Journal of Computer Science

Original Research Paper

Bangla↔English Machine Translation Using Attention-based Multi-Headed Transformer Model

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

[1]Deptartment of Computer Science and Engineering, Khulna University of Engineering and Technology, Khulna 9203, Bangladesh
[2]Graduate School of Science and Technology, Gunma University, Kiryu 376-8515, Japan
[3]School of Computing, Engineering and Intelligent Systems, Ulster University, Northern Ireland, UK

Article history
Received: 22-06-2021
Revised: 10-09-2021
Accepted: 11-09-2021

Corresponding Author:
M. A. H. Akhand
Deptartment of Computer Science and Engineering,
Khulna University of Engineering and Technology,
Khulna 9203, Bangladesh
Email: akhand@cse.kuet.ac.bd

Abstract: Machine Translation (MT) refers to translate texts or documents from the source language into the target language without human intervention. Any MT model is language-dependent and its development requires grammar, phrase rules, vocabulary, or relevant data for the particular language pair. Hitherto, little research on MT for Bangla-English is reported in the literature, although Bangla is a major language. This study presents a deep learning-based MT system concerning both-way translation for the Bangla-English language pair. The attention-based multi-headed transformer model has been considered in this study due to its significant features of parallelism in input processing. A transformer model consisting of encoders and decoders is adapted by tuning different parameters (especially, number of heads) to identify the best performing model for Bangla to English and vice versa. The proposed model is tested on SUPara benchmark Bangla-English corpus and evaluated the Bilingual Evaluation Understudy (BLEU) score, which is currently the most popular evaluation metric in the MT field. The proposed method is revealed as a promising Bangla-English MT system achieving BLEU scores of 21.42 and 25.44 for Bangla to English and English to Bangla MT cases, respectively.

Keywords: Deep Learning, Machine Translation, Neural Machine Translation, Transformer Model

Introduction

Translation of speech or text contents from one natural language to another is often indispensable in politics, business, research and other areas. Translation through human experts is a well-known approach over the centuries. Human translators perform an expert job interpreting conversations between two parties (e.g., country chiefs, tourists, business giants) spoken in different languages. Globalization today requires translating web contents (e.g., website, references, documents) in everyday living. Translating such huge contents (especially text, document and web) persuades machine translation as an emerging research field in recent years (Garg and Agarwal, 2018).

The idea of natural language translation using computer systems appeared in the 1950s (Hutchins, 2000). Machine Translation (MT) has become a research field through the public demonstration of the Georgetown-IBM experiment (Hutchins, 2005). On a fundamental basis, MT was used to conduct direct substitution of words from source language ones to a target language (Hutchins, 1995). However, it is clear that only word-for-word translation does not provide semantic meaning to be useful in real life. Efforts have been made by the research community to develop new methods in the last several decades to improve the quality of MT.

The MT methods are broadly categorized into four approaches: Rule-Based MT (RBMT), Example-Based MT (EBMT), Statistical MT (SMT) and Neural MT (NMT). A number of hybrid methods combining two individual approaches are also available, e.g., RBMT and SMT (Xuan *et al*., 2012). RBMT is basically based on linguistic information and it produces translation through rules generated by human experts considering verbs, phrases, prepositions, etc., of the language pair (Bhattacharyya, 2015). EBMT takes a parallel corpus that contains the source sentence and its translation. After taking help from parallel corpus, the translation mechanism finds similar words/phrases to adopt the previously available word/phrase to translate a new

sentence (Sumita and Iida, 1991). SMT is an MT model which generates translation on the basis of probability generated through statistical analysis of bilingual aligned corpora (Babhulgaonkar and Bharad, 2017). NMT is the most recent method with encoders and decoders in the core; it is a data-driven approach that trains a special Neural Network (NN) model for MT (Kalchbrenner and Blunsom, 2013). NMT has emerged as a powerful approach to MT research with the advancement of deep neural networks over the last decade (Stahlberg, 2020).

A number of remarkable researches are available in the literature with rich resources which achieved good performance for English-French (Luong *et al*., 2014), English-German (Jean *et al*., 2014), English-Chinese (Wang *et al*., 2018) language pairs. In contrast, MT resources on the Bangla language are very limited despite being a major language in the world, the fifth-ranked globally with 228 million native speakers and the first language of Bangladesh (Akhand *et al*., 2016). A number of Bangla-English MT studies are available with different methods, but they are not significant with respect to resource-rich language (Dandapat and Lewis, 2018; Hasan *et al*., 2019a; Siddique *et al*., 2021). Therefore, the aim of this study is to develop an NMT system for the Bangla-English language pair.

A deep learning-based transformer model is investigated in this study to develop an MT, taking advantage of the transformer's parallelism features in the input data processing. A transformer model consists of encoders and decoders, where learnable parameters are tuned to identify the best performing MT model for Bangla to English and vice versa. The proposed model is tested on SUPara benchmark Bangla-English corpus and evaluated the Bilingual Evaluation Understudy (BLEU) score. The proposed method is revealed as a promising Bangla-English MT system while compared with the prominent existing methods on the basis of the achieved BLEU scores.

The rest of the paper briefly reviews existing Bangla-English studies, describes the proposed methodology, reports experimental studies and results. Finally, the paper concludes the findings with a few remarks.

Related Studies

A number of studies have been reported over the last decade for Bangla-English MT with different techniques. Most of the existing studies are only considered Bangla to English (denoted as B2E) or English to Bangla (denoted as E2B) case. Among the existing studies, the E2B method called ANUBAAD (Naskar *et al*., 2004) 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 use the RBMT approach.

Several other RBMT methods have been available for E2B MT in recent years. Dandapat *et al*. (2010) investigated a Translation Memory (TM) based EBMT architecture for E2B. They built two TMs: One is based on phrase pairs alignment and the other is based on a word-aligned file from source to a target language. Finally, they integrated TM with EBMT and compared it with basic EBMT. Salam *et al*. (2013) suggested an EBMT method emphasizing unknown word handling using Word Net and International-Phonetic-Alphabet (IPA) based transliteration with software. Salam *et al*. (2017) proposed another EBMT method for E2B where the unknown words are searched in WordNet using synonyms, antonyms and hypernyms. Francisca *et al*. (2011) proposed an E2B RBMT that divides the words of English sentences based on sentence characteristics like grammar and structure. A lexical analyzer is used to generate the class of sentences utilizing the information of the word from a dictionary. With the help of the partially or fully matched fuzzy rules, output Bangla sentences are generated using a dictionary.

Ashrafi *et al*. (2013) used Context-Free Grammar (CFG) in replacing the tokenized words with the variable in their E2B RBMT. CFG provides grammatical rules according to the English and Bangla language structures. They created an intermittent parse tree to stimulate computational history. The outcome is the substitution of the English words with equivalent Bangla meaning as well as reordering the previous tree to get the actual parse tree by Bangla CFG rules. Muntarina *et al*. (2013) proposed the E2B RBMT model on the basis of tense-based rules. The model constructs a parse tree for input English sentences and then converts it into Bangla parse tree based on production rules for both languages generated by syntactic and morphological analysis.

Rabbani *et al*. (2014) proposed an E2B RBMT approach, which transforms different forms of English sentences (like active, passive, assertive, interrogative, imperative, exclamatory, simple, complex and compound) into 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 and Hasan (2018) proposed an algorithm that takes person, verb root and tense as arguments and finds what should be appropriate verb in the sentence, which later applied to E2B RBMT system architecture.

A few studies have been carried out on B2E RBMT. Anwar *et al*. (2009) used Context-Sensitive Grammar (CSG) rules to analyze a Bangla sentence syntactically. The sentences can be simple, complex, or compound. After analyzing, the sentences are translated into English.

Rahman *et al*. (2010) proposed a method of using root words to translate Bangla to English. Morphological analysis is used to find out the root word. In addition to the root word, parts of speech and grammar of the source sentence are also detected. After combining all, a Bangla sentence is translated into English. Anwar *et al*. (2010) focused on the lexical mappings and structural analysis of Bangla sentences. They introduced a rule-based grammatical approach to perform syntactic analysis on every type of sentence. The system tokenizes the Bangla words based on the lexicon and uses a parser to group the tokenized words according to grammatical rules. Chowdhury (2013) projected a system where Bangla sentences are read from left to right and corresponding English words are generated by using a dictionary and context of the Bangla sentence. In addition to word generation, a set of grammatical rules are used to analyze the source sentence properly. Arefin *et al*. (2015) used CSG rules for translating assertive, interrogative and imperative Bangla sentences into English. The rules are developed based on the mood of the sentence and ignoring sentence structure. Alamgir *et al*. (2016) also used CSG to translate imperative, optative and exclamatory Bangla sentences into English.

Mukta *et al*. (2019) proposed a phrase-based E2B MT using fuzzy rules. The system takes the input of different types of sentences based upon the tense, phrase and affirmative and negative sentences. The system also emphasizes English grammar, verbs, prepositions, inflection and other grammatical rules on Bangla. After tokenizing and matching fuzzy rules, the model translates English sentences to Bangla with the help of a dictionary. Anwar (2018) also used fuzzy logic for B2E MT, which includes syntactic analysis of source language and generation of the target language.

As a data-driven approach, the SMT model has been developed in several Bangla-English MT studies. Roy and Popowich (2010a) presented a phrase-based B2E SMT with a unique transliteration method. In addition, a specialized component for detecting prepositions and Bangla compound words is also used to improve the performance. Roy and Popowich (2010b), in another work, presented a word reordering technique with SMT that had a positive effect on overall performance. Recently, Al Mumin *et al*. (2019a) presented a phrase-based SMT model (called shu-torjoma) for both B2E and E2B. The proposed system excels other developed systems significantly. On the other hand, Rabbani *et al*. (2016) proposed a hybrid phrase-based E2B MT using the concept of RBMT and SMT. The model finds the principle verb from any kind of sentence and then converts it into the simplest form.

Deep learning-based NMT is a recent trend in MT systems in different languages and a few studies are available for Bangla-English. Hasan *et al*. (2019a) used Bidirectional Long Short-Term Memory Network (BiLSTM) and transformer, the two popular deep learning methods, for B2E NMT. In comparison between the methods, BiLSTM based model is found better than the transformer. Hasan *et al*. (2019b), in their study, BiLSTM based methods compared with SMT. They also used different datasets and measured which model worked better for which dataset. Their results showed that the NMT model provides a better result than the SMT model. Dandapat and Lewis (2018) developed an English-Bangla general-purpose MT domain and worked on both SMT and NMT fields. They used Phrasal (Green *et al*., 2015) (for B2E and vice versa) and Treelet (Quirk *et al*., 2005) (for E2B) translation model using different training sets. They also developed a word segmentation model to handle unknown words. They showed that NMT works better than SMT.

Recently, Al Mumin *et al*. (2019b) 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 any other approach. Most recently, Siddique *et al*. (2021) have proposed an NMT architecture for E2B MT based on Recurrent NN (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 Gated Recurrent Unit (GRU) and LSTM are used. The model calculates the error with loss function to improve the model through backpropagation.

Attention-based Multi-Headed Transformer Model for Bangla-English MT

A transformer deep learning model with a multi-headed attention mechanism for both B2E and E2B MT is proposed in this study. The method comprises two major phases: The data preprocessing phase and the transformer model training phase.

Data Preprocessing

For the NMT system, data preprocessing includes tokenization, true-casing, normalizing punctuation and removing non-printable characters from the data. Long sentences and empty sentences may cause a problem so that a fixed sentence length is used in this study like any NMT system. The BPE algorithm (Gage, 1994; Sennrich *et al*., 2015) is applied to the corpus for sub-word segmentation to handle rare words. Preprocessing depends on data to be used and is explained for the selected data in the experimental studies section.

Transformer Architecture and Its Adaptation

The recently proposed deep learning model, Transformer (Vaswani *et al*., 2017), is one of the most significant models in the field of Natural Language Processing (NLP). The significance of the model is that data do not need to feed into the model in a consecutive manner that permits parallelism. So, a transformer model ensures fast training for the NLP tasks. A transformer is widely used in MT, time series prediction (Maxime, 2019), named entity recognition (Davydova, 2017), document generation (Radford *et al*., 2019), biological sequence analysis (Nambiar *et al*., 2020; Rives *et al*., 2021). Another important issue for choosing the model is the open-source model availability and customization facility for a particular task in OpenNMT toolkit (Klein *et al*., 2017).

Figure 1 demonstrates the layers of the transformer model, which has basically four main operating units: Embedding, Encoder, Decoder and Output Generation. In the embedding layer, the words in a sentence are transferred into word embeddings. A word embedding is a fixed-sized vector representing an input word. Then each embedding is added to the positional encoding vector (in the range of -1 to 1) of the same dimension. The resultant vector presents all the necessary information, such as the sequence of words in the input sentence and the distance of different words.

The resultant embedding vector of numeral values is the input to the encoder module. The encoder module contains several encoders in a cascade fashion and the encoded vector is the outcome from the encoder module after the successful operation of individual encoders. The Encoder Output Vector (EOV) is fed to the decoder module and output words are generated sequentially considering decoders' current status and previously generated words. For the sample input sentence in English, 'I Love My Country', Bengali word 'আমি' (phonetic: Ami; means I) is generated first with successive operations on the EOV by the decoders and output generation. To generate the second word, the already generated output word 'আমি' is feed into the decoder module and the word 'আমার' (phonetic: Amar; means my) is generated. The word 'আমার' is used to generate 'দেশকে' (phonetic: Deshke; means country). Finally, the last word 'ভালোবাসি' (phonetic: Bhalobashi; means love) is generated while the third output word is feed into the decoder module.

The separate encoder and decoder modules are the core of the transformer model, which handles the attention mechanism to improve NMT performance. The encoder (or decoder) module is a stack of several encoders (or decoders) and the number of encoders and decoders is generally the same. Figure 2 presents general architectures of an encoder and a decoder illustrating individual layers. An encoder has mainly two sub-layers: Multi Headed Attention (MHA) and Feed Forward NN (FFNN). In each of the sublayers, normalization performs on the vector, adding the input vector of the sublayer and the vector from the MHA/FFNN.

The attention mechanism enables the transformer model to understand how much the other words are relevant to the word that is currently being processed. At first, the attention process multiplies Embedded Input Vector (EIV) with three matrices (such as *Wq*, *Wk* and *Wv*) individually and creates three vectors: A Query vector (*q*), a Key vector (*k*) and a Value vector (*v*). These new vectors are smaller in dimension than the EIV and essential for calculation for attention. The second step in attention is to calculate a score which determines how much focus to put on other parts of the input sentence. To calculate the score for the first word of the shown example (i.e., 'I'), the score of every word in the input sentence is to be calculated against the first word. This can be done by calculating the dot product of the query vector with the key vector while there is *n* number of words in the sentence. Therefore, a score for 'I' is *q1.k1 q1.k2 q1.k3 ..... q1.kn.* The third and fourth steps are to divide the scores by the square root of the dimension of the Key vectors and passing the result through a Softmax operation. Softmax normalizes the scores, so they are all positive and add up to 1. The fifth step is to multiply each value vector by the Softmax score. The sixth step is to sum up the weighted value vectors. This produces the output of the attention layer for the first word.

The transformer model uses MHA with the above-described attention for an individual head. In a multi-headed case, a set of Query, Key and Value vectors are produced for each individual head. It expands the model's ability to focus on different positions. For example, to translate a sentence like "The animal didn't cross the street because it was too tired", the MHA helps to know which word "it" refers to.

A decoder consists of three sub-layers: MHA, Encoder-Decoder Attention (EDA), FFNN. The operations of MHA and FFNN are the same as in an encoder and EDA is significantly different from an encoder. The EDA layer works just like MHA, except it creates its query matrix from the layer underneath it. The output generation mainly consists of the Linear and Softmax layers, which convert the decoder output vector into some probabilistic values. These values help the model generate the next token.

There are several hyperparameters in the transformer model, such as batch size, dropout, learning rate, number of encoder layers, number of decoder layers, number of heads, etc. To achieve better performances, the hyperparameters of the transformer model should be adjusted. Multiple numbers of heads help the self-attention and make the attention layer work better as it increases the model's ability to guess the other words referring to a particular word that is currently being processed.

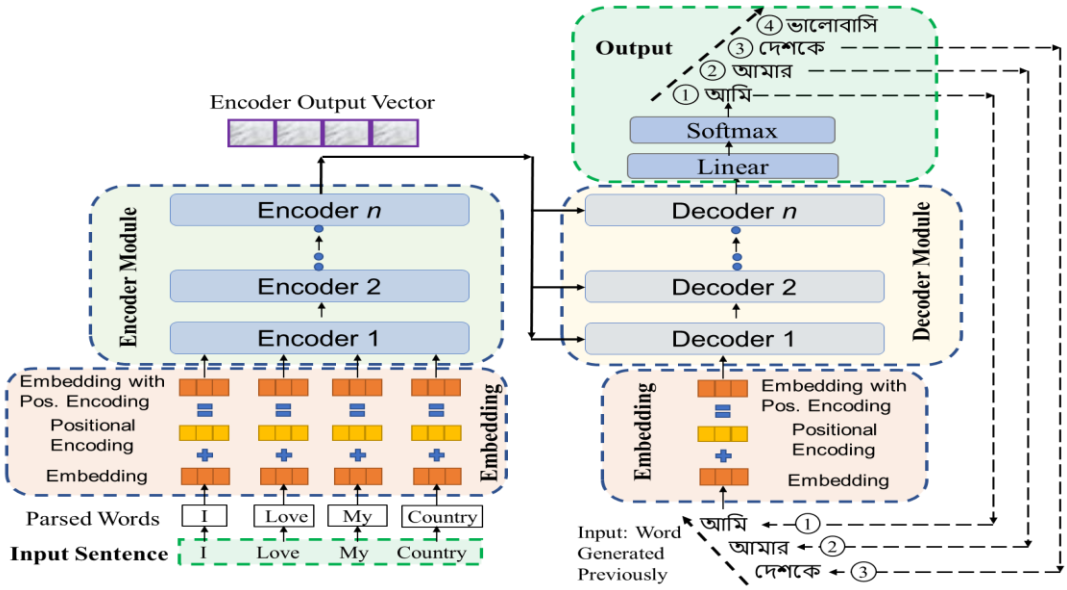

Fig. 1: Architecture of the transformer model for the proposed machine translation. Words of the input sentences are feed to the encoder after embedding with positional encoding. An encoded vector from the encoder module is feed to the decoder and output words are generated sequentially considering decoders' present status and previously generated words

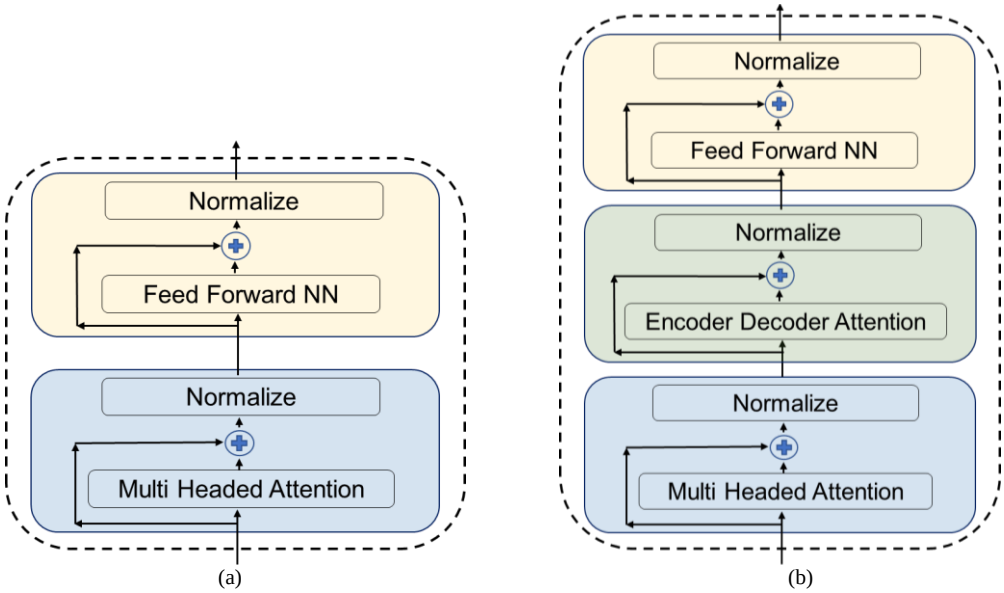

Fig. 2: Architecture of an individual encoder and decoder used in encoder and decoder modules in the transformer model (a) architecture of an encoder; (b) architecture of a decoder

Experiments Studies

This section describes the experimental outcomes of the proposed NMT system on the chosen benchmark dataset. The performance of the proposed method is also compared with existing methods.

Benchmark Data and Preprocessing

A few parallel corpora are available for Bangla-English MT. In this study, SUPara (Al Mumin *et al.*, 2012) dataset is used as a number of recent studies have used this corpus (Al Mumin *et al*., 2019a, 2019b; Hasan *et al*., 2019a, 2019b). The dataset contains 70861, 500 and 500 parallel sentences for training, validation and test sets, respectively. In the data processing step, tokenization, true casing, normalizing punctuation and removing non-printable characters are performed using Moses (Koehn *et al*., 2007), the open-source toolkit for MT. Moses changes the raw sentences into a number of tokens where words and punctuation marks are

separated by a space. As long sentences and empty sentences may cause a problem, the sentence length is limited to 40 in our NMT model. Having a small corpus size results in a poor dictionary which might cause a large number of unknown words in the test case. To handle such scenarios, sub-word segmentation is employed using the BPE algorithm. The algorithm counts the frequency of each word in a corpus and a special stop symbol </w> is added at the end of each token. Characters are then separated. After that, the algorithm finds out the most frequent two consecutive byte pairs and merges the two-byte pairs into one token. As an example, the BPE algorithm can recognize 'r' and 'o' as a consecutive frequent token pair and thus merge them into one token 'ro'. The same explanation is for the 'ses' token. Then the BPE algorithm divides the token 'roses' into two sub-words: 'Ro' and 'ses'; adding '@@' in between them. After all these preprocessing operations, training, validation and test sets contain 65855, 366, 361 parallel sentences, respectively.

Table 1 shows preprocessing effect on few sample sentences in both Bangla (for B2E) and English (for E2B). For the Bangla sentences, English phonetics and meanings are given for better realization to the international community. It is shown in the table that few sentences have some '@@' after preprocessing.

Performance Evaluation and Experimental Setup

For evaluating the performance of the model, the Bilingual Evaluation Understudy (BLEU) score is measured, which is currently the most popular evaluation metric in the MT field (Papineni *et al*., 2001). It is a precision-oriented measurement and evaluates the correctness of system output. BLEU score is measured in three steps. At first, *n*-gram or the number of word matches are calculated in the candidate sentences (system output) and the reference sentences. Then the candidate counts are clipped by their corresponding maximum reference value. Next, the clipped *n*-grams are summed and divided by the number of candidate *n*-grams (Papineni *et al*., 2001). Through this step, the modified precision score (p_n) is found:

$$p_n = \frac{\sum_{Ce\{Candidates\}} \sum_{n-gram \in C} Countclip(n-gram)}{\sum_{C'e\{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')} \quad (1)$$

Here *Candidates* denotes the complete corpus and *C* denotes a hypothesis sentence. The second step is BLEU Brevity Penalty (BP) factor calculation:

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

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 calculated using Eq. (3):

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} W_n \log p_n\right), \quad (3)$$

Here, N is set to 4 as the baseline system and wn is a positive weight that is typically set to $^1/_N$. BLEU score represents the proficiency of an MT system and its higher value indicates better performance. The translation with a score between 20 to 29 is quite understandable (Cloud, 2021).

The proposed NMT model is implemented using the OpenNMT toolkit (Klein *et al*., 2017). To train the model, we have used a batch size of 4096 and neurons in FFNN of 2048. The word embedding size was 512. The encoder-decoder layer size was kept at 6. Adam optimizer (Kingma and Ba, 2015) is used for training the model with a dropout of 0.1. The values of alpha, beta1 and beta2 are 0.00031, 0.9 and 0.998, respectively. The PC in which the experiments were conducted had the following configuration: Processor of 7th Generation Intel® Core™ i5-7400 CPU @ 3.50GHz, GPU of NVIDIA Ge-Force GTX 1070Ti, 8 GB.

Experimental Result and Analysis

A number of experiments have been conducted to improve the performance of the proposed transformer model. Since the number of heads is an important issue, experiments have been performed on varying head numbers to identify the appropriate number. Figure 3 represents B2E and E2B BLEU scores at iteration 10,000 and 20,000 for different heads from 1 to 32. From the figure, it is observed that BLEU scores are different while the number of heads varied, but the scores are not correlated with numbers. Therefore, it is a matter of empirical study to identify the best-suited head number for a dataset. The best BLEU scores for B2E and E2B are 17.24 and 19.09, receptively, at 20,000 iterations when the number of heads is equal to two. It is also observed that the BLEU score at 20,000 iterations is always better than that of 10,000 iterations, which indicates more training steps may provide a better score.

Figure 4 shows the BLEU scores for training, validation and test sets for two heads while training continued for 100,000 iterations. From the figure, it is noticed that initially BLEU score improved rapidly but did not improve much after a certain number of steps. As an example, after 10,000 iterations, the E2B BLEU scores for training, validation and test set are 81.32, 12.57 and 13.16, respectively. On the other hand, at 50,000 iterations, the scores for the three sets are 91.2, 24.15 and 23.51, respectively. A similar observation is also achieved for B2E cases. Notably, the training set BLEU score is much better than validation and test sets for both B2E and E2B cases. The better BLEU score for the training set is logical because its samples are used for training the model and performance on the training set is a kind of memorization. Since the dataset provided a separate validation set, we presented a performance check on it

without using it in the training process. Thus, the act of validation set is similar to the test set in this study and achieved BLEU scores for validation and test set are almost the same. For any MT system, a test set BLEU score is always important, which recognizes the generalization ability of the system. The proposed transformer model with two heads achieved the best B2E and E2B test set BLEU scores (at 100,000 iterations) 21.42 and 25.44, respectively.

Table 2 compares the performance of the proposed transformer model with other prominent Bangla MT models on the basis of the achieved BLEU score on the SUPara test set. The existing methods are a phrase-based SMT and five deep learning-based recent NMT methods. The table has also mentioned a brief description of the datasets used in different methods. The exiting methods reported BLEU scores for the SUPara test dataset, while several methods considered a more extensive training set combining different datasets with the SUPara training set. The training dataset is augmented considering one or more datasets from among Indic Languages Multilingual Parallel Corpus (ILMPC) (Nakazawa *et al*., 2018), Six Indian Parallel Corpus (SIPC) (Post *et al*., 2012), Penn Treebank Bangla-English parallel corpus (PTB), Amader CAT (Hasan *et al.*, 2020) and GolbalVoices (Tiedemann, 2012) those contain ~337K, ~ 20K, 1313, 1,782 and 126,724 sentences, respectively. As an example, Hasan *et al*. (2019a) trained the transformer model with a base configuration with 419,109 sentences combining ILMPC, SIPC, PTB, SUPara and AmaderCAT. On the other hand, our transformer model with optimal heads with BPE is only trained with the SUPara training set (i.e., 70,861 sentences). The results indicate the computational proficiency of the proposed model.

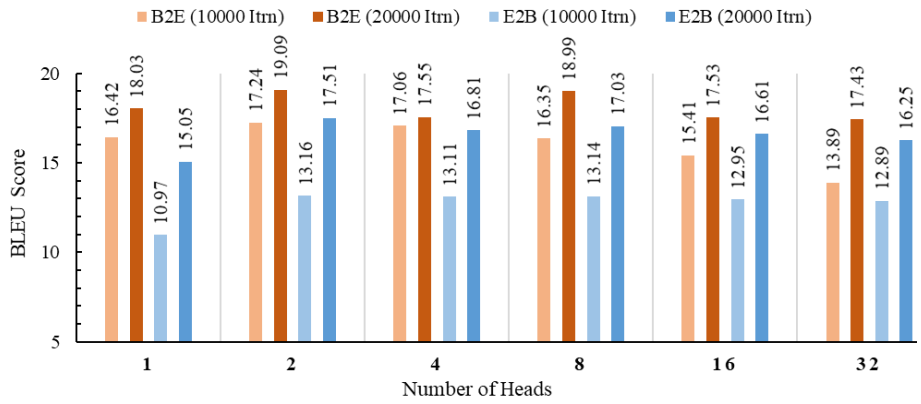

Fig. 3: Bangla to English (B2E) and English to Bangla (E2B) BLEU scores at 10,000 and 20,000 iterations for different heads from 1 to 32

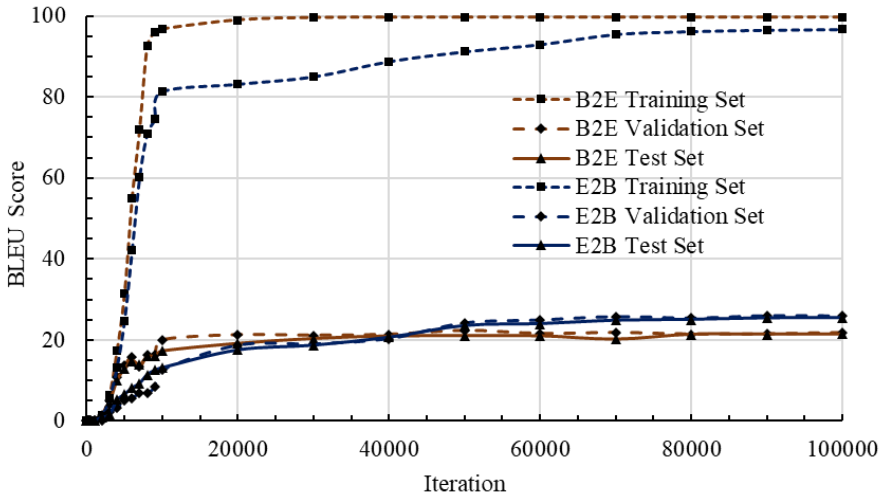

Fig. 4: Bangla to English (marked B2E) and English to Bangla (marked E2B) BLEU Score on training, validation and test sets varying iteration up to 100,000 for two heads

Table 1: Examples of preprocessing effect on sample sentences

	Available in Dataset *(English phonetic and meaning for Bangla sentence)*	Sentence after Preprocessing
Bangla	আমি আমার জন্মভূমিকে ভালবাসি। (English Phonetic: *Ami amar janmavumike bhalobashi* Meaning: *I love my motherland.)*	আমি আমার জন ্ ম@@ ভূমি@@ কে ভালবাসি ।
	তিনি প্রতিদিন সংবাদপত্র পড়েন। (English Phonetic: *Tini pratidin sambadpatra paren* Meaning: *He reads the newspaper every day.*)	তিনি প ্ রতিদিন সংবাদপত ্ র পড়েন ।
	তুমি বই পড়ছ। (English Phonetic: *Tumi bio paracha* Meaning: *You are reading a book*)	তুমি বই পড়@@ ছ ।
English	Please let me go. It's written on the ticket. Ignorance is similar to darkness.	Please let me go . it 's written on the ticket . I@@ g@@ nor@@ ance is similar to darkness .

Table 2: Comparison of the achieved BLEU scores on test set of the benchmark data for Bangla to English (B2E) and English to Bangla (E2B)

Work Ref., Year	Dataset	Training/Validation /Test Samples	Model used	Achieved test set BLEU score		Remarks on Training Set Formation
				B2E	E2B	
Al Mumin *et al*. (2019a)	SUPara, GlobalVoices	197388/500/500	Phrase based SMT	17.43	15.27	SUPara + GlobalVoices
Hasan *et al*. (2019b)	SUPara	70861/500/500	BiLSTM	19.76	-	SUPara
Al Mumin *et al*. (2019b)	SUPara, GlobalVoices	197338 /500/500	GRU	22.68	16.26	SUPara + GlobalVoices
Hasan *et al*. (2019a)	ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109 /500/500	BiLSTM	19.24	-	ILMPC + SIPC + PTB + SUPara + AmaderCAT
	ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109 /500/500	Transformer with base configuration	18.99	-	ILMPC + SIPC + PTB + SUPara + AmaderCAT
	SUPara	70861 /500/500	BiLSTM	19.98	-	SUPara
Proposed transformer based Model	SUPara	70861 /500/500	Transformer with optimal head	21.42	25.44	SUPara

From Table 2, it is observed that any deep learning-based NMT method outperformed the SMT method for both B2E and E2B. The SMT method (Al Mumin *et al*., 2019a) achieved a 17.43 BLEU score for B2E. For B2E, the existing attention-based GRU (Al Mumin *et al*., 2019b) method is shown the best BLEU score among existing deep learning-based methods, which is 22.68. The proposed method has shown the competitive performance having a BLEU score of 21.42. It is notable that the GRU method is trained with a large training set (197,338 samples) combining SUPara and GlobalVoices. In comparison, the proposed transformer model is trained with the SUPara training set with 70,861 samples. The existing transformer model (Hasan *et al*., 2019a) achieved a B2E BLEU score of 18.99 with the base configuration. The outperformance of the proposed model over the existing transformer model indicates the proficiency of model tuning and head selection.

Among the exiting methods presented in Table 2, only two studies considered E2B MT. The achieved E2B BLEU score by existing SMT and NMT methods are 15.27 and 16.26, respectively. On the other hand, the proposed model achieved an E2B BLEU score of 25.44, which is much better than the other two studies. Moreover, both existing methods are trained with samples combining SUPara and GlobalVoices datasets, whereas the proposed model uses only the SUPara training set. The table clearly demonstrates the proficiency of the proposed transformer model for Bangla-English language pair MT.

The reason behind the outperformance of the proposed model is the technique employment and the appropriate setting. In the proposed model, sub-word segmentation helped the model to guess rare words. In addition, a proper number of heads for the proposed model is identified through empirical study, which is two. The two heads enhance the ability of the model to put appropriate attention on different positions of words in a sentence which helps the model to perform better in combination with sub-word segmentation.

Conclusion

In this study, an MT system for the Bangla-English language pair has been proposed using the deep learning technique. Specifically, a standard transformer model is tuned to achieve better performance for B2E and E2B MT. It is identified that two heads in the model have performed better than a larger number of heads while tested on the benchmark dataset. The proposed model considered Byte Pair Encoding in preprocessing. The achieved BLEU scores higher than 20 for both B2E and E2B cases on the benchmark dataset, which indicates the model's translation proficiency.

The proposed model outperformed several leading MT methods in terms of the achieved BLEU score.

The present study also opens several research directions. This study has identified that the performance for Bangla to English is different from English to Bangla, although the training, validation and test sets were the same. Therefore, it will be interesting to investigate the effect of source-target interchange on MT performance. In this study, only SUPara training set is used to train the model; therefore, training with additional samples might improve the performance of the present model.

Acknowledgement

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions to update the manuscript.

Author's Contributions

Argha Chandra Dhar: Participated in design, conducted experiments, performed result analysis and contributed to the writing of the manuscript.

Arna Roy: Participated in the result analysis and prepared the initial the manuscript.

M. A. H. Akhand: Designed the research plan and organized the study, analyzed and interpreted results and prepared the manuscript

Md Abdus Samad Kamal: Participated in conception and design, reviewed the manuscript.

Nazmul Siddique: Participated in conception and design, contributed in model illustration and reviewed the manuscript.

Ethics

It has been testified by the authors that this article has not been submitted to be published in any other journal and contains no ethical issues.

References

Akhand, M. A. H., Ahmed, M., & Rahman, M. H. (2016). Convolutional neural network based handwritten Bengali and Bengali-English mixed numeral recognition. International Journal of Image, Graphics and Signal Processing, 8(9), 40. doi.org/10.5815/ijigsp.2015.09.06

Al Mumin, M. A., Shoeb, A. A. M., Selim, M. R., & Iqbal, M. Z. (2012). Supara: A balanced english-bengali parallel corpus. SUST Journal of Science and Technology, 16(2), 46-51.

Al Mumin, M. A., Seddiqui, M. H., Iqbal, M. Z., & Islam, M. J. (2019a). shutorjoma: An english↔ bangla statistical machine translation system. Journal of Computer Science (Science Publications). doi.org/10.3844/jcssp.2019.1022.1039

Al Mumin, M. A., Seddiqui, M. H., Iqbal, M. Z., & Islam, M. J. (2019b). Neural machine translation for lowresource English-Bangla. Journal of Computer Science, 15(11), 1627-1637. doi.org/10.3844/jcssp.2019.1627.1637

Alamgir, T., Arefin, M. S., & Hoque, M. M. (2016, May). 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) (pp. 932-937). IEEE. doi.org/10.1109/ICIEV.2016.7760137

Anwar, M. M. (2018). Bangla to English Machine Translation using Fuzzy Logic. International Journal of Computer Science and Information Security (IJCSIS), 16(11).

Anwar, M. M., Anwar, M. Z., & Bhuiyan, M. A. (2010, March). Structural Analysis of Bangla Sentences for Machine Translation. In Proceedings of International Conference On Computational Intelligence Applications, India (pp. 230-237).

Anwar, M. M., Anwar, M. Z., & Bhuiyan, M. A. A. (2009). Syntax analysis and machine translation of Bangla sentences. International Journal of Computer Science and Network Security, 9(8), 317-326. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.453.9442&rep=rep1&type=pdf

Arefin, M. S., Hoque, M. M., Rahman, M. O., & Arefin, M. S. (2015, May). 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) (pp. 1-6). IEEE. doi.org/10.1109/ICEEICT.2015.7307534

Ashrafi, S. S., Kabir, M. H., Anwar, M. M., & Noman, A. K. M. (2013). English to Bangla machine translation system using context-free grammars. International Journal of Computer Science Issues (IJCSI), 10(3), 144.

Babhulgaonkar, A. R., & Bharad, S. V. (2017, October). Statistical machine translation. In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) (pp. 62-67). IEEE.. doi.org/10.1109/ICISIM.2017.8122149

Bhattacharyya, P. (2015). Machine translation. CRC Press.

Chowdhury, S.A., (2013). Developing a Bangla to English Machine Translation System Using Parts Of Speech Tagging : A Review. J. Mod. Sci. Technol. 1, 113–119.

Cloud, G., 2021. AutoML: Evaluating Models . URL https://cloud.google.com/translate/automl/docs/evaluate (accessed 6.1.21).

Dandapat, S., & Lewis, W. (2018). Training deployable general domain mt for a low resource language pair: English–bangla. http://rua.ua.es/dspace/handle/10045/76023

Dandapat, S., Morrissey, S., Kumar Naskar, S., & Somers, H. (2010). Statistically motivated example-based machine translation using translation memory. http://doras.dcu.ie/16041/

Davydova, O. (2017). 10 Applications of Artificial Neural Networks in Natural Language Processing . medium.com. URL.

Francisca, J., Mia, M. M., & Rahman, S. M. (2011). Adapting rule based machine translation from english to bangla. Indian Journal of Computer Science and Engineering (IJCSE), 2(3), 334-342. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.5089&rep=rep1&type=pdf

Gage, P. (1994). A new algorithm for data compression. C Users Journal, 12(2), 23-38. https://www.derczynski.com/papers/archive/BPE_Gage.pdf

Garg, A., & Agarwal, M. (2018). Machine translation: A literature review. arXiv preprint arXiv:1901.01122. https://arxiv.org/abs/1901.01122

Green, S., Cer, D., & Manning, C. D. (2014, June). Phrasal: A toolkit for new directions in statistical machine translation. In Proceedings of the ninth workshop on statistical machine translation (pp. 114-121). https://aclanthology.org/W14-3311.pdf

Haque, M., & Hasan, M. (2018, October). English to bengali machine translation: An analysis of semantically appropriate verbs. In 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET) (pp. 217-221). IEEE. doi.org/10.1109/ICISET.2018.8745626

Hasan, M. A., Alam, F., Chowdhury, S. A., & Khan, N. (2019, December). Neural machine translation for the Bangla-English language pair. In 2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE. doi.org/10.1109/ICCIT48885.2019.9038381

Hasan, M. A., Alam, F., Chowdhury, S. A., & Khan, N. (2019, September). Neural vs statistical machine translation: Revisiting the bangla-english language pair. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-5). IEEE. doi.org/10.1109/ICBSLP47725.2019.201502

Hasan M. A., Alam F., Noori S. R. H. (2020). A Collaborative Platform to Collect Data for Developing Machine Translation Systems. In: Uddin M., Bansal J. (eds) Proceedings of International Joint Conference on Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. doi.org/10.1007/978-981-13-7564-4_35

Hutchins, J. (2005). The history of machine translation in a nutshell. Retrieved December, 20(2009), 1-1.

Hutchins, W. J. (1995). Machine translation: A brief history. In Concise history of the language sciences (pp. 431-445). Pergamon. doi.org/10.1016/b978-0-08-042580-1.50066-0

Hutchins, W. J. (Ed.). (2000). Early years in machine translation: memoirs and biographies of pioneers (Vol. 97). John Benjamins Publishing.

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2014). On using very large target vocabulary for neural machine translation. arXiv preprint arXiv:1412.2007. doi.org/10.3115/v1/P15-1001

Kalchbrenner, N., & Blunsom, P. (2013, October). Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1700-1709). https://aclanthology.org/D13-1176.pdf

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810. https://arxiv.org/abs/1701.02810

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions (pp. 177-180). doi.org/10.3115/1557769.1557821

Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206. doi.org/10.3115/v1/P15-1002

Maxime, 2019. An Introduction to Transformers and Sequence-to-Sequence Learning for Machine Learning . Medium. URL. https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04 (accessed 6.1.21).

Mukta, A. P., Mamun, A. A., Basak, C., Nahar, S., & Arif, M. F. H. (2019, February). A phrase-based machine translation from English to Bangla using rule-based approach. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE. doi.org/10.1109/ECACE.2019.8679456

Muntarina, K., Moazzam, M. G., & Bhuiyan, M. A. A. (2013). Tense based English to Bangla translation using MT system. International Journal of Engineering Science Invention, 2(10), 30-38.

Nakazawa, T., Sudoh, K., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W.P., Kunchukuttan, A., Kurohashi, S., 2018. 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. doi.org/10.18653/v1/d19-5201

Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., & Ritz, A. (2020, September). Transforming the language of life: Transformer neural networks for protein prediction tasks. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 1-8). doi.org/10.1145/3388440.3412467

Naskar, S., Saha, D., & Bandyopadhyay, S. (2004). Anubaad–a hybrid machine translation system from english to bangla. simple'04. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.8386&rep=rep1&type=pdf#page=91

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318). doi.org/10.3115/1073083.1073135

Post, M., Callison-Burch, C., & Osborne, M. (2012, June). Constructing parallel corpora for six indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 401-409). https://aclanthology.org/W12-3152.pdf

Quirk, C., Menezes, A., & Cherry, C. (2005, June). Dependency treelet translation: Syntactically informed phrasal SMT. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05) (pp. 271-279). doi.org/10.3115/1219840.1219874

Rabbani, M., Alam, K. M. R., & Islam, M. (2014, May). A new verb based approach for English to Bangla machine translation. In 2014 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 1-6). IEEE. doi.org/10.1109/ICIEV.2014.6850684

Rabbani, M., Alam, K. M. R., Islam, M., & Morimoto, Y. (2016). PVBMT: A Principal Verb based Approach for English to Bangla Machine Translation. Int. J. Comput. Vis. Signal Process, 6(1), 1-9. http://ahadvisionlab.com/cennser/IJCVSP/finalPaper/060102.pdf

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better language models and their implications OpenAI blog. https://openai.com/blog/better-language-models/

Rahman, M. S., Mridha, M. F., Poddar, S. R., & Huda, M. N. (2010, October). Open morphological machine translation: Bangla to English. In 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM) (pp. 460-465). IEEE. doi.org/10.1109/CISIM.2010.5643495

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15). doi.org/10.1101/622803

Roy, M., & Popowich, F. (2010a, May). Phrase-Based statistical machine translation for a low-density language pair. In Canadian Conference on Artificial Intelligence (pp. 273-277). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-13059-5_27

Roy, M., & Popowich, F. (2010b, May). Word reordering approaches for Bangla-English statistical machine translation. In Canadian Conference on Artificial Intelligence (pp. 282-285). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-13059-5_29

Salam, K. M. A., Yamada, S., & Nishino, T. (2013). How to translate unknown words for English to Bangla Machine Translation using transliteration. Journal of computers, 8(5), 1167-1174. doi.org/10.4304/jcp.8.5.1167-1174

Salam, K. M. A., Yamada, S., & Tetsuro, N. (2017, July). Improve example-based machine translation quality for low-resource language using ontology. In International Conference on Applied Computing and Information Technology (pp. 67-90). Springer, Cham. doi.org/10.2991/ijndc.2017.5.3.6

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909. doi.org/10.18653/v1/p16-1162

Siddique, S., Ahmed, T., Talukder, M., Azam, R., & Uddin, M. (2021). English to Bangla Machine Translation Using Recurrent Neural Network. arXiv preprint arXiv:2106.07225. doi.org/10.18178/ijfcc.2020.9.2.564

Stahlberg, F. (2020). Neural machine translation: A review. Journal of Artificial Intelligence Research, 69, 343-418. . doi.org/10.1613/JAIR.1.12007

Sumita, E., & Iida, H. (1991, June). Experiments and prospects of example-based machine translation. In 29th Annual Meeting of the Association for Computational Linguistics (pp. 185-192). doi.org/10.3115/981344.981368

Tiedemann, J. (2012, May). Parallel data, tools and interfaces in OPUS. In Lrec (Vol. 2012, pp. 2214-2218). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.2874&rep=rep1&type=pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, M., Gong, L., Zhu, W., Xie, J., & Bian, C. (2018, October). Tencent neural machine translation systems for wmt18. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers (pp. 522-527). doi.org/10.18653/v1/w18-6429

Xuan, H. W., Li, W., & Tang, G. Y. (2012). An advanced review of hybrid machine translation (HMT). Procedia Engineering, 29, 3017-3022. doi.org/10.1016/j.proeng.2012.01.432