A Phrase-Based Machine Translation from English to Bangla Using Rule-Based Approach

[*a,b]Afsana Parveen Mukta, [a,b]Al-Amin Mamun, [a]Chaity Basak, [a,b]Shamsun Nahar, [a,b]Md. Faizul Huq Arif

[a]*Department of Computer Science and Engineering (CSE), World University of Bangladesh (WUB), Bangladesh*
[b]*Researcher, SenSyss, Bangladesh*
[*]*Email address: afsanacse1206@gmail.com*

***Abstract*—In this paper, a model of transfer architecture has been proposed which represents a Rule-Based Approach. This approach relies on the fuzzy rules. It is a tense and phrase based English to Bangla transfer system. This article represents a knowledge-based technique with a set of data. A rough set technique is used in knowledge representation system for language translation. This technique is used to categorize each English sentence to a particular group using attributes and organized in a pattern. The pattern arranges according to the rules then the system will produce the target sentence Bangla. The whole procedure completes with 6 steps: 1) Collect data 2) Tokenized by word 3) Arrange according to rules 4) Morphological Analyze 5) Reconstruct Bangla sentence using appropriate rule 6) Target sentence. Comparing the experimental result with Google translator, it has been found that the model translation system provides higher accuracy then comparing translator.**

Keywords— Machine Translation, Natural language Processing, Verb Phrase, Noun Phrase, Language Translation.

I. INTRODUCTION

Bangla is a member of Indo-Aryan languages [1], which has come from Sanskrit. Bangla is the state language of Bangladesh where it is spoken as a first language by most of the people. In India, it is permitted provincial language in West Bengal, Tripura and Assam states.

Natural Language Processing is an automatic manipulating system of natural language. People started work on NLP 50 years ago. From the beginning, the programmers found NLP as a complex system. Though there is great scope for research but a limited number of researches have been done on this field. Machine translation system is a part of artificial intelligence.

There are so many approaches which are used for Machine Translation (MT); and rule based approach one of them. Rule based approach is a technique which has been developed first in the field of Machine Translation. This approach is mainly a collection of grammar rules and works in various stages of translation. Parse tree is also used for sentence structure in this approach. Moreover, this paper shows an optimal way by using rule-based Machine Translation (MT) approach from English to Bangla translation to give a better translating system with high accuracy rate. On exploring this paper it is found that some of them used Cockey-Younger-Kasami algorithm for translation where the translation process has occurred through parse tree [2]. Muntarina K. *et al* analyzed all tenses but got 100% success on present indefinite, past indefinite and future indefinite tense [3]. They worked on propositions which are rare concepts [4]. Morphological analysis is implemented with a large number of affixes [5]. Roman characters are used by phonetic mapping [6]. Authors in the article [7] tried to develop a new approach for English to Bangla translation. Francisca J.et al used IF-Then rules for English to Bangla translation [8]. S.A Rahman implements a new NLP Algorithm [9]. Authors in the article [10] proposed a case structure analysis for verb. Here an experiment runs on machine translation system [11].

II. MACHINE TRANSLATION

Machine translation (MT) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another [12]. There are various kinds of machine translation approaches. i) Statistical Machine Translation (SMT) ii) Interlingua approach iii) Corpus-Based approach iv) Example-Based machine translation v) Hybrid Machine Translation approach vi) Rule-Based Approach.

A. Statical Machine Translation (SMT)

SMT models take the view that every sentence in the target language (TL) is a translation of source language (SL) sentence with some probability. SMT systems also presume language and translation models from very large quantities of monolingual and bilingual data using a range of theoretical approaches to probability distribution and estimation [13]. The best translation of the sentence is that which has the highest probability. In SMT there are three major components: language model, a translation model, search algorithm. If target language (a) and source language (b) then we can write,

$$P\left(\frac{a}{b}\right) = P\left(\frac{b}{a}\right) * \frac{P(a)}{P(b)}$$

Where $P(\frac{a}{b})$ depends on the $P(a)$ which is the probability of the kinds of the sentence that are likely to be in the language a. This is known as the language models $P(a)$. The way sentences in b get converted to the sentences b is called translation model $P\left(\frac{a}{b}\right)$.

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

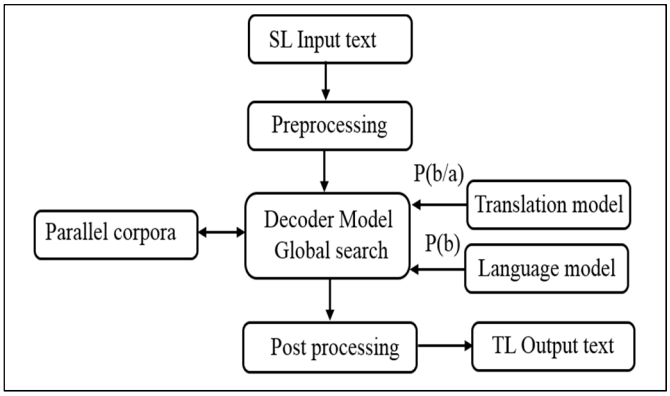

Fig.1. SMT Architecture

B. Interlingua Approach

Interlingua machine translation is the most advanced system. The source text is analyzed in a representation from which the target text is directly generated in Interlingua method. The intermediate representation includes all information necessary for the generation of the target text without 'locking back' to the original text [14]. The Interlingua language creates before creating Interlingua approach. This language shares all the features and makes all the distinctions of all languages. In Interlingua approach, an analyzer is used to put the source language into the Interlingua and convert the Interlingua into the target language using a generator.

Interlingua Approach follows two stages:

1. Extracting the meaning of a source language sentence in a language-independent form.

2. Generating a target language sentence from the meaning

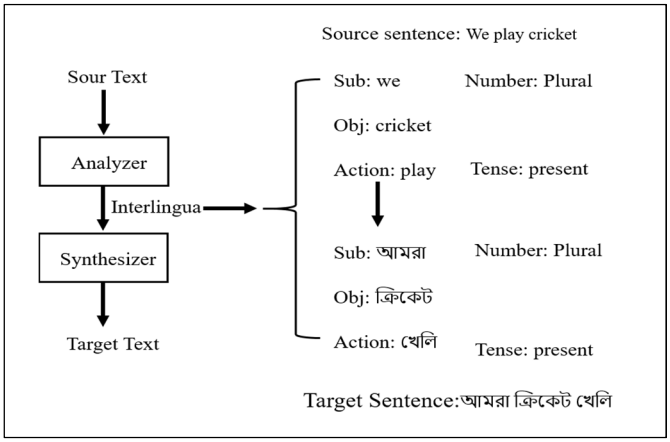

Fig.2:.Interlingua Approach

C. Corpus-Based Approach

In corpus-based MT (CBMT) approach two parallel corpora are available in the source language (SL) and target language (TL) where sentences are aligned. First, it is done by matching fragments against the parallel corpus and then adopting the method to the TL. Finally reassembling these translated fragments appropriately and then translation principle is applied [15]. Fig. 3 shows an example.

Corpus-based Approach has three steps:

1. Matching fragments against the parallel training corpora.
2. Adapting the matched fragments to the target language
3. Recombine these translated fragments appropriately.

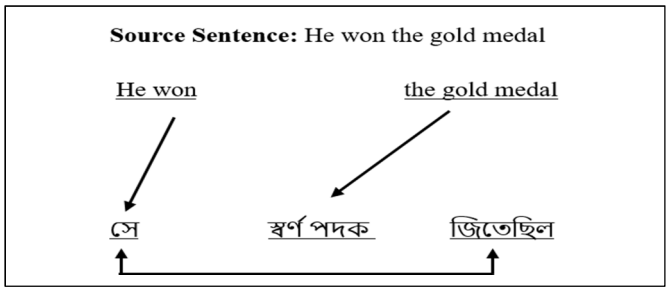

Fig.3. Corpus-Based Approach

D. Example-Based Machine Translation

Example-Based translation is based on recalling similar examples of the language pairs. In 1981 Makoto Nagao first proposed this concept of "Translation by Analogy". The system of Example-Based translation approach can give a set of sentences in the source language. And then using a point to point mapping it translates each sentence in the target language. Memory-Based machine translation is another name of Example-Based Machine Translation. The advantage of an Example-Based system it can train translation program and decodes more quickly. Moreover, this approach works with a small set of data even it will be one sentence pair.

E. Hybrid Machine Translation Approach

Rule-Based translation methodology and statistical translation methodology makes together Hybrid machine translation approach. The hybrid approach to machine translation tries to take the advantages of both frameworks by using available resources. Stat-XFER is such a hybrid machine translation framework, developed to specifically suit machine translation between morphologically-rich and resource-poor language pairs, in this framework, external tools can be provided and used during the process of translation. These include:

a. A Bilingual lexicon, possibly with probabilities per word-pair.
b. A Morphological analyzer of the source language
c. A Morphological disambiguate for the source language
d. A Morphological generator of the target language.

F. Rule-Based Approach

Rule-Based Machine Translation (RBMT) system is formed with a collection of rules. These rules are grammar rules. These rules are made by using a bilingual dictionary and good linguistic knowledge. These rules are processed by the lexicon and software programs. Based on the Chomsky hierarchy inform of computational grammar rules. Basically, these rules consist of an analysis of the source language and generation of the target language in terms of grammar structures. Lexicon provides a dictionary for lookup of words during translation while the software program allows effective

and efficient interaction of the components. The approach depends heavily on language theory hence resource intensive in terms human labor and hours spend when building the rules but easy to maintain, easy to extend to other languages and can deal with varieties of linguistic phenomena.

III. IMPLEMENTED METHOD

This paper is basically focused on English grammar, noun phrase, verb phrase, Bangla grammar Bivokti (Inflection) and also focused on the prepositional phrase. Applying these rules in experiment implements different types of sentence. Like twelve types of tense, three types of phrase, affirmative and also negative sentence. This model input these types of sentence. Then tokenize these sentence by word. For the next step, according to fuzzy rule Bangla translation get from the library. A morphological analyzer analyzing these words for the target sentence. After analyzing Bangla word reconstruct according to the appropriate rules. Finally, the target Bangla sentence is found.

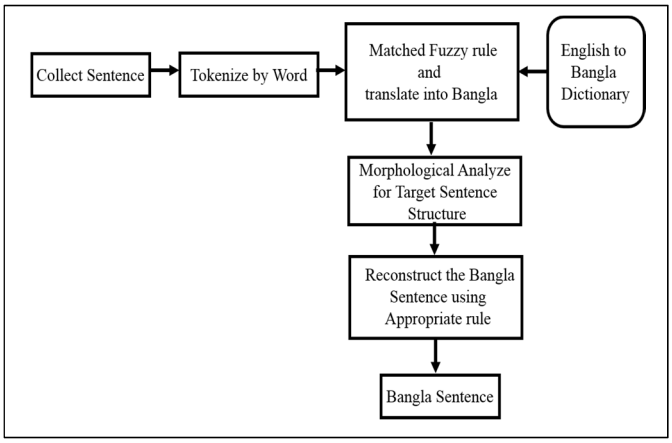

Fig.4. Proposed model for translating English Bangla output sentence

A. Analyzing Grammer for English to Bangla Language

English is a very rich language and English grammar is a large area. Analysis of whole area is a huge task. Analysis has been done for only those things which are indispensable for this language translation, like preposition, noun, verb, and phrase. Proposition and verb are the biggest part of English grammar and also it is indispensable for English grammar. Without a verb, a sentence cannot be identified by the system. A subsequent verb and an auxiliary carries out different meaning in different sentences. In this language translation, it excludes auxiliary verb and joining with the subsequent verb which makes a verb phrase. On the other hand, in Bangla sentence auxiliary verb is not used directly, for an example: "He is playing" here auxiliary verb "is" and the subsequent verb "playing" It will be considered a verb phrase like "is playing". If the auxiliary verb translates into Bangla then it means "হয়" but it is not used in Bangla sentence. That is why the dictionary is developed in this way. In the same way subsequent verb and preposition make prepositional phrase, for example: "I agree with him" here subsequent verb "agree" translate in Bangla as "রাজী" in prepositional phrase "agree with" in Bangla "রাজী হওয়া" In the same way preposition and object make together noun phrase Example: "I am playing in the field" here object "field" in Bangla "মাঠ" but in noun phrase "to field" mean "মাঠে". Here "এ" used as Bivokti. In Bangla grammar, preposition is not used directly. Prepositions are used as Bivokti. If one or two letter used as the suffix after the noun and to make a relation with the other words then it called Bivokti. At the same time, Bivokti is not used in English sentence, different phrase and preposition translate as Bivokti in Bangla sentence. For example: "I am singing at home" in Bangla "আমি বাড়িতে গান গাচ্ছি" here "at" is translated as "তে" and added after object "home" as (বাড়ি + তে = বাড়িতে). Here the preposition is "at". In Bangla grammar, the prepositions "at" translate as Bivokti. The dictionary is built to handle enough auxiliary verb and preposition.

B. English and Bangla Language Structure

Language translation from English to Bangla needs a comparative structure between these two languages. It will be helpful for understanding the major problems of language translation. At first, the sentence structure describes these two languages. With an Example the analysis has been done for the structure, "Subject + Verb + Object". Example – She plays carom (She + plays + carom). Bangla structure: Subject + Object + Verb. Example– {(সে কেরম খেলে) (সে + কেরম + খেলে)}. The main limitation with two language translation is imbalanced sentence structure. And then analyzing grammar to produce a rule for language translation. Because both languages have different grammatical rules. Describe these rules with parse tree with figure 5, 6, and 7.

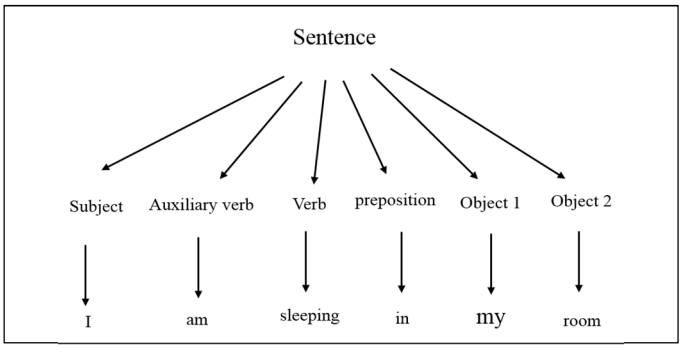

Fig. 5. Parse tree for English sentence

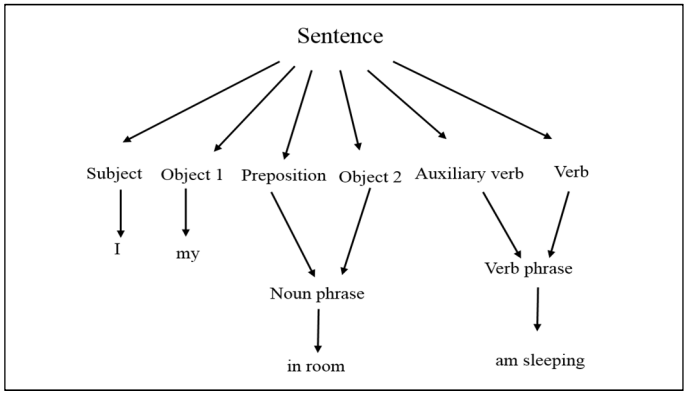

Fig. 6. Parse tree after generating grammatical rule5

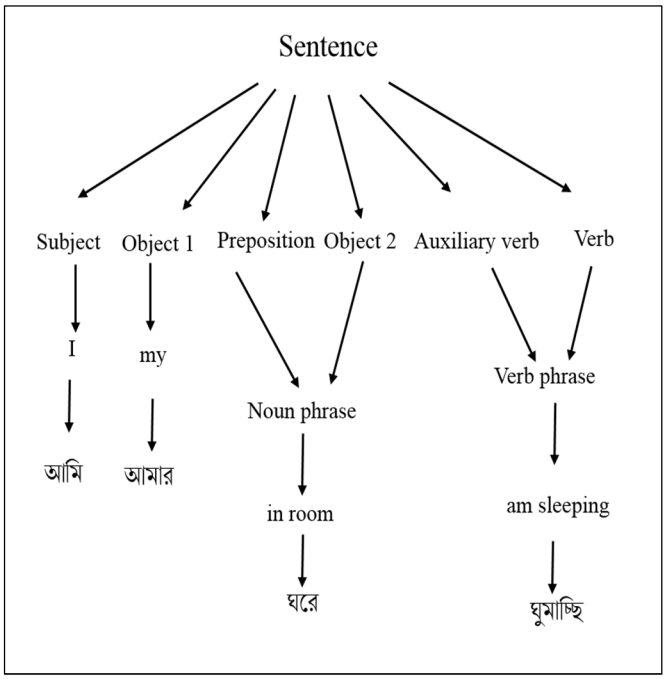

Fig. 7. Parse tree for Bangla sentence

For proper language translation, it is necessary to compare English and Bangla language structure. While making the structure from English to Bangla grammar a morphological analysis is needed. Another challenge of these two languages is to synchronize between Bivokti and preposition. Bivokti use in Bangla grammar and preposition used in English grammar. Bivokti is not existence in English grammar on the other hand preposition is not existence in Bangla grammar. By making verb phrase, noun phrase and prepositional phrase solve this problem. Preposition and noun make together noun phrase, auxiliary verb, and subsequent verb make together make verb phrase, Preposition and subsequent verb make together a prepositional phrase.

C. Sentence comparing with the implemented method and the Google Translator

Comparing with Implemented method and Google Translator. Some difference are showed here a English sentence "I wish peace in the country" accurate Bangla sentence is "আমি দেশে শান্তি কামনা করি" implemented method gives "আমি দেশে শান্তি কামনা করি" and Google Translator gives "**আমি দেশের শান্তি চান**". Another English sentence is "I am sleeping in my room" accurate Bangla sentence is "আমি আমার রুমে ঘুমাচ্ছি" implemented method gives"আমি আমার রুমে ঘুমাচ্ছি" and Google Translator gives "আমি আমার রুমে **ঘুমাচ্ছে".** Here English sentence is "I am ashamed of his conduct" accurate Bangla sentence is "আমি তার আচরণের জন্য লজ্জিত" implemented method gives "আমি তার আচরণের জন্য লজ্জিত" Google Translator gives "আমি তার আচরণের জন্য লজ্জিত". In these three sentences Google Translator gives two incorrect output and implemented method gives three correct output. In this way comparing some sentences with implemented method and Google Translator. It shown that implemented method gives high accuracy more than Google Translator.

IV. Experimental Result

The proposed method finds the accuracy rate compares between two files: one is the original file and the other is implemented output file. First, the program counts sentence and word number from the original file. Then compare sentence by sentence and word by word. If it finds any word mismatch then counts word mismatch and if it finds any sentence mismatch then count sentence mismatch. Finally, the proposed method counts sentence exactness and word exactness rate.

$$SE = \left\{\frac{(TS - MS)}{TS}\right\} * 100\%$$

$$WE = \{\frac{(TW - MW)}{TW}\} * 100\%$$

Here,

SE= Sentence Exactness rate
TS= Total Sentence
MS= Mismatch Sentence
WE= Word Exactness rate
TW= Total Word
MW= Mismatch Word

In above equation total 1113 sentences and 5967 words are applied on Rule-Based Approach. Implemented method and Google translate find different accuracy rates.

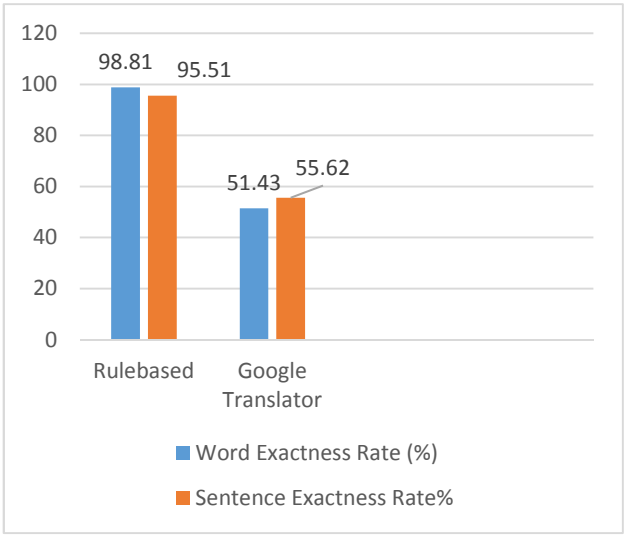

Fig. 8. Exactness rate with compare to Google Translator

By analyzing the Fig. 8, we can see that the sentence accuracy rate and word accuracy rate of Rule-Based approach is higher than Google translator. The rule-based approach shows the higher accuracy rate.

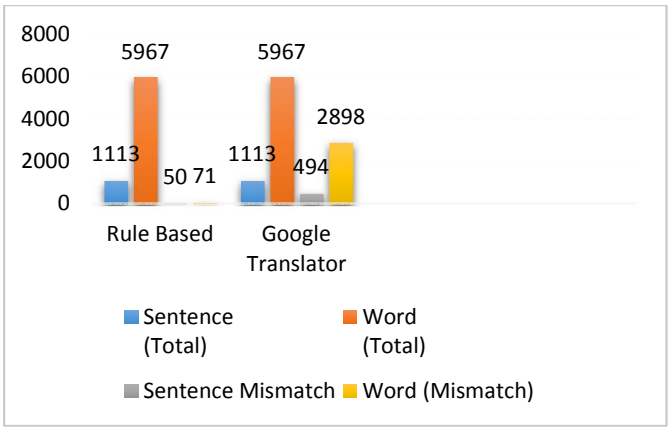

Fig. 9. Word and Sentence count with compare to Google translator

In Fig. 9, there are 1113 sentences and 5967 words are applied. In the Rule-Based approach found 50 mismatch sentence from 1113 sentences and 71 mismatch words from 5967 words. As opposed to Google translator provide 494 mismatch sentence from 1112 sentences and 2898 mismatch words from 5967 words. The rule-based approach shows the lowest number of sentences and words mismatch. After seeing all the evidence it proved that the rule-based approach shows more accuracy than Google translator.

A. Comparison with Related Works

By massive analyzing it found that many types of research had been done by using other approaches but a little work had been done in the Rule-Based Approach from English to Bangla translation. Comparing with others research this paper will provide a high accuracy rate and rich dictionary. This paper also gives a better result on the verb phrase, noun phrase, and preposition phrase.

B. Why is the Rule-Based Approach is Best?

In the rule-based system, grammatical rules are used. Rule-Based Approach can deal with a huge amount of data which is almost difficult with other approaches. It is a trained system that is why it can make decisions much faster without wasting time in calculating the result. In addition, more rules will not make any difficulty for the system. This system creates all possible meaning. The rule-based system picks only the effective one and the bangle meaning change according to the grammar rules. And also can give the proper result on verb phase and noun phase and also preposition phase.

V. CONCLUSION

There are so many methods for Machine Translation but this paper was implemented a Rule-Based Machine Translation system which can translate from English Sentences to Bangla sentences using some sentences pattern. This paper focused on twelve tenses, verb phrase, noun phrase, preposition phrase, affirmative and negative sentences. The implemented method gives the sentence accuracy rate 77.6% and word accuracy rate of 80.88%. This Method is also compared with Google Translator. Rule-Based Machine Translation System is able to use the wisdom of source language, destination language, and grammatical rules. That is why the Rule-Based approach gives the best result comparing to other approaches. The rest two types of phrase and also idioms will be implemented in the future. This system also works on multiple languages.

REFERENCES

[1] "Bengali language Britannica.com." [Online]. Available: https://www.britannica.com/topic/Bengali-language. [Accessed: 31-Jul-2018].

[2] S. Dasgupta, A. Wasif, and S. Azam, "An optimal way of machine translation from English to Bengali," in *Proc. 7th International Conference on Computer and Information (ICCIT)*, 2004, pp. 648–653.

[3] K. Muntarina, M. G. Moazzam, and M. A.-A. Bhuiyan, "Tense Based English to Bangla Translation Using MT System," *International Journal of Engineering Science Invention*, 2013.

[4] S. K. Naskar and S. Bandyopadhyay, "Handling of prepositions in English to Bengali machine translation," in *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, 2006, pp. 89–94.

[5] N. K. Zaman, M. A. Razzaque, and A. A. Talukder, "Morphological Analysis for English to Bangla Machine Aided Translation," in *National Conference on Computer Processing of Bangla, Dhaka, Bangladesh*, 2004.

[6] N. UzZaman, A. Zaheen, and M. Khan, "A comprensive roman (english)-to-bangla transliteration scheme," 2006.

[7] S. Ahmed, M. O. Rahman, S. R. Pir, M. A. Mottalib, and Md. S. Islam. 2003, "A New Approach towards the Development of English to Bengali Machine Translation System", *International Conference on Computer Information and Technology (ICCIT).*

[8] J. Francisca, Md Mamun Mia, Dr. S. M. Monzurur Rahman. 2011, "Adapting Rule B sed Machine Translation from English to Bangla", *Indian Journal of Computer Science and Engineering (IJCSE).*

[9] S. A. Rahman, K. S. Mahmud, B. Roy, and K. M. A. Hasan. 2003, "English to Bengali Translation Using A New Natural Language Processing Algorithm," *in International Conference on Computer Information and Technology (ICCIT).*

[10] M. K. Rhaman and N. Tarannum, "A rule based approach for implementation of bangla to english translation," in *Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on*, 2012, pp. 13–18.

[11] M. M. Asaduzzaman and M. M. Ali, Transfer Machine Translation, "An Experience with Bangla English Machine Translation System", *In the Proceedings of the International Conference on Computer and Information Technology 2003.*

[12] "Science or Fiction: Machine Translation Explained | Blog | Ciklopea." [Online]. Available: https://ciklopea.com/translation/translation-technology/science-or-fiction-machine-translation-explained/. [Accessed: 03-Oct-2018].

[13] A.Way and N. Gough, "Comparing example-based and statistical machine translation," *Natural Language Engineering*, vol. 11, no. 3, pp. 295–309, 2005.

[14] W. J. Hutchins and H. L. Somers, *An introduction to machine translation*, vol. 362. Academic Press London, 1992.

[15] S. Nahar, M. N. Huda, M. Nur-E-Arefin, and M. M. Rahman, "Evaluation of machine translation approaches to translate English to Bengali," in Computer and Information Technology (ICCIT), 2017 20th International Conference of, 2017, pp. 1–5.