News article Summarization

Bansi Mehta
Department of Computer Science
Dalhousie University
bansi.mehta@dal.ca

Deep Patel
Department of Computer Science
Dalhousie University
deep.patel@dal.ca

Janvi Patel
Department of Computer Science
Dalhousie University
janvi.patel@dal.ca

Sanket Shah
Department of Computer Science
Dalhousie University
sanket.shah@dal.ca

1. Introduction

In the big data era, there has been an explosion in the amount of text data from a variety of sources. This volume of text is an inestimable source of information and knowledge which needs to be effectively summarized to be useful. It is necessary to get relevant information within the original content without losing the overall meaning.

2. Problem Statement

It is becoming important to get precise and concise news nowadays due to lack of time going through the entire news article. That generates the need for “Abstractive text summary of news articles” for reducing news reading time. Summarizing the news will provide readers to have succinct overview of interesting details and important information.

3. List of Possible Approaches

There are two broad concepts to text summarization [1]:

1. Extractive: From the original content, sentences which seems most relevant and important are identified and extracted to generate the summary.
2. Abstractive: In contrast to extractive approach, new sentences are generated from original content maintaining and relevance and meaning same as of the original content.

The task before summarization is the preprocessing of the original text, Following steps would be performed as a part of text preprocessing [1][3]:

1. Remove non-alphabet characters.
2. Remove punctuation.
3. Remove special characters.
4. Convert words to lower case.
5. Remove excessive white space.
6. Remove stop words.

The text generated after pre-processing will pass through summarizer to get summarized text.

There are many approaches that could be adopted to generate text summary.

3.1 Seq2seq

This approach is mostly used where the input data is sequential [1]. It is mostly used in areas of natural language translation (input is in one language and output is in another language), named entity recognition (input is text and output is tags). This Seq2Seq ap-proach can also be used in text summarization where to original text is the input (long text) and output is the summarized version of it. This approach as two main components: encoder and decoder which are set in training and testing phase. This method will not work if the input sentences are very long. Reason being it would be difficult for encoder to convert long sentences into fixed length vectors.

3.2 Attention Mechanism

Instead of focusing on entire sentence, attention mechanism only focuses on the parts of it which is then used to generate the output (summarized text) [2].
There are two types of approaches in attention mechanism: Local Attention, Global At-tention
This method has few issues like: Semantic irrelevance, Grammatical error, and Loss rele-vance of main idea.

3.3 PEGASUS

PEGASUS model works in a different manner than the model that randomly select sen-tences. This model eliminates significant lines from the input text. Further, this input texts are to be complied as separate output [3]. This model is widely being used because it sig-nificantly choose only relevant sentences. PEGASUS model uses self-supervised objective GSG to train a transformer model [4]. This improves model's fine-tuning performance on text summarization.

3.4 T5

T5 is the abbreviation for "Text-to-Text Transfer Transformer" [5]. It uses sequence-to-sequence generation method that includes cross-attention layers to the decoder and gen-erates the decoder output autoregressive [5]. One of the approaches mentioned in the research papers [3], is to provide the input to encoder as a series of tokens which are in sequence of embeddings. A self attention layer and feed forward network are the two subcomponents. Same as encoder, decoder follows similar structure with a modification of generalized attention mechanism after every self attention layer. The final layer is a dense layer and uses softmax as activation function [3].

3.5 BERT

BERT is Bidirectional Encode Representation Transformer. Bert includes two separate mechanisms – an encoder that takes text as input and a decoder that produces a pre-diction for the task. It is implemented as a sequence-to-sequence model with a bidirec-tional encoder over corrupted text and a left-to-right autoregressive decoder [5]. It is a pre-trained model that is being used as a transformer encoder to provide a sentence level understanding. To provide the sentence and word level understanding a contextual re-lationship BERT uses Transformer Encoder, but as per the need of making abstractive summary the transformer decoder is also being used to provide some properly structured and meaningful output [6].

4. Project Plan

Table 1: Project Plan for the Term

Task	Start Date	End Date	Status
Research about available methods and approaches	10/09/21	17/09/21	Complete
Setup of project and dependencies	24/10/21	4/11/21	In-Progress
Setting UI for input through tkInter	6/11/21	11/11/21	Pending
Preprocessing of input data	7/11/21	10/11/21	Pending
Generating model through BERT approach (Using BBC news data)	7/11/21	20/11/21	Pending
Fine-tuning with BERT	15/11/21	25/11/21	Pending
Model summary (Accuracy, Precision, Recall, F1 Score, Rouge Score)			Pending
Generating and testing output	25/11/21	30/11/21	Pending
Knowledge transfer and discussion	1/12/21	3/12/21	Pending
Report documentation and Presentation	3/12/21	6/12/21	Pending

References

[1]. A. Pai, "Text Summarization | Text Summarization Using Deep Learning", *Analytics Vidhya*, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/ [Accessed: 05- Nov- 2021].

[2]. L. Gonc ̧alves, "Automatic Text Summarization with Machine Learning—An overview", *Medium*, 2021. [Online]. Available: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25 [Accessed: 05- Nov- 2021].

[3]. Anushka Gupta1 ,Diksha Chugh2 ,Anjum3 ,Rahul Katarya4;"Auto-mated News Summarization Using Transformers", 2021. [Online]. Available: https://www.researchgate.net/publication/353653704_Automated_News_Summarization_Using_Transformers[Accessed: 04- Nov- 2021].

[4]. Ben Goodrich; "GitHub - google-research/pegasus", GitHub, 2021. [Online]. Avail-able: https://github.com/google-research/pegasus [Accessed: 04- Nov- 2021].

[5]. Zhengzhi Lou, Ju Zhang;"Abstractive Summarization on COVID-19 Pub-lications - PDF Free Download",Docplayer.net,2021. [Online]. Available: https://docplayer.net/201739488-Abstractive-summarization-on-covid-19-publications.html [Accessed: 04- Nov- 2021].

[6]. M. Ramina, N. Darnay, C. Ludbe and A. Dhruv, "Topic level summary generationusing BERT induced Abstractive Summarization Model," 2020 4th International Con-ference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 747-752,doi: 10.1109/ICICCS48265.2020.9120997.