Journal of Computer Science

Original Research Paper

Neural Machine Translation for Low-resource English-Bangla

[1]Mohammad Abdullah Al Mumin, [2]Md Hanif Seddiqui,
[1]Muhammed Zafar Iqbal and [1]Mohammed Jahirul Islam

[1]Department of Computer Science and Engineering,
Shahjalal University of Science and Technology, Sylhet, Bangladesh
[2]Department of Computer Science and Engineering, University of Chittagong, Chittagong, Bangladesh

Article history
Received: 25-09-2019
Revised: 20-10-2019
Accepted: 13-11-2019

Corresponding Author:
Mohammad Abdullah Al Mumin
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh
Email: mumin-cse@sust.edu

Abstract: Neural machine translation has recently been able to gain state-of-the-art translation quality for many language pairs. However, neural machine translation has been less tested for English-Bangla language pair, two linguistically distant and widely spoken languages. In this paper, we apply neural machine translation to the task of English-Bangla translation in both directions and compare it against a standard phrase-based statistical machine translation system. We obtain up to +0.30 and +4.95 BLEU improvement over phrase-based statistical machine translation for English-to-Bangla and Bangla-to-English respectively. Due to low-resource and morphological richness of Bangla, English-Bangla translation task produces a large number of rare words. We apply subword segmentation with byte pair encoding to handle this rare words issue. We obtain up to +0.69 and +0.30 BLEU improvement over baseline neural machine translation for English-to-Bangla and Bangla-to-English respectively. We further investigate our system output for several challenging linguistic properties like subject-verb agreement, noun inflection, long distance reordering and rare words translation. We observe that neural machine translation with and without subword segmentation significantly outperform the phrase-based statistical machine translation system, thus establishing itself as the state-of-the-art technology for low-resource English-Bangla machine translation.

Keywords: English-Bangla Machine Translation, Low-Resource, Morphologically Rich, Neural Machine Translation

Introduction

In this era of globalization, every communication becomes gradually international and multilingual. To meet this demand of the globalization, automatic language translation called Machine Translation (MT) has become an attractive area of research. Bangla is the seventh most spoken language all over the world with an estimation of 250 million people in Bangladesh and the Indian subcontinent. As the internet and other communications are predominantly in English, machine translation between English and Bangla languages becomes a much-needed tool to promote this large Bangla spoken community as an active participant of this global world.

Recently, a new paradigm to machine translation, Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Cho *et al*., 2014b; Sutskever *et al*., 2014), has emerged that relies on neural network and has proved itself as competent to the state-of-the-art phrase-based statistical machine translation. Previous works on English-Bangla machine translation are mostly limited to conventional machine translation techniques. Neural machine translation for low-resource English-Bangla has not been explored so intensively yet.

English-Bangla is a low-resource language pair for translation task due to its little training data with only 197K sentence pairs. In addition, Bangla is a morphologically rich language which produces large vocabulary. These two scenarios make an English-Bangla translation system to observe a large number of rare words during training. Observation shows that sentences with many rare words likely to be translated much more badly than sentences containing mainly frequent words (Bahdanau *et al*., 2015; Sutskever *et al*., 2014). To handle this rare words problem, Sennrich *et al*. (2016) suggest NMT models that operate on the subword units during training and practically shows that the subword models improve over other models to handle rare words for the WMT15 translation tasks English-to-German and English-to-Russian significantly. However,

these two translation tasks have been trained on a large amount of training data. Now, it is worthwhile to observe the performance of these subword models on a low-resource translation task.

In this study, our aim is therefore two folds: Firstly, to present the result on the English-Bangla translation using neural machine translation. On both directions, we compare an attention-based neural machine translation system (Bahdanau *et al*., 2015) against a phrase-based statistical machine translation system (Koehn *et al*., 2007). Secondly, to present the result on the low-resource English-Bangla neural machine translation using subword segmentation. We use subword segmentation using byte pair encoding technique proposed in (Sennrich *et al*., 2016). The systems are trained on Shahjalal University Parallel (SUPara) (Mumin *et al*., 2012; 2018b) corpus and GolbalVoices (Tiedemann, 2012) corpus from OPUS (Tiedemann, 2012). In addition, the systems are tuned and evaluated on a balanced development and test dataset, respectively. We have reported our results using automatic evaluation metrics BiLingual Evaluation Understudy, BLEU (Papineni *et al*., 2002), NIST (Doddington, 2002) developed by National Institute of Standard and Technology, and Translation Error Rate, TER (Snover *et al*., 2006). In addition to BLEU and NIST, we have used TER because TER metric helps to perform linguistic error analysis of system output.

The organization of this paper is as follows: after reviewing the literature survey in Section 2, we introduce the theory of the neural machine translation and byte pair encoding in Section 3. Section 4 explains the experimental settings of our system followed by the results and discussion in section 5. Section 6 concludes this paper with some future directions.

Literature Survey

Machine Translation (MT) has been an active research topic since the 1950's (Hutchins, 2005). Since then, there are various approaches adopted by researchers such as rule-based MT (RBMT) (Al-A'ali, 2007; Algani and Omar, 2012; Alsaket and Ab Aziz, 2014; Dwivedi and Sukhadeve, 2010; Mohammed and Aziz, 2011; Shirko *et al*., 2010), corpus-based MT like example-based MT (EBMT) (Nagao, 1984) and Statistical MT (SMT) (Koehn *et al*., 2003), and hybrid-based MT (Costa-Jussa and Fonollosa, 2015), which is a combination of rule-based approaches and corpus-based approaches in order to overcome their limitations.

Since 2015, Neural Machine Translation (NMT) systems have been outperforming SMT for many translation tasks (Cettolo *et al*., 2015). Until now, NMT shows state-of-the-art performance for language pairs having large amounts of parallel corpora such as English-French with 12M-36M sentence pairs (Luong *et al*., 2015; Jean *et al*., 2015a) and English-German with 4.5M sentence pairs (Jean *et al*., 2015b). There are few works examining low-resource translation direction such as Turkish to English (Gülçehre *et al*., 2015) with 160K sentence pairs and English to Vietnamese (Luong and Manning, 2015) with 133K sentence pairs.

Previous works on English-Bangla machine translation have focused on using rule-based machine translation (Sinha *et al*., 1995; Mumin *et al*., 2000; Siddique *et al*., 2003; Asaduzzaman and Ali, 2003; Dasgupta *et al*., 2004), example-based machine translation (Bandyopadhyay, 2001; Saha and Bandyopadhyay, 2005; Naskar and Bandyopadhyay, 2006; Salam *et al*., 2017), phrase-based statistical machine translation (Roy, 2009; Haffari *et al*., 2009; Roy and Popowich, 2010a; 2010b; Islam *et al*., 2010; Pal *et al*., 2014; Pal and Naskar, 2016; Mumin *et al*., 2019), syntax-based statistical machine translation (Pal *et al*., 2016), and hybrid machine translation (Dandapat *et al*., 2010). Neural machine translation has been less examined for low-resource English-Bangla. In (Dandapat and Lewis, 2018), the authors examined neural machine translation between English and Bangla in both directions with synthetic data augmentation using back-translated data and using sub-word representation.

Preliminaries

We use attention-based neural machine translation and subword segmentation based on the byte pair encoding algorithm in our experiments. The discussion of these approaches is given below:

Neural Machine Translation

A neural machine translation system is a neural network that directly models the conditional probability $p(\mathbf{y}|\mathbf{x})$ of translating a source sentence, $\mathbf{x} = x_1,\ldots, x_m$, to a target sentence, $\mathbf{y} = y_1,\ldots, y_n$. It accomplishes such goal through the *encoder-decoder* framework (Sutskever *et al*., 2014; Cho *et al*., 2014b). The *encoder* neural network computes a fixed-length vector z for each source sentence. Based on that encoded vector, the *decoder* produces a translation, one target word at a time and thus, decomposes the conditional probability as:

$$\log p(y \mid x) = \sum_{j=1}^{n} \log p\left(y_j \mid y_{<j}, x, z\right) \quad (1)$$

The entire model is jointly trained to maximize the log-likelihood of the parallel training corpus with back-propagation through time as:

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}\left(y^{(n)} \mid x^{(n)}\right) \quad (2)$$

where, $\left(y^{(n)}, x^{(n)}\right)$ represents the *n*th sentence in parallel corpus of size N and θ denotes the set of all tunable parameters.

Curse of sentence length. An inherent issue with the basic encoder-decoder approach is that an encoder neural network needs to be able to summarize all the necessary information of a source sentence into a fixed-length vector. This makes it difficult for the encoder neural network to tackle with long sentences. Cho *et al*. (2014b) showed that actually the performance of a basic encoder-decoder declines quickly as the length of an input sentence increases.

Attention-Based NMT

Currently, state-of-the-art neural machine translation architecture is based on an attention-based encoder-decoder model (Bahdanau *et al*., 2015), which addresses the long sentences issue of a basic encoder-decoder. This attention-based encoder-decoder model extends an *attention mechanism* to the basic encoder-decoder model, which learns to align and translate jointly (Bahdanau *et al*., 2015). Moreover, the attention-based encoder-decoder model uses a bidirectional RNN (BiRNN) (Schuster and Paliwal, 1997) as encoder instead of RNN as used in a basic encoder-decoder framework. Figure 1 shows the schematic diagram of the attention-based NMT system.

Bidirectional Encoder. The encoder in an attention-based NMT system is a bidirectional RNN that reads an input sequence $x = (x_1,\ldots, x_m)$ and computes a forward sequence of hidden states $\left(\vec{h}_1,...,\vec{h}_m\right)$ and a backward sequence of hidden states $\left(\overleftarrow{h}_1,...,\overleftarrow{h}_m\right)$. The hidden states $\vec{h}_j$ and $\overleftarrow{h}_j$ are concatenated to get the annotation vector $h_j = \left(\vec{h}_j, \overleftarrow{h}_j\right)$. Each annotation h_j summarizes the entire sentence, with a strong focus on word x_j and the neighboring words. For the activation function of an RNN, Gated Recurrent Unit (GRU) (Cho *et al*., 2014a) and Long Shor-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) are usually used.

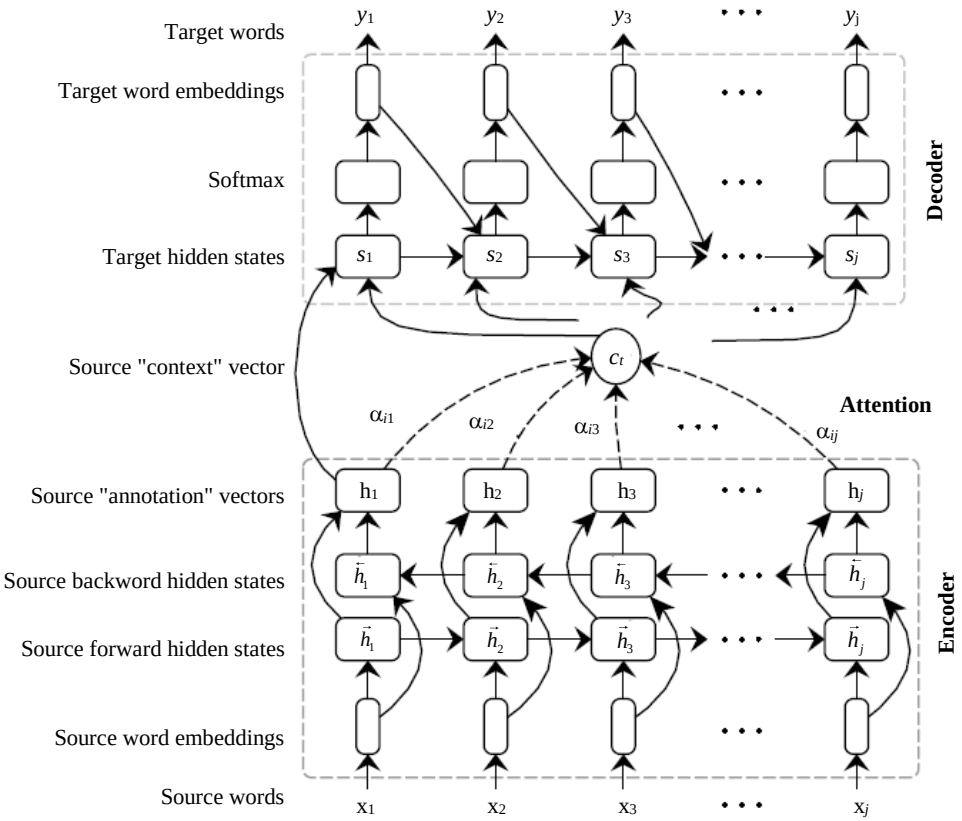

Fig. 1: Attention-based NMT architecture

Attentive Decoder. The decoder is a recurrent neural network that estimates a target sequence $y = (y_1,\ldots, y_n)$. Each word y_i is estimated based on a RNN hidden state s_i, the previously estimated word y_{i-1} and a context vector c_i.

The context vector c_i is computed through an attention mechanism for each target word y_i. Each time the attention-based NMT model produces a word in a translation, the attention mechanism searches for a set of locations in a source sentence where the most relevant information is concentrated. The context vector c_i is, then, computed as a weighted sum of the annotations h_j associated with these source positions:

$$c_i = \sum_{j=1}^{m} \alpha_{ij} h_j. \quad (3)$$

The weight α_{ij} of each annotation h_j is computed by:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})}, \quad (4)$$

where:

$$e_{ij} = a(s_{i-1}, h_j) \quad (5)$$

is an alignment model, which models the probability that y_i is aligned to x_j. The alignment model a is a single layer feed-forward neural network that is learned jointly with the rest of the network through backpropagation.

A detailed description can be found in Bahdanau *et al.* (2015). The training is conducted on a parallel corpus with stochastic gradient descent. For the translation, a beam search with a small beam size is applied.

Byte Pair Encoding

English-Bangla translation task observes a lot of rare words during training due to little training data and the morphological richness of Bangla language. To handle this problem, we segment words using Byte Pair Encoding (BPE) (Sennrich *et al*., 2016). Byte Pair Encoding, originally invented as compression algorithm (Gage, 1994) is transformed to word segmentation as follows:

Learning. First, every word in the training dataset vocabulary is expressed as a sequence of characters, plus an end-of-word symbol. All characters are included to the symbol vocabulary. Then, the most frequent symbol pair is selected and all its occurrences are merged, producing a new symbol that is included to the symbol vocabulary. The previous step is repeated until a defined number of merge operations have been learned.

Training. The defined list of merge operations, learned on the training dataset, can be exploited to any text to segment words into subword units that are in-vocabulary with respect to the training dataset (except for unseen characters).

Experimental Settings

In this section, we present the experimental setup of our systems for English-to-Bangla (En→Bn) and Bangla-to-English (Bn→En) translation task. We describe about the training, development and test dataset used in this experiment. We also describe about the preprocessing techniques applied to our dataset and the core system configuration used in our experiment. Finally, we mention about the evaluation metrics used to evaluate the results of our system.

Dataset

In our experiment, we used *Shahjalal University Parallel (SUPara)* (Mumin *et al*., 2012; 2018b) corpus and *GolbalVoices* (Tiedemann, 2012) corpus from OPUS (Tiedemann, 2012) as a training dataset. SUPara (Mumin *et al*., 2012, 2018b) is a balanced corpus consists of texts from different genres like literature, administrative texts, instructive texts, journalistic texts and texts treating external communication, which are collected from various printed and online media. GolbalVoices (Tiedemann, 2012) corpus consists of only news texts collected form GlobalVoices website[1]. The training dataset contains 197,338 sentences after performing preprocessing techniques on these two corpora. The statistics of the training dataset are given in Table 1.

We used the development dataset, *SUParadev2018* (Mumin *et al*., 2018a) for tuning our system and the test dataset, *SUParatest2018* (Mumin *et al*., 2018a) for evaluating our system's performance. Each of these datasets contains 500 sentences. These two datasets were developed with a vision of using them as a benchmark in English-Bangla MT research. The texts of these two datasets were well-chosen from balanced SUPara (Mumin *et al*., 2012, 2018b) corpus, thus these two datasets are also balanced in genre. In addition, to make these datasets representative in length we selected the texts from 10 subsets of different lengths: 1 to 5 words, 6 to 10 and so forth up to 40 to 45 and finally longer than 45 words. Finally, we tuned these two datasets by correcting misspellings and bad translations by a language expert.

[1] www.globalvoices.org

Table 1: Training dataset statistics-shown are the statistics of the data used in the systems. Data counts shown here are cleaned, normalized and tokenized for English (En) and Bangla (Bn) languages. English data are lowercased additionally.

Dataset	Total Sentences	Language	Total Tokens	Unique Tokens	Average Length
SUPara	70,614	En	980,004	31,215	13.88
		Bn	807,304	58,705	11.43
GlobalVoices	126,724	En	2,533,959	80,520	20
		Bn	2,320,431	124,749	18.31
Total	197,338	En	3,513,963	92,616	17.81
		Bn	3,127,735	154,390	15.85

Preprocessing

We performed several preprocessing techniques on our datasets. We filtered out texts containing foreign language characters, corrected misspellings and bad translations, normalized punctuations, tokenized texts and cleaned sentence pair with length ratio 1:5 and larger than 60 tokens in either side.

We normalized Bangla punctuations in Bangla side of our datasets using our tools and tokenized the normalized datasets using Bangla specific tokenizer [2]. We normalized, tokenized and lowercased English side of our datasets using the standard Moses (Koehn *et al*., 2007) scripts.

SMT System Configuration

We used Moses (Koehn *et al*., 2007) to build a standard phrase-based statistical machine translation system. Word alignment was extracted by GIZA++ (Och and Ney, 2003). We used the following options for alignment symmetrization and reordering model: *grow-diag-final-and* and *msd-bidirectional-fe*. KenLM (Heafield *et al*., 2013) was used as a language model and trained on Bangla monolingual corpus, *SUMono* (Mumin *et al*., 2014), of more than 32 million tokens and English monolingual corpus, *Europarl* (Koehn, 2005) of more than 27 million tokens. The details configuration of this SMT system is discussed in (Mumin *et al*., 2019).

NMT System Configuration

Model. We trained our NMT systems with Nematus (Sennrich *et al*., 2017) which is an implementation of the attention-based encoder-decoder model with small modifications to the attention-based encoder-decoder model in Bahdanau *et al*. (2015). We used Gated Recurrent Units (GRUs) (Cho *et al*., 2014a) for the recurrent neural networks. We extended the model with stacked architecture (Barone *et al*., 2017) by setting both encoder and decoder depths to 4. We used word embedding sizes of 512 and the hidden layer of size 1024. We used layer normalization (Ba *et al*., 2016) in encoder and decoder and tied the input embeddings of the decoder with the softmax output embeddings (Press and Wolf, 2016).

Vocabulary. We replaced words in English-Bangla parallel data whose frequencies are less than 5 by <unk>. As a result, our vocabulary sizes are 33.4K and 47K for English and Bangla respectively.

Training. Our training hyperparameters are: (a) we trained the model using Stochastic Gradient Descent (SGD) with adam (Kingma and Ba, 2014); (b) each update was computed using a minibatch of 80 sentence pairs; (c) we used a maximum sentence length of 100, (d) we set the initial learning rate of 0.0001, reshuffling the training corpus between epochs; (e) we used dropout with probability 0.2; (f) models were saved in every 10000 iterations; It took about 7-8 h to train a model on an NVidia GTX 1080Ti.

Decoding. We used beam search to approximately find the most likely translation given a source sentence. We used a beam width of 12 for decoding.

Subword Segmentation

As mentioned earlier, we used subword segmentation using byte pair encoding for neural machine translation. We applied Byte Pair Encoding (BPE) separately to the already tokenized training corpus and the number of merge rules is set to 59.5K, resulting in vocabularies of size 54.2K and 58.5K tokens for English and Bangla languages, respectively. Fig 2 shows the rare words scenario in the training data without and with subword segmentation using byte pair encoding. We observe that applying byte pair encoding reduces rare words (words with a frequency less than 5) by 27.77% from English side and by 57.12% from Bangla side. During the evaluation, subwords were reassembled. We used the publicly available script released by Sennrich *et al*. (2016).

[2] https://github.com/irshadbhat/indic-tokenizer

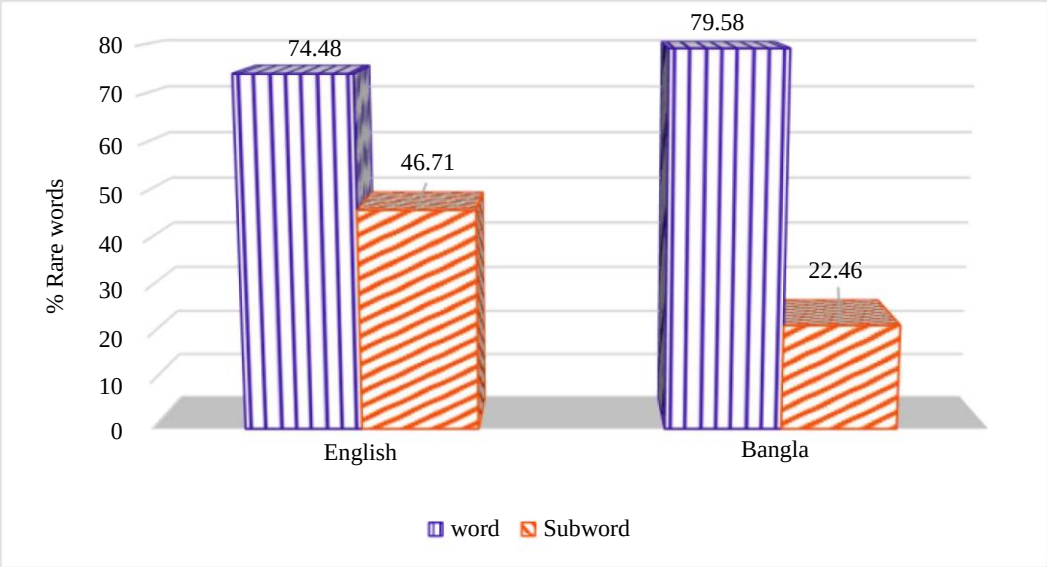

Fig. 2: Rare words scenario: Words having a frequency less than 5 in the training data are considered here as rare words. *word* represents rare words in the training data without segmentation and *subword* represents rare words in the training data with segmentation using *byte pair encoding*.

Evaluation Metrics

We used BiLingual Evaluation Understudy, BLEU (Papineni *et al*., 2002), NIST (Doddington, 2002) developed by National Institute of Standard and Technology and Translation Error Rate, TER (Snover *et al*., 2006) to evaluate the results of our system.

BLEU measures edit distance using n-grams up to length four. A higher BLEU score indicates improvements in translation.

NIST is based on the BLEU metric, but with some modifications. Whereas BLEU simply calculates n-gram precision score by giving equal importance in each n-gram, NIST calculates the score by giving more weight to the rarer correct n-gram. Small variations in translation length do not impact much in the NIST overall score. Like BLEU, a higher NIST score indicates improvements in translation.

TER measures the number of edits required to change a system output that matches a reference translation. It performs four edit operations, namely insertion, deletion, subtraction and phrasal shifts. Contrary to BLEU and NIST, a lower TER score indicates improvements in translation.

Results and Analysis

We have reported and interpreted results of our system from two viewpoints: Overall results and the translation behaviour with respect to the several challenging linguistic properties.

Overall Results

We evaluate low-resource English-Bangla machine translation for our three different systems: Phrase-based SMT, baseline attention-based NMT and attention-based NMT with Byte Pair Encoding (BPE). The results are shown in Table 2.

From Table 2, we observe that the baseline attention-based NMT system performs significantly better than the phrase-based SMT system. We conjecture that the continuous space representation of words and capturing the long distance context of a text through the attention mechanism make attention-based NMT to retain morphological form and syntactic structure of the target text better, thus making the translation quality better. However, baseline attention-based NMT without subword segmentation is poor to translate rare words as evidenced in the sample translation of Table 3c. This is reflected in the poor NIST score in the En→Bn direction as NIST gives more weight to the rarer correct n-gram.

Another important observation is that attention-based NMT system operating on subword unit using byte pair encoding shows significant improvements over baseline NMT system. This reflects that subword segmentation models make rare and unseen words into frequent segments, thus improve the translation quality.

We notice that all systems produce a better translation in Bn→En direction compare to *En→Bn* direction, confirming that it is difficult to generate morphologically rich words which corroborates the results reported by (Koehn, 2005).

Table 2: Translation results - shown are the tokenized BLEU, NIST, and TER scores of various systems on the *suparatest2018* (Mumin *et al*., 2018a) dataset. We highlight the **best** system in bold.

	En→Bn			**Bn→En**		
System	BLEU↑	NIST↑	TER↓	BLEU↑	NIST↑	TER↓
Phrase-based SMT + large LM	15.27	5.13	71.9	17.43	5.76	67.94
Attention-based NMT	15.57	4.72	68.54	22.38	5.98	59.88
Attention-based NMT with BPE	**16.26**	**5.18**	**68.69**	**22.68**	**6.07**	**60.09**

Table 3: *En→Bn*: Sample translations showing behaviour in translating *Subject-Verb agreement*, *Noun inflection*, and *Rare words*. For each example, we show the source (*src*), the human translation (*ref*), and the translation from our three systems: Phrase-based SMT system (*smt*), attention-based NMT (*base*), and attention-based NMT with BPE (*best*). We underlined the focal point in each category.

(a) *Subject-Verb agreement:*		
src	i am watching a nice movie.	
ref	আমি একটি সুন্দর চলচ্চিত্র দেখছি ।	
smt	আমি একটি ভালো চলচ্চিত্র দেখছে ।	×
base	আমি একটি সুন্দর ছবি দেখছি ।	✓
best	আমি একটি সুন্দর চলচ্চিত্র দেখছি ।	✓
(b) *Noun inflection:*		
src	i love my daughter.	
ref	আমি আমার মেয়েকে ভালবাসি ।	
smt	আমি ভালবাসি আমার মেয়ে ।	×
base	আমি আমার মেয়েকে ভালবাসি ।	✓
best	আমি আমার মেয়েকে ভালবাসি ।	✓
(c) *Rare words translation:*		
src	…thought that Samuel Johnson was the first man who published dictionary first	
ref	ধারণা করা হয় যে, স্যামুয়েল জনসন প্রথম মানুষ যিনি অভিধান প্রকাশ করেন	
smt	এটা ভেবেই স্যামুয়েল জনসন ছিলেন প্রথম ব্যক্তি যিনি প্রথম অভিধান প্রকাশ	✓
base	ভাবা হয় যে স্যামুয়েল জনসন, যিনি প্রথম <UNK> লেখা <UNK>	×
best	মনে করা হয় যে স্যামুয়েল জনসন, যিনি প্রথমে অভিধান প্রকাশ করেছিলেন	✓

Linguistic Behaviour

We present three sample translations for En→Bn and Bn→En translated by our three systems: Phrase-based SMT (*smt*), baseline attention-based NMT (*base*) and attention-based NMT using BPE (*best*).

En→Bn. Translating into a morphologically rich language from a morphologically poor language like translating from English to Bangla face difficulty to generate *subject-verb agreement* and *noun inflection* during translation. In Table 3a and 3b, we show the behaviour of our systems in generating these two features. We observe that the phrase-based SMT system is failed to generate correct translation in both cases whereas both NMT systems succeed.

In Table 3c, we show the behaviour of our systems in translating *rare words*. We observe that baseline NMT system without subword segmentation produces *UNK* word and SMT system's translation is very poor. However, the baseline attention-based NMT system overcomes this problem by applying subword segmentation on the training data.

Bn→En. *Long distance reordering* is a challenge during translation for syntactically different language pair, especially translating into fixed word order language like translating from Bangla to English. In Table 4a, we observe that both NMT systems are capable to retain long distance reordering in translation which is a problem in SMT system.

In Table 4b, we show the behaviour of our systems in translating sentences contain *negation*. We observe that the phrase-based SMT system mistakenly generate double negation whereas NMT systems generate correct translation.

In Table 4c, we show the behaviour of our systems in translating *rare words*. We observe that both the phrase-based SMT system and the baseline NMT system without subword segmentation are failed to translate rare words. However, by applying subword segmentation on the training data the attention-based NMT system overcomes this problem.

Table 4: Bn→En: Sample translations showing behaviour in translating *Long distance reordering*, *Negation* and *Rare words*. For each example, we show the source (*src*), the human translation (*ref*) and the translation from our three systems: Phrase-based SMT system (*smt*), attention-based NMT (*base*) and attention-based NMT with BPE (*best*). We underlined the focal point in each category.

(a) *Long distance reordering:*		
src	ভাল স্বাস্থ্য রক্ষার জন্যে শারীরিক ব্যায়াম অত্যাবশ্যক ।	
ref	physical exercise is very necessary to preserve good health.	
smt	good health for the need to protect the physical exercise is essential.	×
base	physical exercise is essential to protect good health.	✓
best	physical exercise is essential to protect good health.	✓
(b) *Negation:*		
src	আমি কখনো তাকে ভুলব না ।	
ref	i will never forget him.	
smt	i never not forget him.	×
base	i will never forget him.	✓
best	i will never forget him.	✓
(c) *Rare words translation:*		
src	যে সকল শিশু দারিদ্র্যসীমার নিচে বাস করে তারা স্বল্প-সুবিধাপ্রাপ্ত শিশু ।	
ref	the children who live under the poverty line are the underprivileged children.	
smt	that all children live under the poverty line they স্বল্প-সুবিধাপ্রাপ্ত children.	×
base	the children who live below the poverty line are of a <UNK> child.	×
best	the children who live below the poverty line are a small privileged child.	✓

Conclusion and Future Direction

In this paper, we investigate the performance of the Neural Machine Translation (NMT) compare to the phrase-based Statistical Machine Translation (SMT) on low-resource English-Bangla translation task in both directions. We also investigate the performance of NMT operating on subword segmentation of the training data to handle the rare word problems in the low-resource and morphologically rich English-Bangla translation. We further investigate our system output for several challenging linguistic properties which pose challenges in English-Bangla translation task, namely subject-verb agreement, noun inflection, long distance reordering and rare words translation.

Result shows that NMT systems improve the translation quality over phrase-based SMT system and NMT system operating on the subword units of the training data exhibits state-of-the-art performance for low-resource English-Bangla machine translation in both directions. We observe that NMT systems are more capable than phrase-based SMT system to generate subject-verb agreement, noun inflection and long distance reordering in their translation. In addition, the NMT system operating on the subword units of the training data can translate rare words efficiently. Thus, we conclude that NMT can be considered as the state-of-the-art model for low-resource language pair as well and NMT benefits from subword segmentation of the training data.

Future work may benefit from investigating hybrid word-char model to handle rare words and OOV issues which is very common in low-resource and morphologically rich language pair translation task.

Acknowledgement

The first author is grateful to Information and Communication Technology (ICT) Division, Government of People's Republic of Bangladesh for the grant to do this research work.

Funding Information

Mohammad Abdullah Al Mumin's work has been supported by ICT Division, Ministry of Posts, Telecommunications and IT, Government of the People's Republic of Bangladesh [Order No: 56.00.0000.028.33.077.17-78, date: 02.04.2018].

Author's Contributions

Mohammad Abdullah Al Mumin: Designed the research plan, organized and ran the experiments, contributed to the presentation, analysis and interperetation of the results, added and reviewed genuine content where applicable.

Md Hanif Seddiqui: Made considerable contributions to this research by critically reviewing the literature review and the manuscript for significant intellectual content.

M Zafar Iqbal: Supervised the study and made considerable contributions to this research by critically reviewing the manuscript for significant intellectual content.

Mohammed Jahirul Islam: Supervised the study and made considerable contributions to this research by critically reviewing the manuscript for significant intellectual content.

Conflict of Interest

The authors declare that they have no Conflict of Interest.

References

Al-A'ali, M., 2007. Pre-editing and recursive-phrase composites for a better English-to-Arabic machine translation. J. Comput. Sci., 3: 410-418. DOI: 10.3844/jcssp.2007.410.418

Algani, Z.A. and N. Omar, 2012. Arabic to English machine translation of verb phrases using rule-based approach. J. Comput. Sci., 8: 277-286. DOI: 10.3844/jcssp.2012.277.286

Alsaket, A.J. and M.J. Ab Aziz, 2014. Arabic-Malay machine translation using rule-based approach. J. Comput. Sci., 10: 1062-1062. DOI: 10.3844/jcssp.2014.1062.1068

Asaduzzaman, M. and M.M. Ali, 2003. Morphological analysis of Bangla words for automatic machine translation. Proceedings of the 3rd International Conference on Computer and Information Technology, (CIT' 03), Dhaka, pp: 271-276.

Ba, J.L., J.R. Kiros and G.E. Hinton, 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bahdanau, D., K. Cho and Y. Bengio, 2015. Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd International Conference on Learning Representations, (CLR' 15), San Diego, CA.

Bandyopadhyay, S., 2001. An example based MT system in news items domain from English to Indian languages. Mach. Tran. Rev., 12: 7-10.

Barone, A.V.M., J. Helcl, R. Sennrich, B. Haddow and A. Birch, 2017. Deep architectures for neural machine translation. arXiv preprint arXiv:1707.07631.

Cettolo, M., N. Jan, S. Sebastian, L. Bentivogli and R. Cattoni *et al*., 2015. The iwslt 2015 evaluation campaign. Proceedings of the International Workshop on Spoken Language Translation, Dec. 3-4, Da Nang.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau and F. Bougares *et al*., 2014a. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (NLP' 14), ALC, Doha, Qatar, pp: 1724-1734. DOI: 10.3115/v1/D14-1179

Cho, K., B. van Merrienboer, D. Bahdanau and Y. Bengio, 2014b. On the properties of neural machine translation: Encoder-decoder approaches. Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, (SST' 14), ACL, Doha, Qatar, pp: 103-111. DOI: 10.3115/v1/W14-4012

Costa-Jussa, M.R. and J.A. Fonollosa, 2015. Latest trends in hybrid machine translation and its applications. Comput. Speech Language, 32: 3-10. DOI: 10.1016/j.csl.2014.11.001.

Dandapat, S. and W. Lewis, 2018. Training deployable general domain MT for a low resource language pair: English-Bangla. Proceedings of the 21st Annual Conference of the European Association for Machine Translation, May 28-30, Universitat d'Alacant, Alacant, Spain, pp: 109-117.

Dandapat, S., S. Morrissey, S. Kumar Naskar and H. Somers, 2010. Statistically motivated example-based machine translation using translation memory. Proceedings of the 8th International Conference on Natural Language Processing, (NLP' 10), Kharagpur, India.

Dasgupta, S., A. Wasif and S. Azam, 2004. An optimal way of machine translation from English to Bengali. Proceedings of the 7th International Conference on Computer and Information, (CCI' 04), pp: 648-653.

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of the 2nd International Conference on Human Language Technology Research, (LTR' 02), Morgan Kaufmann, pp: 138-145.

Dwivedi, S.K. and P.P. Sukhadeve, 2010. Machine translation system in Indian perspectives. J. Comput. Sci., 6: 1111-1111. DOI: 10.3844/jcssp.2010.1111.1116

Gage, P., 1994. A new algorithm for data compression. C Users J., 12: 23-38.

Gülçehre, C., O. Firat, K. Xu, K. Cho and L. Barrault *et al*., 2015. On using monolingual corpora in neural machine translation. CoRR.

Haffari, G., M. Roy and A. Sarkar, 2009. Active learning for statistical phrase-based machine translation. Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, (ACL' 09), pp: 415-423.

Heafield, K., I. Pouzyrevsky, J.H. Clark and P. Koehn, 2013. Scalable modified kneser-ney language model estimation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, (ACL' 13), ACL, Sofia, Bulgaria, pp: 690-696. www.aclweb.org/anthology/P13-2121

Hochreiter, S. and J. Schmidhuber, 1997. Long short-term memory. Neural Comput., 9: 1735-1780.

Hutchins, J., 2005. The history of machine translation in a nutshell.

Islam, M.Z., J. Tiedemann and A. Eisele, 2010. English to Bangla phrase-based machine translation. Proceedings of the 14th Annual Conference of the European Association for Machine Translation, Saint-Raphaël, (TSR' 10), France, pp: 27-28.

Jean, S., K. Cho, R. Memisevic and Y. Bengio, 2015a. On using very large target vocabulary for neural machine translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, (ACL' 15), pp: 1-10. DOI: 10.3115/v1/P15-1001

Jean, S., O. Firat, K. Cho, R. Memisevic and Y. Bengio 2015b. Montreal neural machine translation systems for wmt'15. Proceedings of the 10th Workshop on Statistical Machine Translation, (SMT' 15), Association for Computational Linguistics, Lisbon, Portugal, pp: 134-140.

Kalchbrenner, N. and P. Blunsom, 2013. Recurrent continuous translation models. EMNLP, 3: 413-413. www.aclweb.

Kingma, D.P. and J. Ba, 2014. Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations, (CLR' 14).

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit, 5: 79-86. www. mt-archive.info/MTS-2005-Koehn.pdf.

Koehn, P., F.J. Och and D. Marcu, 2003. Statistical phrase-based translation. Proceedings of the Conference of the NAACL on HLT, (CNH' 03), ACL, pp: 48-54. DOI: 10.3115/1073445.1073462

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch and M. Federico *et al*., 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, (PDS' 07), ACL, pp: 177-180.

Luong, M.T. and C.D. Manning, 2015. Stanford neural machine translation systems for spoken language domains. Proceedings of the International Workshop on Spoken Language Translation, (SLT' 15).

Luong, T., I. Sutskever, Q. Le, O. Vinyals and W. Zaremba, 2015. Addressing the rare word problem in neural machine translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, (ACL' 15), pp: 11-19. DOI: 10.3115/v1/P15-1002

Mohammed, E.A. and M.J.A. Aziz, 2011. English to Arabic machine translation based on recording algorithm. J. Comput. Sci., 7: 120-120. DOI: 10.3844/jcssp.2011.120.128

Mumin, M.A.A., A.A.M. Shoeb, M.R. Selim and M.Z. Iqbal, 2012. Supara: A balanced English-Bengali parallel corpus. SUST J. Sci. Technol., 16: 46-51.

Mumin, M.A.A., A.A.M. Shoeb, M.R. Selim and M.Z. Iqbal, 2014. Sumono: A representative modern Bengali corpus. SUST J. Sci. Technol., 21: 78-86.

Mumin, M.A.A., M.H. Seddique, M.Z. Iqbal and M.J. Islam, 2019. Shu-torjoma: An English-Bangla statistical machine translation system. J. Comput. Sci., 15: 1022-1039. DOI: 10.3844/jcssp.2019.1022.1039

Mumin, M.A.A., M.H. Seddiqui, M.Z. Iqbal and M.J. Islam, 2018a. Supara-benchmark: A benchmark dataset for English-Bangla machine translation.

Mumin, M.A.A., M.H. Seddiqui, M.Z. Iqbal and M.J. Islam, 2018b. Supara0.8m: A balanced English-Bangla parallel corpus.

Mumin, M.A.A., M.I. Ahmed, M.A. Bhuiyan, M.R. Selim and M.Z. Iqbal, 2000. An implementation of machine translation between Bangla and English. Proceedings of International Conference on Computer and Information Technology, (CCI' 00), Dhaka, Bangladesh, pp: 290-294.

Nagao, M., 1984. A framework of a mechanical translation between Japanese and English by analogy principle. Artificial Human Intell.

Naskar, S.K. and S. Bandyopadhyay, 2006. A phrasal EBMT system for translating English to Bengali. Satellite Workshop.

Och, F.J. and H. Ney, 2003. A systematic comparison of various statistical alignment models. Computat. Linguist., 29: 19-51. DOI: 10.1162/089120103321337421.org/anthology/D13-1176

Pal, S. and S.K. Naskar, 2016. Hybrid Word Alignment. In: Hybrid Approaches to Machine Translation, Costa-Jussà M., R. Rapp, P. Lambert, K. Eberle and R. Banchs *et al*. (Eds.), ISBN-13: 978-3-319-21311-8, pp: 57-75.

Pal, S., S.K. Naskar and J. van Genabith, 2016. Forest to string based statistical machine translation with hybrid word alignments. Proceedings of the Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, (NCS' 16), Konya, Turkey.

Pal, S., S.K. Naskar and S. Bandyopadhyay, 2014. Word alignment-based reordering of source chunks in pb-smt. Proceedings of the 9th International Conference on Language Resources and Evaluation, (LRE' 14), European Language Resources Association, Reykjavik, Iceland, pp: 3565-3571.

Papineni, K., S. Roukos, T. Ward and W.J. Zhu, 2002. Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on ACL, (AMA' 02), ACL, pp: 311-318. DOI: 10.3115/1073083.1073135

Press, O. and L. Wolf, 2016. Using the output embedding to improve language models. arXiv preprint arXiv:1608.05859.

Roy, M. and F. Popowich, 2010a. Phrase-based statistical machine translation for a low-density language pair. Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence, (AAI' 10), Ottawa, Canada, pp: 273-273.

Roy, M. and F. Popowich, 2010b. Word reordering approaches for Bangla-English statistical machine translation. Proceedings of the Canadian Conference on Artificial Intelligence, (CAI' 10), pp: 282-285.

Roy, M., 2009. A semi-supervised approach to Bengali-English phrase-based statistical machine translation. Proceedings of the Canadian Conference on AI, (CCA' 09), Springer, pp: 291-294.

Saha, D. and S. Bandyopadhyay, 2005. A semantics based English-Bengali EBMT system for translating news headlines. Proceedings of the MT Summit X 2nd Workshop on Example-Based Machine Translation, (BMT' 05), Phuket, Thailand, pp: 125-133.

Salam, K.M.A., S. Yamada and N. Tetsuro, 2017. Improve example-based machine translation quality for low-resource language using ontology. Applied Comput. Inform. Technol., 727: 67-90. DOI: 10.1007/978-3-319-64051-8_5

Schuster, M. and K.K. Paliwal, 1997. Bidirectional recurrent neural networks. IEEE Tran. Signal Proc., 45: 2673-2681.

Sennrich, R., B. Haddow and A. Birch, 2016. Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the ACL, (AMA' 16), pp: 1715-1725.

Sennrich, R., O. Firat, K. Cho, A. Birch and B. Haddow *et al*., 2017. Nematus: A toolkit for neural machine translation. arXiv preprint arXiv:1703.04357

Siddique, M.H., A.M. S. Rana and A. Al Mamun, 2003. A new approach of Bangla machine translation considering allomorph.

Shirko, O., N. Omar, H. Arshad and M. Albared, 2010. Machine translation of noun phrases from Arabic to English using transfer-based approach. J. Comput. Sci., 6: 350-350. DOI: 10.3844/jcssp.2010.350.356.

Sinha, R., K. Sivaraman, A. Agrawal, R. Jain and R. Srivastava *et al*., 1995. Anglabharti: A multilingual machine aided translation project on translation from English to Indian languages. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, Oct. 22-25, IEEE Xplore Press, Vancouver, BC, Canada, pp: 1609-1614. DOI: 10.1109/ICSMC.1995.538002

Snover, M., B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, 2006. A study of translation edit rate with targeted human annotation.

Sutskever, I., O. Vinyals and Q.V. Le, 2014. Sequence to Sequence Learning with Neural Networks. In Advances in Neural Information Processing Systems, Ghahramani, Z., M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger (Eds.), pp: 3104-3112.

Tiedemann, J., 2012. Parallel data, tools and interfaces in opus.