Bengali-English Neural Machine Translation Using Deep Learning Techniques

Preprint · April 2023

Nipun Paul
Ahsanullah University of Science & Tech
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Ishmam Faruki
Ahsanullah University of Science & Tech
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Mutakabbirul Islam Pranto
Ahsanullah University of Science & Tech
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Md Tanvir Rouf Shawon
Ahsanullah University of Science & Tech
33 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Accepted and Presented in ”2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)”

Bengali-English Neural Machine Translation Using Deep Learning Techniques

Nipun Paul*, Ishmam Faruki[†], Mutakabbirul Islam Pranto[‡], Md. Tanvir Rouf Shawon[§] and Nibir Chandra Mandal[¶]
Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology, Dhaka, Bangladesh
{nipun4338*, ishmamfaruki443[†], pranto1416[‡], shawontanvir95[§]}@gmail.com, nibir.cse[¶]@aust.edu

***Abstract*—Bengali is one of the most widely spoken languages and one of the hardest to translate due to its extensive vocabulary. Earlier, it was fairly difficult to translate from Bengali to English. Using neural machine translation (NMT), it is now possible to translate from Bengali to English quite flawlessly. In order to carry out the task of neural machine translation, four different Seq2Seq models — Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU)—are studied in this work. We combined four distinct datasets and used the resultant dataset in association with the four models. Our study shows that BiLSTM is the most effective model for the Bengali to English NMT task. Here, the resemblance between the generated output and the real one is assessed using two frequently used performance metrics termed BLEU and ROUGE. We have achieved scores of 47.4, 35.8, 32.0 and 22.8 for BLEU-1, 2, 3 and 4 respectively on BiLSTM. Last but not least, our outcomes are among the finest of other studies performed earlier.**

***Index Terms*—Machine Translation, Natural Language Processing, Bengali-to-English, Neural Machine Translation**

I. Introduction

Machine translation is now the most important field in natural language processing. It allows communication between speakers of various native languages and shares information by eliminating the linguistic divide. Word-by-word translation and sentence-by-sentence translation are two different types of machine translation. Compared to word-by-word translation, phrase translation provides more information [1]. The most recent approach to machine translation is Neural Machine Translation (NMT) [2]. It makes use of neural networks trained on enormous amounts of data. Memory requirements for NMT models are way lower than those for SMT models. Also, it surpasses traditional SMT models in terms of accuracy. The key objective here is to translate from Bengali to English and vice versa, as there are approximately 272 million native or second-language speakers. Bengali is the world’s sixth most spoken native language and seventh most spoken language overall. For the Bengali-English language pair, we worked on four seq2Seq (Sequence to Sequence) models: Seq2Seq learning using LSTM, Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU).

Sequence to Sequence [3] LSTM model uses layers of LSTM to generate hidden cells of the input sequence. It uses hidden cells of previous timesteps and then applies this hidden sequence to decode it into a target language. GRU typically acts as a recurrent unit, with a reset and an update gate controlling the amount of information that needs to flow from the history state and to the current input. Bidirectional long-short term memory or BiLSTM, where input flows in two directions: one processes input moving forward, while the opposite receives input in the reverse direction, making it a little different from a typical LSTM. And lastly, a Bidirectional GRU, or BiGRU, is a paradigm for sequence processing that consists of two GRUs, similar to a BiLSTM where only two gates: the input and forget gates consist of. In this paper, we have experimented with the above models and compared their results using the ROUGE score and BLEU evaluation metrics. Our contributions are summarized as follows.

- We have explored four different sequential model (LSTM, GRU, BiLSTM and BiGRU) and compared their capability of doing machine translation on a partial blend of four independent datasets.
- We have evaluated our implemented models using both the BLEU and ROUGE score which are widely used in evaluating the generation task.
- Additionally, we have obtained certain outputs that are comparable to those of earlier studies on neural machine translation.

II. Related Works

The architecture of a machine translation system from English to Bengali is described in the paper [4] by Shaykh Siddique et al. A recurrent neural network with encoders and decoders was used to create the system (RNN). The method presented maps English and Bengali sentences using a knowledge-based context vector. The best results are obtained when employing the linear activation function in the encoder layer and the tanh activation function in the decoder layer to measure the performance of the model. GRU fared better than LSTM when the two layers were being executed. The softmax and sigmoid activation functions are used to activate the attention layers. In terms of cross-entropy loss metrics, the methodology of the model performs better than previous cutting-edge systems. The reader can quickly detect the format of the machine translation from English to Bengali. In the same way, the study [5] by Gaurav Tiwari et al. evaluates and compares two NMT models for the English-Hindi language pair based on different parameters. Both encoders are used in the Sequence to Sequence Learning architecture. (1) LSTM - was used to implement the encoder and decoder. (2) Convolutional Neural Network (CNN) using LSTM. Both models use an attention mechanism to boost their actual performance. The investigation provides multiple insights which identify the best model and settings for this kind of translation. Finally, they declared ConvS2S as the best model according to BLEU. In another study [6] by Levi Corallo et al., the WMT-2021 English-German data set, which contains 400,000 strings from German news stories with parallel English translations, is used as the basis for a GRU-based recurrent neural network (RNN). Here, the system acts as a testbed for converting strings from German news sources into English sentences, laying the groundwork for future applications and research in the field. This approach can be helpful while creating mobile applications for rapid translation in real-world situations where effectiveness is essential.

The problems with training and translation inference where these were computationally expensive, were resolved in one of the most notable studies [7] by Yonghui Wu et al. Rare words have previously been problematic for NMT systems, a problem that was also resolved in this paper. The authors described the architecture of a deep LSTM network that utilized attention and residual connections and included eight encoder and eight decoder layers. In order to better handle uncommon terminology, they divided words into a small number of common sub-word units for both input and output. This approach handles uncommon words naturally, strikes a reasonable balance between the adaptability of ”character”-delimited models and the effectiveness of ”word”-delimited models, and gradually improves the system's overall accuracy. The English-to-French and English-to-German datasets from WMT'14 were employed, and the architecture produced results that were competitive with Google's phrase-based production system while lowering translation errors by an average of 60%. Another work [8] Md. Arid Hasan et al. explored the translation of the Bengali-English language pair using BiLSTM and Transformer-based NMT. The results the author produced using a variety of datasets significantly outperform earlier results on these datasets. This study also examined the factors that affect model performance and define data quality.

The studies like [9], [10] and [11] separately used LSTM, GRU, and BiLSTM in different languages like English to Hindi or Arabic machine translation, whereas paper [12] represents a descriptive survey on uses of NMT to Low-Resource Languages and proposed some possible solutions on different cases. Finally, for both directions of English-Bangla translation, this article [13] by MA Al Mumin et al. compared phrase-based statistical machine translation to neural machine translation. They achieved improvements over phrase-based statistical machine translation of up to +0.30 and +4.95 BLEU for translations from English-Bangla and Bangla-English respectively. They used subword segmentation with byte pair encoding to solve the problem of odd words and obtained up to +0.69 and +0.30 BLEU gains over baseline neural machine translation. The author found that phrase-based statistical machine translation underperformed neural machine translation for a number of difficult linguistic features, including subject-verb agreement, noun inflection, long-distance reordering, and rare word translation.

III. Background Study

This section describes the sequential models and the performance metrics we used in our study.

A. Sequential Models

Long short-term memory (LSTM) [14] networks are a form of recurrent neural network (RNN) [15] that can learn order dependency in sequence prediction challenges. Prior to the introduction of LSTM, RNN had a long-term reliance, which meant that it couldn't predict words stored in long-term memory but could make more accurate predictions based on current input. Apart from single data points such as pictures, LSTM features feedback links that let it to process the entire sequence of data. As a result, by default, LSTM may maintain information for a long time by preventing information from vanishing by managing the flow of information through gates. **Gated Recurrent Unit (GRU)** [16] is an advanced type of RNN and a simpler form of LSTM with fewer gates. Like LSTM, GRU also solves the vanishing gradient problem by allowing the network to have capabilities of both short and long-term memory. GRU only has one channel that carries the information, which is the hidden state. A **Bidirectional LSTM** (BiLSTM) is a sequence processing model which consists of two LSTMs where one takes the input in a forwarding direction, while the other in a backward direction. BiLSTM has the ability to control the flow of information in both directions, which makes it one of the most effective sequence processing models. Like BiLSTM, **Bidirectional GRU (BiGRU)** is also architecture with two GRUs in both forward and backward directions. It is a bidirectional recurrent neural network with only the input and forgets gates which like BiLSTM, control the flow of information.

B. Performance Metrics

Recall-Oriented Understudy for Gisting Evaluation or ROUGE [17] compares an autonomously generated summary or predicted translation to a set of reference or actual summaries. ROUGE scores are classified as ROUGE-1, ROUGE-2, and ROUGE-L where each individual calculates precision, recall and F1-score for each individual generated summary.
ROUGE-N =

$$\frac{\sum_{S\epsilon\{ReferenceSummaries\}gram_n\epsilon S}\sum Count_{match}(gram_n)}{\sum_{S\epsilon\{ReferenceSummaries\}gram_n\epsilon S}\sum Count(gram_n)} \tag{1}$$

The **Bilingual Evaluation Understudy (BLEU)** metric rates translations on a scale of 0 to 1, attempting to assess the sufficiency and fluency of machine translation output. The closer the test sentences score to 1, the greater the overlap with their human reference translations, and hence the better the system is regarded to be [18].

IV. Dataset

The dataset that we are going to use is bn-en tab-delimited bilingual sentence pairs which is a partial combination of four particular datasets: Tatoeba Project[1], Samanantar[2], ALT Project[3] and Tico-19[4].

TABLE I: Frequency table of four datasets.

Dataset	Bn-En Parallel Pairs
Tatoeba Project	4,618
Samanantar	8,605,000
ALT Project	20,000
Tico-19	3,073

We combined four datasets by randomly selecting small portions of single pair of sentences. We've used all the pairs (4618) from Tatoeba Project as the dataset is perfectly balanced. Though Samanantar is one of the largest resources of bn-en pair, the dataset is raw with unnecessary punctuations, symbols, and a mixture of both English words in Bengali sentences. The same can be seen on Tico-19, as they are gathered from multiple resources like online newspapers, blogs etc. So, we've randomly selected a few portions of pairs from Samanantar (350 pairs) and Tico-19 (100 pairs). From the ALT Project, we've chosen 136 pairs of Bn-En sentences. We then combined all the datasets into a single one and shuffled them in between. Table II indicates some examples from the dataset we used the dataset.

TABLE II: Examples from final dataset.

English Sentence	Bengali Sentence
Its a very tough test.	টেস্টে অনেক কঠিন পরীক্ষা দিতে হবে।
See you later, OK?	আবার পরে দেখা হবে, ঠিক আছে?

Finally, our custom combined bn-en dataset consists of 5204 English and Bengali tab-delimited bilingual parallel sentences. Because of our low-end resources the dataset is small, we have used 95% data for the training set and 5% data for the testing set. Among the training set, 10% data is split for validation purposes. The number of samples we have used as train, test and validation maintaining the above mentioned ratio are $4,943$**,** 261 **and** 495.

V. Methodology

In this section the complete workflow of our study is described. The workflow of our proposed methodology can be seen in Fig. 1.

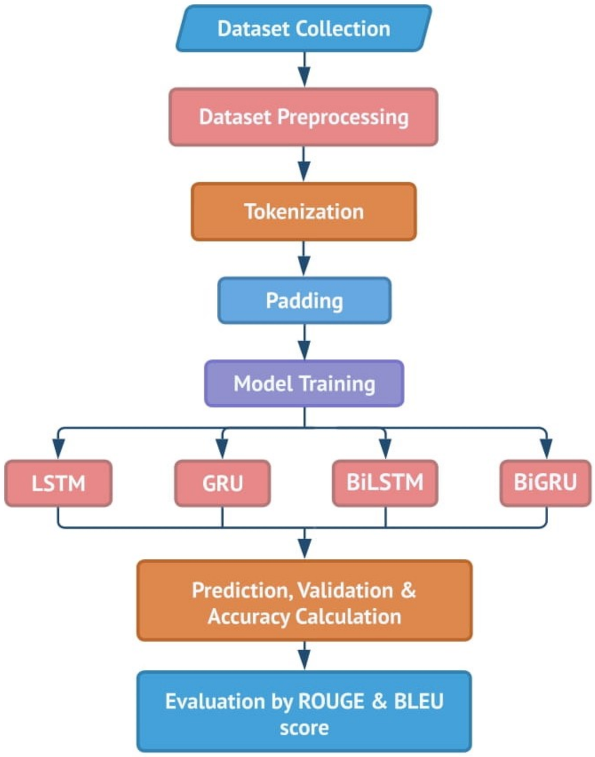

Fig. 1: Workflow of our paper.

A. Preprocessing

Some text preprocessing processes are performed in order to normalize the dataset. All the individual letters of the dataset are transformed to lowercase, and all punctuation is eliminated. Characters that do not relate to Bengali or English letters are also removed.

B. Tokenization and Padding

The maximum length for both English and Bengali sentences are 20 and 18 consisting of 1,835 and 3,400 unique words of vocabulary respectively. Now the dataset must be tokenized in its initial state. So all words in each Bengali and English sentence are tokenized based on their frequency.

As the maximum length of Bengali and English sentences differ from each other, we have padded the remaining blanks with 0 (Zeros). The input data has finally been vectorized and shaped and prepared for usage in our neural network models. The dataset is split between train and test sets and is fed to the neural network.

C. Model Training

As we are using sequential models, after padding, source vocabulary, in this case, Bengali along with the input length and output dimension is being sent as parameters to the

[1] http://www.manythings.org/anki/
[2] https://ai4bharat.iitm.ac.in/samanantar
[3] https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/
[4] https://opus.nlpl.eu/tico-19-v2020-10-28.php

embedding layer. The output of the embedding layer is then passed to the neural network layers: LSTM, GRU, BiLSTM, and BiGRU.

1) Activation Function: The activation function's objective is to add non-linearity to a neuron's output. We have used softmax, which is suggested as the best and most efficient output producer in this paper [4].

2) Loss Function: To train any neural network, errors need to be calculated so that the model gives a more accurate prediction, which is done with a back propagation loss function. We have used Categorical Crossentropy Function, which is best known as a multi-class classification loss function. It is a Softmax activation plus a Cross-Entropy loss.

3) Optimizer: Adam is an RMSprop and Stochastic Gradient Descent hybrid. And the primary benefit of employing Adam in our work is time optimization throughout the training dataset [19].

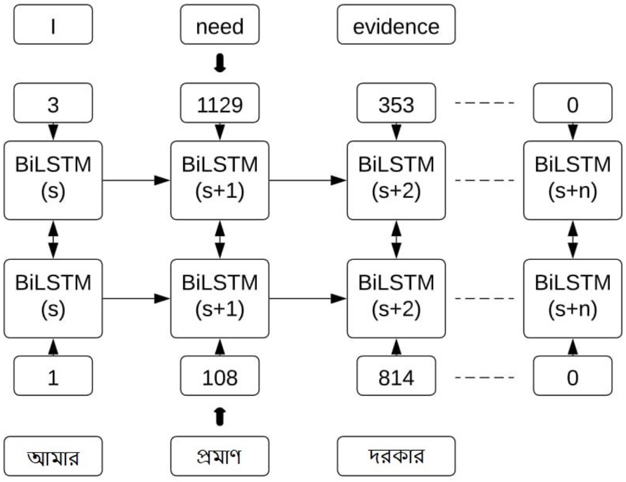

Fig. 2: Training of a sample with BiLSTM.

Fig. 2 shows the training of BiLSTM with a training sample where both Bengali and English tokenized sentences are fed to the neural network. The mapped output is then fed to the NMT model to get the translated sentence. For evaluating model performance, we have used **early stopping** to stop training when the minimization of validation loss has stopped improving. The patience was set to 10 to stop training after 10 epochs without improvement.

D. Prediction, Validation & Accuracy Calculation

After completing model training, accuracy is calculated upon the chosen loss function which is a built-in feature by Keras. We have calculated the accuracy for each of the four models by their prediction and validation results received from the training stage.

E. Evaluation by ROUGE & BLEU score

The last phase is the evaluation phase where evaluation is done upon each prediction using ROUGE and BLEU scores. We have used ROUGE-1, ROUGE-2, and ROUGE-L where for each individual generated sentence we have calculated the ROUGE score and merged it into a mean ROUGE score for calculating combined precision, recall, and F1-score for all of the four models separately. Also we have used BLEU-1, 2, 3, and 4. All the generated translations and reference sentences are passed to the BLEU metrics and in return, we get a BLEU score for the individual model.

VI. Experiments & Analysis

The architecture, optimized hyperparameters, performance, and analysis of the proposed models are all shown in this section.

A. Model Architecture

For both LSTM & GRU, we've used 256 embedding & 256 repeat vector layers surrounded by two 256 neural network layers. Lastly, the output of the neural network layer is fed to a time distributed layer containing a softmax activation function. The same architecture is followed for BiLSTM & BiGRU as well, but the surrounding two 256 neural network layers are bidirectional. Table III contains the optimized hyperparameters used in our models.

TABLE III: Optimized Hyperparameters.

Models	**Max Epoch**	**Batch Size**	**Learning Rate**
GRU	70	64	0.001
LSTM	50	64	0.001
BiLSTM	40	64	0.001
BiGRU	35	64	0.001

B. Experimental Results

We have implemented all four models and trained and tested them with our dataset. For evaluation purposes, we have checked training accuracy, validation loss, and validation accuracy and used ROUGE and BLEU score to measure the translation quality. The Accuracy vs Loss graph of BiLSTM can be seen in Fig. 3. Table IV denotes the ROUGE score and Fig. 4 shows the BLEU score achieved by the model on the testing samples.

C. Result Analysis

We see that BiGRU was way faster than the other three models to achieve a good training accuracy (95.5%) with fewer epochs (35). We stopped training after 35 epochs so that it doesn't overfit the dataset. The same result can be seen in validation accuracy (86.4%). So, undoubtedly BiGRU is the fastest among LSTM, GRU, and BiLSTM. BiLSTM is a bit similar to BiGRU with some extra epochs but with a little bit of higher accuracy rate on both training (96.3%) and validation (86.6%). On the other hand, LSTM is the slowest among others, which takes more than 70 epochs to get a decent accuracy rate in training (93%) and validation (85.3%). Validation loss has the same pattern as accuracy. So as of our study now, it is clear that the average performance of

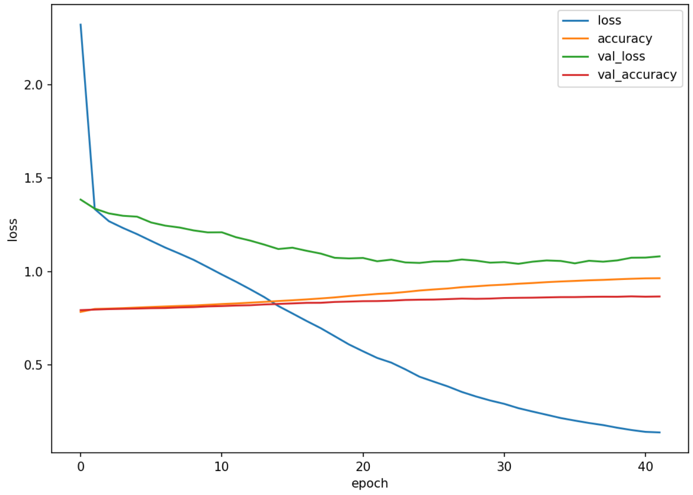

Fig. 3: Accuracy vs Loss graph of BiLSTM.

the BiLSTM layer is better than LSTM, GRU, or BiGRU. Now let's look at Table IV, where ROUGE scores are being calculated for each of the models.

TABLE IV: Average ROUGE score on test samples.

		Models			
		LSTM	GRU	BiLSTM	BiGRU
rouge-1	Recall	44.79%	41.93%	50.78%	46.67%
	Precision	52.87%	49.35%	60.62%	58.82%
	F1 score	47.55%	44.46%	54.37%	50.98%
rouge-2	Recall	23.97%	21.92%	30.08%	24.78%
	Precision	24.78%	22.29%	31.76%	25.98%
	F1 score	24.12%	21.97%	30.70%	25.12%
rouge-L	Recall	44.37%	41.42%	50.33%	45.48%
	Precision	52.26%	48.66%	60.01%	56.88%
	F1 score	47.07%	43.87%	53.86%	49.51%

From the Table IV, BiLSTM, BiGRU, and LSTM, all three of them are competing for the best place to achieve. As GRU is way behind them, our choices narrow down to BiLSTM, BiGRU, and LSTM. Not if we compare LSTM and BiGRU, we can see that in some specific cases, like precision and f1-score, BiGRU is way more accurate than LSTM (about 6-7%). So we can narrow down our list to BiLSTM and BiGRU. For BiLSTM, ROUGE-1, or for Unigram, the recall, precision, and F1-score is 50.78%, 60.62%, and 54.37% whereas for BiGRU it is 2-4% lower. The same scenario can be seen on ROUGE-2 or Bigram and ROUGE-L or for the longest matching sequence also. If we look closely at the difference between them, we will see something interesting here.

We have calculated the difference between BiLSTM and BiGRU on a scale of ROUGE score. The recall difference between BiLSTM and BiGRU in ROUGE-1 or Unigram was **4.11%**, which later changed to **5.3%** for ROUGE-2. But for both precision and f1-score, the difference was increased from **1.8%** to **5.78%** and **3.39%** to **5.58%** respectively. This means when we are considering more consecutive words, the difference between BiLSTM and BiGRU starts to increase and BiLSTM is going to surpass BiGRU at a decent speed. We know that ROUGE focuses on how much the words in the actual references appear in the targeted output [13]. Also, we know that BLEU focuses on how much the words in the targeted outputs appear in the actual reference [14]. To become fully aware of which one of them to choose, we are going to use the BLEU score to be more precise about it.

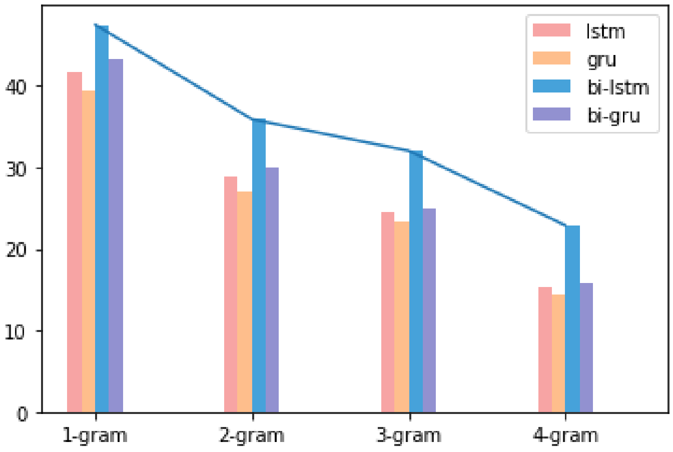

Fig. 4: BLEU score on test samples.

Now in Fig. 4, we see that for BLEU-1, BiLSTM holds the maximum score of 47.46% among all the models, whereas BiGRU has a score of 43.21%. And if we see clearly, gradually the scoring rate of BiGRU is dropping for BLEU-2 (29.9), 3 (25), and 4 (15.72). We know that BLEU-3 and 4 hold the maximum importance to measure whether the translation is merely better or nearly worst. On the other hand, LSTM and GRU both are maintaining a consistent score on each of the BLEU. They both scoring not too good, but also not too bad as well, where LSTM is slightly better (nearly 1-2%) than GRU in each of the cases and the difference between them is decreasing.

But for BiLSTM, there is only consistency. For BLEU-1, BiLSTM had the maximum score (47.46). Gradually, it started improving more and more over the other three models. If we see deeply, in BLEU-1, BiGRU was leg behind BiLSTM by **4.25%**. In BLEU-2, BiLSTM surpassed BiGRU by **5.96%** (35.86). For BLEU-3 and 4, BiLSTM scored **7.01%** (32.01) and **7.15%** (22.87) higher than BiGRU. In fact, the increasing rate of BiLSTM from BiGRU is holding a good consistency like that we have seen on ROUGE also, in the cases when we actually needed it, like BLEU-3 and 4. That means, for more consecutive words, BiLSTM is showing a best result than the other three models here which has been proved by both ROUGE and BLEU scores.

D. Performance Comparison

We have performed at a level that is comparable to past work on this domain. In our work, we have achieved a max BLEU of **22.87** for BLEU-4 in BiLSTM whereas, in this paper [5], the author has reached about 15.20 with 8 layers. We have achieved more than **33%** higher BLEU score with our model where 2 layers of BiLSTM are used with our small dataset compared to their huge dataset. Also in another paper [9] that we have already discussed before, the author used SUParatest2018 Bn-En dataset [20] and achieved 22.68 BLEU using Attention-based NMT with BPE whereas we've achieved **+0.19** more BLEU score than them using BiLSTM in our dataset. The same result can be seen in another paper [8], where we have achieved a **+3.63** BLEU score using BiLSTM where they used BiLSTM + BN-Emb, EN-Emb architecture on the same dataset. In another work [11], without using attention mechanism [21], the author gained the highest BLEU score of 19.98, where we are **+2.89** score ahead.

On the basis of all the information we have gathered, it can be said that BiLSTM is the most accurate in case of translation by measuring ROUGE and BLEU score in our architecture to boost the translation of Bengali to English and vice versa. Finally, Table V shows two sample outputs generated by the BiLSTM model. In the first output, our model perfectly generated the reference sentence without any error. But in the second sentence, our model generated duplicate words which caused an erroneous sentence that might be because of translating rare words. These errors can be solved if more efforts are given.

TABLE V: BiLSTM sample outputs.

Bengali Sentence	**Reference English Sentence**	**Generated Sentence**
তিনি আমাকে লিখতে শিখিয়েছেন	he taught me how to write	he taught me how to write
আমি একদমই ক্লান্ত নই	i m not tired at all	i m not at at tired

VII. Conclusion & Future Works

In this paper, we have used our minimum resources to achieve the best results we can get. We have shown that BiLSTM is performing better than any other model in case of more consecutive words proven by both ROUGE and BLEU score with a decent increasing rate and gained a score of **47.46**, **35.86**, **32.01** and **22.87** on BLEU-1, 2, 3 and 4 respectively. However, with a huge Bengali dataset with rare words, will BiLSTM show the same promising result is still an open research topic. It is difficult to determine a sentence's actual meaning. And in the case of translation, it is the hardest job to maintain grammatical errors, and collect rich vocabulary and huge computational power. All these parameters combinedly help to build a perfect, rich, meaningful translation that can easily understandable by anyone. So in the future, using multiple dense layers, tuning parameters and hyperparameters, combining powerful computational techniques, and using Bengali-English datasets containing huge vocabulary, like the combination of this resource [22] can give us more promising results.

References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," University of Southern California Marina Del Rey Information Sciences Inst, Tech. Rep., 2003.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

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

[5] G. Tiwari, A. Sharma, A. Sahotra, and R. Kapoor, "English-hindi neural machine translation-lstm seq2seq and convs2s," in *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2020, pp. 871–875.

[6] L. Corallo, G. Li, K. Reagan, A. Saxena, A. S. Varde, and B. Wilde, "A framework for german-english machine translation with gru rnn." in *EDBT/ICDT Workshops*, 2022.

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

[8] M. A. Hasan, F. Alam, S. A. Chowdhury, and N. Khan, "Neural machine translation for the bangla-english language pair," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–6.

[9] N. Bensalah, H. Ayad, A. Adib, and A. Ibn El Farouk, "Lstm vs. gru for arabic machine translation," in *Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020) 12*. Springer, 2021, pp. 156–165.

[10] P. Shalu and M. Meera, "Neural machine translation for english to hindi using gru," *Available at SSRN 3851323*, 2021.

[11] A. Roy, A. C. Dhar, M. Akhand, and M. A. S. Kamal, "Bangla-english neural machine translation with bidirectional long short-term memory and back translation," *Int. J. Comput. Vis. Signal Process*, vol. 11, no. 1, pp. 25–31, 2021.

[12] S. Ranathunga, E.-S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, 2021.

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

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] D. E. Rumelhart, "Learning internal representations by error propagation, in parallel distributed processing," *Explorations in the Microstructure of Cognition*, pp. 318–362, 1986.

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

[17] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

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

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "Supara-benchmark: A benchmark dataset for english-bangla machine translation," *IEEE Dataport*, 2018.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] T. Hasan, A. Bhattacharjee, K. Samin, M. Hasan, M. Basak, M. S. Rahman, and R. Shahriyar, "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation," *arXiv preprint arXiv:2009.09359*, 2020.