IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 2, Ser. II (Mar. –Apr. 2022), PP 55-60
www.iosrjournals.org

A Rule Based Approach for Implementation of English to Bangla Translator

Md. Foridul Islam[1], Md. Omar Faruqe[2], Abdullah Al Shiam[3], A. F. M. Mahbubur Rahman[4], Utpala Nanda Chowdhury[5]

[1,2,3,4,5](Computer Science and Engineering, University of Rajshahi, Bangladesh
forid.cse@gmail.com[1], faruqe@ru.ac.bd[2], shiam.cse@shu.edu.bd[3], mmr@ru.ac.bd[4], unchowdhury@ru.ac.bd[5]

Abstract:
Machine translation (MT) is the process of text translation from one language to another using bilingual data sets and corresponding grammatical rules. The majority of Bangladeshi, particularly the distant villagers, cannot read or write English well. So, an efficient English to Bangla language translator can alleviate the inability. In this study, we proposed an English to Bangla MT system using rule based methodology. The proposed system works in three steps. The Scanner reads the input sentence, tokenizes words, checks spelling, populates information and stores the results. For spell checking, the scanner uses a database and word morphology technique of each word and finds lexical information from the dictionary. The Parser parses the input sentence to check syntactic correctness and then identifies its tense category and stores the results. The parser uses a rule based top-down parsing technique for this purpose. And finally, the Bangla Generator generates Bangla sentences which are equivalent to the input English sentence.

Key Word*: Machine Translation (MT); Rule based MT; English to Bangla translation.*

Date of Submission: 02-04-2022 Date of Acceptance: 15-04-2022

I. Introduction

Machine Translation (MT) is the translation of text from one natural language (source language) to another language (target language) using a computerized system with or without human interaction [1]. MT is an automated system, sometimes referred to as Natural Language Processing (NLP), which uses other language resources and bilingual data sets to build language and phrase models for text translation. Ideally, MT is a batch process that is applied to a given text for producing a perfect translated text [2]. The aim is to fill the communication gap between different societies with language diversity. Manual human translation is time consuming for any language. But an efficient MT system can reduce both time and cost involved in the translation. The official language of Bangladesh is Bangla. Nowadays many documents, papers, journals, books, records, news etc. are written in English, which is not understandable among all the people of Bangladesh. Hence it is important for an automatic translation system from English to Bangla. This motivated the development of an automatic system that may be capable of translating sentences from English to Bangla with an intent to help Bangladeshi students or general people usually who are unfamiliar with English. It is very important in our modern civilized society. MT is also important in the field of business, economics and the industrial world. In this translation process the meaning of the source language is fully restored in the target language.

II. Literature Review

Nowadays, Computers are widely used in every part of life. MT is indisputably an important topic socially, politically, commercially, scientifically, intellectually or philosophically. And the importance of MT is likely to increase as the 20th century ends and the 21st century begins [3]. If we can use this machine for the purpose of translating sentences from one human language to another, it would be a nice technology to break linguistic barriers among different nations. MT systems have several approaches: Direct-Based MT, Corpus-Based MT, Knowledge-Based MT, Lexical-Based MT, and Dialogue-Based MT. Direct-Based MT systems translate individual words in a sentence at a time from one language to another using phrasebook [4]. Corpus-Based MT rely on the study of bilingual text corpora. Statistical MT and Example-based MT come under this category. Statistical MT is good for catching exclusions to rules during the translation. The primary advantage of the statistical MT is that it does not require philological information in the translation process [5]. Knowledge-Based MT requires to be formed based on ontology and semantic web [6]. Lexical-Based MT systems translate individual words, with lexical information.

III. Related Works

Currently, several institutions are continuously working on MT system implementation and improvement. Since the mid-80's and 90's, numerous institutions have started research activities in this arena. To improve the accuracy and efficiency, a number of MT groups together with Government and Commercial projects are actively involved in different countries. Some of the important ongoing MT projects are:

a. Asian Projects

Asian nations- Thailand, Malaysia, Indonesia, China and Japan have been working on a multilingual MT system, which adopts an Interlingua approach. It was launched in 1987 and it conducts research on MT electronic dictionary and related areas.

b. Indian Projects

India also conducts research on MT at different universities, institutes and research organizations. Recently Tamil University in Thanjavur, National Center for Software Technology (NCST) in Mumbai, Center for Development of Advanced Computing (CDAC) in Pune, and IIT in Kanpur have undertaken such projects [7].

c. Bangladeshi Projects

In Bangladesh, many institutions are engage in developing MT system, that includes:

Anubadok-MT system: Mukta, Afsana Parveen, et al. proposed a Phrase-Based MT system using Rule-Based approach for English to Bangla translation [8].

MT system for English to Bangla using NLP: Usually, each MT system has three stages but this system which constitutes four stages namely Parts of Speech (POS) tagging, parse tree generation, English parse tree to Bengali parse tree transfer and finally, English to Bangla translation using artificial intelligence.

Texttran-MT system: Department of Computer Science, University of Dhaka, had undertaken an MT project named TEXTTRAN for translation between Bangla and English. Unfortunately, it has been stopped due to some unknown reason.

Google Translator: Google translator facilitates text translation from one to another in 109 different languages including Bangla. It often produces translations that contain significant grammatical errors. For example, some of the nouns like person characters remain frowny while translating in this platform even though its performance for single words is quite satisfactory [9].

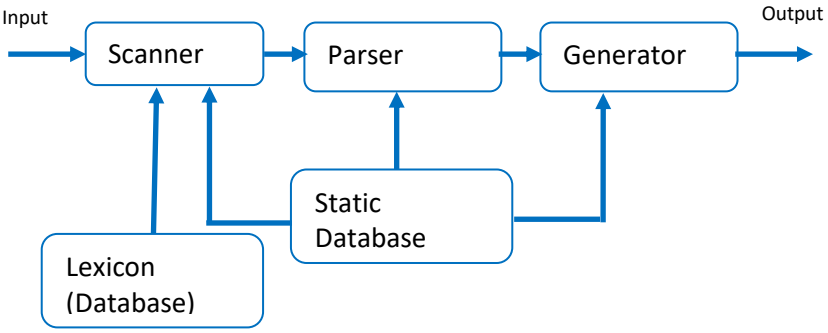

Figure 1: Block Diagram of the Proposed MT System

IV. Proposed Machine Translation System

The proposed model of MT system works in three phases: Scanning, Parsing and Bangla sentence generation. The Scanner reads an English sentence, separates the words of the English sentence and populates the lexical information. After scanning all the words in a sentence, the resulting information is stored into a text file that is fed to the parser as input. The parser parses the input sentence for syntactic correctness using the top-down rule based parsing technique. Finally, the generator finds the Bangla meaning of each English word using the dictionary meaning from the parser output. Then after formatting the Bangla sentence, the output is displayed. The basic architecture of the proposed model of the MT system is depicted in figure 1.

A. The Scanner

The Scanning is the first stage of the proposed MT system. It reads an input sentence, separates the words in the sentence and checks whether the word is in the lexicon as well as every word is correctly spelled. Once the whole sentence is scanned, the next step is to tokenize the sentence into words by eliminating space, comma, dot, and hyphen from the sentence and prepare the words for the next phases.

The Scanner reads a sentence of the English language, separates its words, checks the spelling of each word and populates with lexical information in a file. It uses a lexicon or database for this purpose. The database contains only root words, its parts of speech and other information like person, numbers for nouns and tense for verbs

and so on. When the scanner searches a word in the dictionary, it uses word morphology technique. For example, in the sentence "**They are playing football**", the word playing = play + ing. The scanner searches the word "play" in the dictionary and identifies its validity. After scanning all the words in a sentence, it stores the resulting information into a text file.

i. The Parser

After completing the scanning phase, a rule based top-down parser uses the output of the scanner to check the syntactic correctness for the input sentence. A parser uses several numbers of phrase structure rules(s->Sub + Aux + V + ing +obj etc.). A rule based parser consists of a set of rules which are manually written. The real power of the rule based method is in the capability to learn the rules from a syntactically preprocessed input. The parser consists of several regenerative rules describing the syntax of a sentence. Among all the rules, one is the master rule that defines the sentence and it is called the first level rule. The first level of rules for each kind of sentence (assertive, interrogative, imperative, optative and exclamatory) are defined. Each of these rules have two distinct phases: assertive and negative including subjects and predicates. These are the second level rules. Finally, the third level rules are defined that are used to identify a word, its form and other related information. These rules include noun, pronoun, verb, adverb, adjective, conjunction, preposition and interjection. The first level rules identify the tense, sentence or other information about the input sentence.

During the execution of a rule, it tries to prove the clause from the left to the right. If one clause fails, it tries to prove another alternative clause. This process continues until the clause or the rule is successful. If all the alternative clauses are examined and fail to be proved, it backtracks to its preceding clause or rule in order to prove it with another alternative. If it succeeds, it proceeds to the next clause or rule to prove it. This process continues until the goal and all its sub rules are proved or all the alternatives of all rules and clauses are examined. When all the goals and all its sub rules are proved, the parser produces its output, which consists of subjects, objects, verbs according to the tense and person and tense of the sentence. If all the alternatives are examined and no match is found, the parser gives an error message saying that the given sentence is syntactically incorrect.

Consider the following grammar:

S NP VP
NP ART N
VP (AUX) V (OBJ)
OBJ NP
N boy | football
AUX am | is | are
V play | sleep
ART a | an | the

The top-down parser for the sentence: "The boy is playing football"

S NP VP
ART N VP
the N VP
the boy VP
the boy AUX V OBJ
the boy is V OBJ
the boy is playing OBJ
the boy is playing NP
the boy is playing N
the boy is playing football

b. The Generation of Bangla Sentence

The output of the parser is received by the generator to produce syntactically correct Bangla sentences using the meaning of the words of the input English sentence. This is the final phase of the translation process. It first receives information such as person, tense, form of verb, case of subject etc. from the parser. Then it finds

the meaning of each word from the static database and rearranges these meanings to generate the desired Bangla output sentence. It is notable that the static database contains only root words, their derivatives are formed using parser outputs and morphological techniques as described earlier. The Bangla meaning for each noun or pronoun or adjective or adverb is directly replaced. But for verbs, articles and prepositions, the meaning is generated according to the tense form and form of subjects. The most frequently used morphology is to generate different forms of verbs from the root verb. In order to translate a sentence from English to Bangla, it needs to translate each phrase as a whole and then add appropriate suffixes according to the following process followed by rearranging the words or phrases according to the Bangla grammar. This can easily be done using the tense and person of the subject of the sentence and a suffix table as depicted in Table. 1.

Table no 1: Suffix of Bangla Grammar

Tense/Person	***First***	***Second***	***Third***
1	ই	অ	এ
2	ইতেছি	ইতেছ	ইতেছে
3	ইয়াছি	ইয়াছ	ইয়াছে
4	-	-	-
5	ইয়াছিলাম	ইয়াছিল	ইয়াছিলে
6	ইতাছিলাম	ইতেছিল	ইতেছিলে
7	-	-	-
8	ইবি	ইব	ইবে
9	ইতে থাকবি	ইতে থাকিব	ইতে থাকিবে
10	-	-	-

In the static database, only the meaning of root verb forms is included. To generate the correct form of a verb according to subjects of a given English sentence, the corresponding suffix is added at the end of the meaning of the root verb. For example, for the root verb word "read" the meaning of "read" is stored in static database as "pora (পড়)". For the sentence "I am reading a book", the meaning of verb is concatenating of "পড়" and the suffix of present indefinite tense "ইতেছি". That is "পড়ইতেছি. So, the meaning of the sentence is "ami ekti boi poritesi (আমি একটি বই পড়ইতেছি)". In concatenating a suffix with a root verb, maintaining the word formation rules is required. The vowels in Bangla are used in two forms: as their original symbols and modified symbols with constants. For example, the vowel "আ" is used in this form at the beginning of a word, however when it is added after a consonant the form "আ" is used (e.g., ক + আ = কা). In some case the vowel is implies as পড় (পড় + অ). There are some special cases where the meaning is changed after adding a suffix with the meaning of the root verb. Such as আমি + এর = আমার, সে + এর = তাহার, তুমি + এর = তোমার etc.

V. Results and Discussion

This section provides some sample outputs for some corresponding inputs. The results about the performance of the proposed MT system are tested. Consider the sentence:

"**The boy will eat rice**"

First the scanner tokenizes the sentence into words and check spelling and populates the information of each word. The output of the scanner is as follows:

The	article	definite
boy	noun	third
will	auxverb	future
eat	verb	present
rice	noun	third

From this information, the parser generates the following output:

will	empty	present

Where empty indicates that there is no suffix with a verb. Using this information, the tense rule identifies the tense of the sentence that is.

future indefinite tense

Using the above information, the generator produces the following Bangla sentence:

বালকটি ভাত খাইবে।

The actual output generated by our system for this sentence is given below in figure 2.

Figure 2: Sample output of our MT system for a sentence.

In Google translator, precision for Bangla language is not guaranteed, especially for multiple words. The comparison of the proposed system with the Google Translator is tabulated in Table. 2. This clearly indicates that the accuracy of the translated Bangla sentence is more consistent.

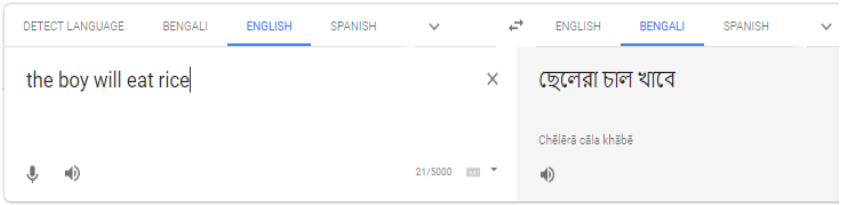

Figure 3: Result of Google translator

Table no 2: Result comparison

Input	***Expected Output***	***Google Translator***	***Proposed System***
He got a loan from bank	সে ব্যাংক থেকে একটি ঋণ পেয়েছে।	তিনি ব্যাংক থেকে একটি ঋণ পেয়েছিলাম।	সে ব্যাংক থেকে একটি ঋণ পেয়েছে।
People have learned from experience	মানুষ অভিজ্ঞতা থেকে শিখেছে।	মানুষ অভিজ্ঞতা থেকে শিখিছে।	মানুষ অভিজ্ঞতা থেকে শিখেছে।

VI. Conclusion

In this paper, an effective methodology for English to Bangla translation using Rule based MT method has been presented. Machine translation is always a challenging job. This technique can give a superior level of effectiveness compared with other approaches over English to Bangla language. For example, the outcome of Google translator shown in Fig. 3 has not got its precision while translation from English to Bangla. Thus, our formulated system gives satisfactorily better or more accurate results. The system uses a rule based top-down parsing approach for English sentences using Left Recursive parsing technique. It incorporates the morphological technique in scanning and parsing phases. Moreover, the system can translate sentences for all the present, past and future (Indefinite, Continuous and Perfect) tenses. Similar strategy can be applied to develop a Bangla to English machine translation system.

References

[1]. Mohamed Amine Cheragui, "Theoretical Overview of Machine Translation", Africcan University, Adrar, Algeria, Icwit 2012.
[2]. Prof. Abdulla H. Homiedan, "Machine Translation" African University, Adrar, Algeria, Proceeding ICWIT, 2012.
[3]. Nayeema Islam, "Implementation of Transfer Based Bangla to English Machine Translation System", Bangladesh, 1998.
[4]. A. Godase and S. Govilkar, "Machine Translation Development for Indian Languages and its Approaches", International Journal on Natural Language Computing, Vol. 4, No. 2, pp-55-74, 2015.
[5]. Mary Priya Sebastian, Sheena Kurian K. and G. Santhosh Kumar: A Framework of Statistical Machine Translator from English to Malayalam. In: Proceedings of Fourth International Conference on Information Processing, Bangalore, India ,2010.

[6]. NIRENBURG, S., CARBONELL, J., TOMITA, M., GOODMAN, K., Machine trans-lation: a knowledge-based approach, Morgan Kaufmann, San Mateo, California, 1992.
[7]. https://www.cdac.in/index.aspx?id=mc_mat_MTS.
[8]. Mukta, Afsana Parveen, Al-Amin Mamun, Chaity Basak, Shamsun Nahar, and Md Faizul Huq Arif. "A Phrase-Based Machine Translation from English to Bangla Using Rule-Based Approach." In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-5. IEEE, 2019.
[9]. https://translate.google.com

Md. Foridul Islam, et. al. "A Rule Based Approach for Implementation of English to Bangla Translator." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(2), 2022, pp. 55-60.