Article Summarization and Title Generation

Rohit Aggarwal (2019474), Kushiluv Jangu (2020076)

March 13, 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 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.

4 Baseline Results

Code - Colab
Code - Github

```
In [32]: df['text'][1]

Out[32]: 'Malaika Arora slammed an Instagram user who trolled her for "divorcing a rich man" and "having fun with the alimony". "Her lif
         e now is all about wearing short clothes, going to gym or salon, enjoying vacation[s]," the user commented. Malaika responded,
         "You certainly got to get your damn facts right before spewing sh*t on me...when you know nothing about me."'

In [33]: df['generated_summary'][1]

Out[33]: 'malaika Arora slams an Instagram user who trolled her for "divorcing a rich man" and "having fun with the alimony" "her life n
         ow is all about wearing short clothes, going to gym or salon, enjoying vacation[s]"'

In [34]: df['generated_title'][1]

Out[34]: 'malaika Arora trolled her for "divorcing a rich man" and "having fun with the alimony" "her life now is all about wearing'

In [44]: df['rouge_scores_summary'][1]

Out[44]: {'rouge-1': {'r': 0.7, 'p': 0.19444444444444445, 'f': 0.30434782268431004},
          'rouge-2': {'r': 0.4444444444444444,
           'p': 0.11428571428571428,
           'f': 0.1818181785640496},
          'rouge-l': {'r': 0.7, 'p': 0.19444444444444445, 'f': 0.30434782268431004}}

In [43]: df['rouge_scores_title'][1]

Out[43]: {'rouge-1': {'r': 0.4, 'p': 0.18181818181818182, 'f': 0.2499999957031251},
          'rouge-2': {'r': 0.2222222222222222,
           'p': 0.09523809523809523,
           'f': 0.13333332913333346},
          'rouge-l': {'r': 0.4, 'p': 0.18181818181818182, 'f': 0.2499999957031251}}
```

Figure 1: 1 : Example of summary and title generated of a random sample

```
In [42]: df['text'][57]

Out[42]: 'Producer of \'Babumoshai Bandookbaaz\' Kiran Shroff, while talking about sexist comments she faced, has said that a Central Bo
         ard of Film Certification (CBFC) member asked her how she could make a film like this being a woman. A male member added, "But
         she is not a woman. Look at what she is wearing." Kiran said that such thoughts were regressive.'

In [40]: df['generated_summary'][57]

Out[40]: "producer of 'Babumoshai Bandookbaaz' Kiran Shroff says she faced sexist comments. a member of the Central Board of Film Certif
         ication asked her how she could make film such as this being the woman."

In [41]: df['generated_title'][57]

Out[41]: "producer of 'Babumoshai Bandookbaaz' says she faced sexist comments."

In [45]: df['rouge_scores_summary'][57]

Out[45]: {'rouge-1': {'r': 0.25, 'p': 0.06896551724137931, 'f': 0.10810810471877293},
          'rouge-2': {'r': 0.0, 'p': 0.0, 'f': 0.0},
          'rouge-l': {'r': 0.25, 'p': 0.06896551724137931, 'f': 0.10810810471877293}}

In [46]: df['rouge_scores_title'][57]

Out[46]: {'rouge-1': {'r': 0.125, 'p': 0.1111111111111111, 'f': 0.11764705384083066},
          'rouge-2': {'r': 0.0, 'p': 0.0, 'f': 0.0},
          'rouge-l': {'r': 0.125, 'p': 0.1111111111111111, 'f': 0.11764705384083066}}
```

Figure 2: 2 : Example of summary and title generated of a random sample

```
df.head()
```

Out[14]:

	author	headlines	text	preprocessed_text	generated_summary	generated_title	rouge_scores_summary	rouge_scores_title
0	Chhavi Tyagi	Daman & Diu revokes mandatory Rakshabandhan in...	The Administration of Union Territory Daman an...	administration union territory daman diu revok...	the order made it compulsory for women to tie ...	order made it compulsory for women to tie rakh...	{'rouge-1': {'r': 0.2222222222222222, 'p': 0.1...	{'rouge-1': {'r': 0.2222222222222222, 'p': 0.1...
1	Daisy Mowke	Malaika slams user who trolled her for 'divorc...	Malaika Arora slammed an Instagram user who tr...	malaika arora slammed instagram user troll div...	malaika Arora slams an Instagram user who trol...	malaika Arora trolled her for "divorcing a ric...	{'rouge-1': {'r': 0.7, 'p': 0.1944444444444444...	{'rouge-1': {'r': 0.4, 'p': 0.1818181818181818...
2	Arshiya Chopra	'Virgin' now corrected to 'Unmarried' in IGIMS...	The Indira Gandhi Institute of Medical Science...	indira gandhi institute medical sciences igim ...	the indiana Gandhi institute of medical scienc...	indiana Gandhi institute of medical sciences c...	{'rouge-1': {'r': 0.25, 'p': 0.090909090909090...	{'rouge-1': {'r': 0.25, 'p': 0.1, 'f': 0.14285...
3	Sumedha Sehra	Aaj aapne pakad liya: LeT man Dujana before be...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	lashkar e taiba kashmir commander abu dujana k...	kabhi hum aage was killed by security forces. ...	kabhi hum aage was killed by security forces. ...	{'rouge-1': {'r': 0.1, 'p': 0.0322580645161290...	{'rouge-1': {'r': 0.1, 'p': 0.0434782608695652...
4	Aarushi Maheshwari	Hotel staff to get training to spot signs of s...	Hotels in Maharashtra will train their staff t...	hotels maharashtra train staff spot sign sex t...	hotels in Maharashtra will train staff to spot...	a mobile phone app called Rescue Me will alert...	{'rouge-1': {'r': 0.7, 'p': 0.1707317073170731...	{'rouge-1': {'r': 0.2, 'p': 0.0645161290322580...

Figure 3: Dataframe head of the result