Recent Progress, Emerging Techniques, and Future Research Prospects of Bangla Machine Translation: A Systematic Review

M. A. H. Akhand[1], Arna Roy[2], Argha Chandra Dhar[3], Md Abdus Samad Kamal[4]*

Dept. of Computer Science and Engineering, Khulna University of Engineering & Technology Khulna-9203, Bangladesh[1, 2, 3]
Graduate School of Science and Technology, Gunma University, Kiryu 376-8515, Japan[4]

Abstract—**Machine Translation (MT), the way of translating texts or documents from a source language to a target language automatically without human intervention, has gained popularity in the growing information technology-based era of globalization. Bangla is a major language, and several MT studies with different tools and techniques have been investigated in the last two decades. Considering the importance of the Bangla language and its prospects in MT studies, this study provides a comprehensive review of existing Bangla MT studies to meet the timely demand. Specifically, at first, the basic ideas of different MT methods (Rule-based, Example-based, Statistical, Neural, and Hybrid) and performance measures of MT are presented as a background study of the present review. Then an overview of the Bangla language and a brief description of the available Bangla-English corpora are provided. Next, a description of the existing Bangla MT studies is provided categorically following the common strategic fashion to create a valuable reference for current researchers in the field that is also suitable for non-expert users. The achieved performances of individual methods are also compared in a tabular form. Finally, a number of future research prospects are revealed from the studies, encouraging researchers and practitioners to develop a better and comprehensive Bangla MT system.**

Keywords—Machine Translation (MT); Bangla language; rule-based MT; example-based MT; statistical MT; neural MT; hybrid MT

I. INTRODUCTION

The task of content (e.g., voice, speech, texts) translation from one natural language to another has become indispensable in politics, business, research, and other areas, and a human expert usually handles such tasks. Human translators perform a masterful job interpreting conversations between two or more persons (e.g., country chiefs, tourists, business giants) who speak in different languages. Due to rapid globalization, translation becomes essential for ordinary people; translation of web content (e.g., website, document) is also necessary for the era of digitalization and the internet. To handle the translation of such huge contents (especially text, document, and web), machine translation (MT) is a promising research area [1]. In general, MT refers to translating texts or documents from the source language (SL) into the target language (TL) without human intervention.

Individual natural languages advocacy inherently dominated respective MT studies as resources (e.g., corpus)

and language-dependent efforts. Specifically, corpus, rules, and other resources of a particular language pair (e.g., English-German) are not usable in MT for another language pair. Moreover, the MT system developed for a particular language pair is not appropriate for other cases as an individual language holds distinct grammar and phrase rules. As the most internationally used language, MT researches are mainly English language concentric. High resource availability and major MT studies with remarkable performance are available for English-French [2], English-German [3], English-Chinese [4] language pairs. Thousands of other natural languages, including several major languages, are remained much behind in MT activities. Specifically, Bangla is one of the most broadly spoken languages, with approximately 228 million native speakers (fifth-most) and 37 million second-language speakers (seventh-most) [5]. However, MT resources and studies are limited for Bangla. Several Bangla-English MT studies are available with different methods, but their achievements are not significant compared to the resource-rich language [6]–[8]. Therefore, it is a timely demand to line up existing Bangla MT studies for the researchers who intend to work in this promising research field and find a motivation to enhance Bangla MT.

This study is a comprehensive review of Bangla MT studies focusing on individual methods and techniques employed, corpus and/or resources used, and performance achieved. As a prerequisite, the fundamentals of various MT methods, the significance of the Bangla language, and MT performance measurements are explained briefly in Section II. Benchmark corpora and resources for Bangla MT; and individual Bangla MT studies are summarized under different MT categories in Section III. The significance of the current review study and prospects of Bangla MT studies are discussed in Section IV and Section V, respectively. Finally, Section VI briefly concludes the present study with few remarks.

II. BACKGROUND

Various techniques and tools were developed in the last few decades through remarkable research efforts for appropriate MT outcomes in different languages. At the basic level, MT executes the translation of atomic words from one language to another using a dictionary. Nowadays, it is possible to translate whole sentences through corpus techniques, rule-based grammatical techniques, or even idioms

*Corresponding Author

and phrase-based techniques. However, no MT system is available currently that can translate as efficiently as a human translator. Therefore, MT has emerged as a rising research field in Artificial Intelligence.

A. Basic MT Approaches

Existing MT methods are broadly categorized into two approaches: rule-based MT (RBMT) approach and data-driven approach. In the data-driven approach, a parallel corpus is the main element to develop the MT model. Three basic methods in this category are example-based MT (EBMT), statistical MT (SMT), and neural MT (NMT). On the other hand, the hybrid MT (HMT) approach is also available, which combines two or more basic methods. Fig. 1 depicts the classification of the available MT approaches. The following subsections briefly discuss the fundamental points of the five basic MT methods to understand different Bangla MT studies easily.

1) Rule-Based MT (RBMT): Based on linguistic information, RBMT generates translations through human expert-produced grammatical rules regarding verbs, prepositions, inflections, etc. [9]. Dictionaries (unilingual, bilingual or multilingual) and collection of rules covering the main semantic, morphological, and syntactic regularities of source and target languages are the basic requirements of RBMT [10]. Roughly RBMT can be divided into three approaches: Direct MT (DMT), Transfer-based MT, and Interlingua MT [11]. Fig. 2 is the well-known Bernerd Vauquois' pyramid of MT, which shows comparative depths of intermediary representation, interlingua MT at the peak, followed by the transfer-based, then direct translation.
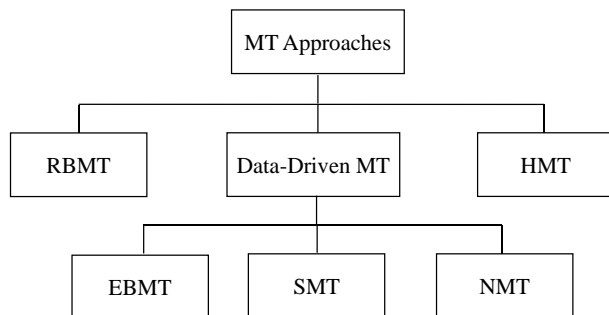

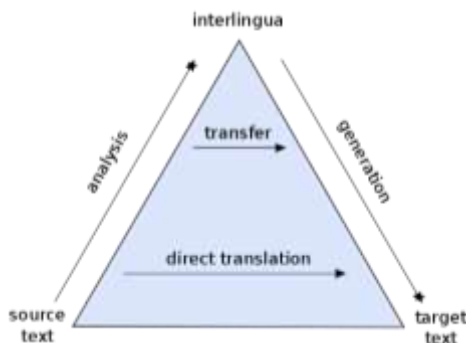Fig. 1.   Classification of MT Methods.

Fig. 2.   Bernerd Vauquois' Pyramid Showing Comparative Depths of Intermediary Representation, Interlingua Machine Translation at the Peak, followed by Transfer-based, then Direct Translation.

DMT is the oldest MT approach based on the dictionary that is used in the pioneer Georgetown–IBM public MT demonstration [12]. DMT attempts to match an SL to a TL, i.e., translating word-by-word directly [13]. The method is quite simple, but the translation quality is very poor due to the lack of syntax and semantic analysis of the source language. Then, the RBMT approach with transfer and interlingua [13] is developed to overcome the limitations of DMT. The interlingua approach was proposed to be language-independent [14]. Fig. 3 shows the block diagram of RBMT with interlingua representation. Interlingua is considered an abstract, homogenous, unambiguous, and independent universal language. For translating using interlingua, the source sentence is converted to the interlingua first, and then the interlingua is converted to the target language sentence [9].

2) Example-Based Machine Translation (EBMT): EBMT is a corpus-based data-driven approach based on human language learning process [15]. The main motivation of EBMT is that human does not translate through deep linguistic analysis. Instead, a human translator first properly decomposes input sentences into specific fragmental phrases, then translates these fragmental phrases into other language phrases, and finally correctly composes these fragmental translations into one long sentence. EBMT was introduced for the English-Japanese language pair as RBMT is complicated for English-Japanese and other language pairs due to structural differences [1].

Fig. 4 shows the basic building block of the EBMT model. Sample sentences from SL and TL are stored as examples in a bilingual corpus (i.e., dictionary), a significant component of this model. The SL sentence is fragmented depending on the granularity of the system and followed by a search for (set of) examples from the dictionary that match (or closely matches) the input SL fragment string, and the relevant fragments are picked. The TL fragments corresponding to the relevant fragments are extracted. If the match is exact, the fragments are recombined to form TL output; else, find the TL portion of the relevant match corresponds to a specific portion in SL and align them. Finally, a combination of relevant TL fragments is performed in order to form a legal grammatical target sentence. Further action to translate the untranslated portions (if happen) using a dictionary (called translation memory) has been investigated recently to improve EBMT performance [16].

3) Statistical Machine Translation (SMT): SMT is proposed presuming that language has an inherent logic that might be helpful to treat language mathematically. In SMT, translations are produced based on probability generated through the statistical analysis of bilingual aligned corpora [17]. SMT does not need much knowledge of the SL and TL like RBMT. Fig. 5 shows a simplified block diagram of SMT using decoder, translation model (TM), and language model (LM). The probabilistic TM assigns a score to every possible translation of source text. The language model measures the fluency of the output and assigns each sentence a probability. In the decoding phase, the translation with the best score is

selected [18]. SMT is a corpus-dependent approach, and the requirement of a large human-translated corpus with various linguistic information is its main drawback.

4) Neural Machine Translation (NMT): NMT is the most recent MT technique based on machine learning with a special neural network (NN) framework called Encoder-Decoder architecture. Fig. 6 shows the basic structure of the NMT model. NMT uses vector representations for words and sequence-model of the input sentence to generate TL words sequentially with encoders and decoders in the core [19]. The input words are first encoded in a one-hot vector and passed through an embedded matrix and hidden layers. In the output layer, the decoder outcome is interpreted as a probability distribution. A softmax activation function is used to ensure proper probability distribution [20]. NMT is a data-driven approach where a NN model is trained with a parallel corpus of SL and TL.

NMT has emerged as a hopeful field in the MT system for showing better performance than other MT systems with different NN models. Early NMT models used a feed-forward NN to develop an MT model, which could not provide sufficiently good results [21]. In the recent NMT studies, different deep learning models [22], such as Recurrent NN (RNN) [23], convolutional NN (CNN) [24], multilayered long short-term memory (LSTM), are is used for encoding and decoding purposes [25]. The recently developed transformer model with many encoder-decoder layers is shown better translation performance [26].

Different techniques are also investigated to improve NMT performance. Normally, NMT uses a parallel corpus of SL and TL for training. However, Sennrich et al. [27] have investigated the use of monolingual data effectively applying back-translation. Back-translation is translating back to SL from TL, and it is a way to train the NMT model for better translation quality [28]. However, NMT has still shown lower performance for low resource words and in word alignment [29].

5) Hybrid Machine Translation (HMT): The aim of HMT methods is to achieve better MT performance overcoming distinct constraints of individual MT methods while integrating individual ones. Fig. 7 shows the building block of the HMT system, which may contain several individual MT models for translation. RBMT and EBMT were brought into HMT by pioneer researchers [30]. Bond and Shirai [31] developed an HMT system that uses the EBMT method but allows to use of RBMT where required. Schwenk et al. [32] used a Statistical Post Editing (SPE) system where SMT is used to correct the errors of RBMT. Several recent HMT methods also performed well for MT in different language pairs. Huang et al. [33] developed an HMT by combining NMT and RBMT: the system used the consistency of RBMT to balance the inadequacy of datasets for the NMT system. Banik et al. [34] proposed an HMT system with NMT and SMT for English-Hindi language pairs. Singh et al. [35] used NMT and RBMT to build an HMT system for Sanskrit to Hindi. Beyala et al. [36] tuned the transformer model's output with phrase-based SMT.
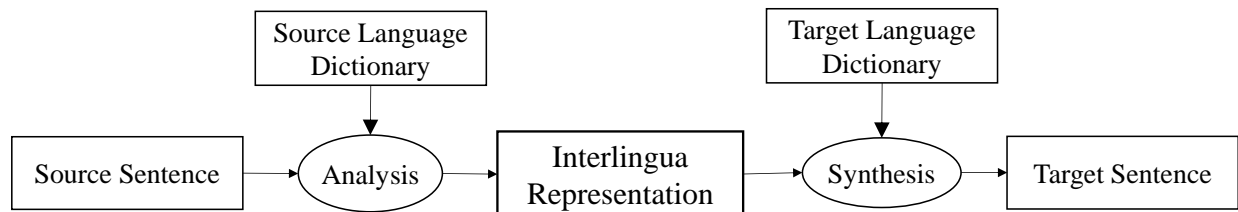
Fig. 3.   Rule-based MT (RBMT) with Interlingua Representation.
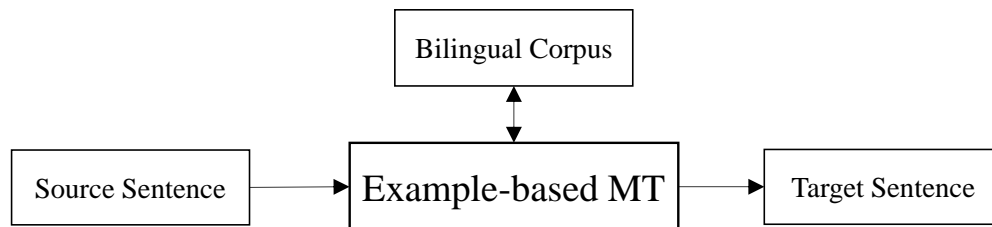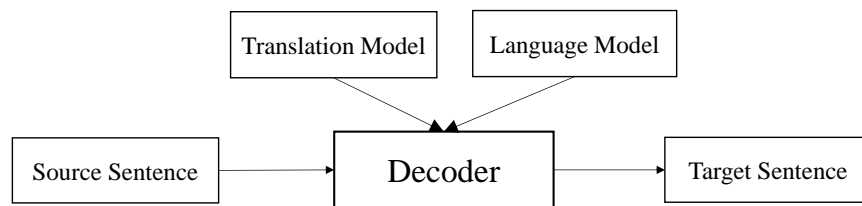
Fig. 4.   Basic Example-based MT (EBMT) Method.

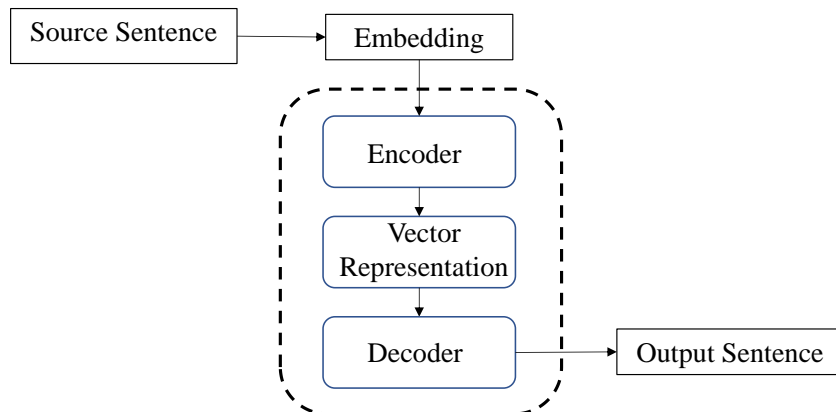Fig. 5.   Basic Statistical MT (SMT) Method.
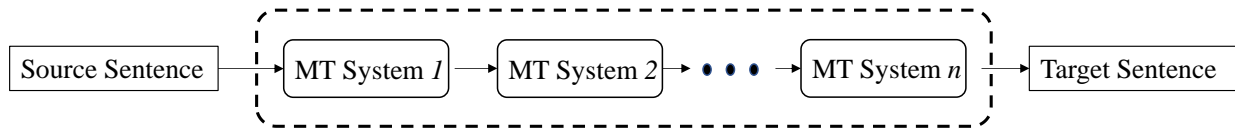
Fig. 6.   Basic Neural MT (NMT) Method.

Fig. 7.   Basic Hybrid MT (HMT) Method.

B. Bangla Language and Its Significance

Bangla, belong to the Indo-European language family, is a major language in the Indian subcontinent and the main language of Bangladesh; and Bangla is bound by different kinds of languages like Oriya, Assamese, etc. The language has come modern phase through a metamorphosis as the territory was under the rule of various administrations [37] for a long time. The basic sentence structure for the Bangla language is *subject + object + verb* in general. As an example, 'আমি সকালে ঘুম থেকে উঠি' for 'I wake up in the morning'. Whereas, English is in the West-Germanic language family [38]. The basic structure of English is *subject + verb + object*; as an example: I wake up in the morning. In the case of adverb, in Bangla sentences, an adverb comes before the verb like "সে আস্তে দৌড়ায়", which can be translated in English "He runs slowly" where adverbs usually come after verb. In Bangla, both masculine and feminine gender share the same form of pronoun like "সে/তিনি" whereas in English the third person singular number pronoun differs in terms of gender such as he/she, him/her, etc.

The form of verb differs in terms of space, time, and person in Bangla language; examples are "তুই এখান থেকে যা", " আপনি এখান থেকে যান", and "তুমি এখান থেকে যাও". These three sentences are translated as "You get out of here" in English. Sometimes the exact meaning of the English word is not used. For an example, "We are playing in the field" which means in Bangla "আমরা মাঠে খেলছি". Here "are" means "hoy/hoi" which is not used in Bangla. So "are playing" is sometimes considered as a verb phrase. Moreover, instead of the preposition, Bivokti (i.e., inflection) is used, mainly joining a letter with a word to relate with other words. In the previous example, "in the field" means "Maathe" in Bangla. "Field" means "Maath" but the word "in", which is a preposition in English, is considered as "e" in Bangla. Bangla has various kinds of inflection in sentences and varies in phonology, also depending on regions.

C. MT Performance Measurement

Performance measurements in the MT system play an indispensable role in determining the efficacy of the existing system and the requirement of optimization. Regarding MT system evaluation, human evaluation and several other matrices are available [39].

Human evaluation is considered as a baseline for MT evaluation. Adequacy and fluency are the most common methodologies of human evaluation, which are measured on each sentence in the output, allotting points from one to five according to translation quality [40]. Adequacy refers to how much meaning and information have been manifested in the source and target languages. It needs the judge to be bilingual in both source and target languages. Fluency indicates how fluent the translation is, and the judge needs to be fluent in the target language. Human evaluation is a very cumbersome and time-consuming process to judge translation quality sentence by sentence. Therefore, automatic evaluation matric is used nowadays instead of human evaluation, and several such methods are briefly described below.

Bilingual Evaluation Understudy (BLEU) is currently the most popular automatic evaluation metric for MT. BLEU considers multiple references, each of which may use a different word choice to translate the same source word. The base of the BLEU metric is a precision measure [41]. At first, a modified n-gram is calculated by counting the number of n-grams or word sequences in the candidate sentences (i.e., system output) alongside the reference sentences. Then the candidate counts are clipped by their corresponding reference maximum value. These clipped n-grams are then summed and divided by the total number of candidate n-grams [41]. Through this step, the modified precision score (p_n) is calculated.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')} \qquad (1)$$

This result is multiplied by an exponential brevity penalty factor where a high-scoring candidate translation must now match the reference translations in length, word choice, and word order. Therefore, the next step is to calculate BLEU Brevity Penalty (BP) factor.

$$BP = \begin{cases} 1, if\ c > r \\ e^{1-\frac{r}{c}}, if\ c \le r \end{cases}, \qquad (2)$$

here c is the length of candidate translation, and r is the length of reference translation. Finally, the BLEU score is the geometric mean of the precision scores and is calculated using Eq. (3).

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n\ log\ p_n) \qquad (3)$$

The MT score NIST comes from the name National Institute of Standards and Technology, and it is an improved version of BLEU. Where BLEU counts all the n-grams equally, NIST takes into account the informativeness of n-grams on the basis of frequency of occurrence [42]. Besides, NIST uses the arithmetic mean of n-gram counts, but BLEU uses the geometric mean of n-gram count. It also tries to minimize the unwanted effects of the brevity penalty factor by BLEU [42], [43]. However, at first, the information weights are calculated by n-grams counts over a set of reference translations to calculate the NIST score.

$$Info(w_1 \dots w_n) = \log_2\left(\frac{the\ number\ of\ occurences\ of\ w_1 \dots w_{n-1}}{the\ number\ of\ occurences\ of\ w_1 \dots w_n}\right) \qquad (4)$$

Finally, the NIST score is calculated by Eq. (5).

$$NIST = \sum_{n=1}^{N}\left\{\frac{\sum_{all\ w_1 \dots w_2\ that\ co-occur}\ Info(w_1 \dots w_n)}{\sum_{all\ w_1 \dots w_n\ in\ system\ output}\ 1}\right\} * exp\left\{\beta \log^2\left[min\left(\frac{L_{sys}}{\overline{L_{ref}}}, 1\right)\right]\right\}, \qquad (5)$$

where β is chosen to make the brevity penalty factor = 0.5 when the number of words in the system output is two-thirds of the average number of words in the reference translation.

N=5 means this formula works with five words at a time.

$\overline{L_{ref}}$ denotes the average number of words in a reference translation; averaged over all the reference translations.

L_{sys} denotes the number of words in the translation which have been scored.

Translation Error Rate (TER) is defined as the minimum number of edits required to change a system output to reference translation [44]. It was designed to reduce the human effort for evaluating an MT method [45]. The general equation is given below:

$$TER = \frac{Number\ of\ edits}{Average\ number\ of\ reference\ words} \qquad (6)$$

here possible edits can be counted as insertion, deletion, and substitution of single words as well as shifts of word sequences.

III. REVIEW OF BANGLA MT SYSTEMS

Several MT systems developed on the Bangla language in the last two decades. The available Bangla studies are Bangla to English (B2E) or English to Bangla (E2B) with different Bangla-English corpora. A few studies are for both B2E and E2B. Bangla-English corpus is an important element to Bangla MT, and therefore, an overview of Bangla corpora is given first. Then existing Bangla MT systems are described briefly in different MT categories.

A. Bangla Corpus

Several Bangla-English parallel corpora are prepared by different research groups and are publicly available for anyone to use. Table I summarizes prominent Bangla-English corpora mentioning significant attributes of individuals. The corpora are varied in sample sizes. The largest corpus is the Indic Languages Multilingual Parallel corpus (ILMPC) consists of 338500 sentences. On the other hand, the small-sized corpora Penn Treebank (PTB) and AmaderCAT consist of 1313 and 1782 sentences, respectively. In several cases, the available sentences are partitioned into training, validation, and test sets. The training set is to train a model, and the test set is dedicated to the final evaluation of the trained model. The validation set samples may use to evaluate the intermediate performance of a model during training. The last column of Table I referred to several studies that used a particular corpus. Based on recent studies, SUPara corpus [46] is the most popular. The corpus holds quite clean 71861sentences having 244539 words in English and 202866 words in Bengali [46].

B. Review of Bangla RBMT Methods

Using RBMT, based on linguistic information and rule production, diverse techniques for B2E and E2B MT have been investigated for rules generation, including fuzzy rules [47], [48], context-sensitive grammar (CSG) rules [49]–[51], etc. Under the umbrella of RBMT, Rahman et al. [52] utilized morphological analysis in finding the root words from the input Bangla sentences. After matching the Bangla grammar, corresponding English grammar is identified; the input sentence is then rearranged according to it. The final output is the English translation of the corresponding Bangla words with the help of a dictionary. They considered only a few types of sentences. The method seems quite efficient but needs a lot of knowledge about both languages and engagement of the dictionary. Chowdhury [53] projected a system where Bangla sentences are read from left to right, and corresponding English words are generated using a dictionary and the context of the Bangla sentence. In addition to word generation, a set of grammatical rules are used to analyze the source sentence properly.

TABLE I.   BENCHMARK BANGLA-ENGLISH CORPORA SUMMARY

Sl.	Corpus Name [Ref.]	Sample Size (Training/Val./ Test Set)	Corpus Data Link	Significance	Study / Works with the Corpus
1	Enabling Minority Language Engineering (EMILLE) [97]	26287 (25287/500/500)	http://catalog.elra.info/en-us/repository/browse/ELRA-W0037/	-	[78]
2	KDE4 [98]	35365 (33365/1000/1000)	https://opus.nlpl.eu/KDE4-v2.php	i) Currently contains words of 60 different languages ii) Already sentence aligned	[78]
3	SUPara [46]	71861 (70861/500/500)	https://ieee-dataport.org/documents/supara08m-balanced-english-bangla-parallel-corpus	i) First free English-Bangla Parallel corpus ii) Balanced and comprehensible	[7], [74], [76], [77], [82], [83]
4	Global Voices [98]	1031725	https://opus.nlpl.eu/GlobalVoices-v2018q4.php	Contains non-printable characters (e.g., Arabic)	[61], [74], [77]
5	Indic Languages Multilingual Parallel Corpus (ILMPC) [99]	338500 (337000/500/1000)	http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.2020.html	Consists of 7 parallel languages	[7], [76]
6	Six Indian Parallel Corpora (SIPC) [100]	219140 (20000/914/1000)	https://github.com/joshua-decoder/indian-parallel-corpora	i) Consists of 6 languages ii) Sentences are collected from internet documents	[7], [76]
7	Penn Treebank Bangla-English (PTB) [76]	1313	https://panl10n.net/ [Original source link, not accessible]	Multilingual parallel corpus	[7], [76]
8	AmaderCAT [101]	1782	https://github.com/AridHasan/Data-Collection-System-for-Machine-Translation/tree/master/data	i) A collaborative platform ii) Sentences are collected from newspapers	[7]
9	Linguistic Data Consortium [102]	12600 (11000/600/1000)	https://www.ldc.upenn.edu/	-	[70], [72]

Lexical analysis is important in RBMT as the attributes related to sentences in English and Bangla can be known through it, which are important in the next phases. The customized process can be used in the phase of rule generation. Alam et al. [54] used a bilingual lexicon that stores information and helps place words and error checking. After semantic analysis, they have categorized the sentences according to the subject, object, and verb, leading to the generation of Bangla sentences. However, it cannot identify gerund (i.e., –ing form of a verb), more than one subject or object, and several grammatical issues. Francisca et al. [47] investigated an RBMT process that accepts an English sentence as input where the lexical analyzer is used to generate the class of the sentence by utilizing the information of the word from the dictionary. The generalization of sentences is used to find outmatched fuzzy rules for English sentences. The rules may be matched partially or fully. Later, the dictionary is used to find the corresponding Bangla words, leading to the next step to reconstruct the Bangla sentence, depending on the related rules for Bangla sentences. The process seems compelling, but they have not covered all kinds of sentences. Mukta et al. [48] proposed a model similar to [47], but it is based on tense and phrase. This system emphasizes English grammar, verbs, prepositions, inflection, and other grammatical rules of Bangla. English word translation to Bangla takes place with the help of a dictionary, and a morphological analyzer analyzes these words for the target language. After the reconstruction of Bangla sentences by proper production rules, the system delivers the output. Prepositions (e.g., to, in, etc.) do not have any definite meaning in Bangla, and the auxiliary verbs (e.g., are, is) have some meaning but are implicitly used in Bangla. Therefore,

they have assumed preposition and object in one phrase and an auxiliary verb and main verb in one phrase for simplicity. The system seems to work well in comparison with the Google translator.

Anwar et al. [55] used CSG rules in their B2E RBMT system, where after tokenizing, a token is searched in a lexicon, and if found, it is matched by the Bangla grammatical rules. The token is tagged by appropriate parts of speech if matching is found. After that, a parser is used to generate a parse tree for the input string. Finally, the corresponding English sentence is generated by the NLP conversion unit through the help of a corpus. They used the basic bi-gram model as the language model and created basic English sentences by replacing the Bangla words with English words from the lexicon. Using 28 basic production rules, much importance is given to the parts of speech, including simple, complex, and compound sentences. The system shows a remarkable accuracy (over 90 percent) while considered limited sentence types. Muntarina et al. [50] proposed their strategy based on tense-based rules using parse trees. They constructed a parse tree for input English language. Then it was converted into a Bangla parse tree based on production rules for both languages generated by syntactic and morphological analysis. The system uses the NLP conversion technique for conversion and lexicon, which helps simplify the knowledge on both languages and provides the target words for input text. They have considered input and output in the form of tense.

Arefin et al. [56] have designed an MT system that has given much importance to assertive, interrogative, and imperative sentences. They proposed a unique method named

"Transfer" where the conversion from Bangla parse tree to the English parse tree is organized. The system drives the parsing method based on 22 context-free grammar (CFG) rules for Bangla. The transfer method has a separate algorithm that uses 24 CFG rules to generate an English parse tree. The processing of the Bangla parse tree starts from the low-level nodes and goes up to the top level by analyzing. The process continues by creating a subtree of English sentences based on checking the grammatical rule generator module. This system showed higher accuracy than Google translator though the data set is small. Alamgir et al. [51] have depicted a model using the same CFG rule and generation of the parse tree for both languages. 31 CFG rules have been included in a table for Bangla parse tree and another 31 rules have been used to construct the English parse tree. They experimented on imperative, exclamatory, and optative sentences. This system also has higher precision than Google Translator on a limited data set.

Ashrafi et al. [49] used CFG, which helps to replace the tokenized words with the variable. A bilingual dictionary gives apposite information about morphological features along with the meaning of English words. CFG provides grammatical rules according to English and Bangla language structure. An intermittent parse tree is reorganized by Stimulate English Parse Tree module in the form of another parse tree to stimulate computational history. The output is available by substituting the English words with equivalent Bangla meaning as well as reordering the previous tree to get the actual parse tree by Bangla CFG rules. They have used CFG rules for both Bangla and English languages. Example for English, S → NP VP |NP VP ADV | …, NP → N | PN |CN | …, VP → MV | CV| AV CV | etc. For Bangla, S → NP VP | NP ADV VP |…, NP → N | PN | CN … etc. Where, S→ Sentence, NP→ Noun Phrase, VP→ Verb Phrase, ADV→ Adverb, PN → Pronoun, CN→ Complex Noun, MV→ Main Verb, CV→ Complex Verb, AV→ Auxiliary Verb. This architecture is very effective when the sentence falls into the rules made from the morphological analysis. The authors stated the method as Approximate Lexical Meaning Mapping (ALMM).

Anwar et al. [57] focused on structural and syntax analysis to generate grammatical rules in their B2E RBMT system. The system tokenizes Bangla words based on the lexicon and forms groups of the tokenized words according to grammatical rules using a parser. This information helps to create a parse tree to portray the syntactic structure of source sentences. Later they have used fuzzy logic to interpret the input Bangla sentences to convert them into English. Finally, they enumerated the probability of each word (termed as Fuzzy membership) to come first and next in English sentences. They gave much importance to finite verbs, whereas other parts of speech and phrases have contributions to form a sentence. The system needs the help of an aligned bilingual corpus. Fuzzy logic has been used further by the model proposed by Anwar [58] in the conversion phase with a basic RBMT model to interpret the input Bangla sentences to output English sentences. In this model, 28 basic rules have been

used to parse a sentence and generate the parse tree. Mainly focused on establishing and using grammatical rules, three main types of sentences, simple, complex and compound, have been used in the experiment.

Rabbani et al. [59] proposed an E2B RBMT approach, which transforms different forms of English sentences (like active, passive, assertive, interrogative, imperative, exclamatory, simple, complex, and compound) into some simplified forms, i.e., subject + verb + object. After identifying the principal verb from the English sentence, it binds the rest of the parts of speech as subject and object. Bangla output sentences are generated by the translation of English words of the newly structured English sentences. Recently, Haque & Hasan [60] proposed an algorithm that takes person, verb root, and tense as arguments and finds what should be the appropriate verb in the sentence, which later applied to E2B RBMT system architecture.

Islam et al. [61] have used the tagging of a token as word, number, person, etc., in their RBMT method to identify the structure of Bangla sentences. Later, it motivates word-by-word translation from Bangla to English and applies necessary suffix and grammatical rules that lead to final output. They investigated three approaches to tackle different forms of verb representation in Bangla sentences. In name identification, they have tried to handle unknown words and names. The names of persons are identified by a method emphasizing with tags.

Table II summarizes the above discussed Bangla RBMT studies mentioning achieved test set accuracies. Notably, most of the studies are related to B2E translation. In few cases, few parameters (e.g., accuracy) are not reported clearly in the corresponding articles mentioned in the comments. Among the B2E studies, Anwar [58] achieved the best accuracy for sample, complex, and compound sentences with 95%, 80%, and 80%, respectively, with their self-prepared dataset. On the other hand, 100% test set accuracy is reported for E2B by Ashrafi et al. [49], although information about the dataset is not provided clearly.

C. Review of Bangla EBMT Methods

Only a few Bangla MT studies are available with EBMT. Dandapat et al. [62] investigated a translation memory (TM) based EBMT architecture. They built two TMs: one is based on phrase pairs alignment (PT), and another is based on word aligned file from source to target language (LT), where these two TMs are used for translation of unmatched parts. At first, the system finds the closest match in the input sentences to be translated and then links with equivalent translation. Later, inapposite fragments are detected, and the main translation is found in the recombination step by adding, substituting, and rearranging fragmented translations. They conducted their experiments on different systems: Basic EBMT, EBMT+ TM (PT) in the recombination step, EBMT+TM (PT+LT), EBMT+SMT in the recombination step, and SMT.

TABLE II.   TEST SET PERFORMANCE COMPARISON AMONG BANGLA RBMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Test Set Size	Model Used	Accuracy on Test Set	Comments
1	M. M. Anwar et al., 2009 [55]	Self-Prepared	450 (Simple Sentence)	RBMT with context sensitive grammar rules	93.33% (B2E)	
		Self-Prepared	540 (Complex Sentence)		92.6% (B2E)	
		Self-Prepared	420 (Compound Sentence)		91.67% (B2E)	
2	M. Anwar et al., 2010 [57]	Self-Prepared	Less than 900 (Simple Sentence)	RBMT with fuzzy logic	About 90% (B2E)	Data size and outcomes are not mentioned precisely.
		Self-Prepared	Less than 800 (Complex Sentence)		About 80% (B2E)	
		Self-Prepared	About 550 (Compound Sentence)		About 80% (B2E)	
3	Rahman et al., 2010 [52]	Self-Prepared	6	RBMT with Morphological approach	-	Statistical method used for performance measure and accuracy not mentioned
4	Francisca et al., 2011 [47]	Self-Prepared	79/-/ 27	RBMT with fuzzy rules	-	Performance is not mentioned
5	Alam et al., 2011 [54]	-	-	RBMT with modified approach	-	Statistical method used for performance measure and accuracy not mentioned
6	Chowdhury, 2013 [53]	-	-	RBMT with Parts of Speech Tagging	-	Performance is not mentioned
7	Ashrafi et al., 2013 [49]	Self-Prepared	Not Stated	RBMT with Approximate Lexical Meaning Mapping (ALMM)	100% (E2B)	Experiment outcomes are not available
8	Muntarina et al., 2013 [50]	Self-Prepared	600	RBMT with Tense Based Approach	86.16% (E2B)	
9	Arefin et al., 2015 [56]	Self-Prepared	420	RBMT with Context-Sensitive Grammar	83.09% (B2E)	Assertive, Interrogative and Imperative sentences are considered
10	Alamgir et al., 2016 [51]	Self-Prepared	400	RBMT with Context Sensitive Grammar	81.5% (B2E)	Imperative, Optative and Exclamatory sentences are considered
11	M. Anwar, 2018 [58]	Self-Prepared	Less than 900 (Simple Sentence)	RBMT with Fuzzy logic	About 95% (B2E)	Accurate data and result are not shown
		Self-Prepared	Less than 800 Complex Sentence		About 80% (B2E)	
		Self-Prepared	About 550 (Compound Sentence)		About 80% (B2E)	
12	Mukta et al., 2019 [48]	Self-Prepared	1113	Phrase-based RBMT	Mismatch 50 (E2B)	

Khan et al. [63] have proposed an E2B model in EBMT using WordNet [64] and International Phonetic Alphabet (IPA) [65] based transliteration. The system begins with taking English sentences as input and then parsing them into chunks which are similar to tokenization in RBMT. The chunks are matched with an example-based English-Bangla parallel corpus by a matching algorithm whose outcomes are Chunk-String Templates (CSTs) and unknown words. CSTs are the combination of chunks in English and Bangla languages and the information of alignment of words. The translation of unknown words uses a transliteration process, a procedure of converting a text or word from one language to another language. It is useful for people to pronounce foreign words. Lastly, the output is produced with the help of WordNet and the generation rules. Unknown word handling is the specialty of the model. For this purpose, the model first tries to find semantically related words in WordNet and the closest meaning of the words from the dictionary. If the process does not work, the system needs the help of IPA-based transliteration and Akkhor Bangla Software. Overall translation quality of the model seems good but some inconsistencies have been found using WordNet and due to the small corpus. Salam et al. [66] proposed another EBMT method where ontology is used to improve the quality. The model is similar to [63] but some changes made this model unique. The unknown words are searched in WordNet using synonyms, antonyms, and hypernyms, which develop a vast option to increase the quality.

Table III summarizes the above discussed Bangla EBMT studies mentioning achieved test set performance scores. Notably, the three studies mentioned above are related to E2B translation with self-prepared corpora. Based on the achieved BLEU scores, Dandapat et al. [62] are achieved the best among the three mentioned methods with the value of 57.56.

TABLE III.   TEST SET PERFORMANCE COMPARISON AMONG BANGLA EBMT METHODS FOR ENGLISH TO BANGLA (E2B). [N.B.: NO EMBT STUDY ON BANGLA TO ENGLISH (B2E)]

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Sample Size: Train./Val./ Test Set	Model Used	Performance Score on Test Set			Comments
					BLEU	NIST	Accuracy	
1	Dandapat et al., 2010 [62]	Self-Prepared Medical data	381/-/41	EBMT with translation memory (Probable Target)	57.47(E2B)	5.92(E2B)	-	
				EBMT with translation memory (Probable Target + Lexical Table)	57.56(E2B)	6.00(E2B)		
				EBMT with SMT	52.01(E2B)	5.51(E2B)		
2	Khan et al., 2013 [63]	Self-Prepared	2000 /-/336	EBMT with unknown word translation mechanism	-	-	41.33%(E2B)	Simple, complex and mixed with various phenomena for testing
3	Salam et al., 2017 [66]	Self-Prepared	2000 /-/336	EBMT with CSTs	-	-	38.69%(E2B)	Simple, complex and mixed with various phenomena for testing
				EBMT with CSTs and unknown word translation mechanism	-	-	36.90%(E2B)	

D. Review of Bangla SMT Methods

SMT is a well-known data-driven approach and SMT models for Bangla-English MT studies are developed in several studies. Uddin et al. [67] have proposed an SMT architecture based on different parameters. Alongside the established parameters like for Bangla and English sentence length, for the various probability of occurrences, etc., they have created new parameters based on few complex sentences: Bi-occurred parameter, Bi-distribution parameter, Absent-Distribution parameter, and Subject-check parameter. The Bi-occurred parameter is for the doubly occurred Bangla verbs. The Bi-distribution parameter works with the Bi-occurred parameter and estimates the appropriate position of English translation for the doubly occurred Bangla verbs. To translate Bangla sentences, sometimes extra words are needed to add in English that are implicit and not connected to any Bangla words. The Absent-Distribution parameter handles this type of problem. The Subject-check parameter handles multiple subjects.

A phrase-based SMT for E2B is proposed by Islam et al. [68], where a 5-gram language model (i.e., five words at a time) has been furnished with different corpora and used in the baseline system along with training data made by the aligner. After finding some English words in the Bangla translation and comparatively low results in the baseline system, they operated a cleaning process of corpora with a sentence alignment process. They achieved improvement in the development after executing this new translation system with an 8-gram language model. They also specialized in preposition handling by assigning inflections to the noun in Bangla (applicable to Bangla corpus) and a transliteration module to identify unknown words. They combined the preposition handling module, transliteration, and new translation system; the combined system outperforms other methods on various dictionaries. They used MOSES, GIZA++, MERT, and SRILM [69] toolkits to construct the whole system.

Roy and Popowich [70] presented a phrase-based B2E SMT with a unique transliteration method. They have designed a module that can handle prepositions and Bangla compound words. Their transliteration module at first finds the untranslated words. Later the best-matched translation is found with the help of a monolingual English dictionary. The preposition handling module, at first, removes the inflections of the Bangla words. Later, appropriate English words with prepositions are applied with the help of the bilingual dictionary. Bangla compound words are handled by a splitting algorithm proposed by Koehn & Knight [71]. In another study, Roy and Popowich [72] applied a different word-reordering approach to the phrase-based SMT model. As an automatic word-reordering approach, they used an algorithm proposed by Crego & Mari [73].

Mumin et al. [74] presented a phrase-based SMT model (called Shu-torjoma) for both B2E and E2B. They used various monolingual, bilingual and parallel corpus to train the model. The preprocessor module processes data into a favorable format at the next step, including punctuation and lexical normalization, tokenization, morphological segmentation, syntactical reordering, etc. The preprocessed resources are then trained and tuned to create various statistical models: 5-gram language model, translation mode using GIZA++, Lexicalized Reordering Model, etc., to refine the system. Then the translated texts are found by MOSES decoder. On the other hand, Rabbani et al. [75] proposed a hybrid phrase-based E2B MT using the concept of RBMT and SMT. The model finds the principal verb from any kind of sentence and then converts it into the simplest form.

Dandapat and Lewis [8] developed an English-Bangla general-purpose domain and worked on both SMT and NMT fields. Using different training sets, they used phrasal (for B2E and vice versa) and Treelet (E2B) translation models and developed a word segmentation model to handle unknown words. They developed a word breaker to handle out of vocabulary words where they have used a linguistic suffix list for partitioning inputs and parallel corpora to rank the

partitioned candidates based on frequency. They also used the transliteration module to transliterate foreign words.

Hasan et al. [76] showed a comparative study between SMT and NMT. They used SRILM as a language model and MOSES decoder to train their SMT system and gather different corpora. They also covered 3-gram and 5-gram language models under training. Al Mumin et al. [77] also depicted a comparative result between SMT and NMT where their preprocessed (correction of spelling, pronunciation normalization, etc.) data has been used in the SMT system using MOSES. The whole architecture of this SMT is much like their Shu-torjoma [74].

Table IV summarizes the above discussed Bangla SMT studies mentioning achieved test set performance scores. Several studies reported performance for both B2E and E2B translations; others are for B2E or E2B. For B2E, Al Mumin et al., 2019 [74] achieved the best accuracy showing a BLEU score of 17.43 with SUPara corpus. On the other hand, the best BLEU score for E2B was 23.30, achieved by Islam et al. [78] with KDE4 corpus. It is also notable from studies with both B2E and E2B that the performance score is slightly different between B2E and E2B.

E. Review of Bangla NMT Methods

Nowadays, NMT is the most studied method with different machine learning and deep learning methods in different languages, and studies with NMT are also popular in Bangla MT. Dandapat and Lewis [8] developed an NMT model combining with an SMT model discussed in the previous

section. The NMT system using only conventional bidirectional RNN failed to exceed the score of SMT. They used Phrasal [79] (for B2E and vice versa) and Treelet [80] (for E2B) translation models using different training sets. They also developed a word segmentation model to handle unknown words. Finally, the introduction of early stopping, byte per encoding (BPE) and backpropagated synthetic data enhanced the performance of the NMT model. It outperformed significantly on low-resource data like Bangla.

Hasan et al. [7] used Bidirectional LSTM (BiLSTM) and transformer, the two popular deep learning methods, for B2E NMT. Their preprocessing includes tokenization of English and Bangla sentences, normalization of punctuation, limitation of sentences length and identification of abbreviation, Email, URLs, etc. They have trained their models and created multiple experimental settings on different schemes like using one corpus and multiple corpora. In comparison between the methods, the BiLSTM-based model is found better than the transformer. Hasan et al. [76], in another study, where BiLSTM based methods are compared with SMT. They used different corpora and identified the best performances of each model with a particular corpus. Their results show that the NMT model offers a better result than the SMT model.

Mumin et al. [77] investigated the attention-based model and Byte Pair Encoding (BPE) in their NMT model. They separately examined the basic attention-based model and attention-based model with BPE for both B2E and E2B. It is shown that the attention-based model with BPE gives comparatively better results than other approaches, e.g., SMT.

TABLE IV. TEST SET PERFORMANCE COMPARISON AMONG BANGLA SMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Sample Size: Train./Val./ Test Set	Model Used	Performance Score on Test Set			Comments
					BLEU	NIST	TER	
1	Uddin et al., 2005 [67]	Not Stated		Baseline SMT	-	-	-	No experiment is conducted
2	M. Z. Islam et al., 2010 [78]	EMILLE	25287/500/500	SMT with Final combined system	5.70(E2B)	3.16(E2B)	0.83(E2B)	
		KDE4	33365/1000/1000		23.30(E2B)	5.18(E2B)	0.63(E2B)	
		EMILLE+KDE4	58652/1500/1500		11.70(E2B)	4.27(E2B)	0.76(E2B)	
3	Roy & Popowich, 2010 [70]	Linguistic Data Consortium	11000/600/1000	Phrase-based SMT with Transliteration	9.1 (B2E)	-	-	
4	Roy & Popowich, 2010 [72]	Linguistic Data Consortium	11000/600/1000	SMT with Lexicalized reordering	8.2 (B2E)	-	-	
				SMT with Manual reordering	8.4 (B2E)	-	-	
				SMT with Automatic reordering	9.3 (B2E)	-	-	
5	Dandapat & Lewis, 2018 [8]	Websites, Webdunia, WMT	976634/3500/6000	Baseline SMT	16.56(B2E) 7.41(E2B)	-	-	
6	Hasan et al., 2019 [76]	ILMPC, SIPC, PTB, SUPara	346845/500/956	SMT+ 3-gram Language Model	14.61(B2E)	-	-	Training set is merged but ILMPC is used for development and test set
				SMT+ 5-gram Language Model	14.82(B2E)	-	-	
7	Al Mumin et al., 2019 [74]	SUPara, Global Voices	197338/500/500	Phrase-based SMT	17.43(B2E) 15.27(E2B)	5.76(B2E) 5.13(E2B)	67.94(B2E) 71.93(E2B)	Training data sets are combined; SUPara for test and development

Recently, Siddique et al. [6] proposed architecture for E2B MT based on RNN. Their process starts with the preprocessing and tokenization of the English and Bangla sentences according to frequency. Later with the help of a context vector, the English and Bangla sentences are mapped where embedded RNN, both GRU and LSTM are used, which is similar to the attention model. The model calculates the error with loss function to improve the model through backpropagation. They also identified that using a large number of parallel sentences in the corpus may improve the result.

Akter et al. [81] investigated an NMT method using pre-trained embedding and synthetic monolingual data for E2B. They considered two modifications with the baseline NMT: a pre-trained word embedding model for source and target languages and a synthetic monolingual data addition model. NMT with a pre-trained word-embedding model reduces workload, brings outside model information, and decreases the number of parameters. It has shown improvement in the BLEU score. The addition of synthetic monolingual data in the NMT model associated with back translation helps to handle out of vocabulary words. This model showed a relatively better BLEU score for E2B over the other existing methods.

The most recent NMT models for Bangla MT are [82] and [83]. Dhar et al. [82] investigated a transformer-based NMT model for B2E MT where different parameters (especially, number of heads) are tuned for a better outcome. The model is tested on a benchmark of Bangla-English corpus, which outperformed some other MT methods. On the other hand, Roy et al. [83] considered BiLSTM in their MT study for both B2E and E2B. Attention mechanism with BiLSTM model and a special data augmentation mechanism, called Back Translation (BT), are the significant features of the proposed model. The model outperformed the existing prominent models for B2E MT while tested on a benchmark corpus.

Table V summarizes the above discussed Bangla NMT studies mentioning achieved test set performance scores. Notably, among the studies that performed B2E translation, only a few recent studies performed both B2E and E2B translation. For B2E, the most recent study by Roy et al. [83] achieved the best performance showing a BLEU score of 23.12. They used data augmentation with a back-translation mechanism considering GlobalVoices corpus with SUPara corpus. On the other hand, the best BLEU score for E2B 27.46

was achieved by Akter et al. [81] with synthetic monolingual data in the NMT model.

F. Review of Bangla HMT Methods

There are a few Bangla studies with HMT. Among the existing HMT studies, the E2B method called ANUBAAD [84] is the pioneering one which is a hybrid MT system using EBMT and RBMT explicitly. ANUBAAD considered noun phrase, adverbial phrase, and verb phrase. The system morphologically analyzes the input sentences and defines some formal grammars. Noun phrases and adverbial phrases are translated through EMBT with a template matching module, whereas verb phrases are translated using the RBMT approach.

Rabbani et al. [85] investigated the principal verb-based MT (called PVBMT), which is a hybrid of RBMT and SMT, belongs to the HMT paradigm. After passing through lexical analysis, the words that are tagged in the previous step are bound. In the next step, PVBMT determines the verbs in a sentence that works with three types of verbs within a sentence: auxiliary verb (AV), finite verb (FV), and non-finite verb (NV). If a sentence has more than one verb, then PVBMT creates different sets for different types of verbs according to their meanings and positions. Then PVBMT defines the Bangla sentence structure corresponding to the English sentence and generates the output. They transformed different English sentences into the simplest forms, e.g., Subject+Verb+object, and then translated the sentences into Bangla.

Islam et al. [61] recently investigated B2E MT blending RBMT with data-driven MT (i.e., SMT and NMT). Specifically, first, they implemented some basic grammatical rules that identified names as subjects and optimized Bengali verbs in their RBMT. Next, they integrated RBMT with each of SMT and NMT separately using different approaches. Besides, they performed rigorous experiments over several datasets to provide a comparison among the approaches in terms of translation accuracy, time complexity and space complexity. They also discussed how their blending approaches could be reused for other low-resource languages.

Table VI summarizes the above discussed Bangla HMT studies mentioning achieved test set performance scores. Notably, three studies in the table are with self-prepared datasets. Based on the achieved BLEU score, the method by Islam et al. [61] is the best, showing a score of 18.73.

TABLE V.   TEST SET PERFORMANCE COMPARISON AMONG BANGLA NMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Sample Size (Train./Val./ Test Set)	Model Used	Performance Score on Test Set BLEU	NIST	TER	Comments
1	Dandapat & Lewis, 2018 [8]	Websites, Webdunia, WMT	976634/3500/6000	NMT with synthesis	20.23(B2E) 9.73(E2B)	-	-	
				NMT with BPE	20.64(B2E) 9.80(E2B)	-	-	
2	Hasan et al.,2019 [7]	ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109/500/500	BiLSTM with Bangla and English Embeddings	19.24(B2E)	-	-	
		ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109/500/500	BiLSTM with Bangla Embeddings	19.40(B2E)	-	-	
		ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109/500/500	Transformer	18.99(B2E)	-	-	
		SUPara	70861/500/500	BiLSTM with Bangla and English Embeddings	19.98(B2E)	-	-	
3	Hasan et al., 2019 [76]	ILMPC, SIPC, PTB, SUPara	346845/500/956	BiLSTM with Bangla and English Embeddings	15.62(B2E)	-	-	Training set is merged but ILMPC is used for development and test set
		SUPara	70861/500/500		19.76(B2E)	-	-	
4	Al Mumin et al., 2019 [77]	SUPara, GlobalVoices	197338/500/500	BiGRU with Attention	22.38(B2E) 15.57(E2B)	5.98(B2E) 4.72(E2B)	59.88(B2E) 68.54(E2B)	
				BiGRU with Attention and BPE	22.68(B2E) 16.26(E2B)	6.07(B2E) 5.18(E2B)	60.09(B2E) 68.69(E2B)	
5	Siddique et al., 2020 [6]	Self-Prepared	4000	GRU and LSTM	-	-	-	Performance on test set is not mentioned
6	Akter et al., 2020 [81]	SUPara, Indic parallel, Open subtitles, OPUS Ububtu, OPUS Gnome, OPUS Tanzil	484131/2000/2000	NMT with pre-trained embedding	26.92(E2B)	-	-	
				NMT with synthetic monolingual data	27.46(E2B)	-	-	
7	Dhar et al., 2021 [82]	SUPara	70861/500/500	Transformer with optimal head and BPE	21.33(B2E)			
8	Roy et al., 2021 [83]	SUPara	70861/500/500	BiLSTM with Attention and BPE	22.88(B2E)			
		SUPara, GlobalVoices	115550 /500/500	BiLSTM with Attention,BPE and BT	23.12(B2E)			

TABLE VI.   TEST SET PERFORMANCE COMPARISON AMONG BANGLA HMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year (Ref.)	Corpus / Dataset	Sample Size: Train./Val./ Test Set	Model Used	Performance Score on Test Set BLEU	Accuracy	Comments
1	Naskar et al., 2004 [84]	Not Stated	-	EBMT and RBMT	-	-	No experiment is conducted
2	Rabbani et al., 2016 [85]	Self-Prepared	9	RBMT with SMT and Principle verb-based approach	-	89.6% (semantic analysis) and 78.3% (syntactic analysis) (E2B)	
3	M. A. Islam et al., 2021[61]	Global Voices	1031725	NMT with RBMT	18.73 (B2E)	-	
				SMT with RBMT	18.02 (B2E)	-	

IV. SIGNIFICANCE OF THE PRESENT STUDY

A comprehensive review on a specific topic is important for the research community to get up-to-date information. Hence, one may get a guideline and/or motivation for further work(s) on it. Due to the language resource dependency, MT studies are scattered on a low resource language (e.g., Bangla), and it is necessary to discuss the studies categorically following a common strategic fashion. Although a few good reviews are available for low-resource languages like Thai [86]; but no such review studies are available for the Bangla language, according to the best of our knowledge. Although several Bangla review studies are available, all are very poor in area and scopes. The pioneer review work by Chowdhury [53] in 2013 considered only B2E RBMT studies emphasizing parts of speech tagging matter. The work by Chopra et al. [87] included only one Bangla SMT in their study. The most recent review by Andrabi and Wahid [88] emphasized Hindi and Urdu, and they considered only a few pioneer Bangla studies. Table VII shows the year-wise projection of Bangla-English MT studies with achieved performance scores summarizing the methods presented in Tables II-VI. It is noticeable from the tables that pioneer Bangla MT studies are with RBMT, and NMT has been explored recently with a relatively better translation score. Considering the importance of the Bangla language and its prospects in MT studies, this comprehensive review on Bangla MT is a timely study with the following significance.

TABLE VII.   YEAR-WISE PROJECTION OF BANGLA -ENGLISH MT STUDIES WITH ACHIEVED PERFORMANCE SCORE

Year	RBMT (12)	EBMT (3)	SMT (7)	NMT (8)	HMT (3)
2004					[84] Naskar et al. (E2B)
2005			[67] Uddin et al. (B2E)		
2009	[55] Anwar et al.; Acc. 93.33% (B2E)				
2010	[57] Anwar et al.; Acc. 90% (B2E)	[62] Dandapat et al.; BLEU: 57.56 (E2B)	[78] Islam et al.; BLEU: 23.30 (E2B)		
2010	[52] Rahman et al; (B2E)		[70] Roy & Popowich; BLEU:9.1 (B2E)		
2010			[72] Roy & Popowich; BLEU:9.3 (B2E)		
2011	[47] Francisca et al. (E2B)				
2011	[54] Alam et al. (E2B)				
2013	[49] Ashrafi et al.; Acc. 100% (E2B)	[63] Khan et al.; Acc. 41.33% (E2B)			
2013	[50] Muntarina et al.; Acc. 86.16% (E2B)				
2013	[53] Chowdhury (B2E)				
2015	[56] Arefin et al.; Acc. 83.09% (B2E)				
2016	[51] Alamgir et al.; Acc. 81.5% (B2E)				[85] Rabbani et al.; Acc. 89.6% (E2B)
2017		[66] Salam et al. ; Acc. 38.69% (E2B)			
2018	[58] Anwar; Acc. 95% (B2E)		[8] Dandapat & Lewis; BLEU: 16.56 (B2E) & 7.41 (E2B)	[8] Dandapat & Lewis; BLEU: 20.64(B2E) & 9.80(E2B)	
2019	[48] Mukta et al;. (E2B)		[76] Hasan et al.; BLEU:14.82 (B2E)	[7] Hasan et al.; BLEU: 19.98 (B2E)	
2019			[74] Mumin et al.; BLEU: 7.43 (B2E) & 5.27(E2B)	[76] Hasan et al.; BLEU:19.76 (B2E)	
2019				[77] Mumin et al.; BLEU: 22.68(B2E) & 16.26(E2B)	
2020				[6] Siddique et al.; (B2E)	
2020				[81] Akter et al.; 27.46(E2B)	
2021				[82] Dhar et al. ; BLEU:21.33 (B2E)	[61] Islam et al.;BLEU:18.73 (B2E)
2021				[83] Roy et al.; BLEU: 23.12 (B2E)	

1) Basic ideas of different MT methods (RBMT, EBMT, SMT, and HMT) and performance measures of automatic MT are presented as background studies of the present Bangla MT review.

2) Overview of Bangla language and a brief description of available Bangla-English corpora are given.

3) Bangla MT studies are briefly described categorically; the achieved performances of the individual methods are compared in a tabular form.

V.   FUTURE PROSPECTS OF BANGLA MT FROM THIS STUDY

This review streamlines the various aspects, techniques, and resources of Bangla MT studies comprehensively to motivate researchers and pave the way for further investigation in this area. It is observed that corpus-based data-driven approaches, especially, NMTs are shown to outperform other methods. Therefore, recent studies with NMT and hybrid methods with NMT might be a way to improve Bangla MT proficiency further. Resources deficiency, especially lack of rich corpus, is the main lagging to build an appropriate NMT model. Therefore, focus on resource development is necessary, although it requires government and non-government efforts. It is noticeable that the Government of Bangladesh has launched a large national project on Bangla language and corpus development for MT, an important component in MT studies [89]. Such efforts might boost Bangla MT studies; however, investigating

innovative modern techniques is also necessary for better performance. Another observation from the present study is that all the Bangla MT studies involve English (i.e., B2E and/or E2B). It is also timely demand to break the boundary of existing study and develop Bangla MT systems for other major languages (e.g., Arabic, Chinese, Japanese) considering global prospects of Bangla language in the coming future.

Recently developed MT methods that are found to be very effective for English and other languages pairs may also be practical approaches for Bangla, subject to appropriate incorporation of relevant linguistic or other features and tuning of parameters. Hence, investigation on the Bangla MT study may perform in different directions. Multiple attention layers, called deep attention, investigated by Zhang et al. [90] perform well for Chinese/Germany/France-English translation tasks. Incorporating such a mechanism with multiple attention layers, an attention-based Bangla NMT model can be developed to improve its performance efficiently. Gated recurrent unit (GRU) employment of [91] and parts of speech tagging of [92] in attention mechanism might also be useful to employ in Bangla MT. In the line of data augmentation, input denoising plus auxiliary decoder investigated in [93] and self-learning, training with synthetically generated data using monolingual a source language corpus, investigated in [94], are also intuitive to improve MT performance for a low-resource language like Bangla. Multi-source translation, an approach to exploit multiple inputs (e.g., in two different languages) to increase performance, and missing data management investigated by Nishimura et al. [95] might also be a way to achieve better Bangla MT performance. Gated recurrent unit (GRU), an advanced LSTM model, and its updated model [96] might perform well for Bangla MT. Moreover, recently developed HMT techniques, such as [34] [35] [33], might bring good motivation for better Bangla MT system development.

VI. Conclusion

In the global era of digitalization, MT studies are much more important than ever. Considering the limited Bangla MT studies despite being a major language, this paper reviewed prominent Bangla-English MT studies. Specifically, the basic MT methods (i.e., RBMT, EMBT, SMT, NMT, and HMT) are explained in short as background knowledge. Bangla MT studies under individual methods are described briefly, and achieved performances are presented in the tabular form in Tables II-VI. A year-wise projection of all the reviewed methods in Table VII gives a timeline hierarchy Bangla MT study. It is noticeable from the hierarchy view that pioneer Bangla MT studies are with RBMT and SMT methods, and the recently developed NMT methods outperformed the pioneer methods.

This study is expected to be a valuable resource and guideline for researchers interested in the Bangla MT system. The brief description of the available Bangla-English benchmark corpus (Table I) helps develop a new MT model. The prospects of the present study are summarized in a separate section (Section V), mentioning different points. At a glance, NMT has an opportunity to develop a better Bangla MT model with recently developed techniques such as various

data augmentations. Moreover, it is time to take Bangla MT studies beyond the involvement of the English language and explore Bangla MT studies involving other languages such as Arabic, Chinese, and Japanese.

References

[1] Garg and M. Agarwal, "Machine translation: A literature review," arXiv. 2018.

[2] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the Rare Word Problem in Neural Machine Translation," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 11–19, doi: 10.3115/v1/P15-1002.

[3] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On Using Very Large Target Vocabulary for Neural Machine Translation," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1–10, doi: 10.3115/v1/P15-1001.

[4] M. Wang, L. Gong, W. Zhu, J. Xie, and C. Bian, "Tencent Neural Machine Translation Systems for WMT18," in Proceedings of the Third Conference on Machine Translation (WMT), 2018, pp. 522–527, doi: 10.18653/v1/w18-6429.

[5] "Bengali language," Wikipedia, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Bengali_language. [Accessed: 08-Jul-2021].

[6] S. Siddique, T. Ahmed, M. Rifayet Azam Talukder, and M. Mohsin Uddin, "English to Bangla Machine Translation Using Recurrent Neural Network," Int. J. Futur. Comput. Commun., vol. 9, no. 2, pp. 46–51, Jun. 2020, doi: 10.18178/ijfcc.2020.9.2.564.

[7] M. A. Hasan, F. Alam, S. A. Chowdhury, and N. Khan, "Neural Machine Translation for the Bangla-English Language Pair," in 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1–6, doi: 10.1109/ICCIT48885.2019.9038381.

[8] S. Dandapat and W. Lewis, "Training deployable general domain MT for a low resource language pair: English–Bangla," in EAMT 2018 - Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018, no. May, pp. 109–117.

[9] P. Bhattacharyya, Machine Translation. CRC Press, 2015.

[10] S. Sreelekha, "Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective," arXiv. 2017.

[11] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges," IJCSI Int. J. Comput. Sci. Issues, vol. 11, no. 5, pp. 159–165, 2014.

[12] L. E. Dostert, "The Georgetown-I.B.M. Experiment," Mach. Transl. Lang. Fourteen Essays, 1955.

[13] W. J. Hutchins, "Machine Translation: A Brief History," in Concise History of the Language Sciences, 1995.

[14] J. Slocum, "A survey of machine translation: its history, current status, and future prospects," Comput. Linguist., vol. 11, no. 1, pp. 1–17, 1985.

[15] M. Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," in Readings in Machine Translation, The MIT Press, 2003, pp. 1–7.

[16] R. Ehab, E. Amer, and M. Gadallah, "English-Arabic Hybrid Machine Translation System using EBMT and Translation Memory," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 195–203, 2019, doi: 10.14569/IJACSA.2019.0100126.

[17] A. R. Babhulgaonkar and S. V. Bharad, "Statistical machine translation," in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 62–67, doi: 10.1109/ICISIM.2017.8122149.

[18] P. Koehn, Statistical Machine Translation. Cambridge: Cambridge University Press, 2012.

[19] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2013.

[20] P. Koehn, Neural Machine Translation. Cambridge University Press, 2020.

[21] F. Stahlberg, "Neural machine translation: A review," Journal of Artificial Intelligence Research. 2020, doi: 10.1613/JAIR.1.12007.

[22] M. A. H. Akhand, Deep Learning Fundamentals - A Practical Approach to Understanding Deep Learning Methods. Dhaka: University Grants Commission of Bangladesh, 2021.

[23] A. Esan et al., "Development of a Recurrent Neural Network Model for English to Yorùbá Machine Translation," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 5, pp. 602–609, 2020, doi: 10.14569/IJACSA.2020.0110574.

[24] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin, "A convolutional encoder model for neural machine translation," in ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2017, doi: 10.18653/v1/P17-1012.

[25] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

[26] X. Liu, K. Duh, L. Liu, and J. Gao, "Very deep transformers for neural machine translation," arXiv. 2020.

[27] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 2016, doi: 10.18653/v1/p16-1009.

[28] M. Graca, Y. Kim, J. Schamper, S. Khadivi, and H. Ney, "Generalizing back-translation in neural machine translation," arXiv. 2019, doi: 10.18653/v1/w19-5205.

[29] P. Koehn and R. Knowles, "Six challenges for neural machine translation," arXiv. 2017, doi: 10.18653/v1/w17-3204.

[30] F. Sánchez-Martínez, M. L. Forcada, and A. Way, "Hybrid rule-based--example-based MT: Feeding Apertium with sub-sentential translation units," Proc. 3rd Work. Example-Based Mach. Transl., 2009.

[31] F. Bond and S. Shirai, "A Hybrid Rule and Example-Based Method for Machine Translation," in Recent Advances in Example-Based Machine Translation. Text, Speech and Language Technology, 2003, pp. 211–224.

[32] H. Schwenk, S. Abdul-Rauf, L. Barrault, and J. Senellart, "SMT and SPE machine translation systems for WMT'09," 2009, doi: 10.3115/1626431.1626458.

[33] J.-X. Huang, K.-S. Lee, and Y.-K. Kim, "Hybrid Translation with Classification: Revisiting Rule-Based and Neural Machine Translation," Electronics, vol. 9, no. 2, p. 201, Jan. 2020, doi: 10.3390/electronics9020201.

[34] D. Banik, A. Ekbal, P. Bhattacharyya, and S. Bhattacharyya, "Assembling translations from multi-engine machine translation outputs," Appl. Soft Comput., vol. 78, pp. 230–239, May 2019, doi: 10.1016/j.asoc.2019.02.031.

[35] M. Singh, R. Kumar, and I. Chana, "Hybrid machine translation system using deep learning," ASM Sci. J., vol. 13, no. 2, pp. 31–45, 2020.

[36] V. L. Beyala, M. J., and P. Li, "Factored Phrase-based Statistical Machine Pre-training with Extended Transformers," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 9, pp. 51–59, 2020, doi: 10.14569/IJACSA.2020.0110907.

[37] R. Wagner, "Chatterji, S. K. The origin and development of the Bengali language," Indogermanische Forschungen, 2015, doi: 10.1515/if-1929-0163.

[38] D. Crystal and S. Potter, "English language," 2020. [Online]. Available: https://www.britannica.com/topic/English-language. [Accessed: 10-Feb-2021].

[39] M. S. Ali, A. Alatawi, B. Alsahafi, and N. Noorwali, "QUES: A Quality Estimation System of Arabic to English Translation," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 245–251, 2020, doi: 10.14569/IJACSA.2020.0110732.

[40] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) evaluation of machine translation," in Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07, 2007, pp. 136–158, doi: 10.3115/1626355.1626373.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 2001, p. 311, doi: 10.3115/1073083.1073135.

[42] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceedings of the second international conference on Human Language Technology Research -, 2002, p. 138, doi: 10.3115/1289189.1289273.

[43] A. Mauser, S. Hasan, and H. Ney, "Automatic evaluation measures for statistical machine translation system optimization," in Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, 2008.

[44] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation, 2006.

[45] K. Wołk and D. Koržinek, "omparison and adaptation of automatic evaluation metrics for quality assessment of respeaking," Comput. Sci., vol. 18, no. 2, p. 129, 2017, doi: 10.7494/csci.2017.18.2.129.

[46] M. Z. Iqbal, "SUPara : A Balanced English-Bengali Parallel Corpus," SUST J. Sci. Technol., vol. 16, no. 2, pp. 46–51, 2012.

[47] J. Francisca, M. M. MIA, and D. S. M. M. RAHMAN, "Adapting Rule Based Machine Translation From English To Bangla," Indian J. Comput. Sci. Eng., vol. 2, no. 3, 2011.

[48] A. P. Mukta, A. Mamun, C. Basak, S. Nahar, and F. H. 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), 2019, pp. 1–5.

[49] S. S. Ashrafi, H. Kabir, M. Anwar, and A. K. M. Noman, "English to Bangla Machine Translation System Using Context-Free Grammars," IJCSI Int. J. Comput. Sci., vol. 10, no. 3, pp. 144–153, 2013.

[50] K. Muntarina, G. Moazzam, and A. Bhuiyan, "Tense Based English to Bangla Translation Using MT System," Int. J. Eng. Sci. Invent., vol. 2, no. 10, pp. 30–38, 2013.

[51] T. Alamgir, M. S. Arefin, and M. M. Hoque, "An Empirical Machine Translation Framework for Translating Bangla Imperative, Optative and Exclamatory Sentences into English," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 932–937, doi: 10.1109/ICIEV.2016.7760137.

[52] M. S. Rahman, S. R. Poddar, M. F. Mridha, and M. N. Huda, "Open morphological machine translation: Bangla to English," in 2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM 2010, 2010, pp. 460–465, doi: 10.1109/CISIM.2010.5643495.

[53] S. A. Chowdhury, "Developing a Bangla to English Machine Translation System Using Parts Of Speech Tagging : A Review," J. Mod. Sci. Technol., vol. 1, no. 1, pp. 113–119, 2013.

[54] M. G. R. Alam, M. M. Islam, and N. Islam, "A New Approach To Develop An English To Bangla Machine Translation System," Daffodil Int. Univ. J. Sci. Technol., vol. 6, no. 1, pp. 36–42, Jan. 2011, doi: 10.3329/diujst.v6i1.9332.

[55] M. M. Anwar, M. Z. Anwar, and M. Al-Amin Bhuiyan, "Syntax Analysis and Machine Translation of Bangla Sentences," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 9, no. 8, pp. 317–326, 2009.

[56] M. S. Arefin, M. M. Hoque, M. O. Rahman, and M. S. Arefin, "A machine translation framework for translating Bangla assertive, interrogative and imperative sentences into English," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2015, no. May, pp. 1–6, doi: 10.1109/ICEEICT.2015.7307534.

[57] M. Anwar, M. Z. Anwar, and A. Bhuiyan, "Structural Analysis of Bangla Sentences for Machine Translation," in International Conference On "Computational Intelligence Applications," 2010, no. April, pp. 3–5.

[58] M. Anwar, "Bangla to English Machine Translation using Fuzzy Logic," Int. J. Comput. Sci. Inf. Secur., vol. 16, no. 11, pp. 156–165, 2018.

[59] M. Rabbani, K. M. R. Alam, and M. Islam, "A new verb based approach for English to Bangla machine translation," in 2014 International

305 | P a g e

Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp. 1–6, doi: 10.1109/ICIEV.2014.6850684.

[60] M. Haque and M. Hasan, "English to Bengali Machine Translation: An Analysis of Semantically Appropriate Verbs," in 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2018, no. October, pp. 217–221, doi: 10.1109/ICISET.2018.8745626.

[61] M. A. Islam, M. S. H. Anik, and A. B. M. A. Al Islam, "Towards achieving a delicate blending between rule-based translator and neural machine translator," Neural Comput. Appl., vol. 33, no. 18, pp. 12141–12167, Sep. 2021, doi: 10.1007/s00521-021-05895-x.

[62] S. Dandapat, S. Morrissey, S. K. Naskar, and H. Somers, "Statistically motivated example-based machine translation using translation memory," in ICON-2010: 8th International Conference on Natural Language Processing, 2010.

[63] M. A. S. Khan, S. Yamada, and T. Nishino, "How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration," J. Comput., vol. 8, no. 5, pp. 1167–1174, May 2013, doi: 10.4304/jcp.8.5.1167-1174.

[64] C. Fellbaum, "WordNet and wordnets," in Encyclopedia of Language and Linguistics, 2005.

[65] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, "The International Phonetic Alphabet," in Linguistics, 2012.

[66] K. M. A. Salam, S. Yamada, and N. Tetsuro, "Improve Example-Based Machine Translation Quality for Low-Resource Language Using Ontology," Int. J. Networked Distrib. Comput., vol. 5, no. 3, p. 176, 2017, doi: 10.2991/ijndc.2017.5.3.6.

[67] M. G. Uddin, M. Murshed, and M. A. Hasan, "A parametric approach to Bangla to English Statistical Machine Translation for complex Bangla sentences -Step 1," in Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh., 2005, no. December 2005, doi: 10.13140/2.1.2720.3526.

[68] M. Z. Islam, J. Tiedemann, and A. Eisele, "English to Bangla phrase-based Machine Translation," in EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation, 2010, no. May.

[69] "SRILM - The SRI Language Modeling Toolkit," 2021. [Online]. Available: https://www.sri.com/case-studies/srilm/. [Accessed: 08-Aug-2021].

[70] M. Roy and F. Popowich, "Phrase-based statistical machine translation for a low-density language pair," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6085 LNAI, 2010, pp. 273–277.

[71] P. Koehn and K. Knight, "Empirical methods for compound splitting," in Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL '03, 2003, vol. 1, p. 187, doi: 10.3115/1067807.1067833.

[72] M. Roy and F. Popowich, "Word Reordering Approaches for Bangla-English Statistical Machine Translation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6085 LNAI, 2010, pp. 282–285.

[73] J. M. Crego and B. Mari, "Syntax-enhanced N -gram-based SMT," MT Summit, 2007.

[74] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "shu-torjoma : An English↔Bangla Statistical Machine Translation System," J. Comput. Sci., vol. 15, no. 7, pp. 1022–1039, Jul. 2019, doi: 10.3844/jcssp.2019.1022.1039.

[75] M. Rabbani, K. M. R. Alam, M. Islam, and Y. Morimoto, "PVBMT: A Principal Verb based Approach for English to Bangla Machine Translation," Int. J. Comput. Vis. Signal Process., vol. 6, no. 1, pp. 1–9, 2016.

[76] M. A. Hasan, F. Alam, S. A. Chowdhury, and N. Khan, "Neural vs Statistical Machine Translation: Revisiting the Bangla-English Language Pair," in 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019, no. September, pp. 1–5, doi: 10.1109/ICBSLP47725.2019.201502.

[77] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "Neural Machine Translation for Low-resource English-Bangla," J. Comput. Sci., vol. 15, no. 11, pp. 1627–1637, Nov. 2019, doi: 10.3844/jcssp.2019.1627.1637.

[78] M. Z. Islam, J. Tiedemann, and A. Eisele, "English to Bangla phrase-based Machine Translation," in EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation, 2010, no. May.

[79] S. Green, D. Cer, and C. Manning, "Phrasal: A Toolkit for New Directions in Statistical Machine Translation," in Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, pp. 114–121, doi: 10.3115/v1/W14-3311.

[80] C. Quirk, A. Menezes, and C. Cherry, "Dependency treelet translation: Syntactically informed phrasal SMT," in ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2005, doi: 10.3115/1219840.1219874.

[81] M. Akter, M. Shahidur Rahman, M. Z. Iqbal, and M. R. Selim, "SuVashantor: English to Bangla machine translation systems," J. Comput. Sci., 2020, doi: 10.3844/jcssp.2020.1128.1138.

[82] A. C. Dhar, A. Roy, A. Habib, M. A. H. Akhand, and N. Siddique, "Transformer Deep Learning Model for Bangla-English Machine Translation," in 2nd International Conference on Artificial Intelligence: Advances and Applications (ICAIAA 2021), 2021, pp. 1–10.

[83] A. Roy, A. C. Dhar, M. A. H. Akhand, and M. A. S. Kamal, "Bangla-English Neural Machine Translation with Bidirectional Long Short-Term Memory and Back Translation," Int. J. Comput. Vis. Signal Process., vol. 11, no. 1, pp. 25–31, 2021.

[84] S. Naskar, D. Saha, and S. Bandyopadhyay, "ANUBAAD – A Hybrid Machine Translation System from English to Bangla," in Symposium on Indian Morphology, Phonology & Language Engineering, 2004, pp. 91–92.

[85] M. Rabbani, K. M. R. Alam, M. Islam, and Y. Morimoto, "PVBMT: A Principal Verb based Approach for English to Bangla Machine Translation," Int. J. Comput. Vis. Signal Process., vol. 6, no. 1, pp. 1–9, 2016.

[86] S. Lyons, "A review of Thai–English machine translation," Mach. Transl., vol. 34, no. 2–3, pp. 197–230, Sep. 2020, doi: 10.1007/s10590-020-09248-8.

[87] D. Chopra, N. Joshi, and I. Mathur, "A Review on Machine Translation in Indian Languages," Eng. Technol. Appl. Sci. Res., vol. 8, no. 5, pp. 3475–3478, 2018, doi: 10.48084/etasr.2288.

[88] S. A. B. Andrabi and A. Wahid, "A Review of Machine Translation for South Asian Low Resource Languages," Turkish J. Comput. Math. Educ., vol. 12, no. 5, pp. 1134–1147, Apr. 2021, doi: 10.17762/turcomat.v12i5.1777.

[89] B. Ministry of ICT, "Enhancement of Bangla Language in ICT through Research & Development," Ministry of ICT, Bangladesh, 2020. [Online]. Available: http://eblict.gov.bd/site/page/07289180-0932-4a40-b5ad-8a163780e16f. [Accessed: 01-Jan-2021].

[90] B. Zhang, D. Xiong, and J. Su, "Neural Machine Translation with Deep Attention," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 154–163, 2020, doi: 10.1109/TPAMI.2018.2876404.

[91] B. Zhang, D. Xiong, J. Xie, and J. Su, "Neural Machine Translation With GRU-Gated Attention Model," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 11, pp. 4688–4698, Nov. 2020, doi: 10.1109/TNNLS.2019.2957276.

[92] L. H. B. Nguyen, H. D. Minh, D. Dinh, and T. Le Manh, "Improving neural machine translation with POS tags," ICIC Express Lett. Part B Appl., vol. 12, no. 1, pp. 91–98, 2021, doi: 10.24507/icicelb.12.01.91.

[93] B. Pan, Y. Yang, Z. Zhao, Y. Zhuang, and D. Cai, "Bi-Decoder Augmented Network for Neural Machine Translation," Neurocomputing, vol. 387, pp. 188–194, Apr. 2020, doi: 10.1016/j.neucom.2020.01.003.

[94] Y. Li, X. Li, Y. Yang, and R. Dong, "A Diverse Data Augmentation Strategy for Low-Resource Neural Machine Translation," Information, vol. 11, no. 5, p. 255, May 2020, doi: 10.3390/info11050255.

[95] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura, "Multi-Source Neural Machine Translation With Missing Data," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 569–580, 2020, doi: 10.1109/TASLP.2019.2959224.

[96] X. Shi, H. Huang, P. Jian, and Y.-K. Tang, "Improving neural machine translation with sentence alignment learning," Neurocomputing, vol. 420, pp. 15–26, Jan. 2021, doi: 10.1016/j.neucom.2020.05.104.

[97] R. Z. Xiao, a. M. McEnery, J. P. Baker, and A. Hardie, "Developing Asian language corpora: standards and practice," in The Fourth Workshop on Asian Language Resources, 2004, pp. 1–8.

[98] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, 2012, pp. 2214–2218.

[99] T. Nakazawa et al., "Overview of the 5th Workshop on Asian Translation," in Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation, 2018, pp. 904–944.

[100] M. Post, C. Callison-Burch, and M. Osborne, "Constructing parallel corpora for six Indian languages via crowdsourcing," 7th Work. Stat. Mach. Transl., pp. 401–409, 2012.

[101] A. Kids, I. N. Learning, B. Mathematics, E. A. R. A. Eva, and R. Rion, "A Collaborative Platform to Collect Data for Developing Machine Translation System," 2018.

[102] LDC, "Linguistic Data Consortium, Philadelphia, PA," 2021. [Online]. Available: https://www.ldc.upenn.edu/. [Accessed: 20-Jul-2021].