A Neural Attention-Based Encoder-Decoder Approach for English to Bangla Translation

Abdullah Al Shiam, Sadi Md. Redwan, Md. Humaun Kabir, Jungpil Shin

Abstract

Machine translation (MT) is the process of translating text from one language to another using bilingual data sets and grammatical rules. Recent works in the field of MT have popularized sequence-to-sequence models leveraging neural attention and deep learning. The success of neural attention models is yet to be construed into a robust framework for automated English-to-Bangla translation due to a lack of a comprehensive dataset that encompasses the diverse vocabulary of the Bangla language. In this study, we have proposed an English-to-Bangla MT system using an encoder-decoder attention model using the CCMatrix corpus. Our method shows that this model can outperform traditional SMT and RBMT models with a Bilingual Evaluation Understudy (BLEU) score of 15.68 despite being constrained by the limited vocabulary of the corpus. We hypothesize that this model can be used successfully for state-of-the-art machine translation with a more diverse and accurate dataset. This work can be extended further to incorporate several newer datasets using transfer learning techniques.

Keywords: Neural Machine Translation (NMT), Machine Translation (MT), Encoder-Decoder Model, Neural Attention.

ACM CCS 2020: Computing methodologies-Artificial intelligence-Machine learning.

MSC 2020: 68T50.

doi:10.56415/csjm.v31.04

1 Introduction

Bangla is one of the most common languages spoken worldwide, with approximately 300 million native speakers and another 37 million as second language speakers. However, reliable machine translation (MT) systems for the language have not been implemented until very recently. Machine Translation (MT) is the translation of text from one natural language (source language) to another language (target language) using a computerized system with or without human interaction [1]. MT is an automated system associated with Natural Language Processing (NLP), which uses other language resources and bilingual datasets to build language and phrase models for text translation. Ideally, MT is a batch process that is applied to a given text to produce a perfect translated text [2]. The aim is to fill the communication gap between different societies with language diversity. Manual human translation is time-consuming for any language. But an efficient MT system can reduce both the time and cost involved in the translation. English being the language of choice internationally, is used in most documents, papers, journals, books, and records in today's world. Subsequently, MT systems like google translate have been popularized by most native Bangla speakers for English-to-Bangla translation. Since there is an abundance of comprehensive English articles online, translation of Bangla to English is much more accurate compared to translating English to Bangla. This is due to the fact that most Bangla translations available online are not refined enough for an NLP-based approach that can capture the nuance and subtlety of the language. The Bangla language has some differences from English and some other languages since Bangla has a large and diverse vocabulary, while the exact words can have different contextual meanings. It is multi-disciplinary research to facilitate machine translation systems to capture the contextual meaning of a sentence while translating to other languages. Hence it is essential for researchers to study the literature describing the differences between specific language pairs to explain the critical mistakes made by the systems and optimize them accordingly. The neural machine translation (NMT) approach for English-to-Bangla translation has been proposed in a recent study [3]. Using the SUPara and Glob-

alVoices corpus, it achieved a BLEU score of +0.30 and +0.69 over previously implemented MT systems. This study has shown NMT to be more efficient and accurate compared to PBMT (phrase-based MT) systems such as shu-torjoma [4] and other SMT (statistical MT) systems [5]. Taking these findings into account, we aim to demonstrate the potential adequacy of NMT systems using a larger dataset, namely the CCMatrix dataset [6],[7]. The proposed encoder-decoder model trained with the CCMatrix dataset has achieved a BLEU score of 15.68 and a NIST (National Institute of Standards and Technology) score of 5.12 in English-to-Bangla translation.

2 Related Works

Traditional MT systems include Direct-Based MT systems that translate individual words in a sentence at a time from one language to another using a phrasebook. Corpus-Based MT relies on the study of bilingual text corpora. Statistical MT (SMT) and Example-based MT fall under this category. SMT is good for catching exclusions to rules. The primary advantage of the SMT is that it does not require philological information in the translation process. Knowledge-Based MT, on the other hand, requires to be formed based on ontology and the semantic web. Lexical-Based MT systems translate individual words with lexical information [8]. The encoder-decoder model has emerged relatively recently and has been successful in many state-of-the-art translation frameworks [9]. With the growing popularity of machine learning models, NMT has been established as the new baseline for MT systems. Contemporary NMT systems have been modernized by Google [10] with phrase alignment [11] and attention mechanisms [12]. A comparison of NMT and SMT systems in recent bilingual studies is given in Table 1.

The first notable Bangla-to-English MT system is the phrase-based MT method proposed in 2010 [15]. Rule-based machine translation (RBMT) was first proposed in a 2012 study to include assertive-affirmative, negative, and interrogative sentences [16]. In the following year, a tense-based (TBMT) system has been proposed [17]. Recurrent neural networks have also seen some success in English-to-Bangla translation over the years [18]. [19] used the neural encoder-decoder

Table 1. Comparison of NMT and SMT Systems for English to other Languages [13], [14]

Language	System	BLEU
DE (German)	SMT, NMT	41.5, 61.2
EL (Greek)	SMT, NMT	47.0, 56.6
PT (Portuguese)	SMT, NMT	57.0, 59.9
PT (Portuguese)	SMT, NMT	41.9, 57.3

model for text normalization. On the basis of recent deep learning work, [20] proposed a bidirectional encoder-decoder model for addressing the problem of Arabic NER, in which the encoder and decoder are bidirectional LSTMs. Character-level embeddings are used in addition to word-level embeddings, and they are combined via an embedding-level attention mechanism. [21] proposes a novel Multimodal Encoder-Decoder Attention Networks (MEDAN). The MEDAN is composed of cascaded Multimodal Encoder-Decoder Attention (MEDA) layers that can capture rich and reasonable question features as well as image features by associating question keywords with important object regions in the image. In conclusion, NMT has become the cornerstone of most recent works on Bangla-English translation [3], [22]. Some of the relevant studies are discussed in Table 2.

3 Corpus Data

3.1 Dataset Description

We obtain the CCMatrix dataset from OPUS [13]. CCMatrix is built in a multilingual sentence space using a margin-based bitext mining technique, resulting in many parallel sentences and tokens in different languages. Training NMT systems for multiple language pairs was used to assess the quality of the mined bitexts. English-Bangla paral-

Table 2. Novel Bangla-English MT Studies

System	Year	Author
Phrase-Based MT	2010	Islam Z et al. [23]
RBMT	2012	Rhaman MK et al. [24]
TBMT	2013	Muntarina K et al. [25]
LSTM-RNN	2019	Islam MS et al. [16]
NMT	2019	Hasan MA et al. [26]
RNN	2020	Siddique S et al. [27]
Attention-based NMT	2021	Abujar S et al. [17]

lel datasets with more than 10M combined tokens available in OPUS and two other corpora relevant to this work are shown in Table 3 for comparison. The CCMatrix dataset is chosen in this work since the dataset is entirely built using the available sentence pairs on the internet, and it is not curated by human experts. The rationale behind choosing the dataset is to enable the MT system to access a wide range of vocabulary without being constrained by a specific corpus or human expertise. Another justification is that the CCMatrix dataset is also one of the most extensive datasets in terms of the number of tokens, as shown in Table 3.

Table 3. Corpus Data

Corpus	Bn Tokens	En Tokens
WikiMatrix v1	35.3M	1000M
wikimedia v2021040	10.3M	349.2M
CCMatrix v1	88.6M	98.7M
CCAligned v1	37.8M	38.9M
Tanzil v1	6.1M	5.6M
GlobalVoices	3.3M	4.9M
SUPara	244K	202K

3.2 Data Pre-processing

After downloading the dataset, the raw text data is preprocessed in several steps as shown in Figure 1. First of all, we perform Unicode normalization and split the punctuations. Then we add a 'start' and 'end' token to each sentence. Then each sentence is cleaned by removing special characters (if any). Then a word index is created and reversed so that each 'token id' points to a unique word.

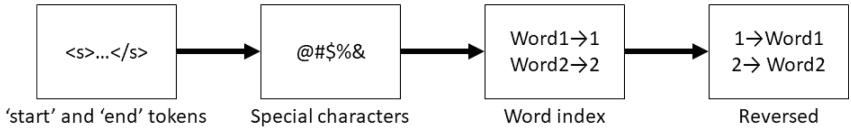

Figure 1. Preprocessing of the raw text

Traditional encoding schemes include sub-word regularization [28] and byte pair encoding (BPE) [29], which have been used in some NMT systems in the past. However, these encoding methods have been shown to be suboptimal for pretraining in some cases [30]. As an alternative, we vectorize the text data after the tokenization step in this work. Vectorization refers to transforming the strings into an array of token indices by using the Tensorflow [31] TextVectorization layer. This layer implements an 'adapt' method that reads the input epoch-by-epoch, much like model training. Importantly, the final dictionary of tokens and words is also used to decode the output of the model.

4 Proposed Framework

4.1 Encoder-Decoder Model

The encoder-decoder model operates such that the conditional probability of the target sentence given the source sentence is maximized. The encoder transforms the input phrase sequence into a dense vector form, and the decoder takes that representation and converts it into subsequent word sequences [23]. This leads to the model's performance being constrained by the maximum sentence length in training data.

Any sentence longer than the maximum length may lead to inaccurate translation. The neural attention mechanism is used to optimize this model with more flexibility. Instead of directly encoding the input sentence into a fixed-length vector, this method converts it into a sequence of vectors. When translating the encoded text, a subset of these vectors is chosen by employing this mechanism [24], [25], as shown in Figure 2.

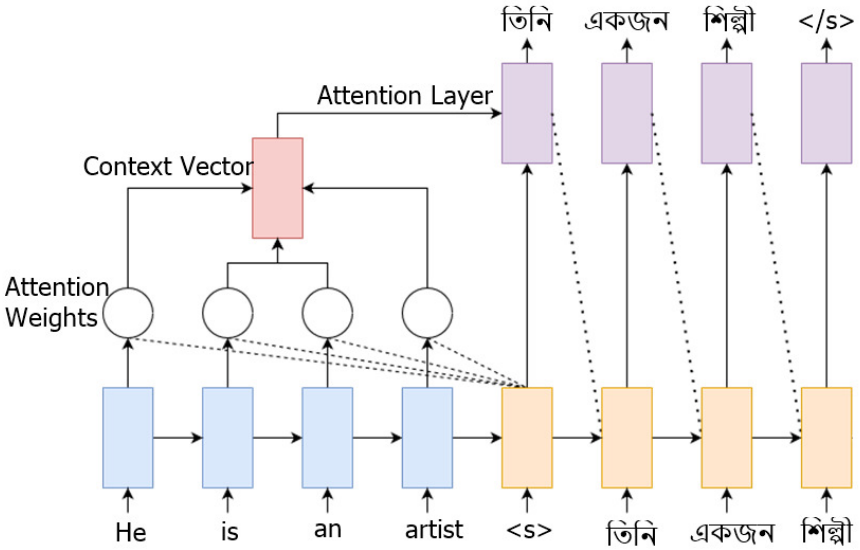

Figure 2. Neural Attention-Based Encoder-Decoder

The encoder part of the model is a Gated Recurrent Unit (GRU). This can also be implemented using bidirectional LSTM. The GRU is initialized with Glorot uniform and acts as the encoder combined with an embedding layer. The uniform distribution U[-a,a,] is defined as follows.

$$a = \sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}, \tag{1}$$

where n_{in} is the number of input neurons in the weight tensor, and n_{out} is the number of output neurons in the weight tensor.

4.2 Neural Attention

The attention or alignment mechanism for the encoder-decoder model retains all the input sentence's hidden states during the decoding phase. The attention model proposed by Bahdanau et al. used in this work produces hidden states for each of the elements in the input sequence from the encoder. The alignment score of each encoder output with respect to the decoder inputs and hidden state is then calculated at each step by multiplying the decoder's hidden state by all of the encoder's hidden states defined as follows.

$$e^t = \left[s_t^T h_1, \ldots, s_t^T h_{T_x} \right]. \tag{2}$$

Then the probability distribution is calculated as follows.

$$\alpha^t = \text{softmax}\left(e^t\right). \tag{3}$$

This gives the context vector as follows

$$a_t = \sum_{i=1}^{T_x} \alpha_i^t h_i. \tag{4}$$

The decoder uses this context vector to generate new hidden states. The decoder produces a tensor, which is then passed through a text vectorization layer to produce the final output.

4.3 Experimental Setup

The encoder-decoder model is implemented using Keras [27] with an embedding dimension of 256 and 1024 units in each layer. The encoder uses the Embedding layer to produce an embedding vector for each token in an input array of tokens and then transforms the vectors using the GRU layer. The processed sequence is passed as the attention inputs, implemented by the Additive Attention layer. The decoder does the exact same thing and uses the output of the GRU layer as the 'query' to the attention layer. The model is trained using an RTX 2060 GPU with Nvidia CUDA 11.2. The loss function used for training is Sparse Categorical Cross-entropy, and the optimizer is the Adam

optimizer [32]. The total number of trainable parameters is 41,650,940. In this paper, the model is trained for 30 epochs and tested on the Tatoeba dataset [33].

5 Experimental Results and Discussion

The model training results are shown in Figure 3.

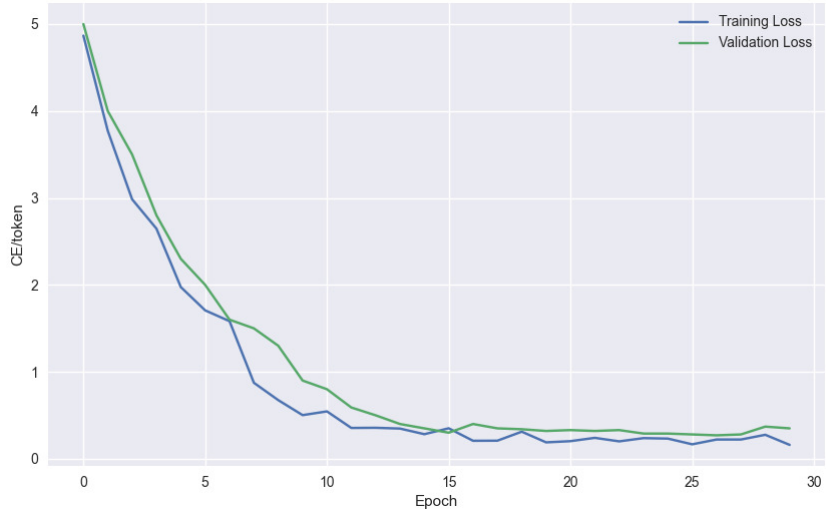

Figure 3. Training and validation loss curves during the model training. CE/token denotes cross-entropy loss per token in training and validation data

The BLEU score [34] is the primary metric used to quantify the model's performance. NIST calculates the score by giving more weight to the rarer correct n-gram [26], whereas BLEU measures edit distance using n-grams up to length four. The geometric average of modified n-gram precisions is used to calculate BLEU. The brevity penalty (BP) is then calculated as

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}, \tag{5}$$

where c is the length of the candidate translation and r is the length of the effective reference corpus. The BLEU score is as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \tag{6}$$

A higher BLEU score indicates improvements in translation. Table 4 shows the BLEU score of the proposed model compared with the current state-of-the-art models [4]. The proposed model's performance is on-par with the current best-performing models, as shown in Table 4. Interestingly, the BLEU score of the model is higher than the Phrase-based SMT, while the NIST score is marginally lower. The primary intuition is that the CCMatrix corpus has a large number of commonly used sentences, while the more complex and nuanced sentences are more frequent in the SUPara and GlobalVoices corpora. Further studies using other datasets built using text mining can lead to a better understanding of the model's performance.

Table 4. Translation Accuracy

System	Dataset	BLEU	NIST
Phrase-based SMT + large LM	SUPara, GlobalVoices	15.27	5.13
Attention-based NMT	SUPara, SUPara	15.57	4.72
Attention-based NMT with BPE	SUPara, GlobalVoices	16.26	5.18
Proposed Model	CCMatrix	15.68	5.12

The attention plot for a sample translation (How are you?-কেমন আছেন আপনি?) shows that the majority of the weights are concentrated on the diagonal of the matrix. This denotes which parts of the input sentence have the model's attention while translating, as shown in Figure 4.

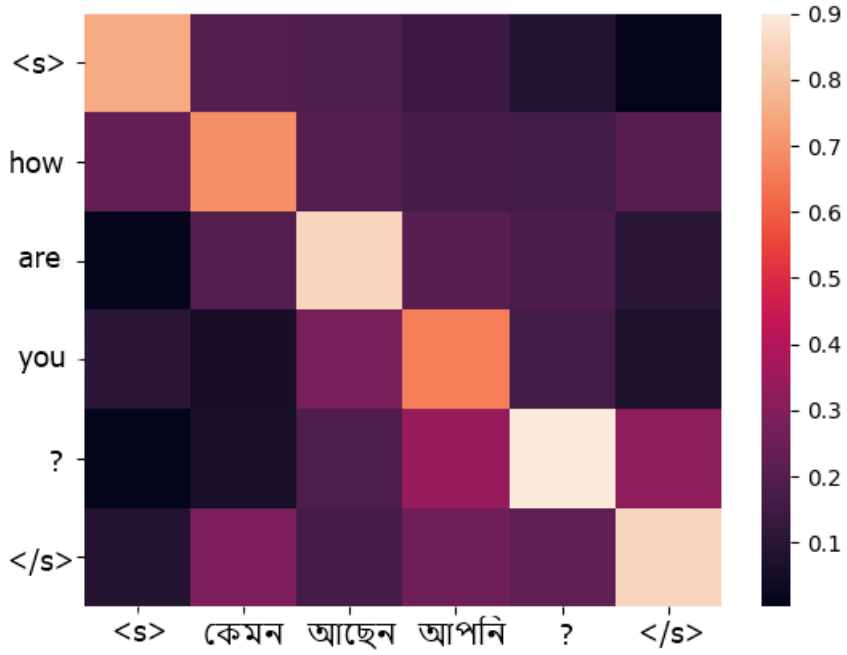

Figure 4. Attention plot of a sample translation

The model can be hypothetically improved upon by occasionally training with its own translations since the attention mechanism facilitates access to the previous output. Another aspect that demands further study is how well the model performs for a specific vocabulary and sentence length. Future works should also consider using transfer learning for a more generalized model that can adapt to multiple datasets. The model was not trained for Bangla-to-English translation, which should also be considered an important criterion for a generalized MT solution.

6 Conclusion

In this work, we have used a relatively newer approach for automated English-to-Bangla translation using the CCMatrix dataset. The system

can be extended further to deal with complex and compound sentences with a more diverse vocabulary for translation. The same approach can also be implemented for building a multi-language translation system. Furthermore, the inclusion of a more extensive database of bilingual dictionaries and adding more and more words to the lexicon can yield an even better BLEU score for uncommon words and sentence structures. Successful implementation can deliver a sound expert system for translating any text document from English to Bangla and vice versa.

7 Source Availability

The source code for this work is available at the following link https://github.com/sadiredwan/cbmt-en-bn under CC by 4.0. Please follow the ODC Attribution-Sharealike Community Norms and publish any derivative works under a similar open license.

References

[1] M. A. Cheragui, "Theoretical overview of machine translation." in *ICWIT*. Citeseer, 2012, pp. 160–169.

[2] A. H. Homiedan, "Machine translation," *African University, Adrar, Algeria*, 2012.

[3] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "Neural machine translation for low-resource english-bangla," *Journal of Computer Science*, vol. 15, no. 11, pp. 1627–1637, 2019.

[4] M. Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "shutorjoma: An english to bangla statistical machine translation system," *Journal of Computer Science (Science Publications)*, 2019.

[5] 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)*. IEEE, 2019, pp. 1–6.

[6] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, and A. Joulin, "Ccmatrix: Mining billions of high-quality parallel sentences on the web," *arXiv preprint arXiv:1911.04944*, 2019.

[7] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," *arXiv preprint arXiv:1911.00359*, 2019.

[8] A. Godase and S. Govilkar, "Machine translation development for indian languages and its approaches," *International Journal on Natural Language Computing*, vol. 4, no. 2, pp. 55–74, 2015.

[9] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, and Y. Liu, "Neural machine translation with explicit phrase alignment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1001–1010, 2021.

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[11] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[12] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[13] S. Castilho, J. Moorkens, F. Gaspari, I. Calixto, J. Tinsley, and A. Way, "Is neural machine translation the new state of the art?" *The Prague Bulletin of Mathematical Linguistics*, no. 108, 2017.

[14] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," *arXiv preprint arXiv:1609.04186*, 2016.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[16] M. S. Islam, S. S. S. Mousumi, S. Abujar, and S. A. Hossain, "Sequence-to-sequence bangla sentence generation with lstm recurrent neural networks," *Procedia Computer Science*, vol. 152, pp. 51–58, 2019.

[17] S. Abujar, A. K. M. Masum, A. Bhattacharya, S. Dutta, and S. A. Hossain, "English to bengali neural machine translation using global attention mechanism," in *Emerging Technologies in Data Mining and Information Security*. Springer, 2021, pp. 359–369.

[18] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 1, pp. 154–163, 2018.

[19] J. Tiedemann, "The tatoeba translation challenge–realistic data sets for low resource and multilingual mt," *arXiv preprint arXiv:2010.06354*, 2020.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[21] 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, pp. 138–145.

[22] J. Tiedemann and L. Nygaard, "The opus corpus-parallel and free: http://logos. uio. no/opus." in *LREC*, 2004.

[23] Z. Islam, J. Tiedemann, and A. Eisele, "English to bangla phrase-based machine translation," in *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, 2010.

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

[25] K. Muntarina, M. G. Moazzam, and M. A.-A. Bhuiyan, "Tense based english to bangla translation using mt system," *International Journal of Engineering Science Invention*, vol. 2, no. 10, pp. 30–38, 2013.

[26] 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)*. IEEE, 2019, pp. 1–6.

[27] S. Siddique, T. Ahmed, M. Talukder, R. Azam, M. Uddin *et al.*, "English to bangla machine translation using recurrent neural network," *arXiv preprint arXiv:2106.07225*, 2021.

[28] Y. Shibata, T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, and S. Arikawa, "Byte pair encoding: A text compression scheme that accelerates pattern matching," 1999.

[29] K. Bostrom and G. Durrett, "Byte pair encoding is suboptimal for language model pretraining," *arXiv preprint arXiv:2004.03720*, 2020.

[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[31] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[32] M. Lusetti, T. Ruzsics, A. Göhring, T. Samardzic, and E. Stark, "Encoder-decoder methods for text normalization," in *Proceedings*

of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), 2018, pp. 18–28.

[33] M. N. Ali and G. Tan, "Bidirectional encoder–decoder model for arabic named entity recognition," *Arabian Journal for Science and Engineering*, vol. 44, no. 11, pp. 9693–9701, 2019.

[34] C. Chen, D. Han, and J. Wang, "Multimodal encoder-decoder attention networks for visual question answering," *IEEE Access*, vol. 8, pp. 35 662–35 671, 2020.

Abdullah Al Shiam, Sadi Md. Redwan,
Md. Humaun Kabir, Jungpil Shin

Received November 30, 2022
Revised 1 – January 25, 2023
Revised 2 – February 22, 2023
Accepted February 28, 2023

Abdullah Al Shiam
ORCID: https://orcid.org/0000-0002-8787-5584
Department of Computer Science and Engineering,
Sheikh Hasina University,
Netrokona-2400, Bangladesh
E–mail: shiam.cse@shu.edu.bd

Sadi Md. Redwan
ORCID: https://orcid.org/0000-0002-1859-1617
Department of Computer Science and Engineering,
University of Rajshahi,
Rajshahi-6205, Bangladesh.
E–mail: sadi.redwan@ru.ac.bd

Md. Humaun Kabir
ORCID: https://orcid.org/0000-0001-7225-0736
Department of Computer Science and Engineering,
Bangamata Sheikh Fojilatunnesa Mujib Science and Technology University,
Jamalpur-2012, Bangladesh.
E–mail: humaun@bsfmstu.ac.bd

Jungpil Shin
ORCID: https://orcid.org/0000-0002-7476-2468
School of Computer Science and Engineering,
The University of Aizu Aizuwakamatsu,
Fukushima, 965-8580, Japan.
E–mail: jpshin@u-aizu.ac.jp