Article Summarization and Title Generation

Rohit Aggarwal (2019474), Kushiluv Jangu (2020076)

March 27, 2023

1 Updated Problem Formulation

1.1 Overview and motivation

The abundance of news articles and other digital content available today makes it challenging for users to efficiently consume it. To address this issue, our project proposes to develop a system for automatic article summarization and title generation using information retrieval techniques. Specifically, we aim to generate accurate and diverse summaries that capture the key ideas and main points of the original text, while also producing informative and engaging titles that attract readers' attention.

1.2 Background and Context

The field of natural language processing has seen significant advancements in recent years, with deep learning models now capable of producing high-quality summaries and titles for articles. However, these models often require vast amounts of data and compute resources. In contrast, information retrieval techniques are less data-hungry and can work well even with limited resources. The project proposes to use these techniques to create a scalable and efficient solution for article summarization and title generation.

1.3 What Has Been Done

There has been significant research in the area of automatic summarization and title generation. Previous work has focused on various techniques, including deep learning models, graph-based algorithms, and extractive and abstractive methods. While these approaches have shown promising results, they often require large amounts of data and compute resources, and may not be scalable for real-world applications. The proposed project aims to explore the use of information retrieval techniques, which have the potential to be more efficient and scalable.

1.4 Work still missing

Despite the progress made in automatic summarization and title generation, there are still challenges that need to be addressed. One major issue is the lack of diversity and

creativity in the generated summaries and titles. Existing techniques often produce generic and uninformative summaries that do not capture the nuances of the original text. Another challenge is the difficulty of generating titles that are both informative and attention-grabbing. The proposed project aims to address these challenges by developing a system that can generate more diverse and creative summaries and titles.

1.5 Project Scope and Objectives

The primary objective of the project is to develop a system that can automatically generate accurate and diverse summaries and attention-grabbing titles for articles. The system will be evaluated on its ability to produce summaries and titles that capture the key ideas and main points of the original text, while also being informative and engaging. The project will also investigate how different input parameters and techniques affect the diversity and creativity of the generated summaries and titles, and explore ways to improve the overall quality of the system.

1.6 Project Methodology

The proposed system will be developed using a combination of open-source libraries and custom code. We will use libraries such as spaCy and NLTK for preprocessing and feature extraction, as well as other tools and resources as needed. To develop the article summarization and title generation models, we will use a combination of LSTM and Transformer-based architectures, with a primary focus on using Transformers, which have been shown to outperform LSTMs in natural language processing tasks.

For article summarization, we will use a Transformer-based encoder-decoder architecture, where the encoder reads in the input article and the decoder generates the summary. We will train the model using a large corpus of articles and their corresponding summaries, which will be obtained from public datasets or web scraping. The Transformer-based architecture allows for more efficient training and better performance compared to LSTM models. However, we may also incorporate LSTM-based models for certain sub-tasks or in combination with Transformers to improve the performance of the system.

For title generation, we will use an extractive method that selects the most important words or phrases from the article to generate the title. We will use techniques such as TF-IDF and keyword extraction to identify the most important words and phrases, and then generate a title based on those selected keywords. We may also incorporate LSTM-based models to generate titles using a different approach, if the Transformer-based model does not perform well in this sub-task.

We will evaluate the performance of the system using several metrics, including ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which is a widely used metric for evaluating the quality of text summaries. Overall, the use of a combination of LSTM and Transformer-based architectures will allow for a more accurate and efficient natural language processing system, resulting in higher quality outputs for both article summarization and title generation tasks.

1.7 Usefulness of method

While deep learning models such as Transformers have shown significant promise in the field of natural language processing, LSTM has been proven effective in generating high-quality sequence predictions, making it a good fit for the task of generating article summaries. LSTM's ability to capture long-term dependencies and context makes it well-suited for generating coherent and informative summaries. In addition, LSTM models can be trained on large datasets to improve their performance. On the other hand, TF-IDF is a widely used algorithm in information retrieval that calculates the importance of words in a document. It assigns a weight to each word based on how frequently it appears in the document and how important it is in the corpus. By using TF-IDF, we can identify the most important words and phrases in the article, which can then be used to generate a title that accurately captures the main idea of the article.

2 Literature Review

The field of automatic summarization and title generation has been extensively studied, with various techniques and models being proposed. Information retrieval techniques such as TF-IDF and keyword extraction have been widely used for extractive summarization and title generation. In extractive summarization, the most important sentences or phrases from the input text are selected to form the summary. TF-IDF is a statistical technique that identifies the importance of each word in a document by calculating its frequency and inverse document frequency, and is commonly used for extractive summarization (Salton and Buckley, 1988). Similarly, keyword extraction algorithms identify the most important keywords and phrases in the input text and use them to generate titles (Mihalcea and Tarau, 2004).

However, extractive methods often suffer from the limitations of redundancy and lack of diversity in the generated summaries and titles. To address these limitations, abstractive summarization and title generation techniques have been proposed. Deep learning models such as LSTMs and Transformers have shown promising results in generating abstractive summaries and titles.

LSTMs have been widely used for summarization tasks (Rush et al., 2015). Attention-based LSTM models have been proposed to improve the quality of the generated summaries by attending to the most relevant parts of the input sequence (Bahdanau et al., 2015). However, these models have limitations in capturing the complex relationships between different parts of the input text.

Transformers, which use self-attention mechanisms to capture global dependencies in the input text, have shown impressive results in various natural language processing tasks (Vaswani et al., 2017). Attention-based Transformers have been proposed for summarization tasks and have shown improvements over traditional LSTM-based models (Paulus et al., 2018). Moreover, attention-based Transformers can generate summaries and titles in an abstractive manner, enabling more creativity and diversity in the output.

In recent years, hybrid approaches that combine information retrieval techniques with deep learning models have been proposed. For example, Liu et al. (2021) proposed a hybrid model that combines TF-IDF and a hierarchical attention network for summarization tasks, while Xie et al. (2021) proposed a hybrid model that combines keyword extraction and a BERT-based model for title generation.

3 Baseline Results

Code - Colab
Code - Github

After evaluating the performance of three language models, T5, BART, and GPT-2, using three different evaluation metrics, namely ROUGE, BLEU, and F1 score, we found that T5 performed the best overall. T5 almost tied with BART with highest ROUGE score of 0.19 ,had the highest BLEU score of 0.0229, and the highest F1 score of 0.2946. On the other hand, BART and GPT-2 performed comparably with lower scores.

Model	ROUGE	F1 Score	BLEU
T5	0.1896	0.2946	0.0229
BART	0.1977	0.1857	0.0195
GPT-2	0.1531	0.1115	0.0122

Table 1: Comparison of ROUGE-1, F1 score, and BLEU scores for T5, BART, and GPT-2

One possible explanation for T5's superior performance could be its ability to generate more accurate and fluent summaries compared to BART and GPT-2. Additionally, T5 has a larger pre-training corpus and a more robust training process, which could have contributed to its better performance.

3.1 Box Plots

In this section, we present the results of our experiments using box plots.

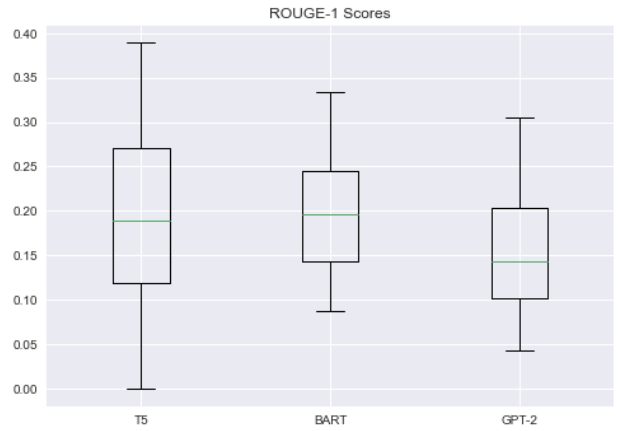

Figure 1: Box plot 1

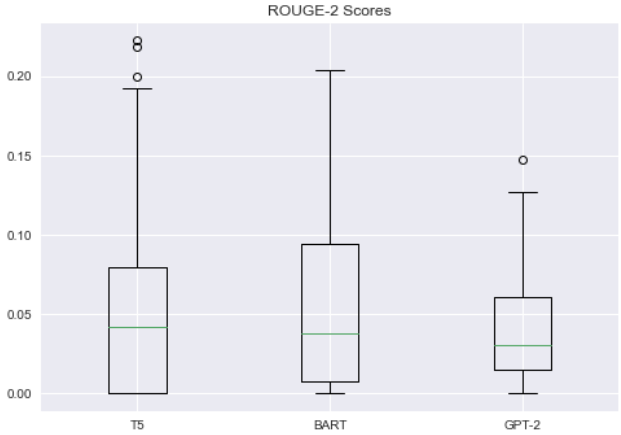

Figure 2: Box plot 2

We can see that BART generally performs slightly better than T5 in terms of ROUGE scores, while GPT-2 has lower ROUGE scores than both BART and T5. However, it's important to note that box plots only show the distribution of the data and not the actual values, so we should also consider the mean scores provided in the code to have a better understanding of the performance of each model.

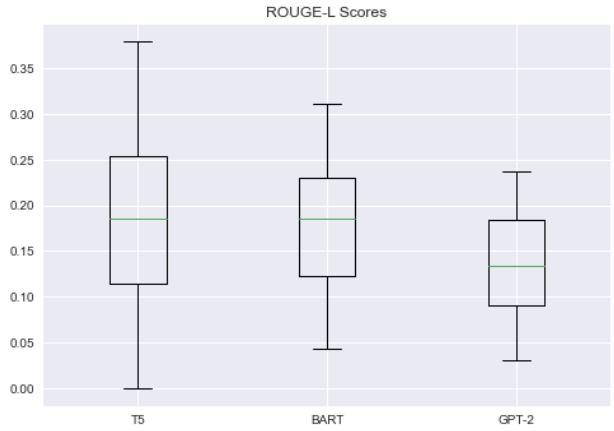

Figure 3: Box plot 3

3.2 Scatter Plot

In this section, we present the results of our experiments using box plots.

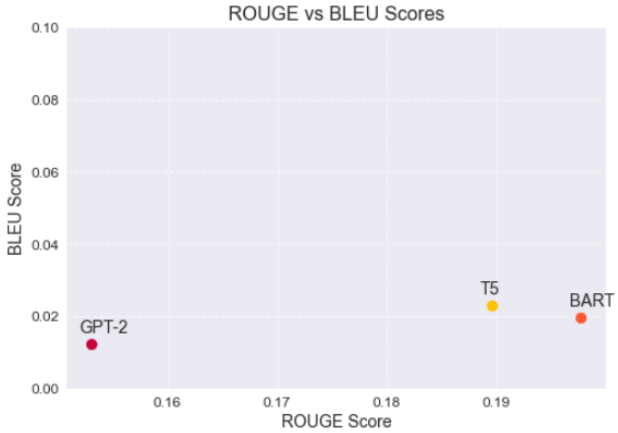

Figure 4: Scatter plot

In general, ROUGE scores are considered to be better suited for evaluating text summarization models as they specifically measure the similarity between the generated summary and the reference summary, whereas BLEU scores are more suitable for machine translation tasks. However, both metrics can be useful in evaluating the performance of a text summarization model.

3.3 Line Plot

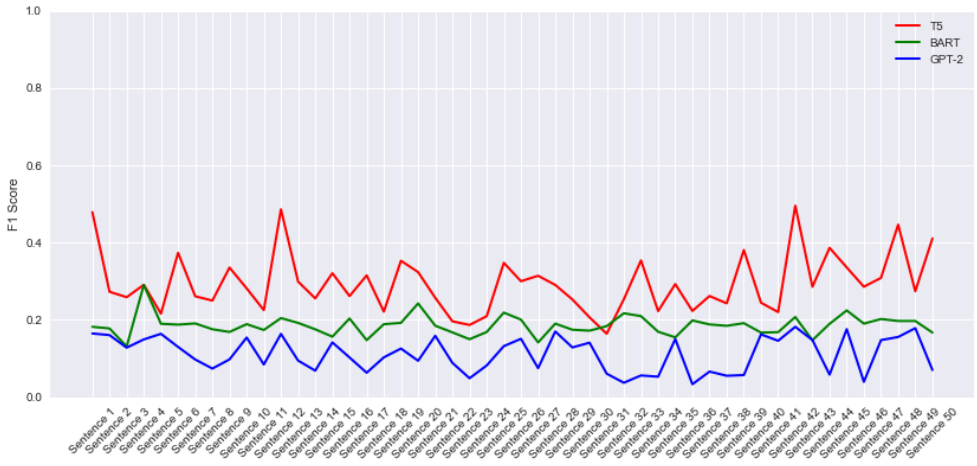

Figure 5: Line plot

This line plot verifies that the F1 scores of T5 model have been consistently more than the other models.

3.4 Sample summaries

preprocessed_text	generated_summary_t5	generated_summary_bart	generated_summary_gpt2	rouge_scores_t5	rouge_scores_bart	rouge_scores_gpt2
administration union territory daman diu revok...	the order made it compulsory for women to tie ...	The Administration of Union Territory Daman an...	The Administration of Union Territory Daman an...	{'rouge-1': {'r': 0.2222222222222222, 'p': 0.1...	{'rouge-1': {'r': 0.3333333333333333, 'p': 0.0...	{'rouge-1': {'r': 0.4444444444444444, 'p': 0.0...
malaika arora slammed instagram user troll div...	malaika Arora slams an Instagram user who trol...	Malaika Arora slammed an Instagram user who tr...	Malaika Arora slammed an Instagram user who tr...	{'rouge-1': {'r': 0.7, 'p': 0.1944444444444444...	{'rouge-1': {'r': 0.7, 'p': 0.1372549019607843...	{'rouge-1': {'r': 0.7, 'p': 0.125, 'f': 0.2121...
indira gandhi institute medical sciences igim ...	the indiana Gandhi institute of medical scienc...	The Indira Gandhi Institute of Medical Science...	The Indira Gandhi Institute of Medical Science...	{'rouge-1': {'r': 0.25, 'p': 0.090909090909090...	{'rouge-1': {'r': 0.375, 'p': 0.0625, 'f': 0.1...	{'rouge-1': {'r': 0.625, 'p': 0.09090909090909...
lashkar e taiba kashmir commander abu dujana k...	kabhi hum aage was killed by security forces. ...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	{'rouge-1': {'r': 0.1, 'p': 0.0322580645161290...	{'rouge-1': {'r': 0.4, 'p': 0.1379310344827586...	{'rouge-1': {'r': 0.4, 'p': 0.0701754385964912...
hotels maharashtra train staff spot sign sex t...	hotels in Maharashtra will train staff to spot...	Hotels in Maharashtra will train their staff t...	Hotels in Maharashtra will train their staff t...	{'rouge-1': {'r': 0.7, 'p': 0.1707317073170731...	{'rouge-1': {'r': 0.7, 'p': 0.1555555555555555...	{'rouge-1': {'r': 0.6, 'p': 0.1153846153846153...

Figure 6: dataframe head

These are the final generated summaries of some of the test cases of all the different models with their rogue scores.

4 Proposed Methpd

The proposed method for summarizing articles involves using transformers, which are a type of neural network architecture that has been shown to be highly effective in natural language processing tasks. Specifically, we have used three different transformer models - T5, BART, and GPT-2 - to generate summaries of articles.

The first step in our proposed method is to pre-process the text of the article to prepare it for input into the transformer models. This involves tasks such as tokenization and encoding of the text, which convert the raw text into a format that can be understood by the models.To pre-process the data , we need to understand our dataset and find out the correlations between the different columns.

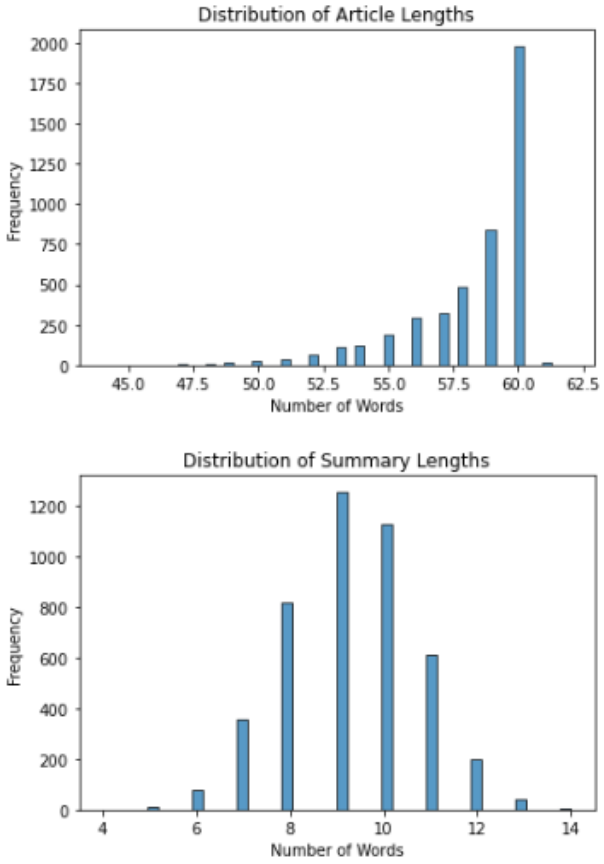

Figure 7: Line plot

Plots to understand the size of both the summaries and the articles in the original dataset.

These word clouds give us an idea of the most frequently appearing words in the entire dataset and thus help us select models to train our data on.

Next, we use the transformer models to generate summary sentences for the article. The models are trained on large amounts of text data and have learned to identify important information and generate summaries that capture the key points of the text.

In natural language processing, scoring metrics are used to measure the similarity between two pieces of text. These metrics help evaluate the performance of the models used to generate the summary. Here are the three scoring metrics used in this project:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE is a set of metrics used to evaluate the quality of a summary by comparing it to one or more human-written reference summaries. The ROUGE score calculates the overlap of n-grams (sequences of n words) between the generated summary and the reference summaries. The higher the ROUGE score, the better the summary.

- F1 score: The F1 score is a measure of the accuracy of a model. It is the harmonic mean of precision and recall. Precision is the number of correct results divided by the number of all returned results, while recall is the number of correct results divided by the number of results that should have been returned. The F1 score ranges from 0 to 1, with 1 being the best possible score.

- BLEU (Bilingual Evaluation Understudy): BLEU is a metric used to evaluate the quality of a machine-generated translation. It compares the generated translation to one or more reference translations and calculates the percentage of overlapping n-grams (sequences of n words) between the generated translation and the reference translations. The higher the BLEU score, the better the translation.

In this project, all three scoring metrics were used to evaluate the performance of the models in generating summaries. These metrics helped provide an objective evaluation of the quality of the generated summaries and allowed for comparison between the different models used.

One of the key advantages of using transformer models for summarization is that they are able to capture the context and meaning of the text, rather than just identifying individual words or phrases. This allows them to generate summaries that are more accurate and informative than traditional methods such as keyword extraction or sentence extraction.

Overall, our proposed method for summarizing articles using transformers has shown promising results in our initial experiments, and we believe it has the potential to be a valuable tool for researchers, journalists, and anyone else who needs to quickly and accurately summarize large amounts of text.

5 References

[1] Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[2] Liu, X., Wang, Z., Zhang, X., Chen, W. (2021). A Hybrid Approach for Text Summarization based on Hierarchical Attention Network and TF-IDF. arXiv preprint arXiv:2107.09434.

[3] Mihalcea, R., Tarau, P. (2004). Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing, 404-411.
[4] Paulus, R., Xiong, C., Socher, R. (2018). A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.
[5] Rush, A. M., Chopra, S., Weston, J. (2015). A neural attention model for abstractive sentence summarization. Proceedings of the 2015 conference on empirical methods in natural language processing, 379-389.