Abstractive Text Summarisation

Billa Amulya
Computer Science and Engineering
National Institute Of Technology Calicut
Calicut, Kerala
billa_b180404cs@nitc.ac.in

Bhukya Vasanth Kumar
Computer Science and Engineering
National Institute Of Technology Calicut
Calicut, Kerala
vasanthkumar_b180441cs@nitc.ac.in

Sevakula Jyothi
Computer Science and Engineering
National Institute Of Technology Calicut
Calicut, Kerala
jyothi_b180359cs@nitc.ac.in

***Abstract*—Natural language processing (NLP) describes the interaction between human language and computers by automatic manipulation of natural language like text and speech by a software. NLP is the thing that makes it workable for PCs to peruse the text and decipher it. In recent years, availability of the data generated has increased to exponential scale. Summarisers make it simpler for users to summarize the data without pursuing it totally. There are two different approaches, namely Extractive Summarisation and Abstractive Summarisation. This work is based on the Abstractive Summarisation which is quite advanced than the former. Abstractive text Summariser generates new sentences, possibly rephrasing or using the words that were not in the original text in a human readable format. This ensures that the core information is conveyed through the shortest text possible. In this work, the main goal was to increase the efficiency and diminish train loss of sequence to sequence model for making a superior abstractive text summariser.**

***Index Terms*—Summary, NLP, Abstraction, Extraction, Encoder, Decoder, Seq2Seq**

I. Introduction

The advancement of natural language processing (NLP) over the years has created more possibilities in the way we can manipulate data. With the phenomenal growth in the internet and technology, everything now in front of us is data that is being used for various purposes. There are massive amounts of information and documents online that serve various tasks. With the increase in availability of documents, demand for exhaustive research in the area of automatic text summarization is expanding.

According to [1], summary can be defined as ”a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that.”. [2] says, The Global Natural Language Processing (NLP) Market was valued at USD 10.72 billion in 2020, and it is expected to be worth USD 48.46 billion by 2026, registering a CAGR of 26.84% during the forecast period (2021-2026).

The automated text summarisation should produce an accurate summary from the given text without losing the key data and comprehensive meaning. The summary generated should help to understand required information in less time, thereby reducing the reading time. The automated text summary enables people to concentrate on key terms of the given text that are worth noting. Since the last few years, various approaches and models have been designed to automate the summary of particular texts. Text summarisation is widely used in areas like news article summary, tracking patient’s health history, search engines, etc. Apps like In-shorts use this method to summarise news and provide headlines. Search engines like Google Chrome use them to generate snippets of products, and to facilitate headlines for news across the globe.

A human can produce a summary of a text in various ways based on how they understand it and the keywords they use. Doing the same with machines is challenging and hence, text summarisation is considered to be limited. Since, the machines have a deficiency of human knowledge, text summarisation turns out to be a non-trivial task[3] and the results vary.

II. Problem Definition

To generate a short, precise summary of a longer text by retaining the key information of the text using Natural language processing (NLP) techniques.

A. Input and Output

Input:
1. Data(Sentence/Paragraph)
2. Original summary of data (Used for calculating accuracy)
Output:
A short summary is generated based on abstractive text summarisation.

B. Diagramatic Representation

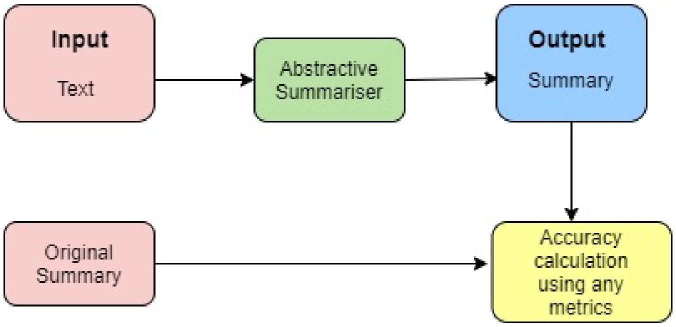

Fig. 1. Input and Output Diagram

Hazari
29-09-2021

III. LITERATURE SURVEY

A summary can be defined as a text that represents the main ideas or key information in an original text in less space. Suppose, if all the sentences in the original text are important, then the process of summarization would be less effective as the size of the summary would effect its informativeness. The main challenge in the process of summarization is identifying the informative segments and finally generating it as a consice text[1].

The evaluation of quality of a summary is one of the most difficult problem this is because there is nothing called ”ideal” summary. In many cases such as news articles, human summarizes they only meet upto an 60% of the time measuring sentence content overlap[13]. In general, summarization can be defined as combination of topic identification, interpretation and generation. For this identification part the main goal is to retain the most important, central topics. For interpretation part, the main goal is to perform compression through re-interpreting and fusing the pulled out topics into more succinct ones. This is quite important because abstracts are usually much shorter than their equivalent extracts. Where as for the generation part reformulation of the extracted and fused material into a text with new phrases is quite challenging.

Abstractive techniques need a more profound examination of the text i.e it needs deeper analysis. These techniques can produce new sentences,and improve the focus of the summary to maintain a decent compression rate[4]. An RNN Encoder decoder based architecture, which is based on sequence to sequence model is applied to process the data in sequential manner such that the input of any state may depend on the output of the previous states [5,6], this scenario is most likely to occur in a sentence, where the meaning of a word is closely related to the previous words meaning.

Teacher forcing is a fast and effective way to train RNNs. However, this approach may result in more fragile/unstable models.In [14], Seq2seq Model with Attention using GRU and Teacher Forcing is implemented,which would work fine on shorter summaries (50 words).

Raphal et al. surveyed several abstractive text summarisation processes in general [15]. Their study differentiated between different model architectures, such as reinforcement learning (RL), supervised learning, and attention mechanism. In addition, comparisons in terms of word embedding, data processing, training, and validation had been performed. However, there are no comparisons of the quality of several models that generated summaries.

Sutskever et al. [7] describes an end-to-end approach to *sequence to sequence* learning using a Multilayer LSTM. The neural network contains encoder and decoder. Encoder uses a fixed length of text as input and Decoder represents the output. In [8], a bi-directional RNN with LSTM’s in encoding layer and attention mechanism in decoding layer and the sequence to sequence model is used to generate a abstractive summary of text, thereby increasing efficiency and reducing the training loss. According to [9], Text-To-Text Transfer Transformer (T5) based abstractive text summarization method shows better performance than baseline attention based seq2seq approach.

Shi et al. presented a comprehensive survey of several abstractive text summarisation models, which are based on sequence-to-sequence encoder-decoder architecture for convolutional and RNN seq2seq models. The focus was the structure of the network, training strategy, and the algorithms employed to generate the summary [16].

Recent approaches that used deep learning for abstractive text summarisation and metrics for evaluating these approaches and the challenges faced while implementing these approaches and their solutions were mentioned in[4]. The RNN and attention mechanism were the most commonly employed deep learning techniques.we can also notice that few methods applied LSTM to resolve the gradient vanishing problem that occured when using an RNN, while other approaches applied a GRU.

IV. DESIGN

A. *Seq2Seq Model Design*

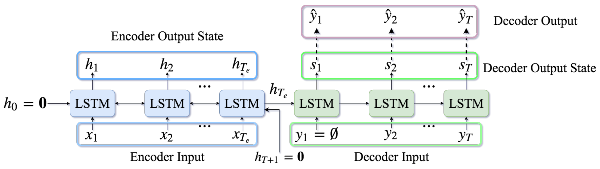

Fig. 2. Seq2Seq Model

B. *Explanation*

The above shown Encoder-Decoder Architecture is a Simple seq2seq model. The encoder has a input range of Te units and after evaluation the decoder has delivered the output of range T units following the condition T¡Te. Every encoder in the state ht, can receive the previous encoder’s hidden state ht-1, this is valid in both unidirectional and bidirectional LSTM but the bidirectional LSTM has an additional feature of accessing the next encoder’s hidden state which is ht+1.

C. Flowgraph

The following Flow Graph displays the flow of control information of the model.

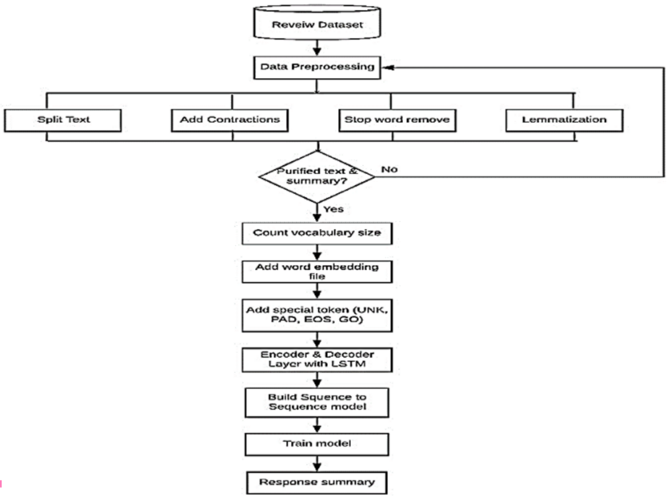

Fig. 3. FlowChart

V. Work Done

1. Studied and examined various papers and completed the Literature work for this topic
2. Inspected some popular text summarising web applications to understand the features they provide [10]
3. Examined open source frameworks of text summarisation works.
4. Briefly discussed the proper parameters, inputs and outputs of the model from our understanding.
5. Explored various methods to generate abstract summary for a given text.

VI. Work Plan

1. Understand the accuracy judging miniatures such as BLUE Score and ROUGE Score
2. Setup local development environments such as repositories, dependencies in local machines (Linux/ Windows) as well as Github open source
3. Explore the Seq2Seq Encoder-Decoder Model in order to generate and deliver the summary.
4. Develop the implementation of the basic design of Seq2Seq Encoder-Decoder Model
5. Designing a new model in-order to increase the accuracy and develop the implementation for it
6. Evaluate the generated summary according to the studied models i.e to [11] and [12]

VII. Summary

Our work deals with the implemenation of a model for abstractive text summarisation. The work uses deep learning in order to increase the efficiency. Decreased train loss is also a major contribution in this the sequence to sequence model.

In this Report, we discussed how exactly abstractive text summarisation is a challenging task, yet much needed in present times. We briefly discussed about the previous works done in this area. We arrived at a model which uses seq2seq along with RNN encoder-decoder architecture. Preliminary works done and the future work plan are also mentioned inorder to implement the abstractive text summarisation model.

VIII. Acknowledgement

It is with great respect that we remember the names of all who have been a great help and guidance throughout our project. With a profound sense of gratitude, we would like to express our heartfelt thanks to our guide and project coordinator, *Dr. Raju Hazari, Assistant Professor, Department of Computer Science and Engineering* for his expert guidance, co-operation, and immense encouragement in pursuing this project. We also express our sincere regards to *Dr. P Arun Raj Kumar, Assistant Professor, Department of Computer Science and Engineering* for their valuable coordination. Our sincere thanks are extended to all the teachers of the Department of Computer Science and Engineering and to all our friends for their help and support.

References

[1] Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. 2002.“ Introduction to the special issue on summarization,“Computational linguistics 28, 4 (2002).

[2] https://www.mordorintelligence.com/industry-reports/natural-language-processing-market.

[3] Mehdi Allahyari, Seyedamin Pouriyeh and Mehdi Assef “Text Summarization Techniques: A Brief Survey“.

[4] Saranyamol C S, Sindhu L, “A Survey on Automatic Text Summarization“, International Journal of Computer Science and Information Technologies, 2014,Vol. 5 Issue 6.

[5] C. L. Giles, G. M. Kuhn, and R. J. Williams,“Dynamic recurrent neural networks: theory and applications,“ IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 153–156, 1994.

[6] A. J. Robinson,“An application of recurrent nets to phone probability estimation,“ IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 298–305, 1994.

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

[8] Abu Kaisar Mohammad Masum, Sheikh Abujar,“Abstractive method of text summarization with sequence to sequence RNNs,“IEEE (2019).

[9] Ekaterina Zolotareva, Tsegaye Misikir Tashu, “Abstractive Text Summarization using Transfer Learning,”2020.

[10] https://quillbot.com/summarize

[11] https://github.com/google-research/google-research/tree/master/rouge

[12] https://github.com/neural-dialogue-metrics/BLEU

[13] Hovy, E. and C.-Y. Lin. 1999. “Automated text summarization in SUMMARIST“. In I. Mani and M. T. Maybury, editors, Advances in Automatic Text Summarization. MIT Press, Cambridge, pages 81–94.

[14] https://medium.com/@dusejashivam

[15] N. Raphal, H. Duwarah, and P. Daniel, “Survey on abstractive text summarization,“in Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 513–517, Chennai, 2018.

[16] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy,Neural Abstractive Text Summarization with Sequence-ToSequence Models: A Survey, http://arxiv.org/abs/1812.02303,2020.

Hazari 29-09-2021

PROJECT GUIDE
Dr. Raju Hazari
Assistant Professor
Dept. of CSE
NIT Calicut