International Journal of Computer Vision and Signal Processing, 11(1), 25-31(2021) ORIGINAL ARTICLE

Bangla-English Neural Machine Translation with Bidirectional Long Short-Term Memory and Back Translation

Arna Roy, Argha Chandra Dhar, M. A. H. Akhand*

Department of Computer Science and Engineering,
Khulna University of Engineering & Technology,
Khulna 9203, Bangladesh

Md Abdus Samad Kamal*

Graduate School of Science and Technology,
Gunma University, Kiryu 376-8515, Japan

Abstract

Machine translation (MT) has recently drawn attention to the automatic translation of the text, documents, or webpages from one language to another. Among various MT approaches, neural MT (NMT) is the most feasible method, a data-driven approach consisting of special neural networks. Among thousands of natural languages, remarkable efforts on MT are concentrated on a few languages only; and the research is very limited for many major languages such as Bangla. The study aims to build an effective NMT system for Bangla-English MT. Bidirectional Long Short-Term Memory (BiLSTM), a popular deep learning method for sequential data operation, is considered in the present study. Attention mechanism with the BiLSTM model and a special data augmentation mechanism, called Back Translation (BT), are the significant features of the proposed model. The proposed model outperforms the prominent models for Bangla to English MT while tested on a benchmark dataset.

Contribution of the Paper: A BiLSTM with attention mechanism is proposed that is trained considering BT and found effective for low-resource Bangla-English MT cases.

Keywords: Machine Translation, Neural Machine Translation, BiLSTM, Back Translation.

ISSN: 2186-1390 (Online)
http://cennser.org/IJCVSP

Article History:
Received: 29 April 2021
Revised: 18 July 2021
Accepted: 25 August 2021
Published Online:1 September 2021

1. INTRODUCTION

Content (e.g., voice, speech, texts) translation from one natural language to another is essential in politics, business, research, and other areas. Translation through human experts has been well known decades back. In line with rapid globalization, developing an intelligent translation system is highly desired. In this context, machine translation (MT) [1] has drawn attention recently for the automatic translation of the text, documents, or webpages from one language to another. Among various MT approaches, neural MT (NMT) is the most feasible method [2], which is a data-driven approach that consists of special neural networks. Due to the fast development of deep learning methods, NMT is becoming the most promising MT field. Among thousands of natural languages around the globe, remarkable efforts on MT are concentrated on a few languages. The task of MT is inherently language-dependent because data preparation (e.g., training data for NMT) is a vital task. Furthermore, a particular MT system might not be effective for other different language pairs. A number of remarkable researches are available with plenty of resources and have achieved reliable performance for English-French [3], English-German [4], English-Chinese [5]. On the other hand, MT resources and works

*Corresponding author

Email addresses: `roy1707018@stud.kuet.ac.bd` (Arna Roy), `dhar1707069@stud.kuet.ac.bd` (Argha Chandra Dhar), `akhand@cse.kuet.ac.bd` (M. A. H. Akhand), `maskamal@ieee.org` (Md Abdus Samad Kamal)

on the Bangla language are very limited, although it is a rich language with approximately 228 million native speakers. Hence, it can be favorable to fill the gap and enhance Bangla-English MT towards its widespread use in the global community.

Several methods have been reported on the Bangla-English language pair in recent years. Pioneer Bangla-English MT schemes can be categorized into the rule-based MT (RBMT) and statistical MT (SMT). Most of the RBMT schemes use the grammatical rules generated by human experts, underscoring verbs, propositions, phrases, and other features along with linguistic information [6, 7, 8, 9]. Among the SMT methods, the phrase-based SMT system [10] is a notable one. Besides the pioneer MT schemes, a few recent studies have been reported on NMT. Hasan et al. investigated the Bidirectional Long Short-Term Memory (BiLSTM) approach and found its performance better than the SMT scheme [11]. In another study [12], they also examined BiLSTM and Transformer-based NMT schemes. Mumin et al. investigated the attention mechanism with Byte Pair Encoding (BPE) using NMT [13] and compared them with the SMT. Attention-based NMT with BPE seems to provide a good result on the benchmark dataset.

This study aims to build an effective NMT system for the Bangla-English language pair. BiLSTM [14], the popular deep learning method for sequential data operation, is considered in the present study. Attention mechanism with BiLSTM model and a special data augmentation mechanism, called Back Translation (BT), are the significant features of the proposed model. The proposed model outperforms the prominent models for Bangla to English MT while tested on a benchmark dataset.

The rest of the paper is organized as follows. Section 2 describes the proposed BiLSTM with the BT model. Section 3 presents the experimental results of the proposed model and compares the performance with prominent existing methods. Finally, Section 4 concludes the paper with a few observed remarks.

2. BANGLA-ENGLISH MACHINE TRANSLATION (BEMT)

BiLSTM with an attention mechanism is proposed for BEMT. Specifically, the BiLSTM is trained on the preprocessed corpus considering the BT technique. The following subsections briefly describe the data preprocessing, the used model, and the BT training mechanism.

2.1. Data Preprocessing

Data preprocessing is an essential task for an MT system. Preprocessing includes tokenization, true-casing, normalizing punctuation, and removing non-printable characters from the data. The maximum number of words in a sentence is limited to 40 in preprocessing, as long sentences may cause problems. Byte Pair Encoding (BPE) algorithm is applied to the corpus to handle the rare words. Preprocessing depends on the specific data used, and the details of the preprocessing method will be explained in the experimental studies section. The dataset SUPara [15] is considered as the principal benchmark dataset in this study. Additionally, GlobalVoices [16] is used as a secondary dataset to use its English sentences for the BT purpose.

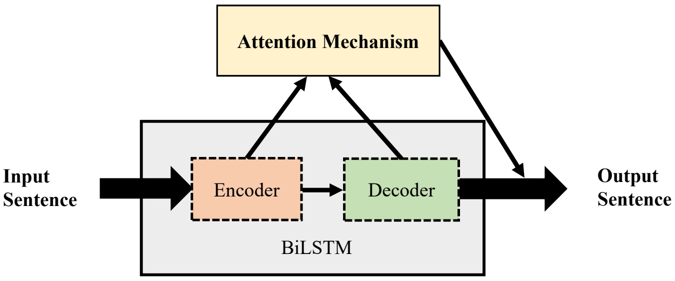


Figure 1: The detailed operational view of BiLSTM with attention mechanism for BEMT.

2.2. BiLSTM with Attention Mechanism

Bidirectional LSTM (BiLSTM) [17] is an improvement of unidirectional LSTM. BiLSTM consists of two LSTMs that receive inputs from both the forward and backward directions. In this way, the output layer receives information from both the backward and forward directions [18]. Thus, the input sequence is fed in the forward order for one network and the reverse order for another. The outputs of the two networks are usually concatenated at each training step, although there are other options, e.g., summation. BiLSTM is useful for the scenarios where the output at a certain time step depends on the previous and future time steps, e.g., machine translation, speech recognition [19].

BiLSTM with attention mechanism is the latest NMT model considered in this BEMT. The BEMT model is roughly divided into three major modules: the encoder, decoder, and attention modules. Fig. 1 depicts the block diagram of the model showing interactions among the individual modules. The word sequence of the source language (i.e., Bangla) goes into the encoder part. Then the outputs of both the encoder and decoder pass to the attention mechanism, which later helps generate the output word sequences in the target language (i.e., English). Fig. 2 shows the detailed operational view of BiLSTM with attention mechanism for BEMT where LSTM cells are the main building blocks (square boxes in the encoder and decoder). Operations among the encoder, decoder, and attention mechanisms are crucial in the model and are briefly described below.

1) Encoder: The encoder reads a sentence, given as $X = (\text{word}_1, \text{word}_2, \ldots, \text{word}_t)$, from both directions. The words are converted into an embedded form of definite size, then fed into the LSTM blocks. The encoder consists of two types of LSTM cells: forward and backward. The forward LSTM cells read the sentence from word_1 to word_t, and the backward LSTM cells do the same in the reverse order.

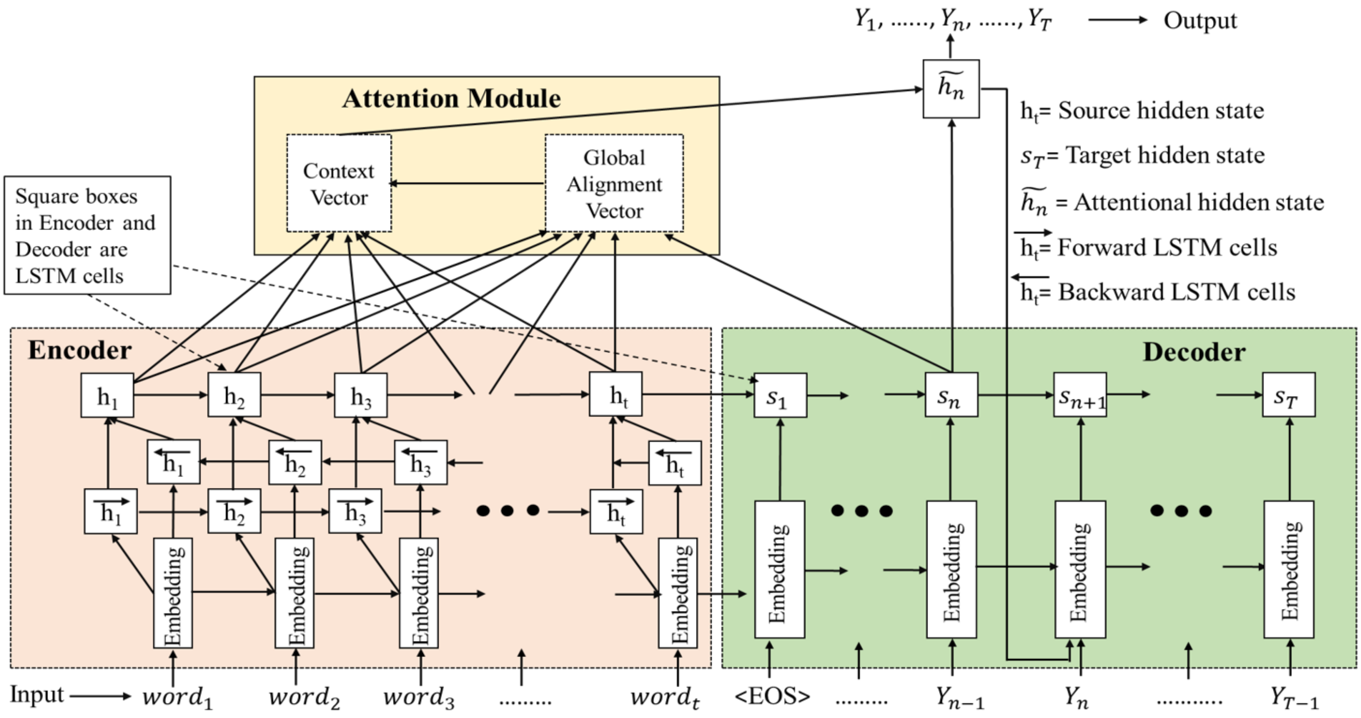

$Y_1, \ldots\ldots, Y_n, \ldots\ldots, Y_T$ → Output
$\widetilde{h_n}$
h_t= Source hidden state
s_T= Target hidden state
$\widetilde{h_n}$ = Attentional hidden state
$\overrightarrow{h_t}$= Forward LSTM cells
$\overleftarrow{h_t}$= Backward LSTM cells
Attention Module
Context Vector
Global Alignment Vector
Square boxes in Encoder and Decoder are LSTM cells
Encoder
Decoder
h_1 h_2 h_3 h_t
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ $\overleftarrow{h_t}$
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ $\overrightarrow{h_t}$
s_1 s_n s_{n+1} s_T
Embedding
Input → $word_1$ $word_2$ $word_3$ .... $word_t$ <EOS> .... Y_{n-1} Y_n .... Y_{T-1}

Figure 2: The detailed operational view of BiLSTM with attention mechanism for BEMT.

After that, the two hidden states are concatenated to get the final source hidden states $h_1, \ldots, h_t$. When the task of the encoder comes to an end, the decoder starts working.

2) Decoder: The decoder decides the appropriate parts and pays attention to them. Here, the last source hidden state of the encoder is fed into the first target hidden state of the decoder. In general, a decoder at a particular time step n takes an attentional hidden state ($\tilde{h}_{n-1}$) and output (Y_{n-1}) of the previous time step as inputs to generate embeddings. The output of the corresponding embedding and the previous time step target hidden state (s_{n-1}) help to generate the current target hidden state (s_n). Later the s_n and the corresponding context vector from the attention module are simply concatenated to compute the current attentional hidden state $\tilde{h}_n$. The $\tilde{h}_n$ is then fed into the Softmax layer to produce output, Y_n.

3) Attention Module: The main task of the attention module is to calculate how much attention should be paid to a particular input word to generate a particular output word. The attention mechanism used in this study is called the global attention mechanism [20]. The attention module consists of two vectors: a context vector and a global alignment vector. The global alignment vector is computed by comparing the current target hidden state s_n from the decoder module with all the source hidden states (i.e., $h_1, \ldots, h_t$). Then the context vector is calculated by taking the weighted average of the overall source hidden states, which contributes to the output of the target hidden state (s_n).

2.3. Training with Back Translation (BT)

The BT technique [21] is employed in training the proposed BEMT. BT is a simple data augmentation method that can be used for a low resource language pair to improve training performance [22]. The BT method requires a unidirectional dataset of the target language. The process is achieved by training the target on the source language at first. After training, the unidirectional target dataset is fed into the model to generate a dataset of the source language. Finally, the generated source language and unidirectional target language dataset are added to the training data. Considering SUPara as the principal dataset and GolbalVoices as a secondary dataset, the training steps in the proposed BEMT with BT are as follows:

Step 1:
The model is trained from English to Bangla using the SUPara training dataset.

Step 2:
The English sentences from GlobalVoices are fed into the model to obtain the outputs as Bangla sentences.

Step 3:
The GlobalVoices English sentences and the output Bangla sentences (from **Step 2**) are added to the SUPara training set (i.e., training set augmented); then, the model's final training is performed.

3. EXPERIMENTAL STUDIES

This section describes the experimental results of the proposed NMT system. The performance has also been

Table 1: Examples of original and preprocessed sentences for both modes (Bangla to English and English to Bangla) of translation.

Mode of Translation	**Sentence Available in the Dataset** *(with English phonetic and meaning)*	**Sentence after Preprocessing**
From Bangla to English	আমি আমার জন্মভূমিকে ভালবাসি। (English Phonetic: *Ami amar janmavumike bhalobashi* Meaning: *I love my motherland*.)	আমি আমার জন ◌্ ম@@ ভূমি@@ কে ভালবাসি ।
	আমরা সমাজে বাস করি। (English Phonetic: *Amra samaje bas kari* Meaning: *We live in society*.)	আমরা সমাজে বাস করি ।
	অজ্ঞতা অন্ধকারের সামিল। (English Phonetic: *Aggota andhakarer samil* Meaning: *Ignorance is similar to darkness*.)	অজ ◌্ ঞতা অন ◌্ ধকারের সা@@ মিল ।
From English to Bangla	I love my motherland.	I love my motherland.
	We live in a society.	We live in a society.
	Ignorance is similar to darkness.	I@@ g@@ nor@@ ance is similar to darkness.

compared with the notable existing methods.

3.1. Benchmark Data and Preprocessing

In this study, SUPara is used as the primary dataset, and GlobalVoices is used as a supportive dataset for BT. GlobalVoices dataset contains 126477 English sentences, where some sentences contain non-printable characters and words from different languages (e.g., Arabic). A total of 49695 sentences are chosen for BT from the dataset by filtering out sentences containing non-printable characters and words from other languages. Using the filtered English sentences, a total of 49695 Bangla sentences were produced by BT. The SUPara dataset contains 70861, 500, and 500 Bangla-English sentences for training, validation, and test purposes, respectively. Therefore, the merged dataset contains 120556 (=70861+49695), 500, 500 sentences for training, validation, and test, respectively. Next, the data tokenization, true casing, normalizing punctuation, and removing non-printable characters are performed on the data using Moses [23]. It changes the raw sentences into the number of tokens where words and punctuation marks are parted by a space.

Comparatively, a small-sized corpus may result in a poor dictionary and can arise a large number of rare word problems. Therefore, subword segmentation is applied using the BPE algorithm. The BPE algorithm counts the frequency of each word in a corpus, and a special stop symbol $< /\mathrm{w} >$ is added at the end of each token. Characters are then separated. After that, the algorithm finds out the most frequent consecutive byte pairs and merges them into one token. Table 1 provides some example sentences before and after preprocessing. For example, in the sentence “Ignorance is similar to darkness,” the BPE algorithm can distinguish ‘n,’ ‘o’, and ‘r’ as consecutive frequent token pairs and thus merge them into one token ‘nor’. The same explanation goes for ‘I’, ‘g’ and ‘ance’ tokens. So, BPE algorithm divides the word ‘ignorance’ into four sub-words: ‘I’, ‘g’, ‘nor’ and ‘ance’ by adding ‘@@’ between them. The maximum sentence length is kept at 40 words, as long sentences may cause problems. After applying all the processes, the dataset retains 115550, 500, 500 training, validation, and test sentences, respectively.

3.2. Performance Evaluation and Experimental Setup

For performance evaluation, the Bilingual Evaluation Understudy (BLEU) [24] score is measured, which is currently one of the popular evaluation methods in the MT field. With a value in the range of 0 to 1, a BELU score indicates the closeness of a translation to the reference ones. BLEU is a precision-oriented measurement implemented in mainly three steps. At the first step, n-gram or the number of word matches are calculated at the system outputs and the reference sentences. Note that in the computational linguistics, n-gram is defined as a contiguous sequence of n words (or items) from a given text or speech corpus. Next, the candidate counts are trimmed by their corresponding maximum reference values. Then the clipped n-grams are summed and divided by the number of candidate n-grams. Through this step, the modified precision score (p_n) is found as

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

where *Candidates* denotes the complete corpus and C denotes a hypothesis sentence. The second step is the BLEU Brevity Penalty (BP) factor calculation, which is give by

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{1-r/c} & \text{otherwise,} \end{cases} \tag{2}$$

where c is the length of the candidate translation, and r is the length of the reference translation. Finally, the BLEU score is the geometric mean of the precision scores and calculated as

$$BLEU = BP \cdot e^{\sum_{n=1}^{N} w_n \log p_n} \tag{3}$$

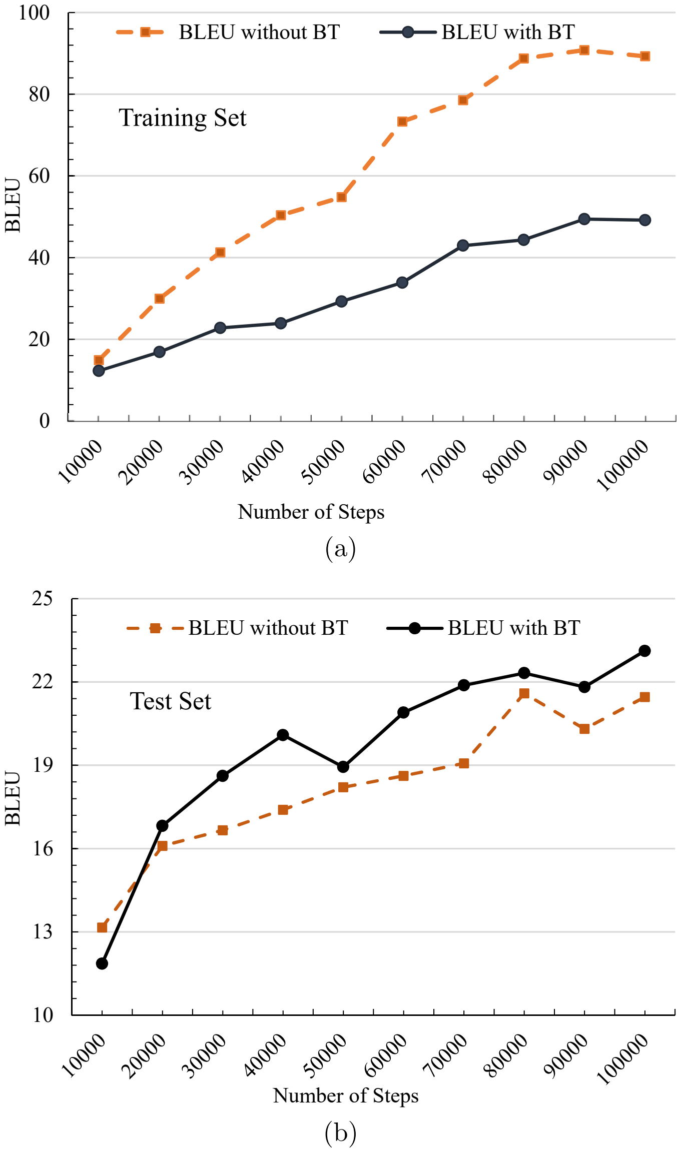
where N is typically 4 and w_n is a positive weight typically set to $1/N$.

The proposed NMT model is implemented using the OpenNMT toolkit [25]. We use BiLSTM having a bidirectional encoder and a unidirectional decoder architecture, each containing two layers. Individual encoders and decoders consist of LSTM blocks. As an attention mechanism, we considered a global one [20]. The Stochastic Gradient Descent [26] is used to train the model. The learning rate was 1, and the dropout was 0.3. The PC used to conduct the experiments has the following configuration: CPU 7th Generation Intel® Core™ i5-7400 3.50GHz, RAM 8 GB, and NVIDIA GPU Ge-Force GTX 1070Ti 8 GB.

3.3. Experimental Results and Performance Comparison

Training with BT is a fundamental task in the proposed BiLSTM based BEMT model. For the effectiveness assessment of BT, experiments were conducted with and without BT. Fig. 3 shows BLEU scores of the model for both training and test sets, for varying training up to 100000 steps with and without BT. For the training set, the achieved BLEU scores at the end of 100,000 training steps are 49.18 and 89.28 for the cases with and without BT, respectively. It is noticeable that the BLEU score is low when BT is included. The reason is that the training is performed on SUPara training sentences additionally to the generated sentences through BT, but the performance is measured on SUPara samples only. The BT inclusion has improved the generalization ability of the model. Therefore, the BiLSTM model with the BT performed better and achieved a better BLEU score than the BiLSTM model without BT on the test set, which is unseen in the training process. The proposed BiLSTM model with BT achieved a BLEU score of 23.12; whereas, the model without BT achieved a 22.88 BLEU score. In any machine learning system, performance on the test set is more desirable as it indicates the ability to work on unnoticed cases (i.e., generalization ability). In this context, a better BLEU score on the test set with BT is promising for the Bangla-English language pair.

Table 2 compares performance on the test set of the proposed model with other prominent works on BEMT with the SMT and NMT approaches. For a fair comparison, the BLEU score of the proposed model is measured on the SUPara test set only since the existing methods BLEU scores are also on the same SUPara test set. An individual model with embedding and dataset information also has been mentioned for better realization of individual studies. Achieved BLEU scores are on 500 SUPara test samples for any method, while the training sets are different (in size and sample) among the methods. From Table 2, it is observed that any NMT method outperforms the SMT method. The SMT method [10] earned a BLEU score of 17.43, which is the lowest value in the table. Among the existing NMT methods, attention-based NMT and attention-based NMT with BPE [13] achieved the most satisfactory BLEU scores, which are 22.38 and 22.68, respectively. Among the existing BiLSTM models, BiLSTM with word embedding [12] achieved the best BLEU score of 19.98. This model consists of an encoder and a decoder [14], and the parameters have been initialized by using a



(a)

(b)

Figure 3: BLEU score on both training and test sets with and without back translation for BEMT.

Table 2: Comparison of the achieved BLEU scores on SUPara test.

Work Ref., Year	Dataset	Train. / Val./ Test Set Size	Model	Embedding	BLEU Score
Mumin et al. [10], 2019	SUPara, GlobalVoices	197388 /500/500	Phrase based SMT	–	17.43
Hasan et al. [11], 2019	SUPara	70861 /500/500	BiLSTM	Bangla and English training dataset	19.76
Hasan et al. [12], 2019	ILMPC, SIPC, PTB SUPara, AmaderCAT	419109 /500/500	BiLSTM	Bangla and English training dataset	19.24
	ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109 /500/500	BiLSTM	Bangla training dataset	19.40
	ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109 /500/500	Transformer	–	18.99
	SUPara	70861 /500/500	BiLSTM	Bangla and English training dataset	19.98
Al Mumin et al. [13], 2019	SUPara, GlobalVoices	197338 /500/500	BiGRU + Attention	–	22.38
			BiGRU + Attention +BPE	–	22.68
The Proposed Model	SUPara	70861 /500/500	BiLSTM + Attention+ BPE	–	22.88
	SUPara, GlobalVoices	115550 /500/500	BiLSTM +Attention + BPE + BT	–	23.12

pretrained word embedding system. On the other hand, the proposed model excels the existing models by achieving a BLEU score of 23.12 for the BiLSTM model with the BT technique. BT helped to increase the training set through data augmentation. Besides, the incorporation of an appropriate attention mechanism and subword segmentation algorithm BPE to handle the rare words have increased the proposed model's performance.

Another critical observation from Table 2 is that the proposed model achieved the best BLEU score for a relatively smaller number of training samples. A few existing models have used huge training datasets along with SUPara training data. For example, the BiLSTM model with embeddings [12] was trained with 419109 sentences combining Indic Languages Multilingual Parallel Corpus (ILMPC), Six Indian Parallel Corpus (SIPC), Penn Treebank Bangla-English parallel corpus (PTB), SUPara, and AmaderCAT. The best performed existing method [13] (BLEU score of 22.68) considered 197338 training samples. On the other hand, the proposed BiLSTM model with attention mechanism has been trained with SUPara and GlobalVoices training samples (i.e., 115550), incorporated with BT. Then again, the proposed model has been trained with only the SUPara training sample (i.e., 70861), without BT. Achieved BLEU scores for both cases are better than any existing method, reflecting the proposed models' computational efficacy over others.

4. CONCLUSIONS

This study has presented the Bangla-English Machine Translation scheme using the BiLSTM by incorporating the attention mechanism. The BilSTM is sequentially trained with the back-translation mode applying the augmented sentences after preprocessing the original sentences obtained from a benchmark dataset. The proposed MT scheme is tested on both the datasets SUPara and GlobalVoices. It is found that the back translation-based training significantly improved the training performance. Furthermore, in the case of smaller training samples, the proposed model achieved the best BLEU score. The proposed MT is also compared with the other NMT methods, and the achieved BLEU scores are found to be the best among the existing methods. Such experimental results demonstrate the computational efficacy and translating performance of the proposed models over others.

References

[1] W. J. Hutchins, Machine translation: A brief history, in: Concise history of the language sciences, Elsevier, 1995, pp. 431–445.

[2] F. Stahlberg, Neural machine translation: A review 69 (2020) 343–418.

[3] T. Luong, I. Sutskever, Q. Le, O. Vinyals, 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), Beijing, China, 2015, pp. 11–19.

[4] S. Jean, K. Cho, R. Memisevic, 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.

[5] M. Wang, L. Gong, W. Zhu, J. Xie, C. Bian, Tencent neural machine translation systems for WMT18, in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 2018, pp. 522–527.

[6] A. P. Mukta, A.-a. Mamun, C. Basak, S. Nahar, F. H. Arif, A Phrase-Based Machine Translation from English to Bangla Using Rule-Based Approach, in: 2019 Int. Conf. Electr. Comput. Commun. Eng., 2019, pp. 1–5.

[7] J. Francisca, M. M. MIA, D. S. M. M. RAHMAN, Adapting Rule Based Machine Translation From English To Bangla, Indian J. Comput. Sci. Eng. 2 (3).

[8] M. Rabbani, K. M. R. Alam, M. Islam, Y. Morimoto, PVBMT: A Principal Verb based Approach for English to Bangla Machine Translation, Int. J. Comput. Vis. Signal Process. 6 (1) (2016) 1–9.

[9] M. Rabbani, K. M. R. Alam, M. Islam, A new verb based approach for English to Bangla machine translation, in: 2014 Int. Conf. Informatics, Electron. Vis., 2014, pp. 1–6. doi:10.1109/ICIEV.2014.6850684.

[10] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, M. J. Islam, shu-torjoma : An EnglishBangla Statistical Machine Translation System, J. Comput. Sci. 15 (7) (2019) 1022–1039.

[11] M. A. Hasan, F. Alam, S. A. Chowdhury, N. Khan, Neural vs Statistical Machine Translation: Revisiting the Bangla-English Language Pair, in: 2019 Int. Conf. Bangla Speech Lang. Process., no. September, 2019, pp. 1–5. doi:10.1109/ICBSLP47725.2019.201502.

[12] M. A. Hasan, F. Alam, S. A. Chowdhury, N. Khan, Neural Machine Translation for the Bangla-English Language Pair, in: 2019 22nd Int. Conf. Comput. Inf. Technol., 2019, pp. 1–6.

[13] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, M. J. Islam, Neural Machine Translation for Low-resource English-Bangla, J. Comput. Sci. 15 (11) (2019) 1627–1637.

[14] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (1997) 1735–1780.

[15] M. Z. Iqbal, SUPara : A Balanced English-Bengali Parallel Corpus, SUST J. Sci. Technol. 16 (2) (2012) 46–51.

[16] J. Tiedemann, Parallel data, tools and interfaces in opus, in: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012.

[17] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR abs/1409.0473.

[18] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (11) (1997) 2673–2681.

[19] A. Graves, N. Jaitly, A. R. Mohamed, Hybrid speech recognition with Deep Bidirectional LSTM, in: 2013 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2013 - Proc., 2013. doi:10.1109/ASRU.2013.6707742.

[20] T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.

[21] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale (2018).

[22] A. Currey, A. V. Miceli Barone, K. Heafield, Copied Monolingual Data Improves Low-Resource Neural Machine Translation, in: Proc. Second Conf. Mach. Transl., 2017, pp. 148–156.

[23] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open source toolkit for statistical machine translation, 2007, pp. 177–180.

[24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, 2002, p. 311–318.

[25] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. Rush, OpenNMT: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations, Vancouver, Canada, 2017, pp. 67–72.

[26] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proc. COMPSTAT 2010 - 19th Int. Conf. Comput. Stat. Keynote, Invit. Contrib. Pap., 2010. doi:10.1007/978-3-7908-2604-$3_1$6.