Text Summarization Using the T5 Transformer Model

Ratan Ravichandran[1], Sri Bharath Sharma P[2], Shriyans Shriniwas Arkal[3], Shubhangee Das[4], Prof. Sasikala Nagarajan[5]

[1-5]Department of Artificial Intelligence and Machine Learning, Dayananda Sagar University, Bangalore, India

---***---

Abstract - *In our information-filled world, it is crucial to focus on the essential content amidst the overwhelming volume of information available. Unfortunately, people often spend a significant amount of time sifting through irrelevant details, inadvertently overlooking crucial information. To address this issue, we present a project that utilizes the T5 transformer model in natural language processing to develop an abstractive text summarization system. By leveraging advanced language modeling techniques, our project aims to enhance efficiency, comprehension, and decision-making processes across various domains.*

***Key Words*: Abstractive summarization, T5 transformer model, Natural language processing.**

1.INTRODUCTION

In our information-filled world, focusing on what truly matters is essential for success. On average, a person spends a significant amount of their lifetime reading useless information, often missing out on significant bits by subconsciously dismissing them. To solve this problem, we built a text summarizer that condenses lengthy text into shorter concise summaries, providing a quick overview of the main information.

Text summarization is a vital tool in today's information-driven world, allowing us to distil the essence of lengthy texts into concise summaries. By employing advanced natural language processing techniques, text summarizers extract key information, enabling readers to grasp the main ideas quickly. In this report, we explore the effectiveness and applications of text summarizers, shedding light on their potential to enhance efficiency, comprehension, and decision-making processes across various domains.

1.1 The T5 Transformer Model

To achieve this, we use the T5 transformer model which is a powerful language model that can understand and generate human-like text. Constructing a text summarizer based on T5 is beneficial because it allows for concise and accurate summarization of lengthy documents. T5's ability to capture contextual relationships and generate coherent summaries makes it an ideal choice for text summarization tasks, enabling efficient information extraction and facilitating quick comprehension of complex texts.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

2. Literature Review

Adhika Pramita Widyassari et al.[1] provides an overview of various techniques and methods used in automatic text summarization, with a particular focus on the Natural Language Toolkit (NLTK). The author explores different approaches, including extractive and abstractive summarization, and discusses how NLTK can be utilized in these techniques.

- Preprocessing: NLTK performs essential text preprocessing tasks like tokenization, stemming, and stop-word removal, aiding in information extraction by breaking text into words or sentences and reducing words to their root form.

- Sentence Scoring: NLTK facilitates extractive summarization by offering tools to calculate sentence similarity (e.g., cosine similarity) and assign scores, enabling the selection of relevant sentences based on their importance.

- Feature Extraction: NLTK's part-of-speech tagging and named entity recognition assist in identifying entities and key terms, enhancing summary accuracy and relevance.

- Language Modeling: In abstractive summarization, NLTK helps build language models (e.g., n-gram models) for generating concise and coherent summaries by predicting probable next words or phrases.

- Evaluation: NLTK includes evaluation metrics (e.g., ROUGE, BLEU) to assess summary quality by comparing them with reference summaries and measuring similarity or effectiveness.

Khilji et al. [2] examines Abstractive Text Analysis, described as a natural language processing (NLP) technique that aims to generate a concise and coherent summary of a given text by understanding its content and generating new sentences. Abstractive summarization involves creating novel sentences that capture the key information and main ideas of the source text in a more human-like manner.

3. PROPOSED ARCHITECTURE

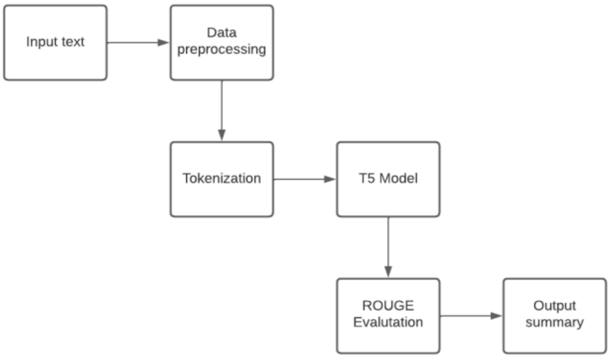

Fig -1: System Architecture

Our system employs the cutting-edge T5 transformer model for effective text summarization. The process starts with data preprocessing, including cleaning and organizing. Tokenization divides data into smaller units for processing. The T5 model is then trained to understand input and generate informative summaries.

Once trained, it condenses key information in new input documents. Evaluation is done using the ROUGE metric, measuring similarity to human-written summaries. Higher scores indicate better summarization. This architecture leverages T5's power to process input, generate concise summaries, and assess quality, streamlining information extraction for quicker comprehension and decision-making.

3.1 Architecture Workflow

This code implements a sequence-to-sequence (Seq2Seq) neural network, leveraging the T5 model, to achieve text summarization. The process encompasses data preparation, where libraries are imported, and the "multi_news" dataset is loaded, split, and organized. Tokenization and preprocessing are employed to adapt the data, utilizing the "t5-small" tokenizer and defining a summarization prefix.

The core of the workflow involves model training, where the pre-trained T5 model is fine-tuned for summarization. The Seq2SeqTrainer facilitates this training, optimizing the model's capacity to generate accurate and concise summaries. After training, the model predicts summaries, and Rouge scores are calculated using the Rouge library to assess the quality of these summaries.

4. EXPERIMENTATION

4.1 Dataset

This dataset, multi_news found on HuggingFace, consists of two columns: a feature column containing news text separated by "|||||," and a target column with human-written summaries. The target column serves as the reference for summaries, while summaries in the feature column provide a condensed overview.

4.2 Model Creation

For model creation, we use a T5 transformer architecture tailored for sequence-to-sequence language tasks. The DataCollatorForSeq2Seq ensures proper tokenization and data collation. The AutoModelForSeq2SeqLM class loads pre-trained T5 weights, to generate coherent sequences, such as text summarization.

4.1 Model Training

The trainer is configured with the necessary components, including training arguments, tokenizer, data collator, and datasets for training and evaluation. By calling the train function, the training process begins, during which the model learns to generate concise summaries from the given input data. Once the training is complete, the trained model is saved to a specified path for later. Additionally, the trained model and a data file are downloaded and copied , enabling further analysis or storage of the results. The model is trained for 10 epochs and 25 epochs and the results are accordingly evaluated.

5. RESULTS AND ANALYSIS

ROUGE, which stands for "Recall-Oriented Understudy for Gisting Evaluation," is a set of metrics used to evaluate the quality of summaries or generated text in natural language processing tasks. It is commonly used in automatic summarization and machine translation evaluation. They are assessed using ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE metrics provide a way to quantitatively assess the quality of summaries or generated text by comparing them to a reference summary. These metrics are widely used in research and evaluation of text generation models to measure their effectiveness in capturing the key information or meaning from the source text.

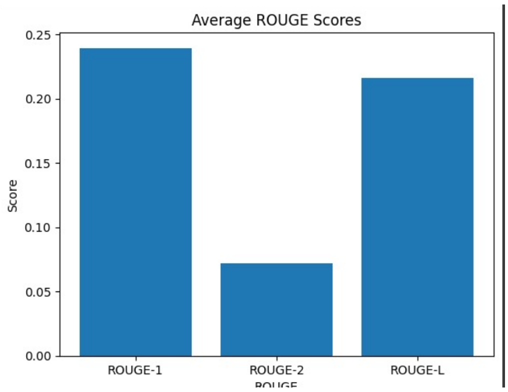

Fig -2: ROUGE Scores for 10 epochs

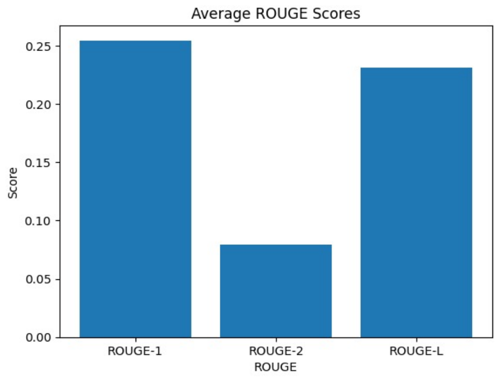

Fig -3: ROUGE Scores for 25 epochs

Fig 1 shows the ROUGE scores, it shows that over the span of 10 epochs, the model's scores get better. Notably, the highest scores are achieved in the order of ROUGE-L, followed by ROUGE-2, ROUGE-1, and ROUGE. This pattern indicates the model's ability to create coherent and fluent summaries while preserving essential information. Despite this progress, the ROUGE scores remain relatively low, which means there is room to improve.

Fig 2 shows the model's progress when it is trained with 25 epochs. Throughout 25 epochs, the model's ROUGE scores demonstrate progressive enhancement. The highest scores are consistently observed in the order of ROUGE-L, followed by ROUGE-2, ROUGE-1, and ROUGE. This pattern highlights the model's capability to generate summaries that are not only coherent but also more fluent compared to the original text, while preserving crucial information.

The model's improvement in ROUGE scores can be attributed to a few key factors. Firstly, longer training exposes the model to a wider range of information, leading to better performance. Additionally, extended training duration enhances the model's grasp of human language, resulting in improved summaries. Furthermore, as the model learns more, its accuracy in producing summaries that align with human-generated content also increases, ensuring factual correctness.

3. CONCLUSIONS AND FUTURE WORK

Our successful project focused on abstractive text summarization introduces a system powered by the T5 transformer language model. The project highlights the utility of abstractive summarization in automating data extraction and elevating decision-making processes. Notably, a comparative analysis reveals that the abstractive model outperforms the extractive counterpart, capturing more comprehensive details.

Looking forward, this technology bears the potential to revolutionize how humans comprehend and utilize textual content, enhancing its accessibility and efficacy across various domains. Future enhancements could include fine-tuning and domain adaptation to tailor models for specific industries, enabling more precise and contextually relevant summaries. Furthermore, addressing the challenge of multi-document summarization is crucial for accommodating scenarios involving related documents, requiring methods to generate coherent summaries from multiple sources.

ACKNOWLEDGEMENTS

We are deeply grateful to our guide, Prof. Sasikala Nagarajan for their support and mentorship throughout the course of this project.

REFERENCES

[1] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi, "Review of automatic text summarization techniques & methods", Journal of King Saud University 2022

[2] Khilji, Abdullah & Sinha, Utkarsh & Singh, Pintu & Ali, Adnan & Pakray, Dr. Partha "Abstractive Text Summarization Approaches with Analysis of Evaluation Techniques", Computational Intelligence in Communications and Business Analytics 2021

[3] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, "Sequence to Sequence Learning with Neural Networks", arXiv Cornell University 2014.

[4]Jakob Uszkoreit, "Transformer: A Novel Neural Network Architecture for Language Understanding", Google Research 2017

[5] Abigail Rai, Study of Various Methods for Tokenization, Applications of Internet things pp 193-200 2020.