Journal of Computer Science

Original Research Paper

shu-torjoma: An English↔Bangla Statistical Machine Translation System

[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: 25-03-2019
Revised: 19-06-2019
Accepted: 16-07-2019

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: An efficient and publicly open machine translation system is in dire need to get the maximum benefits of Information and Communication Technology through removing the language barrier in this era of globalization. In this study, we present a Phrase-Based Statistical Machine Translation (PBMT) system between English and Bangla languages in both directions. To the best of our knowledge, the system is trained on the largest dataset of more than three million tokens each side in English↔Bangla translation task. In the system, we perform data preprocessing and use optimized parameters to produce efficient system output. We analyze our system output from several viewpoints: overall results, comparisons with the available systems, sentence type and length effect, and behaviour of two challenging linguistic properties– prepositional phrase and noun inflection. Our analysis provides useful insights that translating into morphologically richer language is harder than translating from them and this is mainly due to the difficulties of translating noun inflections. Comparisons with the available systems show that our system outperforms the other systems significantly and gain 10.84 BLEU, 2.18 NIST and 19.02 TER points over the next best system. The analysis of the sentence type and length effect shows that simple sentences are easier to translate and the sentences longer than 15 words are harder to translate for English↔Bangla translation task. To foster the English↔Bangla machine translation research, we have developed development and test datasets, which are representative in sentence length and balanced in genre to be used as a benchmark and are made publicly available.

Keywords: English-Bangla Machine Translation, Machine Translation System, Morphologically Rich, Statistical Machine Translation

Introduction

Information and Communication Technology (ICT) has brought a sea change in every aspect of our life. However, in developing country like Bangladesh, people at the grass root level do not get the maximum benefit of this technology due to the language barrier. Because most of the people of the country are monolingual, i.e. they can only speak in their mother-tongue which is Bangla, whereas the language of ICT is English. Machine Translation (MT) research field works with the vision of removing this language barrier. MT translates texts of one natural language into texts of another automatically. An efficient and open source MT system between English and Bangla languages is a dire need to promote English-Bangla MT research and thus boost the socio-economic status of Bangla spoken community of 350 million people worldwide as well as Bangladesh through disseminating ICT at the grass root level of the community and the country.

Since its inception, MT research has gone through major four paradigms: Rule-based MT, Example-Based MT (EBMT), Phrase-based Statistical MT (PBMT) and recent advent– Neural-based MT (NMT). Rule-based MT relies on linguistic rules, whereas Example-based MT, Phrase-based MT, and Neural-based MT extract necessary information from a bilingual corpus with already translated parallel texts. In MT research, PBMT

(Koehn *et al*., 2003) has been considered as the state-of-the-art technology until the advent of NMT (Kalchbrenner and Blunsom, 2013; Sutskever *et al*., 2014; Cho *et al*., 2014) recently, which shows an improved result for many high-resource language pairs (Sennrich *et al*., 2016). However, for low-resource settings, PBMT is still considered as state-of-the-art technology since NMT failed to show improved performance in this scenario (Koehn and Knowles, 2017; Östling and Tiedemann, 2017).

Most of the systems in English↔Bangla MT have involved rule-based or example-based approach. The *E-ILMT* (Garje and Kharate, 2013) and *sata-anuvadak* (Kunchukuttan *et al*., 2014) systems are only based on the PBMT approach, which are English-to-Indian languages translation systems, in which English-to-Bangla translation task is integrated as a part of the systems. The *E-ILMT* (Garje and Kharate, 2013) system is trained on a dataset of 13,015 sentence pairs and the *sata-anuvadak* (Kunchukuttan *et al*., 2014) system is trained on a dataset of 46,277 sentence pairs. Apparently, the systems use a small dataset.

Translation task between English and Bangla is considered as low-resource due to its low training data and lack of language processing resources. Moreover, the contrastive properties of these two languages make the translation task more challenging. English is a morphologically poor language, whereas Bangla language is morphologically rich. Furthermore, English text follows Subject-Verb-Object (SVO) syntactical order, whereas Bangla text follows free syntactical order, though Subject-Object-Verb (SOV) syntactical order is dominant. To capture these contrastive language properties in English↔Bangla translation task, it requires a larger training data. In this regard, an English↔Bangla phrase-based statistical MT system trained on a larger dataset can be an efficient translation system.

In this work, we present an English-Bangla MT system in both directions which uses the log-linear phrase-based SMT approach as its core, which is supported by the other three modules: training resources, preprocessor, and postprocessor. In training resources module, we have developed the largest training data used so far in English↔Bangla machine translation task. Moreover, we have developed a test dataset and a development dataset, which are representative in sentence length and balanced in genre. In the preprocessing module, we have filtered the noise from training data manually and normalized the Bangla punctuations using preprocessing tools. In our core system configuration, we optimize the system parameters to produce an efficient result. We have reported our results using automatic evaluation and human evaluation to justify the effectiveness of the system. To stimulate research in English↔Bangla MT, we have made the models produced by the system and the Bangla preprocessing tools publicly available.

The organization of this paper is as follows: After reviewing the existing research works and systems for English↔Bangla machine translation task in Section 1, we introduce the theory of the phrase-based statistical machine translation in Section 2. We then describe the architecture of our system, *shu-torjoma*, in Section 3. Section 4 explains the experimental setup of our system followed by the results and discussion in section 5. Section 6 mentions about translation resources that are made publicly available, while Section 7 concludes this paper with some future directions.

Related Work

In 1991, the first attempt was made to develop an English-to-Bangla MT system as part of the English-to-Indian Languages system, *Anglabharti* (Sinha *et al*., 1995). The system is developed based on pseudo-interlingua approach. The successor of this system is *Anglabharti-II* (Sinha and Jain, 2003), which uses the combination of the example-based approach and the traditional rule-based approach. A part of the system is English-to-Bangla translation task. This system is not available online.

Anubaad (Bandyopadhyay, 2001; Naskar *et al*., 2004) system translates news headlines from English-to-Bangla using example-based machine translation approach. This system translates headlines using Knowledge bases and Example structure. This system is not available online.

Banganubad (Saha, 2005a) or *EB-ANUBAD* (Saha, 2005b) is an English-to-Bangla translator which uses the hybrid architecture of transformer and rule-based natural language engineering methods along with various linguistic knowledge components. This system is not available online.

E-ILMT (Garje and Kharate, 2013) is a statistical machine translation system for English-to-Indian languages. As a part of this system, English-to-Bangla MT system has been developed using standard statistical MT system and the tools including parts-of-speech tagger, parser etc. This system is not available online.

Akkhor Bangla Software (Salam *et al*., 2011) is an example-based English-to-Bangla machine translation system. This system uses a novel approach for example-based machine translation using WordNet and International-Phonetic-Alphabet (IPA)-based transliteration. This system is not available online.

Anubadok is a transfer-based English-to-Bangla translator. This system uses parts-of-speech information to determine sentence type, subject, object, verb and tense of the source text and then using dictionary translate the corresponding components into the target components and finally, combine these target

components in required syntactical order to generate the target text. This system is available online[i]. However, resources are not open to use for the researchers.

Sata-Anuvadak (Kunchukuttan *et al*., 2014) is a collection of 110 statistical machine translation systems for handling multiway translation of Indian languages. A part of the system is English-to-Bangla MT system, which is based on phrase-based statistical MT system. This system has two versions: one is the baseline system and another one is the baseline system extended with source side reordering. This system is available online[ii].

Google Translate has offered machine translation service for Bangla language since June 2011. Since then *Google Translate* used state-of-the-art statistical-based approach for its MT system until the advent of neural-based approach. Recently the system declares to switch into the neural-based MT system (Wu *et al*., 2016). This system is available online for users[iii]. However, resources are not open to use for the researchers.

Currently, machine translation systems are being developed for many language pairs in a large number of academic and commercial research labs worldwide. (Bojar *et al*., 2018) discuss the state-of-the-art system architectures and performances of the 14 MT systems for European languages, between English and each of Chinese, Czech, Estonian, German, Finnish, Russian, and Turkish in both directions. There are many works that present the machine translation systems between English and Arabic languages using various approaches (Algani and Omar, 2012; Mohammed and Aziz, 2011; Shirko *et al*., 2010; Hatem and Omar, 2010; Badr *et al*., 2009; A'ali, 2007). The authors (Alsaket and Aziz, 2014) also present a machine translation system between English and Malay languages in both directions. The authors (Dwivedi and Sukhadeve, 2010) discuss the machine translation system in Indian perspectives. Commercial machine translation systems are also being developed by large software companies such as IBM, Microsoft, Baidu, and Google. The state-of-the-art technology of these commercial systems has switched into neural machine translation from statistical machine translation.

Phrase-based Statistical Machine Translation

In Statistical Machine Translation (SMT), the main goal is to find the target sentence $\mathbf{y} : y_1,y_2,\ldots,y_m$ given a source sentence $\mathbf{x} : x_1,x_2,\ldots,x_n$ for which the conditional probability $p(\mathbf{y}|\mathbf{x})$ is maximum. We can reformulate the translation probability $p(\mathbf{y}|\mathbf{x})$ using Bayes rule:

$$\hat{y} = argmax_{\mathrm{y}}\, p(\mathbf{y}|\mathbf{x}) = argmax_{\mathrm{y}}\, p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y}) \quad (1)$$

This allows for a language model $p(\mathbf{y})$, a translation model $p(\mathbf{x}|\mathbf{y})$ and a decoder $argmax_{\mathrm{y}}$. *Language model* is trained on monolingual corpora of the target language and takes care of the fluency in the target language. *Translation model* is trained on parallel corpora of source and target languages and extracts lexical correspondences between source and target languages with their probabilities. A *decoder* is used to stitch the information extracted from the language model and the translation model and search for the best probable translation in the space of possible translations.

Word-based Models

Most of the earlier SMT systems were based on word-based models, where each word in the source language is aligned to exactly one word in the target language in the translation model. In the word-based model, a word correspondence model, called alignment **a** is introduced to represent the positional correspondence between the words of the target sentence and the words of the source sentence:

$$p(\mathbf{x}|\mathbf{y}) = \sum_a p(\mathbf{x},\mathbf{a}\,|\,\mathbf{y}) \quad (2)$$

The three fundamental models developed to calculate the probability in Equation 2 are decomposed as follows:

- *fertility model*, which accounts for the probability that a target word y_m generates ϕ_i words in the source sentence.
- *lexicon model*, which models the probability to produce a source word x_n given a target word y_m.
- *distortion model*, which tries to explain the phenomenon of placing a source word in position *n* given that the target word is placed in position *m* in the target sentence. This is also used with inverted dependencies and is known as the alignment model

The different combinations of these three models are commonly known in the literature as IBM models 1-5 (Brown *et al*., 1993). Currently, word alignments based on IBM and HMM (Vogel *et al*., 1996) models are considered to be the state-of-the-art.

Expectation Maximization (EM) Training: The word-based IBM models are all estimated from the training data that consists of bilingual sentence pairs. The estimation procedure assumes that an alignment exists between the words of the sentence pairs, but that the alignment is unknown. If the word alignments were known, the word translation probabilities can be estimated. On the other hand, if the word translation probabilities were known, probabilistic word alignments could be determined.

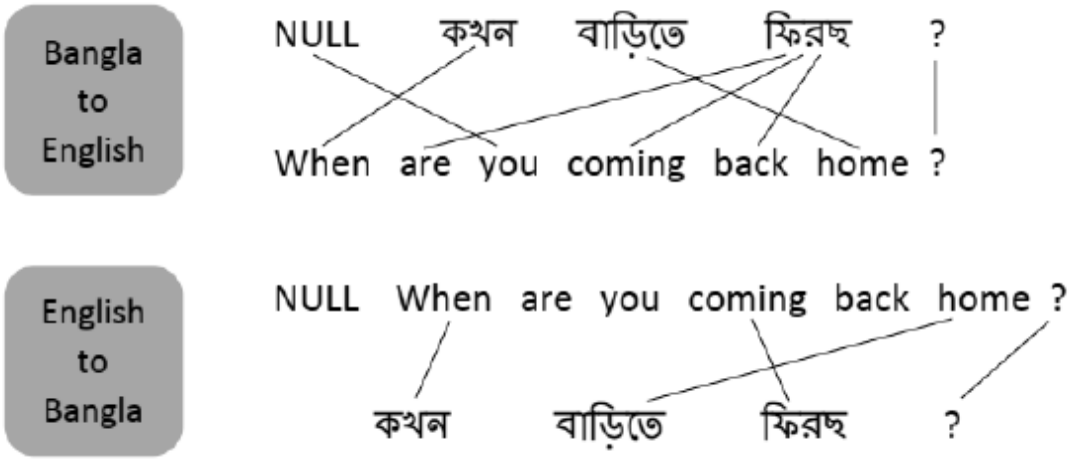


Fig. 1: Asymmetric IBM model alignments: For each Bangla-to-English and English-to-Bangla direction, IBM model allows for many-to-one alignments, not for one-to-many alignments.

To learn the model from this kind of incomplete data, an unsupervised learning technique namely the *Expectation Maximization* (EM) algorithm (Dempster *et al*., 1977) is trained over the data. The EM algorithm first initialize the model, typically with uniform distribution. In the *expectation* step, the algorithm apply the model to the data and estimate the most likely alignments. In the *maximization* step, the algorithm learn the model from the data and augment the data with guesses for the gaps. The algorithm iterate through the two steps until convergence.

Asymmetric Alignments: A by-product of the word-based IBM models is that they establish a word alignment for each sentence pair. However, there is one fundamental problem with the IBM models: They establish asymmetric word alignments, i.e., for a chosen translation direction, they allow for many-to-one alignments, but not for one-to-many alignments. Figure 1 shows a visualization of asymmetric alignments for a parallel English-Bangla sentence in both directions.

Phrase-based Models

Currently, the best performing SMT systems are based on phrase-based models, which translate small word sequences at a time instead of translating each word in isolation like word-based models. Phrase-based models allow to translate from several to several words and not only from one to several, which help to incorporate the context information into the translation model by learning the whole phrases where a phrase can be any sequence of words, even if they are not a linguistic constituent.

The process of the phrase-based models follow the same strategy as for the word-based models with few modifications: segment the source sentence into phrases, then translate each phrase into a target phrase, and finally reorder the target phrases in the output. Figure 2 illustrates the process of the phrase-based models.

Thus, for the phrase-based models, we decompose the translation model, $p(\mathbf{x}|\mathbf{y})$ further into:

$$p(\bar{x}_1^I \mid \bar{y}_1^I) = \Pi_{(i=1)}^{I} \ \varnothing(\bar{x}_i \mid \bar{y}_i) \ \ d(start_i - end_{i-1} - 1) \qquad (3)$$

The source sentence **x** is segmented into a sequence of I phrases $\bar{x}_1^I$ where each segmentation is equally likely. Each source phrase $\bar{x}_i$ in $\bar{x}_1^I$ is translated into a target phrase $\bar{y}_i$. Phrase translation is modeled by a probability distribution $\varnothing(\bar{x}_i \mid \bar{y}_i)$. Reordering of the target phrases is modeled by a relative distortion probability distribution $d(start_i - end_{i-1} - 1)$, where $start_i$ denotes the start position of the source phrase that was translated into the ith target phrase and end_{i-1} denotes the end position of the source phrase that was translated into the $(i-1)$th target phrase.

In the phrase-based translation model, we need to extract the phrase probability translation table that maps source phrases to target phrases with probabilities. Extracting a phrase translation table from a parallel corpus start with a word alignment, which is established as a by-product of the IBM models. However, these models establish asymmetric word alignments as discussed earlier. To resolve this, word alignments are symmetrized by applying some transformations.

Alignment Symmetrization. One approach to symmetric word alignment is to align the parallel corpus bidirectionally using an asymmetric word alignment method. The two resulting word alignments can then be merged by, for instance, taking the *intersection* or the *union* of alignment points of each alignment or a number

of heuristic methods, which usually begin with the intersection and proceed by iteratively adding links from the union (Och *et al*., 1999; Koehn *et al*., 2003). This process is called *symmetrization of word alignments*. Figure 3 illustrates the symmetrization of word alignments.

Extracting Bilingual Phrases (BP). For extracting bilingual phrases from a symmetric word-aligned training corpus, the following two constraints are considered:

1. The words are consecutive and
2. They are consistent with the word alignment matrix, meaning that words inside the phrase are only aligned to words inside the phrase

The phrase-based approach was first presented in (Och, 1999) and named *Alignment Templates*, consisting of pairs of generalized phrases which allow for word classes and include internal word alignments. A simplification of this model is the so-called phrase-based SMT presented in (Zens *et al*., 2002). This approach does not use word classes but instead uses bilingual phrases without internal alignment. The following criterion defines the set of Bilingual Phrases (BP) of the sentence pair $\left(x_1^I; y_1^I\right)$ that is consistent with the word alignment matrix A:

$$BP\left(x_1^I; y_1^I, A\right) = \left(x_j^{j+m}, y_1^{i+n}\right) : \forall\left(i', j'\right) \in A \\ : j <= j' <= j+m \leftrightarrow i <= i' <= i+m \tag{5}$$

Figure 4 shows all the bilingual phrases that are collected according to this definition for the alignment from our example of Fig. 3.

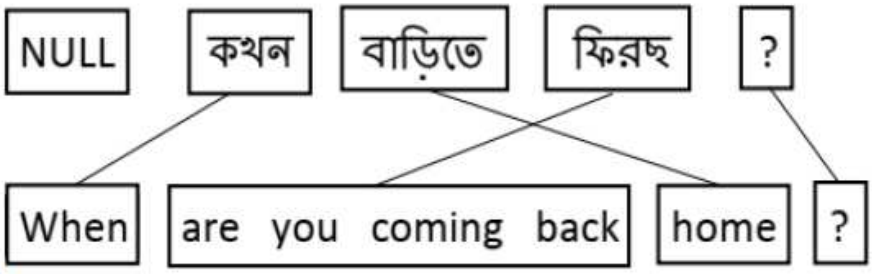


Fig. 2: Phrase-based model: The input is segmented into phrases (not necessarily linguistically motivated), translated one-to-one into target phrases and possibly reordered

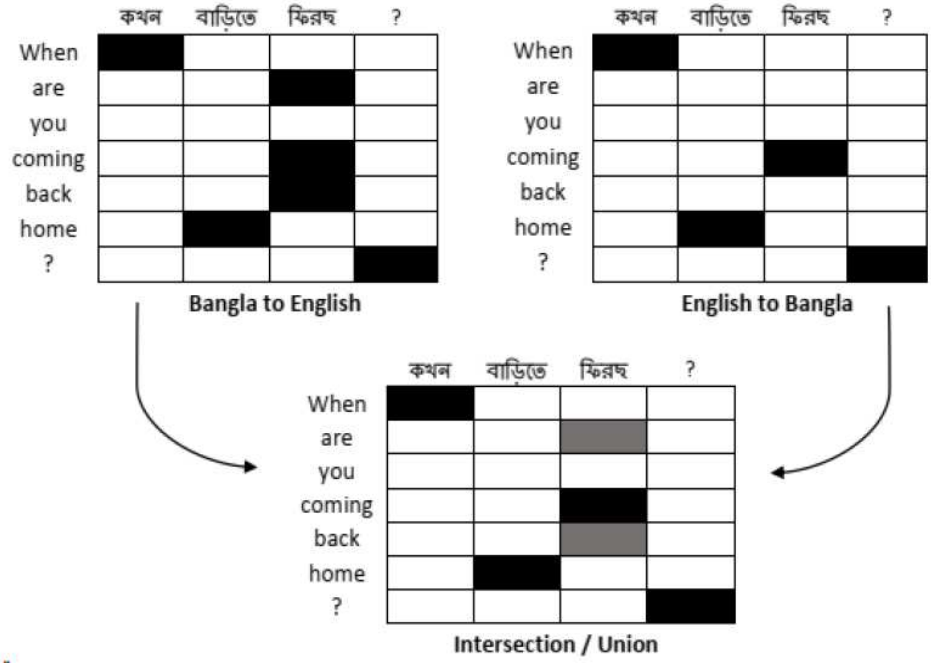


Fig. 3: Symmetrization of IBM model alignments: Bangla-to-English and English-to-Bangla asymmetric alignments can be merged by taking the *intersection* (black) or *union* (gray) of the sets of alignment points.

Fig. 4: Consistent bilingual phrases: The consistent bilingual phrases are extracted from the symmetric word alignment in Figure 3. For some English phrases, no mappings can be found (e.g., *you* or *When are you*).

Estimating Phrase Translation Probabilities. Given the collected bilingual phrases, the phrase translation probability distribution is commonly estimated by relative frequency:

$$\phi\left(\bar{x} \mid \bar{y}\right) = \frac{count\left(\bar{x} \mid \bar{y}\right)}{\sum_{\bar{f}i} count\left(\bar{y}, \bar{x}_i\right)} \tag{6}$$

where, $\text{count}\left(\bar{y}, \bar{x}\right)$ represents the count in how many sentence pairs a particular bilingual phrase is extracted.

Log-Linear Models

State-of-the-art Phrase-Based SMT (PBMT) systems use *maximum entropy* or *log-linear* method as a framework (Berger *et al*., 1996; Och and Ney, 2002) in order to introduce several feature functions that are required to improve the translation process.

Thus, the translation probability $p(\mathbf{x}|\mathbf{y})$ is directly modeled as a log-linear combination of features:

$$\hat{y} = argmax_{\mathbf{y}}\, p\left(\mathbf{y}|\mathbf{x}\right) = argmax_{\mathbf{y}} \exp\{\sum \lambda_k h_k(x_1^n, y_1^m)\}$$

The feature functions h_k can be easily added as necessary into the overall system. Also, the weighting of the different feature functions may lead to the improvement in translation quality. The corresponding weights λ_k are optimized using an optimization algorithm to maximize a scoring function on a development set, which is a small parallel corpus. Figure 5 presents this log-linear framework of an SMT system graphically.

The standard phrase-based model described so far consists of three feature functions: the phrase translation model $\phi\left(\bar{x} \mid \bar{y}\right)$ the reordering model *d,* and the language model $p_{\text{LM}}(y)$. These three feature functions are multiplied together to form our *phrase-based statistical machine translation model*:

$$\hat{y} = argmax_{\mathbf{y}} \prod_{i=1}^{I} \varnothing\left(\bar{x}_i|\bar{y}_i\right)^{\lambda_\phi} d\left(start_i - end_{i-1} - 1\right)^{\lambda_d} \prod_{i=1}^{|y|} p_{LM}(y_i \mid y_1 \cdots y_{i-1})^{\lambda_{LM}} \tag{7}$$

Below we discuss some other feature functions such as direct phrase translation probability, direct and inverse lexical weighting, a word penalty, and a phrase penalty, which are usually used in the state-of-the-art phrase-based SMT systems.

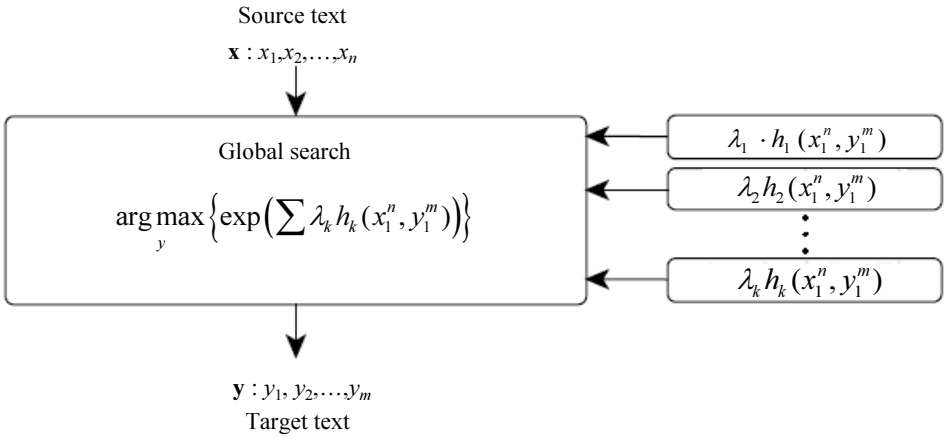


Fig. 5: Log-linear framework of a Statistical Machine Translation (SMT) system. Adapted from (Och and Ney, 2004).

Bidirectional Translation Probabilities

According to Bayes rule, we used an inversion of the conditioning of translation probabilities: $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})p(\mathbf{x})^{-1}$. However, there may arise some situations like that in the training data an unusual source phrase $\bar{x}$ exists that is mistakenly mapped to a common target phrase $\bar{y}$ and $\phi(\bar{x}|\bar{y})$ is very high, maybe even 1. Hence, in the test data, this erroneous phrase translation will almost certainly be used to produce the highest probability translation. In this case, it would be better to use the conditioning of phrase translation probabilities in the actual translation direction, i.e., $\phi(\bar{y}|\bar{x})$. It is even possible to use both translation directions, $\phi(\bar{y}|\bar{x})$ and $\phi(\bar{x}|\bar{y})$, as feature functions.

In practice, a model using both inverse and direct phrase translation probabilities, with the proper weight setting, often outperforms a model that uses only the Bayes-motivated inverse translation probabilities, or only the direct translation probabilities.

Bidirectional Lexical Weighting

Sometimes, infrequent phrase pairs may cause problems due to their overestimation of translation probabilities, especially if they are collected from noisy data. For reliable estimation of a rare bilingual phrase, it is decomposed into its word translations. This is called *lexical weighting*; it is basically a smoothing method.

Based on the word alignment established by word-based IBM models, we can compute the lexical translation probability of a phrase given the phrase by, for instance:

$$lex(\bar{y}|\bar{x},a) = \prod_{i=1}^{lenght(\bar{y})} \frac{1}{|j|(i,j)\in a|} \sum_{\forall(i,j)\in a} w(y_i|x_j) \quad (8)$$

In this lexical weighting scheme, each of the target words y_i is generated by aligned source words x_j with the word translation probability $w(y_i|x_j)$. The lexical translation probabilities $w(y_i|x_j)$ are estimated from the word-aligned corpus. Counts are taken and relative frequency estimation yields the probability distribution. Unaligned words are taken to be aligned to NULL. In practice, it is useful to use both inverse and direct lexical translation probabilities in the model: $lex(\bar{y}|\bar{x},a)$ and $lex(\bar{x}|\bar{y},a)$.

Word Penalty

In the phrase-based model, the system prefers shorter translations due to the language model. To protect against translation output that is too short (or too long), a *word penalty* parameter that adds a factor ω for each produced word is introduced. If ω<1, the scores of shorter translations are increased and if ω>1, longer translations are preferred. The *word penalty* parameter is very effective in tuning output length and often improves translation quality significantly.

Phrase Penalty

Analogous to the word penalty discussed above, a *phrase penalty* parameter that adds a factor ρ for each phrase translation is introduced. If ρ<1, longer phrases are preferred and if ρ<1, shorter phrases are preferred.

Decoding

The task of *decoding* in machine translation is to construct the possible translations for an input text by combining all feature functions of a given model and look for the translation with the highest probability. In principle, decoding corresponds solving the maximization problem in Equation 7. The phrase-based decoding uses the *beam-search stack decoder*, which is the most commonly used decoding algorithm.

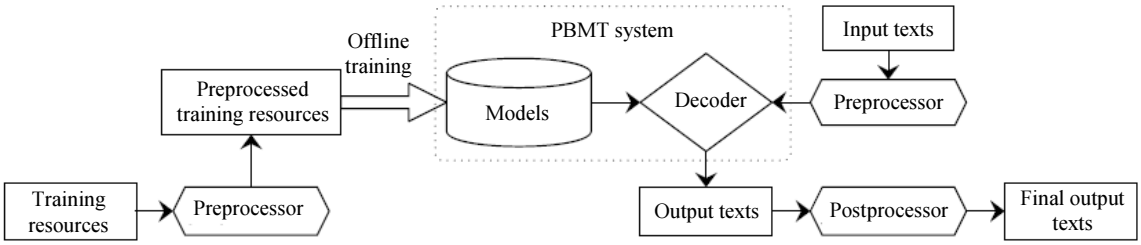


Fig. 6: The *shu-torjoma* system architecture

The *shu-torjoma* System

The *shu-torjoma* system is a modular statistical-based translation system. At the core of the system's architecture is a *log-linear phrase-based SMT* system. The other three modules of the system are *training resources*, *preprocessor,* and *postprocessor*. Figure 6 shows an overview of the system architecture. A brief description of these modules is given below. The experimental setup of these modules is discussed in the next section.

The training resources module works with creating, compiling and enhancing resources required to train the system. Training resources usually include parallel corpus, monolingual corpus, and bilingual dictionary. At this moment, we have developed the largest English-Bangla parallel dataset. We have also developed a test dataset and a development dataset, which are representative in sentence length and balanced in genre.

In the preprocessor module, we preprocess the training resources into a convenient format so that the system performance improves. The preprocessing step usually includes punctuation and lexical normalization, tokenization, morphological segmentation, syntactical reordering among others. At this moment, we have developed punctuation normalization tool and cleaning tool for Bangla language.

The core PBMT system takes preprocessed training resources and performs training and tuning algorithm on those resources to create several statistical models. Using these statistical models, the decoder of the PBMT system finally decode input texts of one language into output texts of another language. In this module, we have used log-linear phrase-based SMT approach as the core of our system.

The postprocessor or post-editor module performs error analysis of the system's output in comparison with reference translation and applies several processing algorithms to make output further better. The postprocessing step includes re-ranking N-best lists, detokenization, desegmentation, transliteration among others. In this module, we have performed an error analysis of our system's output for some linguistic properties.

Experimental Setup

In this section, we present the experimental setup of our system 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 PBMT 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*., 201; 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, journalistic texts, instructive texts, administrative 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[iv]. 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.

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
Global Voices	126,724	En	2,533,959	80,520	20.00
		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

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.

For the language model, we used *Shahjalal University monolingual (SUMono)* (Mumin *et al*., 2014) corpus in En→Bn task and *Europarl* (Koehn, 2005) corpus in Bn→En task. *SUMono* (Mumin *et al*., 2014) is a representative modern Bangla corpus of more than 32 million tokens and *Europarl* (Koehn, 2005) contains more than 27 million tokens.

Data Preprocessing

We performed following preprocessing tasks on our datasets in sequence:

Data filtering. The *GolbalVoices* (Tiedemann, 2012) corpus used in our training dataset contains a considerable amount of noise. We filtered out many sentences due to bad translations. Many translations were not in the correct order and sometimes in other languages, such as Arabic and Japanese. We filtered them out manually. Due to this filtering, the corpus size reduced to 128570 sentences from the original size of 130319 sentences.

Punctuation Normalization. Bangla punctuations suffer from the problem of multiple Unicode codepoints for representation of the same punctuation. This causes data sparsity. We normalized Bangla punctuations in Bangla side of our datasets using our tools by maintaining only one standard of quotation marks, apostrophes, semicolon, colon, Bangla full stop called *dari*. We also normalized the English side of our datasets using the standard *Moses* (Koehn *et al*., 2007) scripts.

Tokenization. We tokenized the normalized Bangla side of our datasets using Bangla specific tokenizer[v]. We also tokenized normalized English side of our datasets using the standard *Moses* (Koehn *et al*., 2007) scripts. After tokenization, Bangla unique tokens reduced to 157,784 from 221,513 and English unique tokens reduced to 94,960 from 198,914.

Data Cleaning. Finally, we cleaned sentence pair with length ratio 1:5 and larger than 60 tokens in either side of our datasets. This cleaning reduces the sentences of our datasets to 197,338 from 199,431. We cleaned our datasets using the standard Moses (Koehn *et al*., 2007) scripts.

PBMT System Configuration

Our core PBMT system is implemented using the *Moses* (Koehn *et al*., 2007) SMT toolkit. We trained our system on English-Bangla parallel training dataset which is a combination of *SUPara* (Mumin *et al*., 2012; 2018b) and *GlobalVoices* (Tiedemann, 2012) corpus. We extracted symmetrized word alignments from this training dataset using GIZA++ (Och and Ney, 2003) and *grow-diag-final-and* heuristic. From the extracted symmetrized word alignments, using maximum Likelihood Estimation (MLE) we estimated the phrase-based translation model with *six* feature functions and lexicalized reordering model with *seven* feature functions. We also trained our system on target monolingual dataset, *SUMono* (Mumin *et al*., 2014) for En→Bn translation task and *Europarl* (Koehn, 2005) for Bn→En translation task, to estimate language model with *one* feature functions. Thus, we got *fourteen* generative models. The weights of these *fourteen* generative models were learned using minimum error rate training (MERT) (Och, 2003) on development dataset, *SUParadev2018* (Mumin *et al*., 2018a), so as to maximize the BLEU (Papineni *et al*., 2002) score. These calculated weights with their generative models produced *fourteen* discriminative models. Finally, the *Moses* decoder exploited these discriminative models to search for the best target text of each text in test dataset, *SUParatest2018* (Mumin *et al*., 2018a). Figure 7 shows the flow of data, models, and processes of our system. In the following, we detail several aspects of our system configuration.

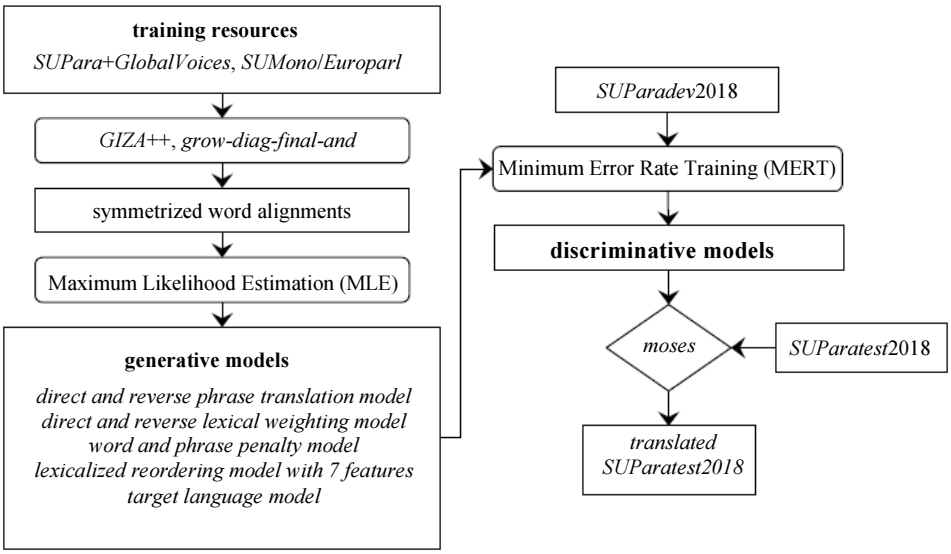


Fig. 7: Phrase-Based Statistical Machine Translation (PBMT) core system: the flow of data, models, and processes. Adapted from: (Lopez, 2008).

Translation Model

In order to estimate the translation model, we first trained En→Bn and Bn→En word-alignment models fully independently from each other by aligning words for each sentence pairs of the English-Bangla parallel training data in both directions using GIZA++ (Och and Ney, 2003). GIZA++ (Och and Ney, 2003) is a freely available implementation of the IBM models (Brown *et al*., 1993). We obtained these word-alignment models in two steps: first, using *mkcls* (Och, 1999) utility in GIZA++ (Och and Ney, 2003), we clustered source and target vocabularies into *50 classes*. Second, using GIZA++ (Och and Ney, 2003), we ran a sequence of IBM word alignment model (Brown *et al*., 1993) training with *five* iterations of Model 1, *five* iterations of HMM, *three* iterations of Model 3 and *three* iterations of Model 4. The obtained two word-alignment models are asymmetric, i.e., for a chosen translation direction, they allow for many-to-one alignments, but not for one-to-many alignments.

Now, we established symmetrized alignments by applying the *grow-diag-final-and* heuristic to the obtained two asymmetric word-alignments using GIZA++ (Och and Ney, 2003). From these ymmetrized alignments, we extracted all bilingual phrases that are consistent with the word alignment. Then, we estimated the phrase translation probabilities using the Maximum Likelihood Estimation (MLE) on these extracted bilingual phrases. Thus, we got a standard phrase table that includes computations of *six* phrase translation scores: *direct and inverse phrase translation probability*, *direct and inverse lexical weighting*, *phrase penalty*, *word penalty*. We allowed a maximum phrase length of 7 tokens in the phrase table.

Lexicalized Reordering Model

We applied a commonly used *msd-bidirectional-fe* setting to estimate lexicalized reordering model. We determined the orientation of two phrases based on word alignments at training time and based on phrase alignments at decoding time. We used three orientation classes: *monotone*, *swap* and *discontinuous*. These orientations were modeled based on both the previous and next phrase and conditioned on both the source and target languages. We allowed distortion limit up to 6 words. Thus, we used *seven* features for the lexicalized reordering model.

Language Model

We trained 5-gram language models with modified Kneser-Ney smoothing (1995; Chen and Goodman, 1999). *KenLM* (Heafield *et al*., 2013) was employed for language model training and scoring.

Tuning and Decoding

We incorporated above mentioned *fourteen* features of our system in a log-linear combination. We tuned the corresponding weights of these features using Minimum Error Rate Training (MERT) (Och, 2003) to maximize BLEU (Papineni *et al*., 2002) on the *SUParadev2018* (Mumin *et al*., 2018a) development set.

We used *Moses* (Koehn *et al*., 2007) decoder which implements a beam search in its decoding process. The decoder exploits the 14 features of the system and their corresponding weights to decode input texts of the *SUParatest2018* (Mumin *et al*., 2018a) into *translated SUParatest2018* texts.

Multiple Run

Due to the instability of the MERT (Och, 2003) tuning algorithm, (Foster and Kuhn, 2009) suggest to run it many times (at least 7) and then choose the weights that give best results. In this regard, we performed tuning and decoding steps 9 times for each system and choose the result with best BLEU score.

Bidirectional Translation

The *Moses* (Koehn *et al*., 2007) SMT toolkit, which is used to implement our PBMT system, is designed to translate one direction at a time. Therefore, it is required to perform training, tuning, and decoding separately for En→Bn and Bn→En translation tasks.

En→Bn. For En→Bn translation task, we trained our system on English-Bangla parallel training dataset which is a combination of *SUPara* (Mumin *et al*., 2012; 2018b) and *GlobalVoices* (Tiedemann, 2012) corpus and on Bangla monolingual dataset, *SUMono* (Mumin *et al*., 2014). Then, we tuned our system using Minimum Error Rate Training (MERT) (Och, 2003) on the Bangla side of the development dataset, *SUParadev2018* (Mumin *et al*., 2018a), so as to maximize the BLEU (Papineni *et al*., 2002) score. Finally, we used the decoder to translate the texts of the English side of the test dataset, *SUParatest2018* (Mumin *et al*., 2018a).

Bn→En. For En→Bn translation task, we trained our system on English-Bangla parallel training dataset which is a combination of *SUPara* (Mumin *et al*., 2012; 2018b) and *GlobalVoices* (Tiedemann, 2012) corpus and on English monolingual dataset, *Europarl* (Koehn, 2005). Then, we tuned our system using Minimum Error Rate Training (MERT) (Och, 2003) on the English side of the development dataset, *SUParadev2018* (Mumin *et al*., 2018a), so as to maximize the BLEU (Papineni *et al*., 2002) score. Finally, we used the decoder to translate the texts of the Bangla side of the test dataset, *SUParatest2018* (Mumin *et al*., 2018a).

Translation Evaluation

We used both automatic evaluation metrics and human evaluation metrics to evaluate the performance of our systems effectively.

Automatic Evaluation Metrics

For the automatic evaluation, 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, substitution and phrasal shifts. Contrary to BLEU and NIST, a lower TER score indicates improvements in translation.

Human Evaluation Metrics

For the manual evaluation, we used *adequacy* and *fluency* metrics according to a 1 to 5 quality scale. As human evaluators, we assigned three native Bangla speakers who are graduates of the English department as well as professionally involved in the translation task. These evaluators were asked to rate *adequacy* and *fluency* to the system's output as follows:

Adequacy measures how much meaning is retained in the translation, with the following scores: 5 for all of the information, 4 for most of the information, 3 for much of the information, 2 for little information, and 1 for none of it.

Fluency indicates how natural the translation sounds to a native speaker of the target language, with the following scores: 5 for Flawless, 4 for Good, 3 for Non-native, 2 for Disfluent, and 1 for Incomprehensible.

Results and Discussion

We have reported and interpreted the results of our system from several viewpoints: overall results, comparisons with the available MT systems, the effect of the sentence type and the sentence length in translation quality, and the translation behaviour with respect to the two challenging linguistic properties– prepositional phrase and noun inflection. We have also discussed how to deal with the linguistic divergences between English and Bangla languages in the English↔Bangla PBMT task.

Overall Results

We present here the overall results of our system according to automatic evaluation metrics and human evaluation metrics as discussed below:

Automatic evaluation. Table 2 shows the overall results of various systems in terms of automatic evaluation metrics BLEU, NIST, and TER for both the En→Bn and Bn→En translation tasks.

For our experiments, we first develop a *baseline system* which uses an almost similar configuration to our proposed system, *shu-torjoma*, with two exceptions. First, the baseline system is trained on the same training dataset as the proposed system but with minimal preprocessing which is common in a standard phrase-based SMT system. The minimal preprocessing include tokenizing both sides, lowercasing English side, and cleaning the sentence pairs with length ratio 1:9 and larger than 80 tokens in either side. Second, the baseline system used the same system configuration as the proposed system but trained with a 3-gram language model.

We compared different n-gram language models for En→Bn and Bn→En translation tasks and observed the best BLEU score for 5-gram Language Model (LM). Accordingly, we trained the baseline system with 5-gram LM and got an improvement over the baseline system as shown in Table 2.

Finally, we trained the baseline system with 5-gram LM and data preprocessing as mentioned in section 4.2. We denote this system as our proposed system, *shu-torjoma*. From Table 2, we observe that our proposed system provide improved performance over the other two systems. This implicates that removing noise from the data and normalizing punctuation of the data have positive impacts on the MT performance.

From Table 2, we also observe that for all systems the best scores are obtained in the Bn→En translation task, which thus confirms to be easier than En→Bn translation task. It corroborates the results reported by (Koehn, 2005) which is translating into morphologically richer languages is more difficult than translating from them.

Human Evaluation. Since the human evaluation is a costly and time-consuming process, we evaluated only the output of the *shu-torjoma* system by human evaluators. We used the system's output of all 500 sentences of the *suparatest2018* (Mumin *et al*., 2018a) dataset for human evaluation. Table 3 shows the individual and average ratings of three evaluators for *adequacy* and *fluency* of the target languages for both the En→Bn and Bn→En translation tasks.

Regarding the human assessment of *adequacy*, Bn→En translation task was rated as more capable of translating meaning than En→Bn translation task for all human evaluators. This supports the result produced by the automatic evaluation metrics as shown in Table 2 and thus, corroborates the results reported by Koehn (2005) which is translating into morphologically richer languages is more difficult than translating from them.

Regarding *fluency*, En→Bn translation task was rated as more fluent than Bn→En translation task for all human evaluators. Apparently, this result contradicts with the result of automatic evaluation metrics as shown in Table 2. However, this can be explained as: in En→Bn translation task, the target language, Bangla, is flexible as a relatively free word order language. A human evaluator can detect this flexibility whereas, in the automatic evaluation, detection of this flexibility is not possible due to the given fixed word order of the target reference translation.

Comparisons with Available MT Systems

We compared the performance of our proposed phrase-based SMT system with two available online English-Bangla MT systems: *Anubadok* and *Sata-anuvadak*. *Anubadok* is a transfer-based English-to-Bangla translator, whereas *Sata-anuvadak* is an English↔Bangla phrase-based SMT system and uses a small dataset of 46,277 sentence pairs in its training.

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

	En→Bn			**Bn→En**		
System	BLEU↑	NIST↑	TER↓	BLEU↑	NIST↑	TER↓
Baseline	14.21	4.93	73.29	16.91	5.75	68.37
Baseline + 5-gram LM	15.13	5.07	72.84	17.21	5.75	68.05
shu-torjoma	**15.27**	**5.13**	**71.9**	**17.43**	**5.76**	**67.94**

Table 3: Human evaluation: The performance of the *shu-torjoma* system on the *suparatest2018* (Mumin *et al*., 2018a) dataset in both language directions evaluated by human evaluators in the 5-scale measure. H1, H2 and H3 denote the human evaluators

	Adequacy				**Fluency**			
Language Direction	H1	H2	H3	Avg.	H1	H2	H3	Avg.
En→Bn	3.79	3.57	3.70	3.69	2.43	1.95	2.29	2.22
Bn→En	3.85	3.63	3.74	3.74	2.23	1.75	1.91	1.96

The review of these two systems is given in section 2. Table 4 shows the performance of these systems along with our proposed system, *shu-torjoma*. From Table 4, we observe that our phrase-based SMT system which is trained on a balanced dataset and on approximately 5 times larger dataset than *Sata-anuvadak* system outperforms the other two systems significantly.

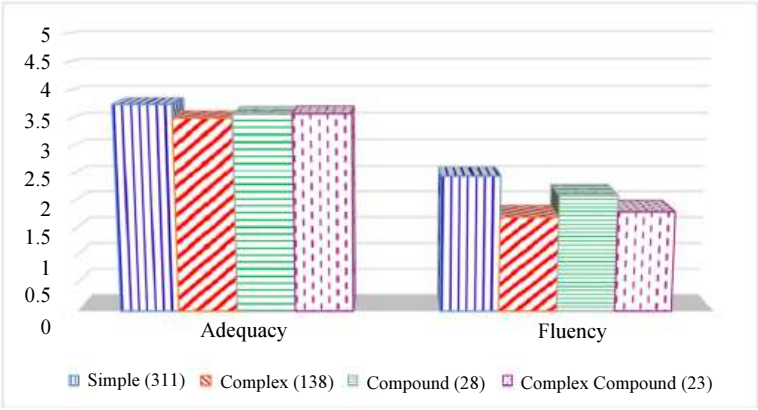
Sentence Type Effect

We investigate in our experiment how translation quality is affected by the type of the source sentence. In this regard, we categorize our test set by human evaluators into four sentence types: s*imple*, *complex*, *compound,* and *complex-compound.* We then evaluate the output of our system for those subsets with the human evaluation metrics: *adequacy* and *fluency*. Figure 8 presents the *adequacy* and *fluency* scores on subsets of different source sentence type for both En→Bn and Bn→En translation tasks.

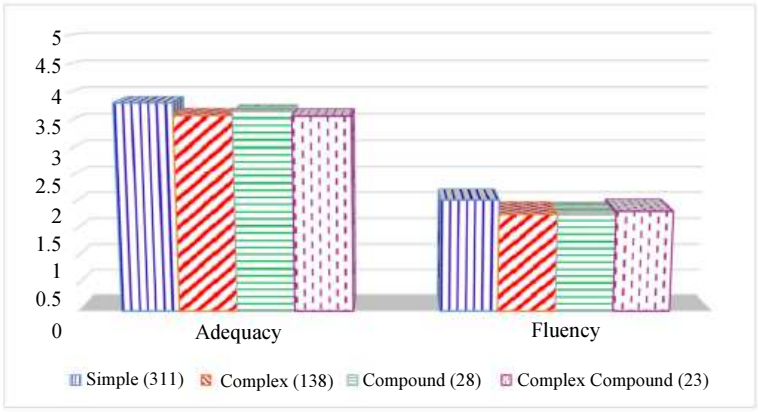
From the Fig. 8, we observe that simple sentences are a little bit easier to translate and complex sentences are a little bit harder to translate in terms of *adequacy* and *fluency*. This suggests that we need to handle complex syntax and long-distance dependencies of clauses of complex sentences between English and Bangla. A solution may be to preprocess the source sentences so that syntactic structure of the source language more closely resembles the target language.

Table 4: System comparisons: The performance of various English-Bangla machine translation systems available online on the *suparatest2018* (Mumin *et al*., 2018a) dataset

	En→Bn			**Bn→En**		
System	BLEU↑	NIST↑	TER↓	BLEU↑	NIST↑	TER↓
Anubadok	4.43	2.95	90.92	–	–	–
Sata-anuvadak	1.31	1.77	103.89	4.18	3.18	83.80
shu-torjoma	**15.27**	**5.13**	**71.9**	**17.42**	**5.76**	**67.94**



(a)



(b)

Fig. 8: Sentence type effect: Human evaluation of the system's output on subsets of the different source sentence type in the *suparatest 2018* (Mumin *et al*., 2018a) dataset. Count of each sentence type is mentioned in the parentheses next to the respective sentence type (a) **En → Bn** (b) **Bn → En**

Sentence Length Effect

We further investigate in our experiment how translation quality is affected by the length of the source sentence. In this regard, we develop our test set by taking texts from 10 subsets of different lengths as discussed in section 4.1. We then evaluate the output of our system for those subsets with the human evaluation metrics: *adequacy* and *fluency*. Figure 9 presents the *adequacy* and *fluency* scores on subsets of different source sentence length for both En→Bn and Bn→En translation tasks.

From the Fig. 9, we observe that sentences longer than 15 words face problem for both En→Bn and Bn→En translation tasks according to both *adequacy* and *fluency* metrics. This suggests that we need to use special mechanism to translate longer sentences. One mechanism may be split longer source sentences into convenient segments, translate these segments, and finally, merge these translated segments. For Bn→En translation task, sentences longer than 30 words are considered less reliable due to small sample counts.

Linguistic Behaviour

We closely observed the system output manually to understand that how the system behaves with different linguistic properties of texts. We explore here two linguistic properties among others, which pose challenges in English-Bangla translation task: *Prepositional phrase* and *Noun inflection*. Table 5 and Table 6 show some samples which exhibit these two linguistic properties for En→Bn and Bn→En tasks, respectively. Each sample contains a source sentence (src), its reference translation (ref), and the system output (*shu-torjoma*). The reference translation (ref) is the translation of the source sentence translated by a human expert. We have chosen the same sample sentences for both En→Bn and Bn→En tasks for making the comparison convenient.

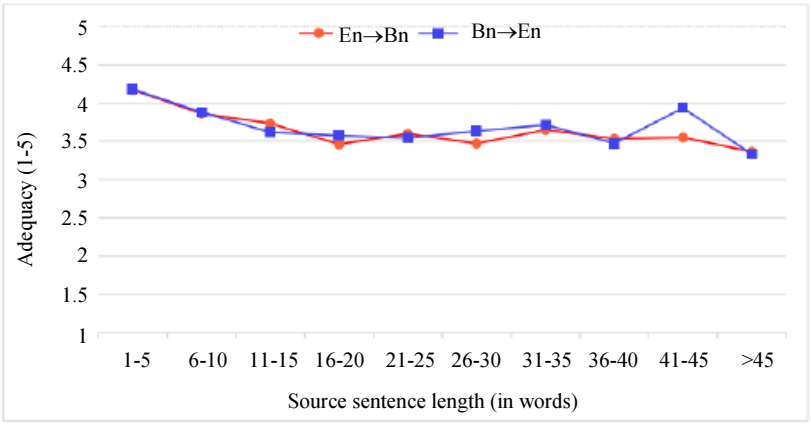
Prepositional Phrase. In Tables 5a and 6a, we show the behaviour of our system for En→Bn and Bn→En tasks in translating the prepositional phrase. We are focusing in particular on the English prepositional phrase *from the account* and its counterpart Bangla prepositional phrase **আপনার অ্যাকাউন্ট থেকে**. We observe that our statistical system translate a prepositional phrase correctly for English-Bangla in both directions. Besides this, our close observations on system's output in translating prepositional phrases reveal that our statistical system is capable of translating prepositional phrases at a satisfactory level for English-Bangla in both directions. English prepositions are translated in Bangla using inflections to the reference objects and/or post-positional words after the reference objects. In Bangla, there are only a few in flections as well as post-positional words for English prepositions (Naskar and Bandyopadhyay, 2006). Moreover, in English, the inventory of prepositions is a close set. For this reason, phrase-based alignments of our system are capable to extract English prepositional phrases and their corresponding correct Bangla inflected or post-position phrases at a satisfactory level. For example, in our current sample prepositional phrase, the translation of preposition, *from*, is **থেকে**. This translated word becomes post-positional word after the reference objects (*your account*: **আপনার অ্যাকাউন্ট**).

Noun Inflection. Tables 5b and 6b show the behaviour of our system in translating the noun inflection for En→Bn and Bn→En tasks. We are focusing in particular on the English noun *daughter* and its counterpart Bangla inflected noun form **মেয়েকে (মেয়ে+কে)**. Our statistical system failed to generate the inflected form of English noun in En→Bn task, while correctly translate of Bangla noun in Bn→En task. In English, a noun phrase remains the same regardless of being subject or object. However, Bangla noun phrases become inflected based on their role in the sentence. For example, in the above case, the noun phrase *daughter* is an object and its Bangla translation becomes the inflected form of Bangla noun **মেয়ে**, which is**মেয়েকে (মেয়ে+কে)**. On the other hand, when the noun phrase *daughter* is a subject as in the sentence '*my daughter loves me'*, the Bangla translation of *daughter* does not produce any inflected form of Bangla noun **মেয়ে**. Thus, translating English noun phrases to Bangla noun phrases suffers from the lack of information about its role in the sentence, making it hard to choose the right inflected forms.

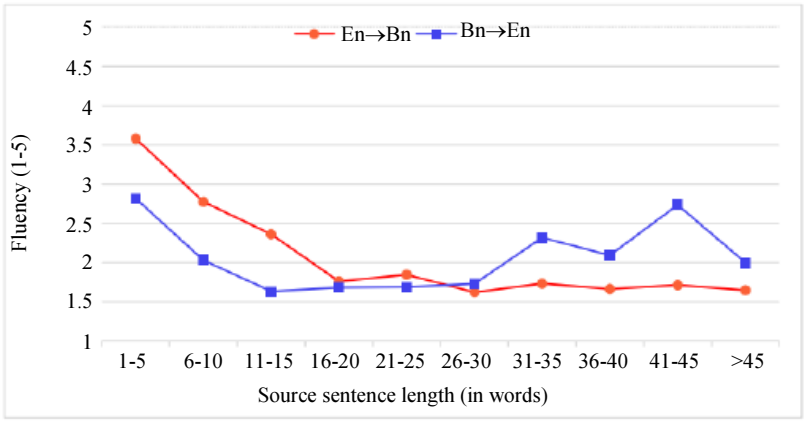
Translation Resources

We have made the *shu-torjoma* system available for public access[vi]. The system enables users to translate texts between English and Bangla in both directions. The development dataset and the test dataset are made publicly available (Mumin *et al*., 2018a), which can be used as a benchmark in the English↔Bangla MT task.

To stimulate research in English↔Bangla MT, we have also released the training models, tuning models, binarized tuning models, language models, binarized language models, and the Bangla preprocessing tools for academic use. Users can generate their own tuning models by tuning the weights of our training models based on their own development dataset. Furthermore, anyone can translate their test dataset straightforward using our binarized tuning models and binarized language models.



(a)



(b)

Fig. 9: Sentence length effect: Human evaluation of the system's output on subsets of the different source sentence length (in words) in the *suparatest2018* (Mumin *et al*., 2018a) dataset (a) Adequacy (b) Fluency

Table 5: En→Bn: Source sentence (src), reference translation (ref) and system output (*shu-torjoma*) samples showing behaviour in translating *prepositional phrase* and *Noun inflection*. We underlined the focal point in each category

(a)	*Prepostional phrase:*		
	src	… do this to avoid fraudulent users to send sms from your account .	
	ref	**প্রতারক ব্যবহারকারীদের আপনার অ্যাকাউন্ট থেকে এসএমএস পাঠানো ঠেকাতে আপনি ...**	
	shu-torjoma	**তুমি অবশ্যই এড়াতে এই প্রতারণাপূর্ণ ব্যবহারকারীরা আপনার অ্যাকাউন্ট থেকে ...**	√
(b)	*Noun inflection:*		
	Src	i love my daughter .	
	Ref	**আমি আমার মেয়েকে ভালবাসি ।**	
	shu-torjoma	**আমি ভালবাসি আমার মেয়ে ।**	×

Table 6: Bn→En: source sentence (src), reference translation (ref) and system output (*shu-torjoma*) samples showing behaviour in translating *prepositional phrase* and *Noun inflection*. We underlined the focal point in each category.

(a)	*Prepostional phrase:*		
	src	**প্রতারক ব্যবহারকারীদের আপনার অ্যাকাউন্ট থেকে এসএমএস পাঠানো ঠেকাতে আপনি ...**	
	ref	… do this to avoid fraudulent users to send sms from your account .	
	shu-torjoma	a scammer users to prevent sent to sms from your account of course ...	√
(b)	*Noun inflection:*		
	src	**আমি আমার মেয়েকে ভালবাসি ।**	
	refi	i love my daughter .	
	shu-torjoma	i love my daughter.	√

Conclusion and Future Direction

We have developed an English↔Bangla phrase-based statistical machine translation system, which has been trained on the largest dataset of more than three million tokens and 197,338 sentence pairs. We also tuned and evaluated our system on two different and disjoint datasets, which are representative in sentence length and balanced in text genre. These two datasets are fine-tuned by correcting misspelling and bad translation by a language expert with the vision of using them as a bench mark in English↔Bangla machine translation research and are made publicly available. The models generated by our system are made publicly available, which can be used to translate texts and generate different tuning models based on different development datasets. Moreover, the proposed MT system can be applied to translate technical reports, legal and financial documents, user manuals, meeting minutes, tourism information, and website contents between English and Bangla languages.

The overall result of our system shows that Bangla-to-English translation task is easier than English-to-Bangla translation task. This corroborates the fact that translating into morphologically richer languages is more difficult than translating from them. The data preprocessing and parameter optimization of our system provide 1.06 BLEU, 0.2 NIST, and 1.39 TER points improvements for English→Bangla and 0.52 BLEU, 0.01 NIST, and 0.43 TER points for Bangla→English over the baseline system. Our comparisons with the available English-Bangla MT systems, *Anubadok* and *Sata-anuvadak*, show that our system outperforms these systems significantly and gain 10.84 BLEU, 2.18 NIST, and 19.02 TER points over the next best system, *Anubadok*. Our investigation of the translation effect on sentence type and sentence length reveals that simple sentences are a little bit easier to translate and the sentences longer than 15 words are harder to translate for English-Bangla translation task in both directions. We have further investigated two linguistic properties among others which pose challenges in English↔Bangla translation task, namely *prepositional phrase* and *noun inflection*. Our statistical system shows satisfactory performance in translating a *prepositional phrase* for English-Bangla translation task in both directions. However, in translating *noun inflection*, our system succeeds in Bangla-to-English direction but failed in English-to-Bangla direction.

While our statistical system shows satisfactory results in translating *prepositional phrases*, the system has weakness in translating *noun inflections*, particularly for English-to-Bangla. In this area, our system requires further attention. Since it is tempting to observe how neural-based approach behave with these linguistic properties, we also plan to explore the neural machine translation system for English-Bangla translation task in both directions.

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 Banglades [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

A'ali, A.M., 2007. Pre-editing and recursive-phrasecomposites for a better English-to-Arabic machine translation. J. Comput. Sci., 3: 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.

Alsaket, A.J. and M.J.A. Aziz, 2014. Arabic-malay machine translation using rule-based approach. J. Comput. Sci., 10: 1062-1062.

Badr, I., R. Zbib and J. Glass, 2009. Syntactic phrase reordering for English-to-Arabic statistical machine translation. Proceedings of the 12th Conference of the European Chapter of the ACL, Mar. 30, Athens, Greece, pp: 86-93.

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

Berger, A.L., P.F. Brown, S.A.D. Pietra, V.J.D. Pietra and A.S. Kehler *et al*., 1996. Language translation apparatus and method using context-based translation models. US Patent, 5: 510-981.

Bojar, O., R. Chatterjee, C. Federmann, M. Fishel and Y. Graham *et al*., 2018. Proceedings of the third conference on machine translation. Brussels, Belgium.

Brown, P.F., V.J.D. Pietra, S.A.D. Pietra and R.L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. Computat. Linguistics, 19: 263-311.

Chen, S.F. and J. Goodman, 1999. An empirical study of smoothing techniques for language modeling. Comput. Speech Language, 13: 359-394.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau and F. Bougares *et al*., 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation learning. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Oct. 25-29, Doha, Qatar, pp: 1724-1734.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the em algorithm. J. Royal Stat. Soc. Series B Stat. Methodol., 39: 1-38.
DOI: 10.1111/j.2517-6161.1977.tb01600.x

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, Mar. 24-27, San Diego, California, pp: 138-145.

Dwivedi, S.K. and P.P. Sukhadeve, 2010. Machine translation system in Indian perspectives. J. Comput. Sci., 6: 1111-1116.

Foster, G. and R. Kuhn, 2009. Stabilizing minimum error rate training. Proceedings of the 4th Workshop on Statistical Machine Translation, pp: 242-249.

Garje, G. and G. Kharate, 2013. Survey of machine translation systems in India. Int. J. Nat. Language Comput., 2: 47-67.

Hatem, A. and N. Omar, 2010. Syntactic reordering for Arabic-English phrase-based machine translation. In: Database Theory and Application, Bio-Science and Bio-Technology, Springer, Berlin, Heidelberg, ISBN-10: 978-3-642-17621-0, pp: 198-206.
DOI: 10.1007/978-3-642-17622-7_20

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, Aug. 4-9, Sofia, Bulgaria, pp: 690-696.

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

Kneser, R. and H. Ney, 1995. Improved backing-off for M-gram language modeling. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, May 9-12, IEEE Xplore Press, Detroit, MI, USA, pp: 181-184.
DOI: 10.1109/ICASSP.1995.479394

Koehn, P. and R. Knowles, 2017. Six challenges for neural machine translation. Proceedings of the 1st Workshop on Neural Machine Translation, Aug. 4, Vancouver, Canada, pp: 28-39.
DOI: 10.18653/v1/W17-3204

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit, 5: 79-86.

Koehn, P., F.J. Och and D. Marcu, 2003. Statistical phrase-based translation. Proceedings of the HLT-NAACL Main Papers, May 1, Edmonton, 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 ACL Demo and Poster Sessions, Jun. 2, ACL, Prague, pp: 177-180.

Kunchukuttan, A., A. Mishra, R. Chatterjee, R. Shah and P. Bhattacharyya, 2014. Shata-anuvadak: Tackling multiway translation of Indian languages. Proceedings of the 9th International Conference on Language Resources and Evaluation, May 1, Rekjyavik, Iceland, pp: 1781-1787.

Lopez, A., 2008. Statistical Machine Translation. 1st Edn., Cambridge University Press, ISBN-10: 0521874157, pp: 488.

Mohammed, E.A. and M.J.A. Aziz, 2011. English to Arabic machine translation based on reordring algorithm. J. Comput. Sci., 7: 120-120.

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: Arepresentative modern Bengali corpus. SUST J. Sci. Technol., 21: 78-86.

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.

Naskar, S., D. Saha and S. Bandyopadhyay, 2004. Anubaad-a hybrid machine translation system from English to Bangla. Simple'04.

Naskar, S.K. and S. Bandyopadhyay, 2006. Handling of prepositions in English to Bengali machine translation. Proceedings of the 3rd ACL-SIGSEM Workshop on Prepositions, (SWP' 06), pp: 89-94.

Och, F.J., 2003. Minimum error rate training in statistical machine translation. Proceedings of the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Jul. 07-12, ACM, Sapporo, Japan, Pages: 1-7. DOI: 0.3115/1075096.1075117

Och, F.J. and H. Ney, 2002. Discriminative training and maximum entropy models for statistical machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Jul. 07-12, ACM, Philadelphia, Pennsylvania, pp: 295-302. DOI: 10.3115/1073083.1073133

Och, F.J. and H. Ney, 2003. A system aticcomparison of various statistical alignmentmodels. Comput. Linguistics, 29: 19-51.

Och, F.J. and H. Ney, 2004. The alignment template approach to statistical machine translation. Comput. Linguistics, 30: 417-449.

Och, F.J., 1999. An efficient method for determining bilingual word classes. Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics, Jun. 08-12, ACM, Bergen, Norway, pp: 71-76. DOI: 10.3115/977035.977046

Och, F.J., C. Tillmann and H. Ney, 1999. Improved alignment models for statistical machine translation. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and very Large Corpora. (PLC' 99).

Östling, R. and J. Tiedemann, 2017. Neural machine translation for low-resource languages.

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 Association for Computational Linguistics, July 07-12, ACM, Philadelphia, Pennsylvania, pp: 311-318. DOI: 10.3115/1073083.1073135

Saha, G.K., 2005a. English to Bangla translator: The Banganubad. Int. J. Comput. Process. Languages, 18: 281-290.

Saha, G.K., 2005b. The eb-anubad translator: A hybrid scheme. J. Sci., 6: 1047-1050.

Salam, K.M.A., Y. Setsuo and T. Nishino, 2011. Translating unknown words using WordNet and IPA-based-transliteration. Proceedings of the 14th International Conference on Computer and Information Technology, Dec. 22-24, IEEE Xplore Press, Dhaka, Bangladesh, pp: 481-486. DOI: 10.1109/ICCITechn.2011.6164838

Sennrich, R., B. Haddow and A. Birch, 2016. Edinburgh neural machine translation systems for WMT 16.

Shirko, O., N. Omar, H. Arshad and M.A lbared, 2010. Machine translation of noun phrases from Arabic to English using transfer-based approach. J. Comput. Sci., 6: 350-350.

Sinha, R. and A. Jain, 2003. AnglaHindi:An English to Hindi machine-aided translation system. Indian Institute of Technology, Kanpur, India.

Sinha, R., K. Sivaraman, A. Agrawal, R. Jain and R. Srivastava *et al*., 1995. Anglabharti: A multilingual machine aided translation projection translation from English to Indian languages. Cybernetics, Intel. Syst., 2: 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. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Aug. 1-6), Cambridge, pp: 223-231.

Sutskever, I., O. Vinyals and Q.V. Le, 2014. Sequence to sequence learning with neural networks. Proceedings of the Advances Neural Information Processing Systems, (IPS' 14), pp: 3104-3112.

Tiedemann, J., 2012. Parallel data, tools andinterfaces in opus. Linguistics Res. Eng. Comput., 2: pp: 2214-2218.

Vogel, S., H. Ney and C. Tillmann, 1996. Hmm-based word alignment in statistical translation. Proceedings of the 16th Conference on Computational Linguistics, Aug. 05-09, Copenhagen, Denmark, pp: 836-841. DOI: 10.3115/993268.993313

Wu, Y., M. Schuster, Z. Chen, Q.V. Le and W. Macherey *et al*., 2016. Google's neural machine translation system: And bridging the gap between human machine translation.

Zens, R., F.J. Och and H. Ney, 2002. Phrase-based statistical machine translation. In: Advances in Artificial Intelligence. Jarke, M., G. Lakemeyer, J. Koehler (Eds.), Springer, Berlin, Heidelberg, ISBN-10: 978-3-540-45751-0, pp: 18-32

[i] www.anubadok.sourceforge.net
[ii] 2www.cfilt.iitb.ac.in/indic-translator
[iii] https://translate.google.com
[iv] www.globalvoices.org
[v] https://github.com/irshadbhat/indic-tokenizer
[vi] https://banglasketch.org/shu-torjoma