Group Report - Group Project

Group - 2
Meet Daxini
Mohiddin Bacha Shaik
Sri Sankeerth Koduru

News Article Summarization using Sequence-to-Sequence Models

DATS 6312 - Natural Language Processing
Spring - 2024
Dr. Amir Jafari

Table of Contents

1. Introduction:

In the digital age, staying updated with news is crucial but time-consuming. The vast amount of information available online often overwhelms readers, making it challenging to extract key insights efficiently. This project addresses this issue by exploring Sequence-to-Sequence (Seq2Seq) models for automating news article summarization using advanced Natural Language Processing (NLP) techniques. By condensing lengthy articles into concise summaries, our system aims to provide efficient access to essential information, enabling users to stay informed without spending excessive time on reading.

1.1. Objective

Through the exploration of Seq2Seq models, we aim to develop a robust text summarization system capable of accurately capturing the essence of news articles. By leveraging these models, we seek to generate summaries that maintain coherence, relevance, and accuracy, thereby enhancing the accessibility of news content for readers, researchers, and professionals. Additionally, this project contributes to the advancement of NLP technologies, particularly in the domain of text summarization, by pushing the boundaries of what is achievable in creating informative and concise summaries from large bodies of text.

Additionally, we intend to create a user-friendly application that enables individuals to submit either a direct link to an article, multiple articles, or a text prompt outlining their specific interests. This application will then leverage the trained Seq2Seq model to generate concise and informative summaries of the provided content.

2. Dataset

2.1. CNN Daily News Dataset

The CNN Daily News dataset comprises a collection of news articles from CNN and Daily News, covering various topics such as politics, business, sports, and entertainment. It contains a vast array of articles published over a long period, providing rich and diverse content for analysis. Metadata, including the publication date, title, text body, and text summaries accompanies each article.

2.2. Multi-News Dataset

The Multi-News dataset is a comprehensive collection of news articles and summaries from multiple sources, curated specifically for text summarization tasks. It features articles from various news outlets, ensuring a wide range of perspectives and topics. Each article in the Multi-News dataset is paired with a human-written summary, serving as a reference for summarization algorithms.

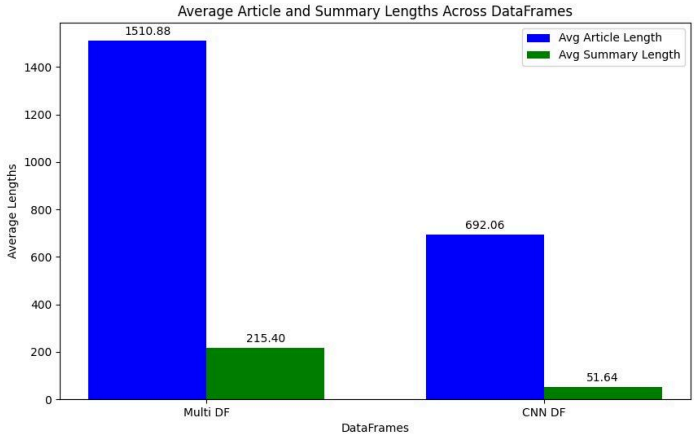

Fig 1 - Average number of words per article and summary

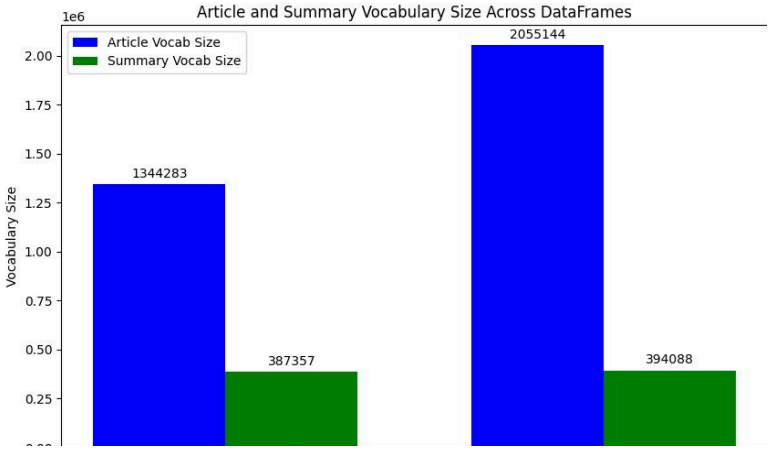

Fig 2 - Size of vocabulary

3. Model

3.1. Data Processing

Normalization of Datasets

Prior to training, both datasets are normalized to identical maximum lengths. This standardization ensures consistency in input size for the models, facilitating more efficient learning and prediction processes.

Removal of Duplicate Summaries

Each dataset is individually purged of duplicate summaries. This step is critical in preventing model bias and promoting diversity within the training data, which can enhance the model's ability to generalize across different types of content.

Completeness Check

Any rows missing either an article or a summary are removed, ensuring that only complete data entries are used in training. This is essential for maintaining the quality and reliability of the training process.

3.2 Model Training and Testing

3.2.1. Encoder-Decoder Using LSTM - Failed to implement

An encoder-decoder architecture using LSTM networks was implemented. While the training phase yielded promising results with significant reductions in loss, challenges arose during the implementation of the summary generation component. The intention was to utilize the encoder-decoder model to maintain consistency in learning and output generation. However, despite efforts to initialize the decoder with states from the encoder and incorporate embeddings and a dense softmax layer for word prediction, functional integration for accurate summary generation was not achieved. This difficulty underscores the complexity of LSTM-based encoder-decoder frameworks and emphasizes the importance of precise state management to ensure coherent text generation

3.2.2. CNN-Daily Mail Trained Transformer Model

Used the CNN/DailyMail dataset from the Hugging Face datasets library and after applying the preprocessing steps mentioned, fine-tuned the BART-large-xsum model which is a transformer-based seq2seq model pre-trained on a large corpus of text data and then fine tuned for summarization using Xsum dataset for this task. Tokenizing the articles and summaries using the BART-large-xsum tokenizer. Setting the hyperparameters such as batch size (Tried different bach sizes of 8 was the max that could be handled by the gpu g5@xlarge instance), learning rate(2e-5), and 3 number of epochs(it took more than 20 hours just for 3 epochs).

Used the Hugging Face Trainer class to handle the training process. Implemented a custom callback to save the evaluation results at each epoch.

3.2.3. Multinews Trained Transformer Model

The Multinews model implemented a robust data processing and model training pipeline utilizing the facebook/bart-large-xsum transformer. A preprocess_data function was used to clean, tokenize, and encode textual data extracted from DataFrames, ensuring compatibility with the selected language model. The create_tensors function converts the data into PyTorch tensors (input IDs, attention masks, labels, and label attention masks) to provide the model with the necessary input structure. Tensor datasets and data loaders with appropriate samplers were constructed to streamline the training process.

Hyperparameter configuration was carefully considered, including the number of epochs, learning rate, and the use of the AdamW optimizer. A linear scheduler with a warm up phase was implemented to dynamically adjust the learning rate. During the iterative training process, ROUGE scores were computed to assess summarization performance, while training and validation losses were monitored. Early stopping was incorporated to prevent overfitting and preserve the best-performing model state, supporting generalization to unseen data.

We implemented a streamlined model testing process for a Multinews summarization model to test the model's performance on unseen data. This involved loading the best-performing model created during the training called 'best_model_Multi_News_final.pt'. Test data was loaded, and a was preprocessed (cleaned, tokenized, encoded) for compatibility. PyTorch tensors and a DataLoader were created for efficient input to the model. During the evaluation phase, we generated summaries for the test set and computed ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) to measure the model's ability to create accurate and concise summaries.

To perform summaries on new data we loaded the best-performing summarization model called 'best_model_Multi_News_final.pt'. To evaluate the model's summarization capabilities, we supplied it with generated article text. This text was pre-processed through tokenization, and a maximum length parameter was utilized to regulate the output length. The intended outcome was the production of a summary that accurately reflects the essential elements of the original input text.

3.3. Results

3.2.1. CNN-Daily Mail Trained Transformer Model

The test results for the CNN-Daily Mail Trained Transformer Model are

- ROUGE-1 (0.440234): This metric assesses the overlap of individual words between the generated summaries and their corresponding reference texts. The substantial score of 0.440234 signifies the model's proficiency in selecting and incorporating key terms and concepts from the source materials. This indicates a robust basis for summarization tasks with respect to content preservation.
- ROUGE-2 (0.20731): This metric evaluates the presence of overlapping two-word phrases (bigrams). While exhibiting a lower score than ROUGE-1, it nonetheless demonstrates the model's capacity to construct reasonably coherent phrases within the summaries. However, the score suggests potential for refinement in grammatical accuracy and structural flow to enhance overall quality.
- ROUGE-L (0.303239): This metric analyzes the longest matching word sequences, offering insights into the similarity of sentence structure and flow in comparison to the reference texts. The score indicates a moderate degree of coherence, while also highlighting the potential for improvement in the overall fluidity and readability of the generated summaries.

3.3.2. Multinews Trained Transformer Model

The test results for the Multinews Trained Transformer Model are:

- ROUGE-1 (0.3710001712954484): This score focuses on the single-word overlap between the generated summary and the reference. Our model achieved a score of 0.371, indicating it captures many important words and concepts from the source articles. While this is a positive starting point, there is potential to improve the comprehensiveness of the summaries.
- ROUGE-2 (0.12235801956537162): This score measures the overlap of two-word phrases. Our model's score of 0.122 suggests it might struggle to generate grammatically correct and well-structured phrases within the summary. This highlights the need for further development in the model's ability to produce fluent and coherent text.
- ROUGE-L (0.2109042732591439): This score considers the longest matching sequence of words between the summaries. Our model's score of 0.211 indicates some level of similarity in sentence structure and flow compared to the references. However, there's room for improvement to enhance the overall flow and coherence of the generated summaries.

4. App

4.1. The app is provided a Search Query

- The user inputs a topic or area of interest.
- The system locates the top 5 most relevant articles on the topic using external API News API.
- The CNN-Daily Mail Transformer Model generates individual summaries for each of the top 5 articles.
- Then after summarizing all the articles we use Multi news transformer to generate summaries of all the articles together

4.2. The app is provided a link or multiple links

- The user inputs a link or a multiple links that they would like to summarize.
- The CNN-Daily Mail Transformer Model generates individual summaries for each of the articles in the provided links.

- If there are multiple links then all the content from all the provided links is summarized by Multi news transformer.

4.3. The app is provided with text

- The user inputs the text that they would like to summarize.
- The CNN-Daily Mail Transformer Model generates summaries for the provided text.

4.4. Summarizing large content

- Both the transformers have been fine tuned using facebook/bart-large-xsum transformer. Bart has a limit of 1024 tokens and we can get articles larger than 1024 tokens and when we are doing a summary of multiple articles it is almost always more than 1024 tokens. So for summarizing we have created a utility to just do a summary of only 1024 tokens or less at a time. We do this by using the NLTK package. It has a sentence tokenizer. We combine sentences until the 1024 token limit is not reached and then we do each group summarization sequentially and combine. This also has a problem that sometimes the summary is not cohesive because we have scraped the data of the article and some of the scraped data might not even be related to the article which gets summarized separately and makes it into the final summary.

4.5. App also gives control over summary generation parameters

- The user might want to generate long or short summaries based on their preference so we have given them sliders to control
- Also we have provided a slider for length penalty to control the verbosity of the output.

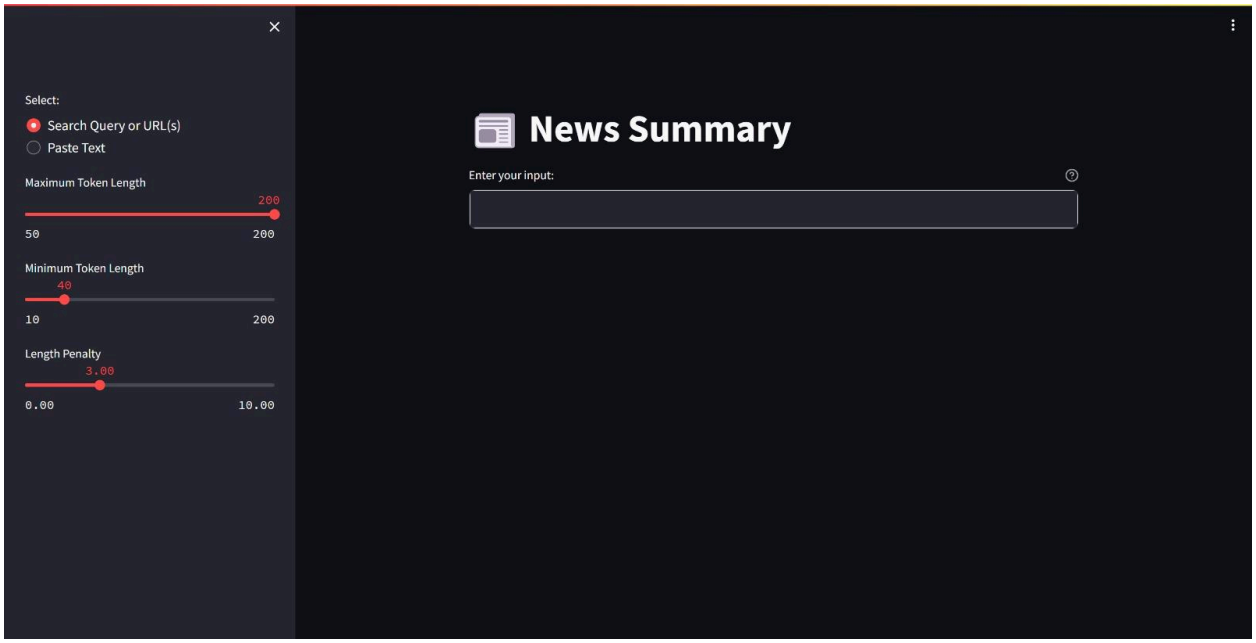

Image 3: App Interface

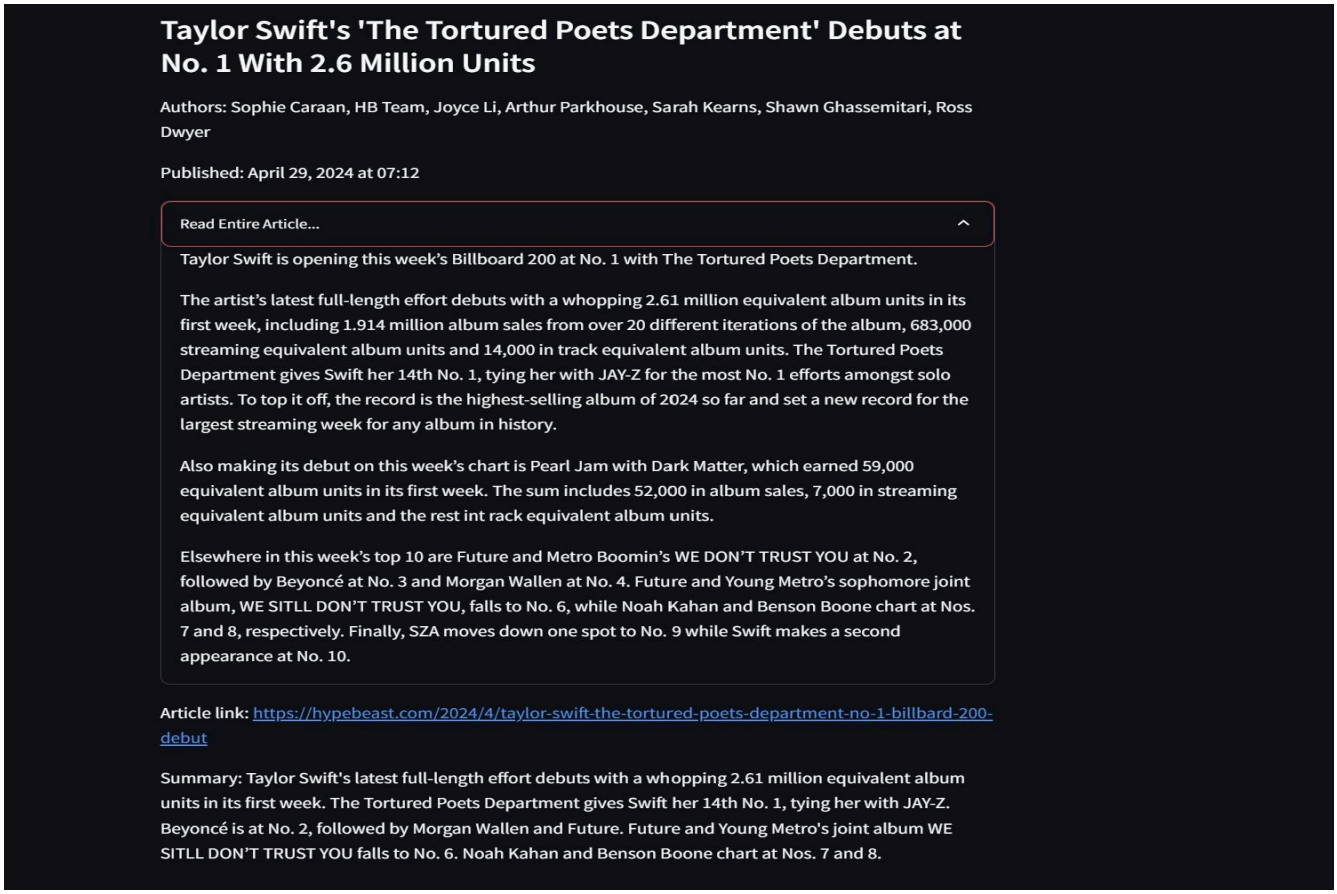

Image 4: Example summarization

5. Conclusion

In conclusion, this project demonstrates the effectiveness of sequence-to-sequence transformer models for news article summarization. Both the CNN-Daily Mail and Multinews trained models successfully extract core information from news articles, presenting it concisely. The CNN-Daily Mail model's higher ROUGE scores indicate its superior summarization capabilities. The Multinews model's performance, while still effective, is likely affected by limitations in the number of tokens it can process. This suggests that its training may not have incorporated the full breadth of news article summaries, an issue to be addressed in future work.

While both models offer value, there is room to improve the fluency and coherence of longer summaries, as reflected in the lower ROUGE-2 and ROUGE-L scores. Future work will outline strategies to specifically target these areas.

6. Future Work

- Enhance model performance on larger articles by increasing the token processing capacity during training. Limiting the model to the first 1024 tokens can lead to a loss of valuable context and insights.
- Optimize model output for larger articles by employing KNN clustering for the removal of irrelevant information. This will improve focus and the model's ability to extract the relevant points.
- Incorporate a multi-news summarization model into the application to provide a comprehensive perspective from diverse sources. This will enhance the understanding of complex topics or events.

7. References

Pytorch Documentaion: https://pytorch.org/docs/stable/index.html

TensorFlow Documentation: https://www.tensorflow.org/guide

Huggingface Documentation: https://huggingface.co/docs

Chat Summarizer: https://medium.com/@ferlatti.aldo/fine-tuning-a-chat-summarizer-c18625bc817d

Long Document Summarizer: https://discuss.huggingface.co/t/summarization-on-long-documents/920/7

ChatGPT: https://chat.openai.com/

Gemini: https://gemini.google.com/app

Claude: https://claude.ai/login?returnTo=%2F%3F