Bengali Dialect Conversion To Standard Bengali Language Using Deep Learning

Md. Ashab Mohiuddin, Md. Younus Hossain Ahsan, Md. Abid Rahman, and Fahim Ahmed
Department of Computer Science And Enginnering
Ahsanullah University Of Science And Technology
{190104128, 190104131, 190104141, 190104149}@aust.edu

I. Introduction

The term dialect is often used to characterize any way of speaking that differs from the standard variety of a language which is largely considered to be dialect-free. Dialects are usually formed around particular regions [1]. However, they may also be used within certain groups of people. Dialect has its roots in both time and space, and this is why it is always connected to geographical, cultural, and social points [2]. A few people in the world actually speak in standard language while the majority of the people speak in dialect[1].There are 2,700 languages with over 7,000 individual dialects spoken around the world today. Chinese Language has the most dialects in the world[2].

Dialects and Accents are two different things though the difference looks confusing sometimes. Accents are the change or the variety of pronunciation of words based on regions while Dialects are a complete change of words or ways of communicating altogether. It goes beyond mere pronounciation [3].

Bengali is the seventh most spoken language and the seventh most spoken language by the total number of speakers in the world [4]. We, Bengalis, have a rich variety of dialects. Approximately 38 dialects are spoken in the Bengali language. People from Rajshahi, Noakhali, Sylhet, Chattogram have the richest dialect in Bangladesh [5]. For example, In original Bengali: ‘ভাইয়ের কী খবর?’ In English it means "How are you, Brother?". In Noakhali Dialect: ‘ভাইসা খবর কিয়া?’ In Rajshahi Dialect: ‘কী খবর বে?’.

Understanding dialects used in different districts is tough and sometimes confusing for the people that are used to listening and speaking in standard Bengali languages. And also people who are used and comfortable with their regional dialect often find it difficult to communicate with the people who use standard bangla language. Corporate sectors, job sectors, banks, immigration centers and even restaurants, often find people struggling to understand each other especially foreigners whose have no idea about the language. A translator like human or even a machine can be a helping hand to resolve these issues.

Noakhali dialect is one of the most famous and richest dialects in Bangladesh. Approximately 7 million people speak in Noakhali dialect primarily in the Greater Noakhali region of Bangladesh as well as southern parts of Tripura in India [6]. But no notable work has been done yet that deals with the conversation of the Noakhali dialect.

Three different models have been tried with available datasets which are seq2seq, fine tunning bangla bert, transformer. Among three models seq2seq has been used but seq2seq does not get used in mainstream language translation. Transformer has performed to give a close translation of the language. It couldn’t reach the expectation due to having a comparatively small dataset.

II. Related Works

Already lots of work has been done in the language conversion field. Islam et al. [7] has used Encoder-Decoder based RNN method with attention and sequence to sequence learning model in their work translating Bangla to English and they have achieved 75% precision.

Dhar et al. [8] has used a transformer model consisting of NN based encoders and decoders and has achieved a BLEU score of 21.33.

Zahurul et al. [9] has used a phrase-based Statistical Machine Translation (SMT) system that translates English sentences to Bangla and they have achieved a BLEU score of 11.7 for long sentences and for short sentences it is 23.3.

Parida et al. [10] has developed a Multimodal Machine Translation (MMT) utilizing bi-lingual text for English to Bangla and achieved 51.1 BLEU score.

Jawaid et al. [11] has built baseline phrase-based MT (PBMT) and hierarchical MT systems where hierarchical MT significantly outperformed PMT. The highest single-reference BLEU score is achieved by the hierarchical system and reaches 21.58%.

III. Dataset

A. Data Collection

As mentioned above, no significant works have been done on this topic and no dataset is available. So, we have to make our own dataset from scratch. We have already collected 685 sentences/words both in Noakhali dialect and also its translation in standard bengali language. We have collected those data from people talking in Noakhali dialect, social media posts [12], blogs [13], online news portal [14].

[1] https://www.thoughtco.com/dialect-language-term-1690446.

[2] https://en.wikipedia.org/wiki/List$_{o}f_{v}arieties_{o}fchinese$

B. Data Description

There are two labels in our dataset: Standard Bengali Language and Noakhali Language. So far, we have collected 420 sentences with a minimum of four words per sentence in our dataset.

IV. Methodology

A. Data Preprocessing

The CSV file is loaded. 'Noakhali language' column as the input and the 'Standard Bangla language' as the target. Sentences have been tokenized using spaces. Every sentence has been turned into an array of tokens. The special token has been added in the beginning '<sos>' and in the end '<eos>'. The dataset has been split into train_data, valid_data, test_data according to 70%, 20%, 10%. Train data is turned into vocab with index numbers. <eos>, <sos>, <pad> have the indices of 1, 2 and 3 in the vocab. Then, each batch with 64 sentence tensors with each tensor having padding of the largest sentence is made. For the transformer model, any sentence having more than 30 words is excluded. Any sentence having a word that does not exist in the dictionary is excluded from the dataset.

B. Encoder

The encoder class is taking a tensor of size (sequence length x batch size). Embedding is happening by taking input as vocab size and turning it into an embedded vector of dimension 300. LSTM is taking the output of the embedding as input which has the size of (sequence length batch size x embedding size). It returns hidden and cell. Each having size of (a layer size, batch size, hidden size) hidden size is 1024.

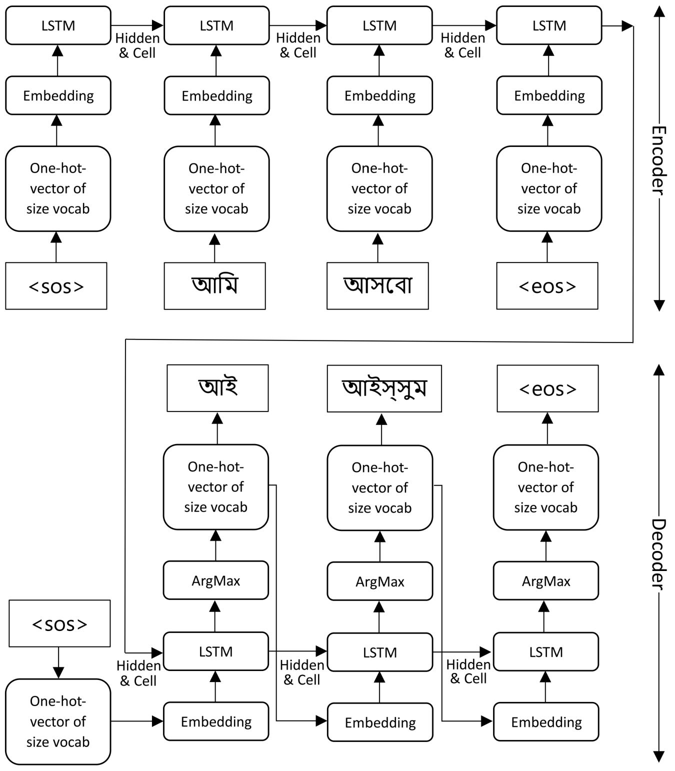


Fig. 1. Seq2seq Model

C. Decoder

The decoder is almost like an encoder but takes a single-word tensor with an input size of (1, batch-size).

D. Seq2seq Model

The seq2seq class merges the encoder and decoder. The decoder goes in a for loop of the size of the largest token array of the batch.

E. Bangla Bert (Fine Tuning)

Banglabert has been built by a community from the BUET CSE department. It is used for fine-tuning on our small datasets.

F. Transformer Model

5 layers of encoder have been used. 30 as the max sequence length of each sentence. Number of attention-heads is eight each of 3 times 64 since embedding output is 512 for each word which goes through a feed-forward network and output is 1536 each . Each of the head is divided into query, key and value vectors.

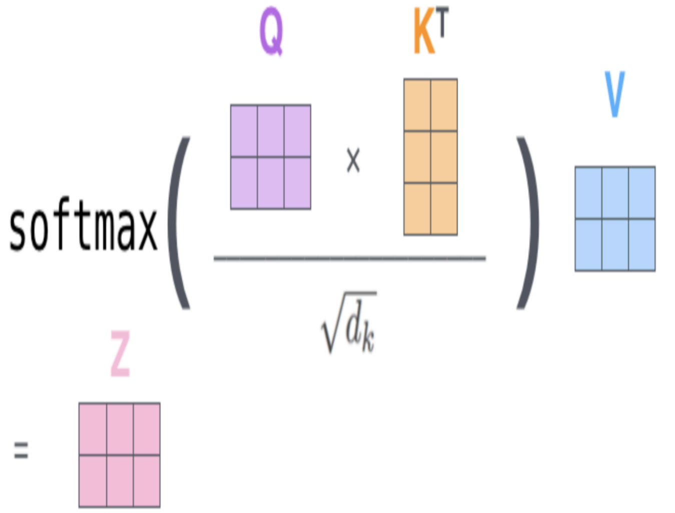


Fig. 2. Softmax Activation Function

Hyper-parameters: 500 epochs with data-set of 100 with 100 batch-size(after splitting 75 for training 15 for validating and 15 for testing where batch size is set to 5). 0.01 learning rate. Feed-forward network after normalization and addition of embedded added with positional encoding outputs(Feed-Forward) 2048. Drop out probability of 0.1. Number of layers encoder and Decoder each of 5. Maximum sentence length 30. Dictionary size 142. Self-attention and cross-attention matrix size 30x30 for each attention-head. Padding Mask is added to Encoder and Padding Mask and Attention Mask has been added to the Decoder.

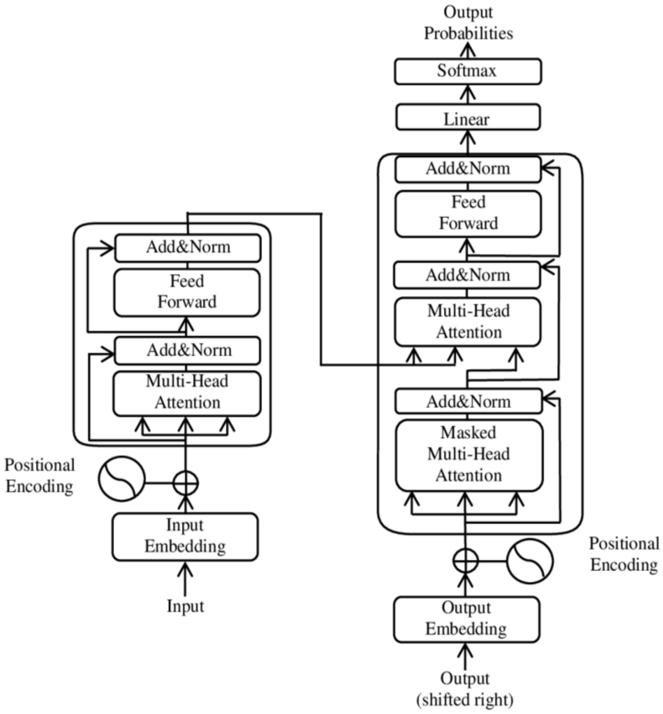


Fig. 3. Transformer Model Architecture

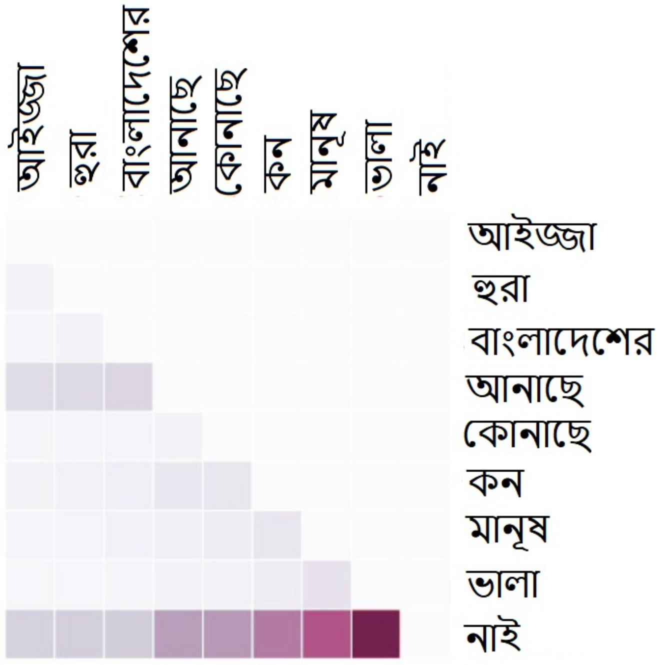


Fig. 4. Self-Attention and Cross-Attention

V. Result

A. Seq2Seq

Since, the model is being trained from scratch hence lack of enough data could not reach the expectations.

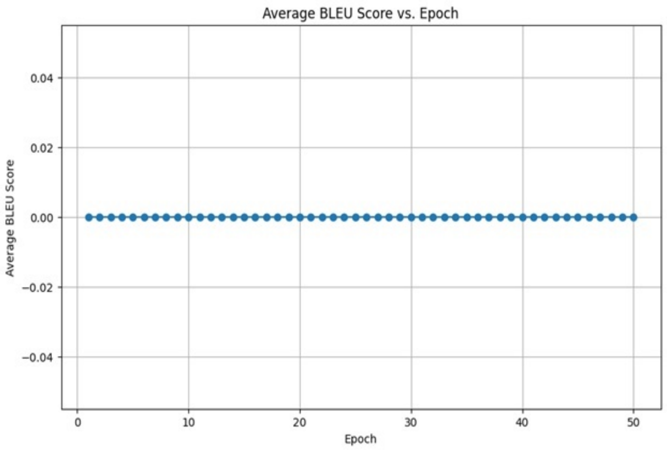


Fig. 5. Seq2seq Model

B. Transformer Model

Validating on train dataset for 100 epochs:

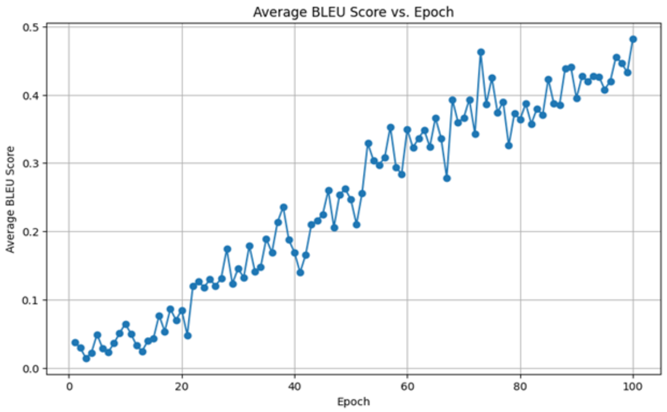


Fig. 6. Validating on train dataset for 100 epochs

For 500 epochs average BLEU score per batch graph follows:

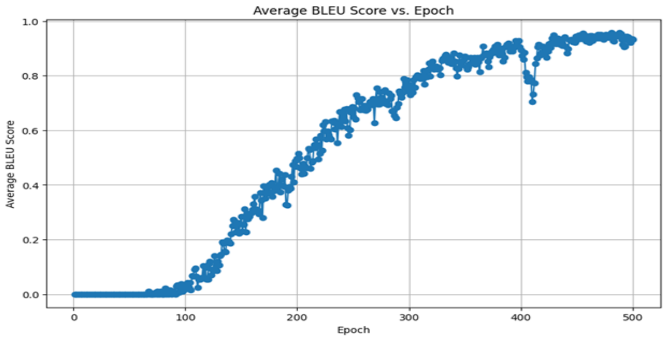


Fig. 7. Average BLEU score per batch graph For 500 epochs

Splitting the dataset into validation and train sets then validating on validation set average BLEU score per batch for 500 epoch statistic is in the following graph:

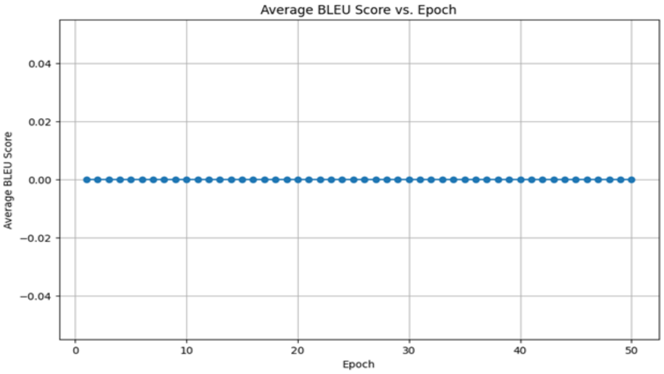


Fig. 8. Average BLEU score per batch over 500 epochs for splitting the dataset into train and validation set

Loss over the training process:

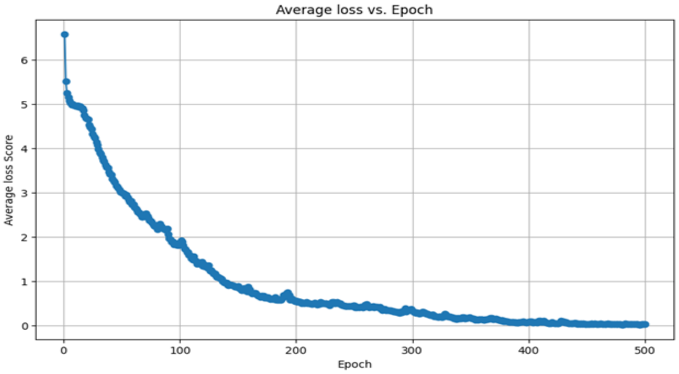
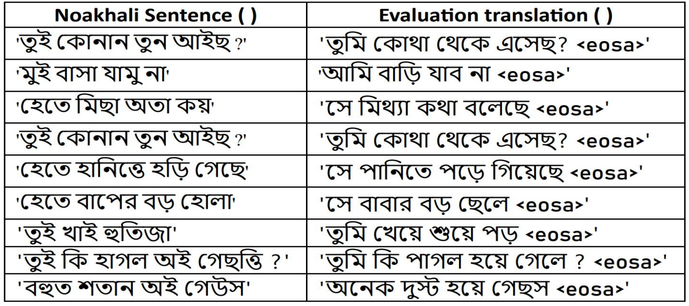


Fig. 9. Loss over the training process

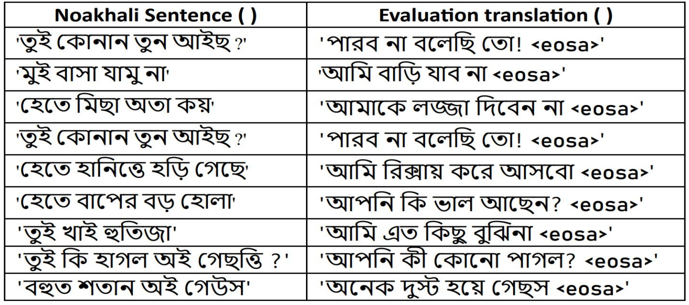
Some outputs on train set:



Noakhali Sentence ( )	Evaluation translation ( )
'তুই কোনান তুন আইছ ?'	'তুমি কোথা থেকে এসেছ? <eosa>'
'মুই বাসা যামু না'	'আমি বাড়ি যাব না <eosa>'
'হেতে মিছা অতা কয়'	'সে মিথ্যা কথা বলেছে <eosa>'
'তুই কোনান তুন আইছ ?'	'তুমি কোথা থেকে এসেছ? <eosa>'
'হেতে হানিত্তে হড়ি গেছে'	'সে পানিতে পড়ে গিয়েছে <eosa>'
'হেতে বাপের বড় হোলা'	'সে বাবার বড় ছেলে <eosa>'
'তুই খাই হুতিজা'	'তুমি খেয়ে শুয়ে পড় <eosa>'
'তুই কি হাগল অই গেছত্তি ?'	'তুমি কি পাগল হয়ে গেলে ? <eosa>'
'বহুত শতান অই গেউস'	'অনেক দুস্ট হয়ে গেছস <eosa>'

Fig. 10. Some outputs on train set

Some outputs on test set:



Noakhali Sentence ( )	Evaluation translation ( )
'তুই কোনান তুন আইছ ?'	'পারব না বলেছি তো! <eosa>'
'মুই বাসা যামু না'	'আমি বাড়ি যাব না <eosa>'
'হেতে মিছা অতা কয়'	'আমাকে লজ্জা দিবেন না <eosa>'
'তুই কোনান তুন আইছ ?'	'পারব না বলেছি তো! <eosa>'
'হেতে হানিত্তে হড়ি গেছে'	'আমি রিক্সায় করে আসবো <eosa>'
'হেতে বাপের বড় হোলা'	'আপনি কি ভাল আছেন? <eosa>'
'তুই খাই হুতিজা'	'আমি এত কিছু বুঝিনা <eosa>'
'তুই কি হাগল অই গেছত্তি ?'	'আপনি কী কোনো পাগল? <eosa>'
'বহুত শতান অই গেউস'	'অনেক দুস্ট হয়ে গেছস <eosa>'

Fig. 11. Some outputs on test set

VI. Conclusion

Noakhali Dialect to Standard Bangla Language Translation is one of the works that has not begun on a much larger scale. This paper is one of the very few, if not the only, done on this work. A dataset has been prepared from scratch. Few models have been run on the dataset. Seq2seq is not meant for language translation, and the result shows how poorly it performed. Fine-tuning on BanglaBert has been tried out with the small amount of data available. It did not work out due to technical confinement and limitations. The transformer model built from scratch performed all the other methods, which were giving a 90% BLEU score on train data and translating them accurately. But on the validation set, it could not do well due to a lack of enough data. In the future, research work will be extended with many other different approaches.

References

1 https://www.facebook.com/thoughtcodotcom/, “Do You Know What a Dialect Is? — thoughtco.com,” https://www.thoughtco.com/dialect-language-term-1690446, [Accessed 17-Jul-2023].

2 “Dialect: Development and Significance - 4112 Words | Research Paper Example — ivypanda.com,” https://ivypanda.com/essays/dialect-development-and-significance/, [Accessed 17-Jul-2023].

3 “Language vs. Dialect Vs. Accent: Learn The Differences — dictionary.com,” https://www.dictionary.com/e/language-vs-dialect-vs-accent/, [Accessed 17-Jul-2023].

4 “The world facebook.” [Online]. Available: www.cia.gov

5 “Dialect.” [Online]. Available: https://en.banglapedia.org/index.php/Dialect

6 null, n. and null, n., “Noakhali dialect.” [Online]. Available: https://en.everybodywiki.com/Noakhali_Dialect

7 Islam, R., Hasan, M. M., Rashid, M., and Khatun, R., Bangla to English Translation Using Sequence to Sequence Learning Model Based Recurrent Neural Networks, 06 2023, pp. 458–467.

8 Dhar, A., Roy, A., Akhand, M. A. H., Kamal, M., and Siddique, N., “Bangla english machine translation using attention-based multi-headed transformer model,” Journal of Computer Science, vol. 17, pp. 1000–1010, 10 2021.

9 Islam, Z., Tiedemann, J., and Eisele, A., “English to Bangla phrase-based machine translation,” in Proceedings of the 14th Annual Conference of the European Association for Machine Translation. Saint Raphaël, France: European Association for Machine Translation, May 27–28 2010. [Online]. Available: https://aclanthology.org/2010.eamt-1.21

10 Parida, S., Panda, S., Biswal, S. P., Kotwal, K., Sen, A., Dash, S. R., and Motlicek, P., “Multimodal neural machine translation system for English to Bengali,” in Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021). Online (Virtual Mode): INCOMA Ltd., Sep. 2021, pp. 31–39. [Online]. Available: https://aclanthology.org/2021.mmtlrl-1.6

11 Jawaid, B., Kamran, A., and Bojar, O., “English to Urdu statistical machine translation: Establishing a baseline,” in Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 2014, pp. 37–42. [Online]. Available: https://aclanthology.org/W14-5505

12 “Correct word practice.” [Online]. Available: https://shorturl.at/dlAK4

13 “English-i bangla.” [Online]. Available: https://shorturl.at/bhsI9

14 “English-i bangla.” [Online]. Available: https://shorturl.at/dnv12