Team 11 CSE431 Final Paper

by MD. SHAJIB HOSSAIN

Submission date: 14-Dec-2023 09:02AM (UTC+0530)
Submission ID: 2258507900
File name: Final_Paper_Team_11.pdf (475.69K)
Word count: 2158
Character count: 10994

Abstractive Bangla News Summarization Using LSTM With Attention to Encoder-Decoder Component

1st Shajib Hossain
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.shajib.hossain@g.bracu.ac.bd

2nd Muhimenul Mubin
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.muhimenul.mubin@g.bracu.ac.bd

***Abstract*—Social media and the internet being so easy to access gives us the opportunity to share information and news with any of us who is connected to the internet and social media directly or indirectly. Therefore the velocity and the flow of information make it more difficult for us to cope with because it is much harder to get the exact information within a short time. New technologies and techniques have been developed to make our lives easier which is why, document summarization is in great demand. On the other hand, the text summarization of the Banglali language has not come to bright light. Therefore, in our paper, we are proposing a novel approach of text or document summarization using the long-short-term machine(LSTM) learning model of Natural Language Processing.**

***Index Terms*—summarization, LSTM, Natural Language Processing**

I. Introduction

We all are familiar with the unprecedented flow of information and its ease of access using the internet and social media. Here considering the flow of information both the internet and social media contribute to our social and daily life. The ease of access and the availability of the internet has increased the flow of information even people just because of the ease of the internet and the being able to share and see any sort of information on social media people can even contribute to the flow of information in a great way and then we all are getting benefit from the flow of information somehow directly or indirectly. So to the exact information in this vast amount of information we as human beings tend to get a summary to save our time and to get to understand what the document is all about. Also, The process of summarizing a document or text plays a vital role in creating a general idea to make a report on in many media houses too. However, it has become very tough to cope with the velocity of the flow of information. That is why automated text summarization techniques have been developed by many researchers. In our research paper, this is what we are going to work on too. While doing our research we have come across several research papers with solid work on different aspects of text summarization using Machine learning techniques of Natural Language processing. However, creating an automated text summarization on Bengali text and documents has not come in bright light yet. Though some significant work has been done still there are some areas in which we need to get improvement on. Therefore, in our paper, we are proposing a technique of NLP(Natural Language Processing) as a new solution for text summarization on Banglai text and documents. Mainly the techniques of text summarization have two different categories such as Extractive and Abstractive. To be more specific, Extractive text summarization actually finds the main and important sentences of a given document and finds the connections and features among the sentences, and based on the features it gives a summary of the text. On the other hand, the Abstractive text summary technique finds the most important and critical information and based on that gives a summary with a completely new sentence that is not present in the given text or document. As we have already discussed only a few researches have been done on Bangla language and most of them had few datasets. Yet, in our research project, we have collected publicly available Bangla news data with proper descriptions and human-generated summaries as well as with the title of the news. For our proposed research aspect we have used the LONG-SHORT-TERM-MEMORY encoder-decoder technique to achieve our goal. And we have received good results using the encoder-decoder technique.

Identify applicable funding agency here. If none, delete this.

II. Related Wroks

Abstractive text summarization approaches out of different kinds exist. Different types of abstractive approaches are described by [2] Yeasmin et al. . Later we tried to focus our attention towards abstractive text summarization on the Bengali language context. We found in total of 14 approaches of Bengali text summarization in regard of both the abstractive and extractive approaches in Haque et al. [1]. Bengali extractive summarization based on document indexing and keyword-based information retrieval was first introduced by Islam et al. [3] in the year 2004. Then by Uddin et al. [4], the extractive technique of text summarization approach of the

English language was applied to the Bengali language. Then theme identification, page rank algorithms, etc. for extractive summarization are used in 2010 by Das et al. [5]. Kamal Sarkar [6], a researcher who first proposed the idea of sentence ranking and stemming process-based Bengali extractive summarization. And after a period Efat et al. [7] proposed a better way of doing that.

A key-phrase-based extractive approach and a pronoun replacement-based sentence ranking approach was shown by Haque et al. [8], [9]. The later years mainly in 2017, Abujar et al. [10] proposed a heuristic approach. And Akther et al. In our study we have also seen that later on a foundation on Bengali extractive summarization approaches have been proposed using deep neural netwrok. Considering our study it can be said text summarization on Bengali text has seen several significant researches but still we are hoping to get a new and more suitable appraoch on this matter.

III. Dataset

As mentioned earlier not a very vast amount of work has been done on Bangla text summarization as much as done in the English language. In most of the research works we have been through during our research work on the Bengali language we have come to know that there was not a good amount of datasets had been used. On the other hand, we know how crucial it is to use a dataset that has a good number of instances and is also cleaned. For our paper we have collected such a dataset that has a good number of instances to create a good automated text summarization model. Here the news data samples have been collected from various online sources like bdnews24.com the dataset consists of almost 20000 news instances as well as summaries for each of the news present in the dataset. Since the dataset was publicly available we have gone through the dataset and it was already much cleaned. In our selected dataset it has 6 different columns as shown in the figure:1. However we are to use only the description and the Summary.

Fig. 1. Few Instances of the Selected Dataset.

IV. Proposed Methodology

For our research paper as mentioned earlier, we are proposing a novel approach to achieve our goal of generating abstractive Bangla news summarization using the LSTM model. Before we dig into the model architecture and how we have built our proposed model first we need to handle some issues and pre-process our data. The process and the techniques that we are going to use for our data pre-processing are discussed below.

A. Data Pre-Processing

- For our proposed model we are going to use only two columns named "Description" and "Summary" as we have built our model based on these two columns by following commonly used Natural Language techniques.
- Then we have removed all the special characters such as (*, , ", @, —, !) and many more using regex which has no significance while finding the important sequence of tokenized words in order to generate a summary of a given text or document.
- While working on this project we have come to know that just like the special characters we stop words and put zero importance on any sequence of sentences while constructing an idea regarding the sentence present in any document. This is why in order to be more precise and clean with our dataset we have removed commonly used stop words in the Bangla language by importing nltk package and its corpus.

B. Proposed Model Architecture and Building

LSTM is commonly known as LONG-SHORT-TERM-MEMORY which is also a better version than ordinary recurrent neural networks. By saying better version of any recurrent neural network we meant to say that in terms of remembering long sequences, this LSTM performs better. Also, the LSTM uses all the hidden layers to get a better understanding of the given training dataset. As per the LSTM model, the model actually uses the seq2seq approach and gives more attention to encoder-decoder components. To be more precise an encoder component of the LSTM model feds each word within the sequence and find the best possible information that is in the sequence and then loads them into the LSTM. Then the model uses a decoder component which finds the best possible sequence on the time stamp considering the information stored in the LSTM while the encoder component was working.

- According to the documentation of the LSTM model "start" and "end" these two keywords must be added in order to generate the tokennized sequence. Therefore, for our project, we have used "sostok" to indicate starting of the sentence and "eostok" to indicate end of the sentence.
- We have also found that after tokenizing the instances in our dataset 18 percent of unique words have been present in the vocabulary.
- After that we built our model by adding an encoder, decoder component, and lastly a dense layer with the help of tensorflow and keras by importing the packages. And after we found the summary on our built model it is seen that we have in total "8190659" trainable parameters.

V. Results and Evaluation

After building the Model our first work to do was to train our model with the proper given parameters. For our proposed model we have split the dataset and have used 10 percent of

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 100)]	0	[]
embedding (Embedding)	(None, 100, 200)	4063400	['input_1[0][0]']
lstm (LSTM)	[(None, 100, 300), (None, 300), (None, 300)]	601200	['embedding[0][0]']
input_2 (InputLayer)	[(None, None)]	0	[]
lstm_1 (LSTM)	[(None, 100, 300), (None, 300), (None, 300)]	721200	['lstm[0][0]']
embedding_1 (Embedding)	(None, None, 200)	591800	['input_2[0][0]']
lstm_2 (LSTM)	[(None, 100, 300), (None, 300), (None, 300)]	721200	['lstm_1[0][0]']
lstm_3 (LSTM)	[(None, None, 300), (None, 300), (None, 300)]	601200	['embedding_1[0][0]', 'lstm_2[0][1]', 'lstm_2[0][2]']
time_distributed (TimeDistributed)	(None, None, 2959)	890659	['lstm_3[0][0]']

Total params: 8190659 (31.24 MB)
Trainable params: 8190659 (31.24 MB)
Non-trainable params: 0 (0.00 Byte)

Fig. 2. Summary of our proposed model

the data for testing and the rest 90 percent of the data for training purposes only. Also we have used 20epoch till now with the batch size of 144 and with no usage of early stopping. After implementing all these the validation and training loss curve we have got is not satisfactory. Therefore our model and the technique that we haev used for our research paper need much improvement. This is evident if we take a closer look at the curve shown in Figure 3.

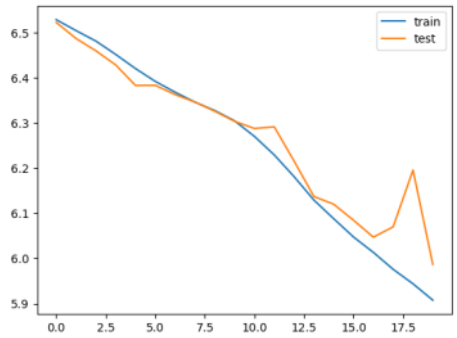

Fig. 3. Few Instances of the Selected Dataset.

VI. Limitations and Future Works

It is certainly evident that using the configurations the model we have proposed by following the LSTM model configuration has not performed up to the mark and has several limitations. Yet here for our paper now we have only used 20 epochs with batch size of 144. I must say we can increase our total trainable parameter and use better configuration for the model so it is possible to come up with better results. For Future works, as we have said by utilizing better configuration and better cleaning techniques of the chosen dataset we are planning to get better results for abstraction text summarization.

References

[1] Haque, M.M., Pervin, S., Hossain, A., Begum, Z.: Approaches and trends of automatic bangla text summarization: Challenges and opportunities. International Journal of Technology Diffusion (IJTD) 11(4), 1–17 (2020)

[2] Yeasmin, S., Tumpa, P.B., Nitu, A.M., Uddin, M.P., Ali, E., Afjal, M.I.: Study of abstractive text summarization techniques. American Journal of Engineering Research 6(8), 253–260 (2017)

[3] Islam, M.T., Al Masum, S.M.: Bhasa: A corpus-based information retrieval and summariser for bengali text. In: Proceedings of the 7th International Conference on Computer and Information Technology (2004)

[4] Uddin, M.N., Khan, S.A.: A study on text summarization techniques and implement few of them for bangla language. In: 2007 10th international conference on computer and information technology. pp. 1–4. IEEE (2007)

[5] Das, A., Bandyopadhyay, S.: Topic-based bengali opinion summarization. In: Coling 2010: Posters. pp. 232–240 (2010)

[6] Sarkar, K.: Bengali text summarization by sentence extraction. arXiv preprint arXiv:1201.2240 (2012)

[7] Efat, M.I.A., Ibrahim, M., Kayesh, H.: Automated bangla text summarization by sentence scoring and ranking. In: 2013 International Conference on Informatics, Electronics and Vision (ICIEV). pp. 1–5. IEEE (2013)

[8] Haque, M.M., Pervin, S., Begum, Z.: Enhancement of keyphrase-based approach of automatic bangla text summarization. In: 2016 IEEE Region 10 Conference (TENCON). pp. 42–46. IEEE (2016)

[9] Haque, M., Pervin, S., Begum, Z., et al.: An innovative approach of bangla text summarization by introducing pronoun replacement and improved sentence ranking. Journal of Information Processing Systems 13(4) (2017)

[10] Abujar, S., Hasan, M., Shahin, M., Hossain, S.A.: A heuristic approach of text summarization for bengali documentation. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–8. IEEE (2017)

Team 11 CSE431 Final Paper

ORIGINALITY REPORT

6%	4%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	dokumen.pub Internet Source	1%
3	"Innovations in Computer Science and Engineering", Springer Science and Business Media LLC, 2020 Publication	1%
4	export.arxiv.org Internet Source	1%
5	Md Abrar Hamim, Moyin Talukder, Afraz Ul Haque, Md. Selim Reza, Md. Samiul Alim, Asif Iquebal Niloy. "Bangla E-mail Body to Subject generation using sequence to sequence RNNs", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023 Publication	1%
6	Nomi Baruah, Shikhar Kr. Sarma, Surajit Borkotokey. "Evaluation of Content	1%

Compaction in Assamese Language",
Procedia Computer Science, 2020
Publication

7	dspace.bracu.ac.bd Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Team 11 CSE431 Final Paper

Article Error You may need to use an article before this word. Consider using the article **the**.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word. Consider using the article **the**.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Missing "," Review the rules for using punctuation marks.

Article Error You may need to use an article before this word.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Prep. You may be using the wrong preposition.

Missing "," Review the rules for using punctuation marks.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Missing "," Review the rules for using punctuation marks.

Article Error You may need to use an article before this word.

Prep. You may be using the wrong preposition.

Missing "," Review the rules for using punctuation marks.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Article Error You may need to use an article before this word. Consider using the article **the**.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Confused You have used either an imprecise word or an incorrect word.

Article Error You may need to use an article before this word.

P/V You have used the passive voice in this sentence. You may want to revise it using the active voice.

Article Error You may need to use an article before this word.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Confused You have used either an imprecise word or an incorrect word.

Missing "," Review the rules for using punctuation marks.

Missing "?" Review the rules for using punctuation marks.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to remove this article.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to remove this article.

Missing "," Review the rules for using punctuation marks.

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

Article Error You may need to use an article before this word.

Missing "," Review the rules for using punctuation marks.

Dup. Did you mean to repeat this word?

Missing "," Review the rules for using punctuation marks.

Missing "," Review the rules for using punctuation marks.

Article Error You may need to use an article before this word. Consider using the article **the**.

Article Error You may need to remove this article.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word. Consider using the article **the**.

Article Error You may need to remove this article.

Article Error You may need to remove this article.

Article Error You may need to remove this article.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to remove this article.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Missing "," Review the rules for using punctuation marks.

Dup. Did you mean to repeat this word?

Missing "," Review the rules for using punctuation marks.

Missing "," Review the rules for using punctuation marks.

Article Error You may need to remove this article.