Text Summarization Tool Using Deep Learning

Dr. S.K. Shinde
Department of Computer Engineering
VPKBIET
Baramati, India
santaji.shinde@vpkbiet.org

Sanika Sahuji
Department of Computer Engineering
VPKBIET
Baramati, India
sanikasahuji12@gmail.com

Vaishnavi Sakhare
Dept. of Computer Engineering
VPKBIET
Baramati, India
vaishnavisakhare0312@gmail.com

Udaysingh Pawar
Dept of Computer Engineering
VPKBIET
Baramati, India
udaysinghpawar0302@gmail.com

***Abstract*—Abstractive summarization is difficult to accomplish with standard NLP techniques. Hence, with the advancements in the field of Deep Learning, various deep learning models can be used to generate and improve the quality of abstractive summaries. One such model is the sequence to sequence model mainly used in neural machine translation. Documents which are often written by professionals are well suited for neural networks to understand the syntax and to generate well structured summaries. This study investigates an encoder-decoder-based model with a self-attention function to generate abstractive text summaries for an input document. It introduces a self-attention function for keeping track of the important keywords and the out-of-vocabulary words presented in the input document. It employs CNN/DailyMail dataset to achieve abstractive text summarization. Specifically, the three different pre-trained models, namely google/Pegasus, T5-small, facebook/BART-large are used to develop a robust text summarization tool.**

***Index Terms*—Abstractive summarization, Pegasus, BART-large, T5 model**

I. Introduction

The rapid growth of the Internet has resulted in a massive increase of the amount of information available, especially regarding text documents (e.g. news articles, electronic books, scientific papers, blogs, etc). Due to the paramount volume of information in the Internet, it has become unfeasible to efficiently sieve useful information from the huge mass of documents. Thus, it is necessary to use automatically methods to “understand”, classify and present all information in a clear and concise way, allowing users to save time and resources. One solution to this problem is using text summarization techniques.

Text summarization aims at automatic creating a compressed version of one or more documents, extracting the essential information in them. In other words, the main goal of a summary is to present the main ideas in a document in less space. It includes two main techniques: Extractive and Abstractive summarization.

Extractive summarization involves selecting sentences or phrases directly from the original text to construct the summary. The chosen sentences are typically the ones that are considered the most informative, relevant, and important in conveying the central ideas of the document. It does not involve generating new sentences or altering the wording. It simply identifies and pulls out existing sentences from the source text. The disadvantages of extractive summarization are: It is limited to using only the sentences present in the source document, it may not capture a cohesive narrative, as sentences are selected individually. Extractive summarization can sometimes result in redundancy if similar information is present in multiple sentences.

On the other hand, abstractive summarization takes a more creative approach by generating new sentences that might not appear in the original text. This method involves understanding the source text, interpreting its meaning, and then rephrasing and reorganizing that meaning into a coherent and concise summary. The main advantage is that it is not restricted to the sentences present in the source text, allowing for more flexibility. It captures the central idea in a more unified manner.

Thus, when compared with the extractive summarization model, the abstractive summarization model more closely resembles the process of human summarization, giving it important research significance.

II. Literature Survey

A. Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey, in IEEE Access 2022 [1]

This study delves into the automatic text summarization, specifically focusing on a feature-based approach. In this method, ”features” act as textual clues that aid computers in identifying crucial content for summarization. The study commences by emphasizing the significance of text summarization and categorizes the two primary approaches: one that extracts existing sentences and another that generates entirely new sentences.

It explores methods for identifying textual clues, such as identifying keywords, determining the most salient sentences, and comprehending the overall context of the text. The study underscores the importance of selecting the most relevant clues and effectively integrating them. Furthermore, it discusses how computers can learn from examples to enhance their ability to discern these clues, employing measures like ROUGE and BLEU to evaluate the quality of computer-generated summaries.

Challenges in the field are also acknowledged, with a focus on potential future improvements. These enhancements may involve incorporating cutting-edge techniques from deep learning and merging diverse summarization methods. There are various techniques used including Sequence-to-Sequence models and Pretrained transformers. The study provides real-world examples of where these methods have proven beneficial in addressing challenges across various domains.

B. BRIO: Bringing Order to Abstractive Summarization, Yale University, Carnegie Mellon University, March 2022 [2]

”Bringing Order to Abstractive Summarization” is a focused and aimed at enhancing the field of abstractive text summarization. This endeavor primarily revolves around the utilization of two prominent algorithms, BART and Pegasus, to introduce structure and efficiency to the summarization process.

BART, known for its strong performance in various NLP tasks, and Pegasus, recognized for its prowess in generating coherent and concise abstractive summaries, stand at the forefront of this project. Their capabilities play a pivotal role in redefining how information is distilled from complex textual content.

By harnessing the power of these algorithms, this project aspires to transform abstractive summarization into a more systematic and effective process. It seeks to imbue summaries with coherence and context, making them more informative and reader-friendly. This project's potential applications span across numerous domains, from news aggregation and content curation to knowledge extraction and automated content generation. The models used here are Pegasus and BART.

C. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, July 2020 [3]

In this research, the effectiveness of pre-training large Transformer-based models for Natural Language Processing (NLP) tasks, including text summarization, has been widely recognized. However, this study highlights two significant gaps in this area. Firstly, it points out the lack of pre-training objectives specifically tailored for abstractive text summarization, indicating a potential avenue for improvement. Secondly, it notes the absence of a systematic evaluation across diverse domains, highlighting the need for comprehensive assessments.

To address these gaps, the study introduces a novel approach called PEGASUS, where important sentences are removed or masked from an input document and then generated collectively as one output sequence from the remaining sentences. The study proceeds to evaluate the PEGASUS model on 12 different summarization tasks spanning various domains, such as news, science, stories, instructions, emails, patents, and legislative bills.

The experimental results are highly promising, as the PEGASUS model attains state-of-the-art performance on all 12 downstream datasets, measured by ROUGE scores. Human evaluation further validates the model's capabilities, demonstrating that its summaries achieve human-level performance on multiple datasets.

This research presents a groundbreaking approach to abstractive text summarization, with the PEGASUS model showcasing remarkable proficiency across a wide range of domains and establishing itself as a powerful solution in the field of NLP.

III. Background Study

The project's workflow for our proposed summarization system is well structured and entails several key steps. In this project's well-structured framework, we begin with the essential step of data collection, where we source the CNN/Daily Mail dataset to serve as our training and evaluation data source. This dataset provides a rich foundation for our text summarization system.

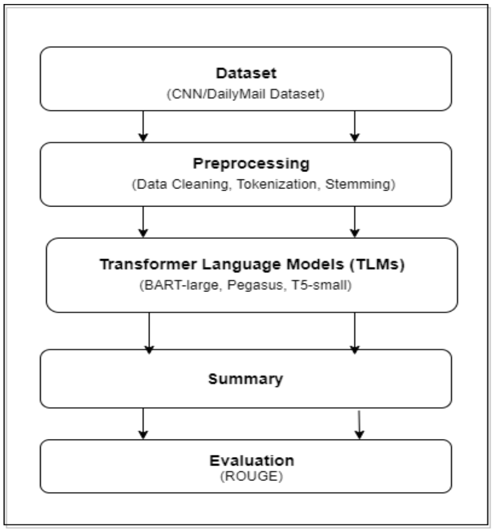
Dataset
(CNN/DailyMail Dataset)
Preprocessing
(Data Cleaning, Tokenization, Stemming)
Transformer Language Models (TLMs)
(BART-large, Pegasus, T5-small)
Summary
Evaluation
(ROUGE)

Fig. 1. Flowchart

Following data collection, we transition to data preprocessing, a critical phase that encompasses data cleaning and text preprocessing. This stage ensures that our dataset is refined and ready for analysis. Furthermore, we meticulously split the data into training, validation, and test sets to support comprehensive model training and evaluation.

With our preprocessed data in hand, the subsequent step involves model selection and loading. We access pre-trained weights for leading models like BART, Pegasus, and T5 from relevant libraries, establishing a strong foundation for our summarization system.

The fine-tuning stage is pivotal, where we train these models on the designated training dataset. The trained models are subjected to evaluation using robust metrics such as ROUGE, performed on the validation dataset. This step enables us to assess the quality and efficacy of the generated summaries.

Upon successful evaluation, we proceed to base model selection, where we identify the model that achieves the highest ROUGE scores on the validation dataset. This model is chosen as the base model for our summarization system, signifying its superior performance.

In the testing phase, the selected base model is subjected to a rigorous assessment on the test dataset to gain insights into its real-world applicability and performance.

The deployment phase follows, wherein the selected model is deployed as an accessible API, making it readily available for various summarization tasks.

To enhance user experience and interaction with our summarization tool, we embark on developing a user-friendly web application. This application will serve as an intuitive platform for users to efficiently interact with the system and harness its text summarization capabilities.

This comprehensive project framework ensures a methodical approach to creating an effective and user-friendly text summarization system, leveraging a well-structured dataset, advanced models, and rigorous evaluation to deliver high-quality summarization results.

IV. Dataset Collection

The CNN/DailyMail dataset reveals a dataset of news around the world, extracted from the news of CNN website. One advantage of this spare corpus is that, it is based on the high quality of writing, beyond the highlights, which consists of sentences provided for each text, a good quality summary composed of 3 or 4 sentences. In fact, it becomes interesting for evaluation purposes, since there is already a summary of the editor himself. In addition, it created a new evaluation set, selected by a number of researchers from the sentences of the text itself, which were the most suitable for the formation of an abstract with a good quality, that contains one more option summary for evaluation purposes. This corpus presents 3,000 texts in different categories, including: business, health, justice, opinion, sports, tech, travel, and world news. The current version supports both extractive and abstractive summarization. For each instance, there is a string for the article, a string for the highlights, and a string for the id.

V. Architecture

The Transformer is a deep learning model known for its encoder-decoder architecture, featuring the innovative Attention mechanism, eliminating the need for recurrent neural networks (RNNs) and significantly enhancing training speed. It comprises two main components: the Encoder and the Decoder.

A. Encoder :

An encoder's primary function is to handle the input data and transform it into a condensed and meaningful representation. This representation, is commonly referred to as a 'context vector' or 'thought vector'. This vector contains the crucial details from the input sequence in a more concise format. The key role of the encoder is to capture the relevant features, relationships, and context within the input data. Then, it generates an encoded representation that encapsulates this comprehension. This encoded representation serves as the foundation for generating the output, which could be a summary.

B. Decoder :

The decoder takes input the hidden states generated by the encoder and the previously generated output tokens and uses them to predict the next output token. At each step, the decoder attends to different parts of the input sequence using its attention mechanism, allowing it to capture complex relationships between the input and output sequences. Its attention mechanism lets it capture complex relationships between the input and output sequences, enabling it to generate high-quality output.

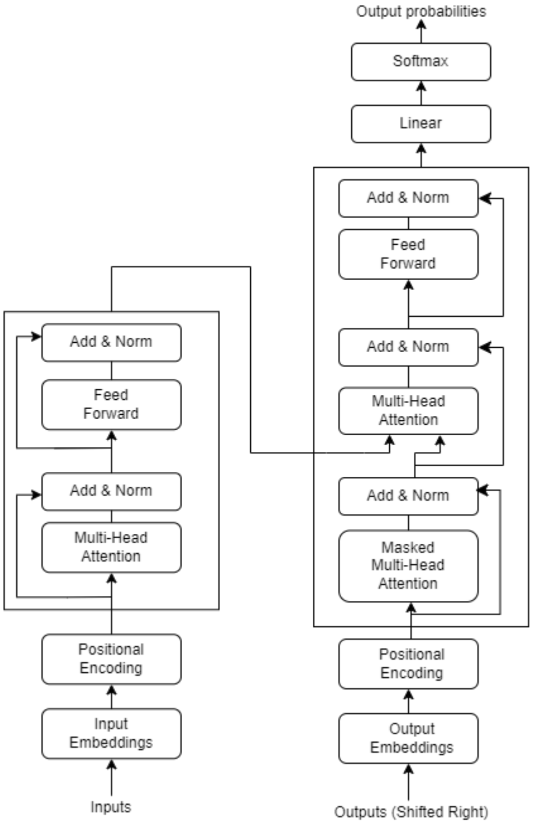

Fig. 2. Encoder-Decoder Architecture

VI. Models

A. PEGASUS :

PEGASUS(Pre-training with Extracted Gap-sentences for Abstractive Summarization) is a state-of-the-art model for abstractive text summarization in the field of natural language processing (NLP). It was developed by Google Research and is part of the broader family of transformer-based models.

PEGASUS is designed to generate human-like summaries that capture the key information and meaning of a given document while producing coherent and fluent text.

PEGASUS follows a two-step training process. In the first step, it is pre-trained on a large corpus of text data using a denoising autoencoder objective. During this step, some parts of the input text are masked, and the model learns to predict the missing words. After pre-training, PEGASUS is fine-tuned on a specific summarization dataset. PEGASUS utilizes an encoder-decoder architecture. The encoder reads the source document and generates a context-rich representation. The decoder takes this representation and generates the summary step by step. During decoding, PEGASUS generates each word of the summary sequentially while considering the context of previously generated words. This autoregressive process allows the model to capture dependencies between words and generate fluent and coherent summaries.

The idea is that instead of masking out only words as it happens in BERT, they mask out entire sentences, and ask the model to guess the removed sentences. As the model tries to get the underlying missing context, it gets better at abstractive summarization-like tasks. Here is the algorithm used for sentence selection

B. BART :

BART (Bidirectional and Auto-Regressive Transformers) is a model used for abstractive text summarization in the field of natural language processing (NLP). It is a sequence-to-sequence model that combines elements from both bidirectional and autoregressive approaches. The first part of BART uses the bi-directional encoder of BERT to find the best representation of its input sequence. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that is designed to capture the contextual information and relationships between words in a text sequence. BERT reads text in a bidirectional manner, taking into account both the preceding and succeeding words for each word in a sentence.

After obtaining token and sentence-level representations for an input text sequence, the decoder's role is to align these representations with the desired output target. However, employing a decoder with a similar design might result in suboptimal performance for tasks like next sentence prediction or token prediction. This is because these tasks necessitate a stronger reliance on a comprehensive input context. To address this, it becomes imperative to adopt model architectures capable of training by predicting the next word solely based on preceding words in the sequence. This is where causal or autoregressive models prove advantageous, as they focus exclusively on past data to make predictions about the future.

Thus, BART attaches the bi-directional encoder to the autoregressive decoder to create a denoising auto-encoder architecture.

C. T5 :

T5, which stands for ”Text-to-Text Transfer Transformer,” is a powerful model in the field of natural language processing (NLP). While it is not specifically designed for abstractive summarization, T5 is a text-to-text model, which means it can be adapted for a wide range of NLP tasks, including abstractive summarization.

It is a transformer model that is trained in an end-to-end manner with text as input and modified text as output, in contrast to BERT-style models that can only output either a class label or a span of the input. It tackles diverse natural language processing tasks by treating them as text-to-text conversions, where both input and output are transformed into text sequences. Operating on the transformer architecture, T5 undergoes pre-training to grasp language nuances and relationships, enhancing its understanding. It then fine-tunes on specific tasks, refining its outputs based on input data.

T5's encoder-decoder structure empowers it to generate coherent and contextually relevant text. It employs an autoregressive decoder for sequential token generation, ensuring output quality. Its strength lies in transfer learning – the model can leverage its pre-trained knowledge across tasks, reducing the need for extensive modifications.

VII. Preliminary Conclusion

In conclusion, this text summarization project will employ advanced Natural Language Processing techniques to elevate information retrieval and comprehension through abstractive summarization. The system will intricately understand the source text, generating concise and contextually coherent summaries that will capture the essence of the information.

Among all the three models, the PEGASUS model gives the best accuracy. It underscores the commitment to leveraging state-of-the-art technologies for effective summarization. The seamless integration of this summarization model into a user-friendly web application will be built using Next.js, which will accessibility and an intuitive interaction experience with the user.

VIII. Future Scope of the Project

The future of text summarization tool holds significant potential for innovation and development. As technology and research continue to advance, there are several directions for future work in this field. It can extend abstractive summarization to incorporate multiple modalities, such as text, images, and audio, to provide more comprehensive summaries. It can support cross-lingual summarization to improve the capabilities of abstractive summarization tools to work with multiple languages. The system can adapt to non-standard text formats to improve the ability of model to handle text that deviates from chat logs, social media posts, and conversations.

Acknowledgment

In closing, we wish to extend our heartfelt acknowledgments to those who contributed to the completion of this review paper. Additionally, our appreciation goes to the co-authors and collaborators who played integral roles in shaping this work. We also acknowledge the data providers and research participants whose contributions were invaluable. This acknowledgment is a tribute to the collaborative spirit that drives research and innovation in our field.

References

[1] D. Yadav, R. Katna, A. K. Yadav and J. Morato, ”Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey,” in IEEE Access, vol. 10, 2022

[2] Yixin Liu, Pengfei Liu, Dragomir Radev, Graham Neubig, ”BRIO: Bringing Order to Abstractive Summarization”, Yale University, Carnegie Mellon University, March 2022

[3] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, ”PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”, July 2020

[4] Y. Chen, ”Research on Abstractive Summarization Technology Based on Deep Learning,” 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China 2022

[5] Z. Hao, J. Ji, T. Xie and B. Xue, ”Abstractive Summarization Model with a Feature Enhanced Seq2Seq Structure,” 2020 5th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), Singapore, 2020

[6] Magdum, P. G., and Sheetal Rathi ”A survey on deep learning-based automatic text summarization models.” International Conference on Artificial Intelligence and Data Engineering. Singapore: Springer Nature Singapore, 2019

[7] W. S. El-Kassas, C. R. Salama, A. A. Rafea and H. K. Mohamed, ”Automatic text summarization: A comprehensive survey”, Expert Syst. Appl., vol. 165, March 2021.

[8] Nazari, Narges, and M. A. Mahdavi. ”A survey on automatic text summarization.” Journal of AI and Data Mining, (2019)

[9] Arun Krishna Chitturi and Saravanakumar Kandaswamy, ”Survey on Abstractive Text Summarization using various approaches”. International Journal of Advanced Trends in Computer Science and Engineering, December 2019.

[10] Yang Liu and Mirella Lapata, ”Text Summarization with pre-trained encoders”, Institute for Language, Cognition and Computation School of Informatics, University of Edinburgh, 2019.