Bengali abstractive text summarization using sequence to sequence RNNs

Md Ashraful Islam Talukder
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
islam15-7100@diu.edu.bd

Sheikh Abujar
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
sheikh.cse@diu.edu.bd

Abu Kaisar Mohammad Masum
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
mohammad15-6759@diu.edu.bd

Fahad Faisal
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
fahad.cse@diu.edu.bd

Syed Akhter Hossain
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
aktarhossain@daffodilvarsity.edu.bd

***Abstract*— Text summarization is one of the leading problem of natural language processing and deep learning in recent years. Text summarization contains a condensed short note on a large text document. Our purpose is to create an efficient and effective abstractive Bengali text summarizer what can generate an understandable and meaningful summary from a given Bengali text document. To do this we have collected various texts such as newspaper articles, Facebook posts etc. and to generate summary from those text we will be using our model. Our model works with bi-directional RNNs with LSTM in encoding layer and attention model at decoding layer. Our model works as sequence to sequence model to generate summary. There are some challenges we have faced while building this model such as text pre-processing, vocabulary counting, missing words counting, word embedding, unknown words find out and so on. In this model, our main goal was to make an abstractive summarizer and reduce the train loss of that. During our research experiment, we have successfully reduced the train loss to 0.008 and able to generate a fluent short summary note from a given text.**

Keywords*— *Natural Language Processing, Deep Learning, Text Pre-processing, Word-Embedding, Missing Word Counting, Vocabulary Counting, Bi-directional RNNs, Attention model, Encoding, Decoding.

I. INTRODUCTION

Conversation in natural language is one of the challenging problem in artificial language and summarization is the crucial part of that. Human can easily summarize any passage or text automatically. They select the salient words from a given text and make the summarization. In the computer system the task is far difficult. Because summarization making is involved with language understanding, reasoning and utilization of common sense knowledge like human.

We can summarize any text or passage with two different methods and those are extractive and abstractive methods. Most of the summarization methods are working as extractive method where first needs dragging the salient words or lines from the given passage. Then to make summary needs to combine them. In abstractive method, it produces a bottom-up summary from the passage where every words may not be in the main passage. This means, abstract method can generate summary of a given passage from itself.

There are very few works happened for Bengali text summarization. In this paper, our proposed method works with abstractive text summarization to create summary. We collected a lots of data and trained the model with those data. After accomplishing the training, we have got a satisfactory result.

Working with text is always challenging. To produce a proper summary with abstractive method we need to undergone some procedure i.e. text pre-processing, vocabulary counting, missing word counting, word embedding and counting, using some special tokens for word encoder and decoder. In our method we work with all these steps. We have applied sequence to sequence model with a two layered bidirectional RNN. On the input text and two layers RNN, each with LSTM using Bahdanau attention model [1] on the target text to produce an efficient summary. Encoder encodes all the input sentences into a fixed-length vector from which decoder makes an output sequence. This model was originally used to resolved relevant

text such as machine translation and we have modified that method to Bengali text summarization.

We have discussed various necessary factors concerning in text summarization to improve the efficiency to generate a more fluent and effective summary. Major processes are mentioned in the details and more importantly deep learning methods and models are explained here.

II. Literature Review

Neural machine translator is alike a traditional machine translator but recently a great approach to machine translation (Kalchbrenner et al, 2013) [8]. An individual neural network that is able jointly turned translation mostly associated with encoder and decoder. For Improving the performance of basic encoder and decoder, using a fixed length of text as input that generates output by the decoder (Bahdanau et al 2014) [1]. Abstractive text summarization creates a summary of a text document using it's intrinsic and chooses the key content of the text document using potential vocabulary. A sequence of aim words as input text in a source text document and predicted aim words of a sequence is called a summary of a text document. For a short text, summarizer has established a self-encoder, decoder RNN attention model on machine translation to text summarization (Ramesh Nallapati, Bowen Zhou, et al,2016) [2].

Sutskever et al. [4] describe an end-to-end approach to sequence to sequence learning used a Multilayer LSTM. The neural network contains encoder and decoder. Encoder used a fixed length of text using as input and Decoder represents the output.

The sequence to sequence learning has done improved neural machine translation. There are two ways of attention model (Minh-Thang Luong et al .2015) [5] one global and another is local. Global approach accepts all of the source text words and local accept the only subset of source text at a time. Both approaches are effective for machine translation.

Neural network response to producer short text conversion. Lifeng Shang et al [10] proposed Neural Responding Machine which proposed Neural Responding Machine which follows normal encoding and decoding.

III. Methodology

In this section, we will represent our model to making an abstractive Bengali text summarizer. Previously there are very few works have been done for Bengali text summarization and that's why we tried to make a text summarizer that can generate a proper summary from a given text. To build up this model and training, we have used tensorflow CPU version-1.13.1. Figure-1 shows the workflow of our model.

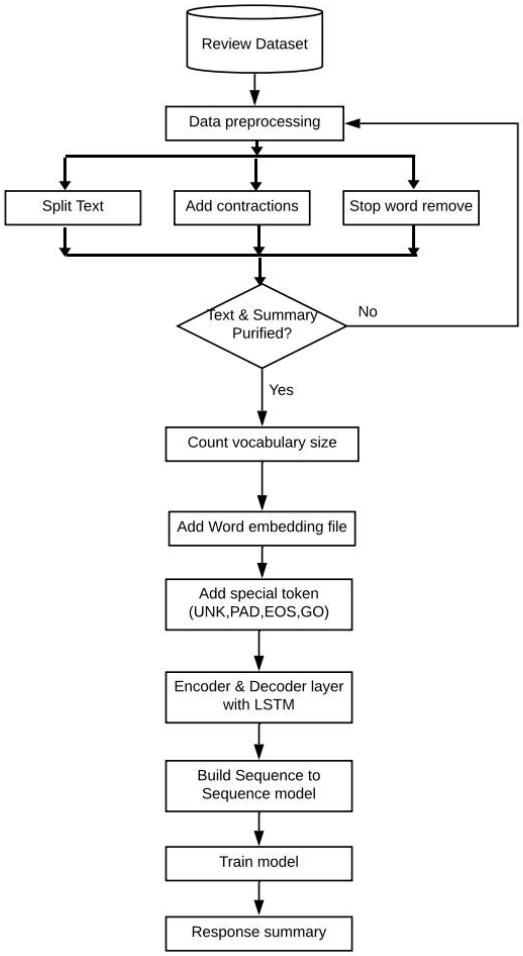

Figure-1: work flow

A. Problem Assertion

We have got a kind of big data set which consists of texts and same number of summary of those text. Let us consider the input sequence of those text has D words, therefore, $x_1, x_2, \dots, x_d$ is coming to the vocabulary size V which generated output sequence that will be as similar of $y_1, y_2, \dots, y_s$ where S<D. That means a summary sequence is less than the text description sequence. Consider that all of the output sequence are coming from the same vocabulary.

B Data Collection

Every deep learning algorithms needs a huge number of data. As large is the dataset is the result is that much better. For our model we need a handsome number of data too. But there is not enough dataset available in online, therefore, we have collected various types data such as news, Facebook posts etc. and then made summary of those texts for our research purpose. In our dataset, there is only two columns available which is necessary and those are text descriptions and summaries.

C. Data Preprocessing

For pre-processing of data, we have undergone of some procedures. First we have added contractions in the both summary and texts descriptions. There are several

contractions available such as, "বি.দ্র.", "ড.", "মো." etc. So, we have remove these and added the full form of those contractions. Then we have cleaned the texts. That means we have removed all the unwanted characters. We have used regular expression to remove those unnecessary component from the texts. After that, we have removed stop words.

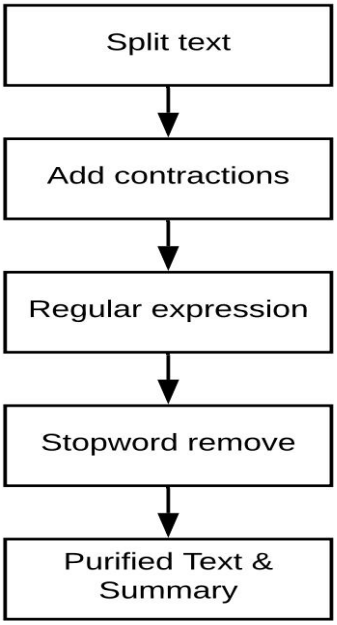

Figure2: Data pre-processing

D. Vocabulary Count & Word embedding

The purport of words does not depend only on frequency but also depends on word similarity. So, we need to count the total number of vocabulary from the purified text descriptions and summary. After the vocabulary counting, we have checked word occurrence such as we have tested the word "দেয়া" and the occurrence of this word was 4.

We have used a pre-trained word to vector file to improve our model. We have used "bn_w2v_model" word to vector file.

E. Model

There are a lots of models in deep learning and different types of model used for different types of purpose. While we are working with text, longest short term memory (LSTM) will be very helpful for text modeling. Machine translation is very important for learning a machine about text sequence. Every translator uses encoder and decoder such as Google translator. The translator translates a sequence of text in a language to another language.

i. Neural Machine Translation

Neural machine translation is an approach to translate one language into another language. Most of the machine translation use encoder and decoder to translate one language into another language. Encoder takes the input sequence and decoder predicts the output sequence and shows that. Neural machine translation uses a target sentence x that maximizes that conditional probability of x. If y is the source sentence then, $arg\left(max_y p(x|y)\right)$.

ii. RNN Encoder–Decoder

The very first two layers RNN encoder-decoder architecture introduces by Cho et al [11]. Later this was exaggerated by Bahdanau et al [1]. Those encoder and decoder model was used only for machine translation.

We have used this a neural network which contains 2 layers RNNs. Encoder contains a fixed length of a sentence and decoder contains the sequence of output. The 2 layers RNNs network are trained unitedly and keep the maximum conditional probability of target text sequence. Hidden unit used to improved memory capacity and training. We train our model to learn the probability of a Bengali sentence to the corresponding Bengali sentence.

If encoder read a target Input sentence $X = (x_1, \dots \dots, x_{T_x})$, in table 1 & 2 Input words are the input of the model. Where c is a context vector, so

$$h_t = f(x_t, h_{t-1}) \quad (1)$$

and

$$c = q(\{h_1, \dots . . h_{T_x}\})$$

Where, h_t= hidden state at the time t. c = Context vector which is generated from hidden state sequence. f and g is non- linear function.

If the decoder predicted word sequence $\{y_1, \dots . . , y_{T_y}\}$, Response Summary of table 1 & 2 are predicted sequence then the probability will be,

$$p(y) = \prod_{t=1}^{T} p(y|\{y_1, \dots, y_{t-1}\}, c) \quad (2)$$

Where, $(y_1, \dots . y_{T_o})$ .Now conditional probability is modeled by,

$$p(y|\{y_1, \dots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \quad (3)$$

Where, g = non-linear function, y_t = output of probability, s_t= hidden sate.

$$c_i = \sum_{j=0}^{T} a_{ij}\, h_j \quad (4)$$

We have used Bi-directional RNN's. Which is consists of forward and backward recurrent neural network. Forward recurrent neural network sequence order is $(x_1\ to\ x_{T_x})$ and the hidden state is $(\overrightarrow{h_1}, \dots, \overrightarrow{h_{T_x}})$. Backward recurrent neural network sequence order is $(x_{T_x} to\ x_1)$ and hidden state is $(\overleftarrow{h_1}, \dots, \overleftarrow{h_{T_x}})$. So,

$$h_j = [\overrightarrow{h_j{}^T}; \overleftarrow{h_j{}^T}]^T \quad (5)$$

Where, h_j=Summary of predicting and following words.

Here, a_{ij}= is softmax of e_{ij} which is normalize exceptional function and show how input position j align with output at position i,

$$e_{ij} = a(s_{i-1}, h_j) \quad (6)$$

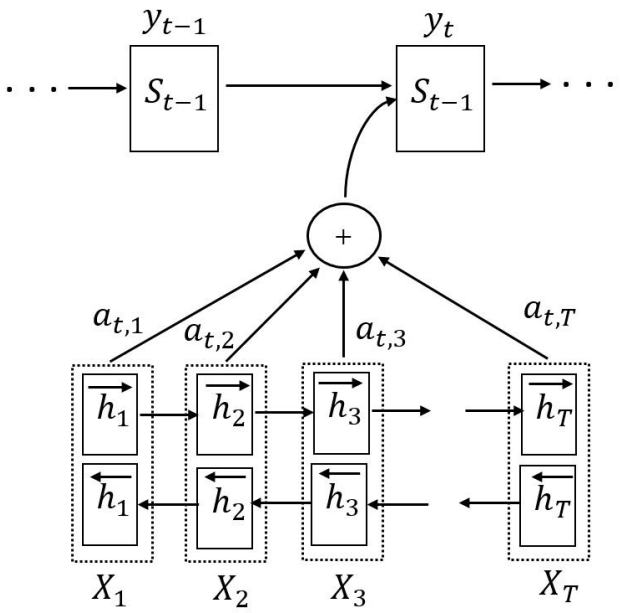

Figure3: View of model

iii. Sequence to Sequence Model

Every sequence to sequence model has encoder and decoder with LSTM cell. In our text summarize method, we have used a word embedding file. Then we did count the vocabulary size of those files which will be used our model input.

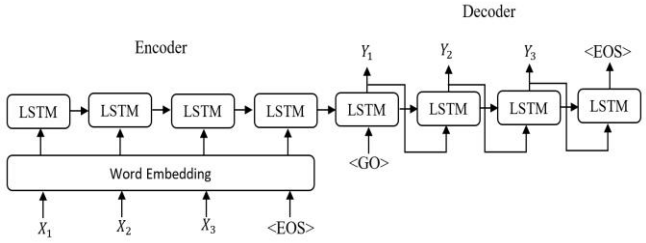

Figure4: Sequence to Sequence model

We have included some special tokens in vocabulary such as <UNK>, <PAD>, <EOS>, <GO>. Vocabulary is limited for some reasons. Some word has not been replaced. Those words are replaced by UNK token. PAD token adds a batch size of each sentence has the same length. EOS token contains the end of the sequence which gives signals to encoder when receiving the input. GO token give instruction to start the process of output sequence in the decoder. In data preprocessing stage, we add UNK and replace the vocabulary. Before train data, we chose GO and EOS in data which contains words id used sequence translation. In this, the sequence to sequence mode *x* is input sequence of encoder and *y* are generated output or response output sequence.

IV. Experiment and Output

We have used tensorflow1.13.1 and sequence to sequence model. When stop training we'll able to build a machine's own summary. For making summary we will take input sentence from the dataset and define the summary length randomly. For the parameter, we have used attention based encoder. We took epoch=70, batch size=2, rnns size=256, learning rate=0.001, keep probability=0.75 and we used Adam Optimizer, which calculated the learning rate of each parameter. For faster converges used vanilla gradient descent optimizer.

Here is given positive response from machine after few hours of training with our model and dataset:

Table1: sample result 1

Original Text:	অনেক অনেক দিন পর তোর হাতের আলতো ছোঁয়া টা আজ আবার পেলাম,অনেক ইচ্ছে ছিল তোর হাতটা ধরে সূর্য উঠার ঠিক আগ মুহূর্তে আর একবার এই ব্যাস্ত শহরের অলিগলি রাস্তা গুলোতে ঘুরবো,কিন্তু ভাবিনি আজি সেই দিনটা হবে!!!অনেক ধন্যবাদ "নীলাম্বরী" এই সকাল টা আমার করে দেওয়ার জন্য।
Original Summary:	ধন্যবাদ "নীলাম্বরী"সকালটা আমার করে দেওয়ার জন্য।
Input Words:	অনেক অনেক দিন পর তোর হাতের আলতো ছোঁয়া টা আজ আবার পেলাম অনেক ইচ্ছে ছিল তোর হাতটা ধরে সূর্য উঠার ঠিক আগ মুহূর্তে আর একবার এই ব্যাস্ত শহরের অলিগলি রাস্তা গুলোতে <UNK> কিন্তু ভাবিনি আজি সেই দিনটা হবে অনেক ধন্যবাদ নীলাম্বরী এই সকাল টা আমার করে দেওয়ার জন্য
Response Summary:	ধন্যবাদ নীলাম্বরী সকালটা আমার করে দেওয়ার জন্য

Table2: sample result 2

Original Text:	আমি এটুকুই চাইব এই অবহেলিত জনগোষ্ঠী যেন আর অবহেলার শিকার না হয়। - প্রধানমন্ত্রী শেখ হাসিনা পর্যায়ক্রমে দেশের আটটি বিভাগীয় শহরে বৃহৎ পরিসরে পিতামাতা ও অভিভাবকহীন নিউরো-ডেভেলপমেন্টাল প্রতিবন্ধী মেয়েদের জন্য পরিচর্যা কেন্দ্র স্থাপন করা হবে। এসব কেন্দ্রে তাদের শিক্ষা, প্রশিক্ষণ, চিকিৎসা, খেলাধুলাসহ সব সুবিধা অন্তর্ভুক্ত থাকবে।
Original Summary:	প্রতিবন্ধী মেয়েদের জন্য পরিচর্যা কেন্দ্র স্থাপন করা হবে।
Input Words:	আমি এটুকুই চাইব এই অবহেলিত জনগোষ্ঠী যেন আর অবহেলার শিকার না হয় প্রধানমন্ত্রী শেখ হাসিনা পর্যায়ক্রমে দেশের আটটি বিভাগীয় শহরে বৃহৎ পরিসরে পিতামাতা ও অভিভাবকহীন

	নিউরো ডেভেলপমেন্টাল প্রতিবন্ধী মেয়েদের জন্য পরিচর্যা কেন্দ্র স্থাপন করা হবে এসব কেন্দ্রে তাদের শিক্ষা প্রশিক্ষণ চিকিৎসা খেলাধুলাসহ সব সুবিধা অন্তর্ভুক্ত থাকবে
Response Summary:	প্রতিবন্ধী মেয়েদের জন্য পরিচর্যা কেন্দ্র স্থাপন করা

V. Conclusion and Future work

We have presented our model through this paper to making Bengali to Bengali summary from a given text using encoding and decoding with LSTM. No model gives us hundred percent accurate predicted results and our model also do so. But our model can provide the maximum accurate predicted summary. There are some drawbacks in our model hence, we are successfully able to build an understandable, meaningful, fluent and short summary with reducing the training loss.

The prime limitation was about the dataset during our research experiment. Because we have not get any dataset available in online therefore we had to create our own dataset. It's not easy to create dataset for summarization and we know that as large the dataset is the deep learning algorithms gives that much better output. Therefore, we are still collecting and making our dataset bigger. Another limitation is our model could make summary with a limited words and we will try to make it big so that it can generate summary from a text which contains unlimited words. There is also not enough better word to vector and even no lemmatizer available for Bengali language. In future, we will try to solve these problems and hope after that we will able to make better text summarization model for Bengali language.

VI. Acknowledgment

We are very thankful to our Daffodil International University Natural Language Processing (DIU-NLP) and Machine Learning Research lab for providing all of the research facilities. We would like to thanks our supervisor for his patience and supports to overcome various obstacles.

References

[1] Dzmitry Bahdanau et al. "Neural Machine Translation by Jointly Learning to Align and Translate". International Conference on Learning Representation (ICLR), 19 May 2014.

[2] Ramesh Nallapati, Bowen Zhou, et al "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond".The SIGNLL Conference on Computational Natural Language Learning (CoNLL), 26 Aug 2016.

[3] K.Cho, B .van Merrienboer, D.Bahdanau, Y.Bengio " On the Properties of Neural Machine translation: Encoder-Decoder Approaches". Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8),7 oct 2014.

[4] Sutskever et al "Sequence to Sequence Learning with Neural Networks". Conference on Neural Information Processing Systems (NIPS,2014).

[5] M. Luong, H. Pham, Christopher D. Manning "Effective Approaches to Attention-based Neural Machine Translation". Conference on Empirical Methods in Natural Language Processing (EMNLP 2015).

[6] Peter J. Liu et al. "Generating Wikipedia by Summarizing Long Sequences". International Conference on Learning Representation (ICLR), 2018.

[7] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba "Addressing the Rare Word Problem in Neural Machine Translation". Association for Computational Linguistics (ACL, 2015)

[8] Kalchbrenner et al (2013) "Recurrent continuous translation models". InProceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709. Association for Computational Linguistics.

[9] Rico Sennrich, Barry Haddow, Alexandra Birch "Neural Machine Translation of Rare Words with Subword Units". Association for Computational Linguistics (ACL, 2016).

[10] Lifeng Shang, Zhengdong Lu, Hang Li "Neural Responding Machine for Short-Text Conversation". Association for Computational Linguistics (ACL 2015)

[11] Cho, K. et al. (2014) Learning Phrase Representations using RNN Encoder Decoder for Statistical Machine Translation. Proceeding softhe2014 Conferenceon Empirical Methodsin Natural Language Processing (EMNLP)