News Article Summarization using Sequence-to-Sequence Models

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

Group 2
Meet Daxini
Mohiddin Bacha Shaik
Sri Sankeerth Koduru

Introduction

- The project aims to address this issue by utilizing Sequence-to-Sequence (Seq2Seq) models
- Develop a robust text summarization system using Seq2Seq models to accurately capture the essence of news articles.
- Generate summaries that maintain coherence, relevance, and accuracy, thereby enhancing the accessibility of news content for readers, researchers, and professionals.
- Create a user-friendly application allowing individuals to submit links to articles, multiple articles, or text prompts, and generate concise summaries using the trained Seq2Seq model.

Datasets

CNN Daily News Dataset

The CNN Daily News dataset is a collection of news articles from CNN and Daily News, covering various topics such as politics, business, sports, and entertainment. It provides a rich and diverse content for analysis, with metadata accompanying each article, including the publication date, title, text body, and text summaries.

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.

- Both datasets are normalized before training to the same maximum lengths to ensure consistency in input size for models.
- Duplicate summaries are removed in both datasets individually to avoid bias and ensure diversity in the training data.

CNN-Daily Mail Transformer Model

- Fine tuned facebook/bart-large-xsum on CNN DailyMail dataset from hugging face.
- Trained for 3 epochs with CrossEntropyLoss which measures the dissimilarity between the predicted token distribution and the actual token distribution.
- Mixed precision training uses a combination of 16-bit and 32-bit floating-point operations to reduce memory usage and improve training speed.

Limitations

- Constraining the summary length to a fixed value of 200 tokens may result in generated summaries that sound abrupt or unnatural
- Bart model has 1024 token limit so articles having more than 1024 token are truncated so not the entire context is used.

CNN-Daily Mail Transformer Model

ROUGE-1 Score: 0.440234, Measures single-word overlap between generated summaries and reference texts.
Result: A substantial score of 0.440234, indicating strong proficiency in capturing and integrating key terms and concepts from the source material.
Insight: Demonstrates a robust foundation for summarization, excelling in content preservation which is crucial for accurate summarization tasks.

ROUGE-2 Score: 0.20731, Evaluates the overlap of two-word phrases between summaries and references.
Result: Though lower than ROUGE-1, the score of 0.20731 shows the model's ability to construct coherent phrases.
Insight: Suggests a need for refinement in grammatical accuracy and structural flow to further enhance the quality of summaries.

ROUGE-L Score: 0.303239, Assesses the longest matching sequence of words in the summaries.
Result: The score of 0.303239 suggests moderate coherence.
Insight: Highlights potential areas for improvement in the fluidity and readability of the summaries to better reflect the original texts' style and structure.

Multinews Transformer Model

- Utilized the facebook/bart-large-xsum transformer for summarization tasks.
- AdamW chosen for its adaptive learning rate capabilities and weight decay regularization to efficiently update model parameters and mitigate overfitting by penalizing large parameter values.
- Employed a linear scheduler with no warm-up steps to gradually decrease the learning rate over training. This stabilizes the training process and improves convergence towards optimal parameter values.

Limitations

- Multi-News contains similar news from multiple sources, and the summary is based on all these sources. Bart models have a token limit (1024 tokens), which may cut off information from additional sources in the dataset, potentially affecting the quality of the summary.

Multinews Transformer Model - Test Results

ROUGE-1 Score: 0.371, Measures single-word overlap between generated summaries and reference texts.
Result: The model achieved a score of 0.371, capturing key words and concepts effectively.
Insight: While the results are promising, there is potential to enhance the depth of the summaries to cover more content comprehensively.

ROUGE-2 Score: 0.122, Evaluates the overlap of two-word phrases between summaries and references.
Result: With a score of 0.122, the model shows a challenge in forming grammatically correct and structured phrases.
Insight: This indicates a need for further development in the model's capability to generate fluent and coherent summaries.

ROUGE-L Score: 0.211, Assesses the longest matching sequence of words in the summaries.
Result: The score of 0.211 suggests moderate alignment with the structure and flow of the reference texts.
Insight: Improvement is needed to enhance the flow and overall coherence of the summaries for better readability.

Features

News Article Summarization

- Generates concise summaries of news articles based on input text or news article links.

Latest News Search

- Provides the latest news updates about any keywords you search for.
- Summarizes the latest news articles based on the searched keywords.

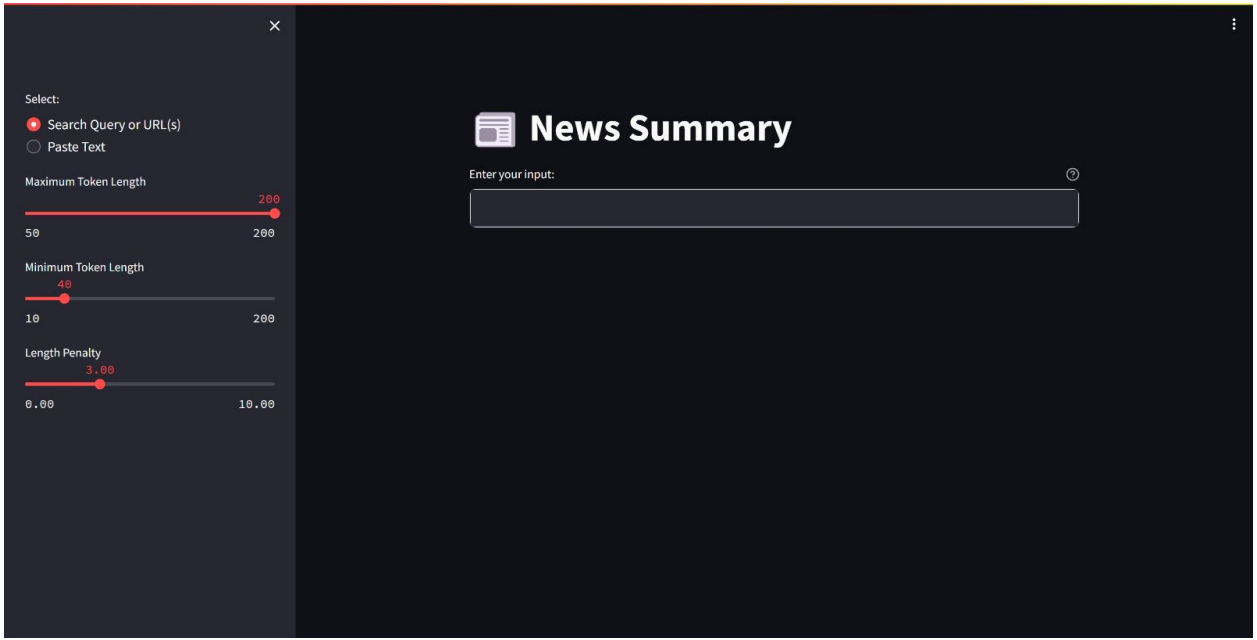
App - News Summarization

Prompt-Based Summarization
User Action: Input a topic or interest area.
System Function: Uses a search engine or news aggregator to find the top five relevant articles.
Output: The CNN-Daily Mail Transformer Model creates individual summaries for each article.

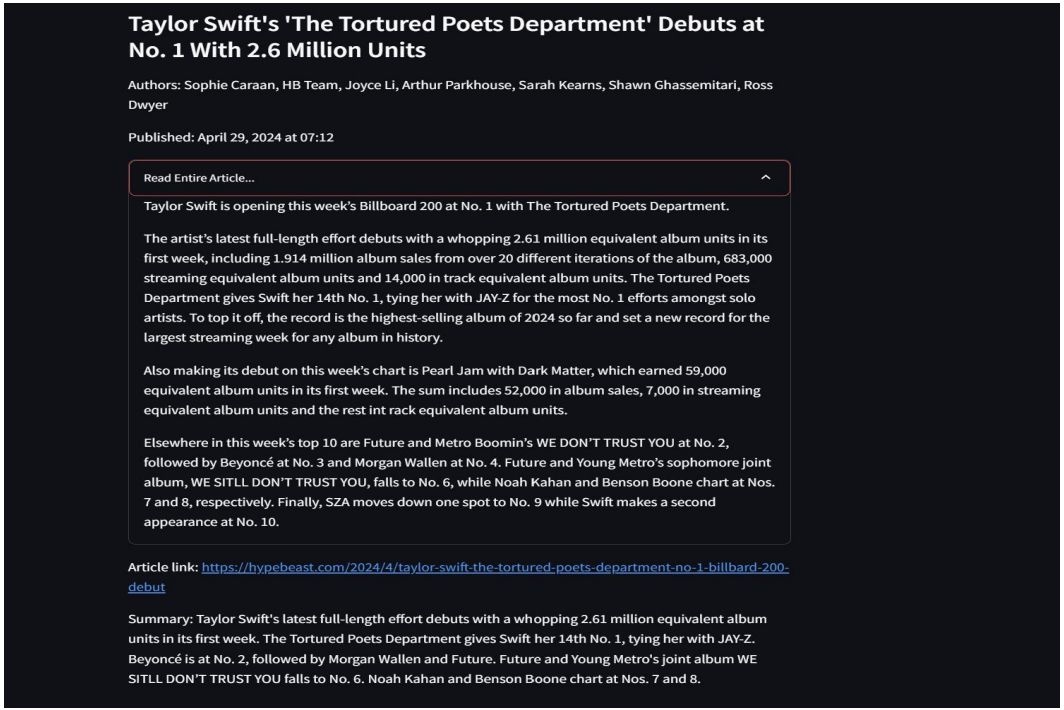
Link-Based Summarization
User Action: Submit one or more links for summarization.
System Function: The CNN-Daily Mail Transformer Model processes each link.
Output: Generates a summary for each linked article.

Text-Based Summarization
User Action: Enter the text to be summarized.
System Function: The CNN-Daily Mail Transformer Model analyzes the input text.
Output: Produces a summary based on the provided text.

App - News Summarization



App - News Summarization



Conclusion

This project highlights the effectiveness of transformer models in news summarization, with both CNN-Daily Mail and Multinews models performing well in extracting key information concisely.

CNN-Daily Mail Model: Showcases superior summarization with higher ROUGE scores, indicating better performance.
Multinews Model: Effective, yet impacted by limited token processing, suggesting an area for future improvement.

Both models have potential for improvement in fluency and coherence for longer summaries. Future efforts will focus on enhancing these aspects to elevate summary quality further.

Model	ROUGE-1	ROUGE-2	ROUGE-L
CNN-Daily Mail Transformer Model	0.440234	0.20731	0.303239
Multinews Transformer Model	0.371	0.122	0.211

Future Work

Increase Token Processing Capacity

Objective: Enhance the model's performance on larger articles.

Action: Expand the token processing limit during training, currently capped at the first 1024 tokens.

Benefit: Captures more context and insights by processing full articles, avoiding loss of critical information.

Optimize Model Output for Large Articles

Objective: Improve focus and relevance in summaries for larger articles.

Action: Implement KNN clustering to filter out irrelevant content.

Benefit: Streamlines content by focusing only on pertinent information, enhancing summary accuracy.

Incorporate Multi-News Summarization

Objective: Provide a more rounded perspective on complex topics by summarizing multiple sources.

Action: Integrate a multi-news summarization model within the application.

Benefit: Offers comprehensive overviews and diverse viewpoints, enriching user understanding of topics

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