NP (PRP me)) (NP (DT a) (NN book))) (..)))

Fig. 4 shows the parse tree for “He gave me a book.”

Fig. 3 Chunk set used by the OpenNLP [2, 3]

Fig. 4 English parse tree for “He gave me a book.”
2.1.3 Transfer English to Bangla Parse Tree

Most of the Machine Translation uses Chomsky Normal Form (CNF) for defining grammar. Here we have used a sorting method for defining Bangla grammar. Let’s see the structure of English sentence and Bangla sentence according to CNF. After that we will describe our sorting method that means English to Bangla parse tree transfer algorithm.

English grammar in CNF
1. \( S = NP + VP + PP + NP \)
2. \( S = NP + AP + NP + PP \)

And corresponding Bangla grammar in CNF
1. \( S = NP + NP + PP + VP \)
2. \( S = NP + PP + NP + AP \)

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

From the above example, for each rule in English there must be a Bangla rule for generating Bangla parse tree. In Context Free Grammar (CFG) we have to define grammar for different sentences.

But here we have used a sorting method for generating Bangla parse tree. After generating English parse tree we found whole sentence with phrase name and also tag name.

We have defined index value for each phrase/chunk and also defined index value for each tag.

These index values are used for sorting. Now consider English parse tree (Fig. 2.1.2-b) and corresponding Bangla parse tree (Fig. 2.1.2-c) which are generated from following chunk string.

Chunk string for English Sentence:
\[ NP(\text{PRP}) + VP(\text{VBD}) + NP(\text{PRP}) + NP(DT,NN). \]

Chunk string for Bangla sentence:
\[ NP(\text{PRP}) + NP(\text{PRP}) + NP(DT,NN) + VB. \]

Here we need to generate equation (2) from equation (1). If we can generate equation (2), then easily we can generate Bangla sentence. Now we will use our sorting method to get Bangla parse tree that means equation (2).

Consider following example.

**Step 1: English Sentence**
\( S = \text{He} + \text{gave} + \text{me} + \text{a pen}. \)

**Step 2: Tagging and Parsing**
\( S = NP(\text{PRP}) + VP(\text{VBD}) + NP(\text{PRP}) + NP(DT,NN). \)

**Step 3: After Indexing**
\( S = 0(1) + 3 + 0(1) + 0(1.5, 2) \)

Fig. 6 and 7 shows index value for phrases and tags.

<table>
<thead>
<tr>
<th>PhraseName</th>
<th>PhraseIndex</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>0</td>
</tr>
<tr>
<td>ADVP</td>
<td>1</td>
</tr>
<tr>
<td>ADJP</td>
<td>2</td>
</tr>
<tr>
<td>VP</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TagName</th>
<th>TagIndex</th>
<th>AlterTagIndex</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>PRP$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PRP</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>WP</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
After getting index value we will perform sorting based on index.

**Step 4: After Sorting**

$$BS = 0(1) + 0(1) + 0(1.5, 2) + 3$$

$$BS = NP (PRP) + NP (PRP) + NP (DT, NN) + VP (VBD).$$

**Step 5: After dictionary search**

$$BS = \text{He} + \text{me} + \text{a pen} + \text{gave}.$$ 

Consider another example

**Step 1: English Sentence**

$$S = \text{what} + \text{is} + \text{your} + \text{name}.$$ 

**Step 2: Tagging and Parsing**

$$S = NP (WP) + VP (VBZ) + NP (PRP$, NN)?$$

**Step 3: After Indexing**

$$S = 0(2) + 3 + 0(1.2, 2)?$$

After getting index value we will perform sorting based on index.
After Sorting we will get.

**Step 4: After Sorting**

$$BS = 0(1.2, 2) + 0(2) + 3?$$

$$BS = NP (PRP$, NN, WP) + VP (VBZ)?$$

**Step 5: After dictionary search**

$$BS = \text{your, name, what} + \text{is}?$$

By considering above steps we may transfer English sentence to Bangla sentence. We have also used following algorithm for transforming

BanglaParseTreeGeneration(English chunk set) {
  1. Searching for conjunctional phrase or tag named CC.
     a. If found split sentence in two segment
     b. Else Don’t need to segment.
  2. searching for Preposition.
     a. If found
        i. At the end of the sentence or segment then place it before next noun and pronoun.
        ii. Not at the end of the sentence or segment then place it before next noun and pronoun.
  3. Retrieve index no corresponding to Phrase and Tag name.
  4. For each index find out corresponding Phrase name and then search that phrase from chunk set
     a. If found then put in SortedChunk = SortedChunk + searched Phrase(Chunk)
  5. Convert SortedChunk to SortedChunkArray[] by Splitting
  6. For each chunk perform tag sorting
     a. SortedChunk[.] = TagSorting(SortedChun k[.]’s tags)
  7. Retrun SortedChunk; // This is the sorted sentence with tag and chunk used to generate Bangla parse tree.
}

GetBanglaSen(BangSortedChunkList){
  String Banglasen ="";
  For each tag in sorted chunk
  Banglasen := Banglasen + GetBanlaWar(ward/tag); GetBanlaWar(English ward/tag)
  { 
    If (tag == IN||TO) // for preposition 
    { Return PrepositionMapping(English ward, reference object); }
    If (tag == VB||VBD||VBG||MD||VBN)
    { Return VerbMapping(….,
    EnglishWard); }
    Search Dictionary for Enlish Ward.
    If found then set BanglaWard := searched ward.
    else if tag == NN then
    set BanglaWard := Phonatic(Engward)
    else BanglaWard = EngWard;
    return Banglaward;
  }
}

Fig. 8 Algorithm for Bangla parses tree generation

2.1.4 Translate English to Bengali using AI

The next process is performing dictionary lookup. After getting Bangla Parse tree means Bangla chunk string we perform dictionary look. If the word is found then we
write the corresponding Bangla meaning on the contrary if not then return that English word. But the technique is different for verbal, prepositional phrases. For translating verb and preposition we have used verb mapping and preposition mapping.

2.1.4.1 Verb Mapping
Verb mapping is the process of creating Bangla verb corresponding to English verb. Bangla has come from an origin of Sanskrit. In Bangla final form of verb can be generated by knowing the base form of the root verb. More precisely construction the final verb in Bangla can be viewed as method of multiplication between a scalar quantity and matrix.

Our Verb Mapping takes four parameters
1. Reference object (Noun/Pronoun)
2. Auxiliary verb
3. Been
4. Main verb

Based on these parameters we generate Bangla verb. Consider following sentence.
I am doing math.

Here Reference object = I.
Auxiliary verb = am
Been =null
And Main verb = doing.

The above example has two verbs am with lemma be and doing with lemma do. Here verb is acting actively. To construct the final verb we first performs a dictionary lookup based on verb tag whose base form in Bangla is “কর।” Then it looks up for the element in the verb mapping table corresponding to person and tense. Here person is generated from reference object and tense is generated from combination of auxiliary and main verb. Let’s see once more.

Here “I” is first person. Auxiliary verb “am” represents present tense. Main verb “doing” represents continuous.
We hope this following table will be more helpful for realizing the Verb mapping based on above information.

<table>
<thead>
<tr>
<th></th>
<th>Indefinite</th>
<th>Continuous</th>
<th>Perfect</th>
<th>Perfect Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>ই</td>
<td>ই</td>
<td>ই</td>
<td>ই</td>
</tr>
<tr>
<td>Past</td>
<td>ইডিয়াম</td>
<td>ইডিয়াম</td>
<td>ইডিয়াম</td>
<td>ইডিয়াম</td>
</tr>
<tr>
<td>Future</td>
<td>অ্যায়</td>
<td>অ্যায়</td>
<td>অ্যায়</td>
<td>অ্যায়</td>
</tr>
</tbody>
</table>

From this table we generate suffix. After generating suffix, it will add with base verb. Here “কর।” is generated as কর * ই = কর।[5]

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.

![Fig. 9 Verb mapping (Suffix table)](image)

Above verb form generating equation can be abstractly written as

\[
\text{Root verb form} \ast \text{Universal suffix matrix} = \text{Final verb forms matrix}
\]

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. [10]

2.1.4.2 Preposition Mapping
Preposition is usually placed before noun/pronoun and shows the relationship among other nouns, pronouns and verbs in the sentence. The noun that follows a preposition, i.e., the reference object is in the accusative case and is governed by the preposition. Prepositions can also be defined as words that begin prepositional phrases (PP). A PP is a group of words containing a preposition, an object of the preposition, and any modifiers of the object.

The translation of the three English prepositions 'in', 'on', and 'at' has involved for identifying the possible inflection to be attached to the head noun of the PP. No postpositional words are placed after the head
noun for these prepositions. The three prepositions ‘in’, ‘on’, and ‘at’ (which are both spatial and temporal in nature) can be translated into the Bengali inflections এ (e), এ (te), -এ (ete) and স (y). Any of these 4 Bengali inflections can be placed after the reference object for any of these 3 English spatial and temporal prepositions. The choice depends on the spelling of the translated reference object. The rule is:

If the last letter of the Bengali representation of the reference object is a consonant, এ (e) or -এ (ete) is added to it (e.g., at/in market), else if the last letter of the Bengali word is a matra (vowel modifier) and if the matra is অ (a), any of এ (et), or স (y) can be added to the Bengali reference word (e.g., in evening 

Web translation performs following operations. 1. Send to machine translator 2. Replace source text to target text 3. Replace source text to target text

### 2.2.1 HTML Parsing

Web pages are shown in browser and web pages are viewed as an HTML page and consist of HTML code. Here we have used HTML parser [1] for parsing the HTML code. In this stage we have filtered the text from HTML source code. For filtering the
HTML source we considered the HTML document as a tree. Consider following HTML code

```
<html>
  <head>
    <title>This is tested by ER</title>
  </head>
  <body>
    <h1>Hi I am hasirul.</h1>
    <a href="#" title="has
    ha">doing test.</a>
  </body>
</html>
```

Fig. 11 HTML source code

If we format this source code according to tree then we will find following tree.

```
<html>
  <head>
  </head>
  <body>
    <h1>
    </h1>
    <a>
    </a>
  </body>
</html>
```

Fig. 12 Tree view of HTML source code

Our algorithm finds those nodes which have no child. Then those nodes are sending to machine translator for translating.

We have used following algorithm for finding the text node.

```
Step 1: Retrieve HTML code into a string
Step 2: While(HTML string !=EOF)
  {
    Searching for initial tag If found
    {
      If(tag is node or element)
        Nodes[]element := element
        Nodes[].parent := parent element
        Nodes[].child := child element
      If(tag is script)
        Put into Script[]:= script
      If(tag is style)
        Put into Style[]:= style
    }
  }
Step 3: foreach(node in nodes)
  {
    If(node.child == false)
    {
      Text[] = node.text;
    }
  }
```

Fig. 13 Algorithm for retrieving text from HTML source.

**2.2.2 Send to Machine Translator**

This is simple stage. Here we just send each text to machine translator as an input parameter of machine translator and machine translator returns target language.

**2.2.3 Replace source text to target language text**

This is the last stage of web translation. Now we have source text, corresponding target text and source HTML code. For replacing source text to target text we need to perform find and replace operation. Find operation searches the source text in HTML source code and replace operation replaces that source text to target text. After replacing source code is written in memory as a HTML file and shown in a browser. We have used algorithm in Fig. 14 for finding and replacing.

```
// we have a string with
source HTML code  strSource
// we have a array of text
text[]
Foreach (string strText in
text[])
{
  Set MSource :=
  replace(strSource, strText, MT(strText));
}

// this function write on disk
Write(MSource);
//MT() returns target text
that means
It takes English text as a
parameter and returns Bangla text
```

Fig. 14 Algorithm for replacing source text to Target text

**3 Results**

As we described our translator perform machine and web translation so we can divide result finding in two parts.

**3.1 Machine Translation**

For every translation it requires minimum 2.40 sec because it requires 2.40 sec for initializing service.

**3.1.1 Performance**

Following table shows performance based on above (Table 3) input and output.

**3.2 Web Translation**

Web translator receives an URL of an English web page and produce corresponding Bangla page temporarily.
### Table 3 Input and corresponding output.

<table>
<thead>
<tr>
<th>Case</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>The boy is playing.</td>
<td>ব্য্যক্তি খেলছে।</td>
</tr>
<tr>
<td>Case 2</td>
<td>This is test. I am waiting for him.</td>
<td>এটি পরীক্ষা। আমি তার জন্য অপেক্ষা করছি।</td>
</tr>
<tr>
<td>Case 3</td>
<td>This is the boy who came yesterday. Who are you? I am waiting for you. I love you.</td>
<td>এটি থাকবে গজাল যে ওখুনি। আপনি কে? আপনার জন্য অপেক্ষা করছি। আমি আপনাকে ভালবাসি।</td>
</tr>
</tbody>
</table>

### Table 4 Performance of sample input and output.

<table>
<thead>
<tr>
<th>Case</th>
<th>Words/Sentences</th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>W: 4, S:1</td>
<td>2.40 Sec</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>W: 8, S:2</td>
<td>2.41 Sec</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>W: 18, S:4</td>
<td>2.45 Sec</td>
<td>94%</td>
</tr>
</tbody>
</table>

### Table 5 Input and output page generated by the system.

<table>
<thead>
<tr>
<th>Input URL</th>
<th>Output URL</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Input Page (English)</th>
<th>Output Page (Bangla)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Input Page" /></td>
<td><img src="image2" alt="Output Page" /></td>
</tr>
</tbody>
</table>

### 3.2.1 Performance

It is quite difficult to create a translator which provides 100% efficiency. Our translator unable to translate compound and big sentences because lack of AI, rule and also lexical resource. If we reach our lexical resource and introduce more grammatical rule then it will show good performance.

### 4 Conclusion

Though a complicated work is done, our proposed system requires more improvement in some cases like detecting multiple Bangla meaning for an English word and improving the artificial intelligence to detect phrases and idioms. Moreover the sentence structure of Bangla and English varies for different sentences. As why the sorting method we have proposed fails to translate all sentences with 100% efficiency. In future we would like to improve our proposed system for detecting multiple meanings, phrase and idioms. Besides we will develop a strong data dictionary.

In this paper we tried to provide transfer architecture in MT and also WT (web translation) which is quite efficient. The grammars and examples we used here are simple. We didn’t consider semantic analysis or other disambiguation related with Bengali. It has been implemented in such a way to make it possible for extending in the future successfully and efficiently.

### References


Mohammad Hasibul Haque is working as a software developer, Software Department, Daffodil International University. He has obtained his Bachelor from Daffodil International University. His research interests are natural language processing, Software architecture.

Dr. Md. Fokhray Hossain was born on 12th November 1965 at Rangpur, Bangladesh. He obtained his B.Sc. (Hons) and M.Sc. in Physics from and subsequently appointed as a research fellow at Dhaka University 1991- Jahangirnagar University 1993. Dr. Hossain obtained his Ph.D. from University of Glamorgan, UK back 1998 through Overseas Development Administration Shared Scholarship Scheme (ODASSS). He worked for Daffodil Institute of IT (DIIT) as a faculty as well as an Academic Director from 1999 to 2008. Currently, he has joined Daffodil International University (DIU) as Registrar on 3rd November 2009. Earlier, he was the Head of the Department of Computer Science and Engineering of DIU. Before that he had also served at various local and foreign universities as academician and researcher including London Metropolitan University, UK. He also achieved different professional certificates such as Time Management and Introduction to HRM (2008) in collaboration with Netherlands Management Programme (PUM) and DIU; Academic Standards and Quality (2005) from National Computing Center (NCC), UK; e-business Model and Strategy (2004) from S. Korea through DGF-KTC scholarship; Class Management & English Language Training (2003) from British Council; Digital Telecommunication Technology (2001) through AOTS scholarship, Japan.

A.N.M Fauzul Hossain is working in International Organization for Migration (IOM). He has obtained his Bachelor from Daffodil International University.