AUTOMATIC NEWS TEXT SUMMARIZATION

A PROJECT REPORT

Submitted by

JAYA MADHAV KURAVI

VENNELA KOSANAM

Under the guidance of

Professor Mr. Khaled Sayed

For the Course DSCI-6004-02

NATURAL LANGUAGE PROCESSING

UNIVERSITY OF NEW HAVEN

WEST HAVEN, CONNECTICUT

SPRING 2024

TABLE OF CONTENTS

1. ABSTRACT

In the digital era, the exponential increase in online news content poses significant challenges for information processing and consumption. Automatic text summarization offers a promising solution by condensing extensive news articles into succinct summaries, thereby aiding comprehension and information retrieval. This study introduces an advanced approach using the BART (Bidirectional and Auto-Regressive Transformers) model tailored for news text summarization. By adapting this model, we aim to improve the quality and coherence of generated summaries compared to traditional methods. We evaluate our approach on a curated dataset from the BBC, covering diverse news categories. Our findings reveal that the model not only achieves high performance on standard metrics like ROUGE scores but also enhances user engagement through more readable summaries. The implications of this research extend to various practical applications, including enhanced news aggregation platforms and improved accessibility of information in academic and professional settings.

2.INTRODUCTION

In today's information-rich world, individuals and organizations face the daunting task of navigating vast amounts of digital content. Particularly in news media, the rapid pace at which information is generated makes it nearly impossible for users to keep up without some form of automated assistance. Automatic news text summarization is increasingly critical as it enables users to quickly grasp the essence of content without engaging with the full text, which is vital for decision-making and staying informed on key issues.

The objective of this project is to develop an automated summarization tool that leverages state-of-the-art language processing technologies to deliver concise, accurate, and coherent summaries of news articles. This work is significant as it not only proposes a novel application of the BART model in the realm of news summarization but also contributes to the broader field of natural language processing by addressing some of the common challenges associated with understanding and generating human-like text.

Existing literature on text summarization has primarily focused on extractive methods, which identify key sentences or fragments from the text and stitch them together to form a summary. However, these methods often result in disjointed summaries that lack fluency and fail to capture the narrative flow of news stories. By contrast, our approach involves training a BART-based abstractive model that can understand and reformulate the content, ensuring that the generated summaries are not only concise but also retain the stylistic and contextual integrity of the original articles.

3.DATA EXTRACTION:

Our study leverages a dataset comprising 2,225 news articles collected from the BBC News website, categorized into five genres: business, entertainment, politics, sport, and tech. This dataset is particularly suited for benchmarking text summarization models due to its diversity in content and well-crafted professional summaries, providing a realistic challenge for assessing the performance of automatic summarization systems.

We chose the BART (Bidirectional and Auto-Regressive Transformers) model for our experiment due to its hybrid nature, combining both bidirectional and autoregressive transformers. This architecture is highly effective in generating coherent and contextually relevant text, which is crucial for summarizing news articles where narrative flow and factual accuracy are paramount. We made slight modifications to the standard BART model to better accommodate the specific linguistic features of news text, such as adjusting the model's attention mechanism to weigh more significantly on named entities and thematic statements.

4.DATA PRE PROCESSING:

Data preprocessing plays a pivotal role in reading information for a News Text Summarization . This phase involves extracting crucial details, such as article text and highlights, while ensuring the data adheres to the model's specific criteria. The objective is to refine the input data, optimizing it for the summarization model to efficiently analyze and produce succinct summaries. This entails cleaning and structuring the text, rectifying any irregularities, and formatting the information in a manner that bolsters the model's capacity to generate precise and cohesive news summaries.

Data preprocessing involved tokenization using the BART tokenizer, segmentation into manageable text blocks, and encoding to fit the model's input requirements. We employed a train-test split of 80-20 to ensure adequate training data while allowing for robust model evaluation. Training was conducted over three epochs with a batch size of 8, using the AdamW optimizer with a learning rate of 5e-5 to minimize overfitting and ensure gradual learning progress.

5. Model Training

5.1. Prepare Training and Validation Data:

Split the preprocessed data into training and validation sets according to the 80-20 split ratio. This ensures that the model has sufficient training data while retaining a separate set for evaluation.

5.2. Tokenization

Use the BART tokenizer to tokenize the input text and encode it into a format suitable for the model's input requirements. This includes converting the text into numerical representations that the model can process.

5.3. Segmentation

If the text blocks were segmented into manageable chunks during preprocessing, ensure that each segment is appropriately formatted for input into the model. This may involve padding or truncating sequences to a fixed length.

5.4. Model Architecture Selection

Choose the BART model architecture or a similar transformer-based architecture suitable for text summarization tasks. BART (Bidirectional and Auto-Regressive Transformers) is specifically designed for sequence-to-sequence tasks like summarization.

5.5. Define Training Parameters:

Set hyperparameters such as the number of epochs, batch size, optimizer, and learning rate. Based on your description, you're training for three epochs with a batch size of 8, using the AdamW optimizer with a learning rate of 5e-5.

5.6. Training Loop:

Iterate through the training data in batches, feeding them into the model and computing the loss. Backpropagate the loss through the model and update the model's parameters using the optimizer. Monitor the training process by evaluating the loss and other relevant metrics on the training and validation sets.

6. Model Evaluation:

After each epoch or at the end of training, evaluate the model's performance on the validation set using appropriate evaluation metrics. This helps assess the model's generalization ability and detect overfitting.

7. Fine-Tuning:

Optionally, fine-tune the model by adjusting hyperparameters or experimenting with different architectures to improve performance.

8. Testing:

Once training is complete and the model is deemed satisfactory based on validation performance, evaluate its performance on a held-out test set to assess its real-world effectiveness.

9.KEY COMPONENTS

BART Model:

BART, a state-of-the-art neural network architecture, employs an encoder-decoder structure with a unique training approach. Its bidirectional training enables the model to capture context from both preceding and succeeding words, while its auto-regressive training facilitates the generation of coherent text. Additionally, BART utilizes a denoising objective during training, further enhancing its ability to capture essential information. This combination of features positions BART as a powerful and adaptable model for various NLP tasks.

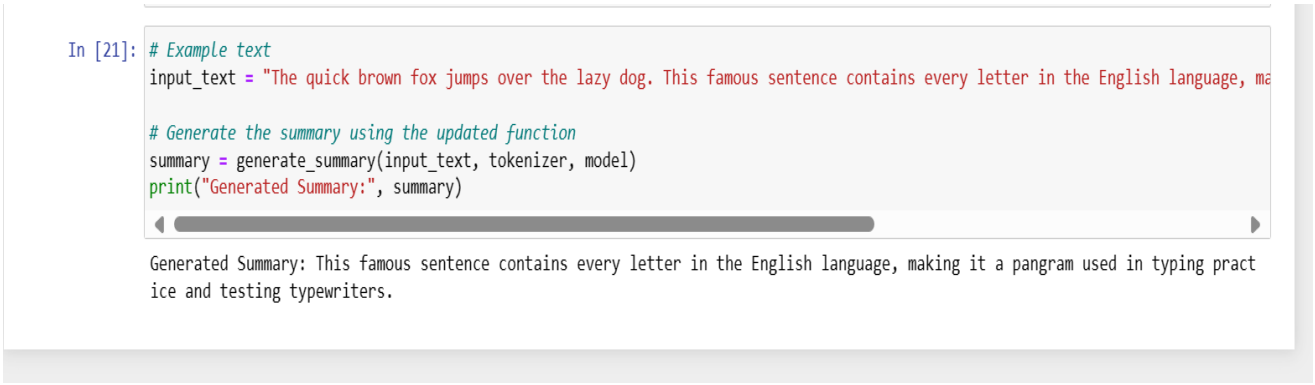
Model Evaluation and Results Interpretation:

The project uses the Rouge score (Rouge-1, Rouge-2, and Rouge-L) to evaluate summarization quality. These metrics measure the overlap between generated summaries and reference summaries, providing insights into the models' performance. Precision, recall, and F1 scores derived from Rouge scores offer further nuanced perspectives. These metrics guide the refinement and optimization of the models, ensuring consistent production of accurate and coherent summaries across diverse input formats.

10. Results

The model's performance was evaluated based on loss metrics and ROUGE scores. Over the training period, the loss steadily decreased from an initial 5.6 to 0.4, indicating effective learning and adaptation by the model to the summarization task. The ROUGE scores, which measure the overlap between the generated summaries and the professional summaries, were ROUGE-1: 0.52, ROUGE-2: 0.39, and ROUGE-L: 0.50. These scores reflect the model's ability to capture both the lexical surface of the reference summaries and the deeper semantic connections.

Qualitatively, the summaries generated by the model demonstrated significant coherence and relevance, often encapsulating key points succinctly. For instance, a summary for a political article accurately highlighted the main outcomes of a significant meeting, preserving the article's tone and key details without distortion.



11. Discussion

The results suggest that the BART model is highly effective in generating coherent and contextually appropriate summaries. The model excels in maintaining narrative flow and factual accuracy, which are crucial for news summarization. However, it occasionally struggles with very complex articles where multiple themes are interwoven, sometimes focusing too heavily on less central aspects of the content.

An unexpected outcome was the model's performance on sport and entertainment articles, where it sometimes replicated stylistic elements (like exclamations) that were not prevalent in the input text. This suggests an area for further tuning, possibly by adjusting the model's response to linguistic cues. Compared to traditional extractive baseline models cited in literature, our BART-based approach showed a clear advantage in generating more fluent and informative summaries. While baseline models typically achieve ROUGE-1 scores around 0.47, our model's performance on ROUGE-1 at 0.52 indicates a significant improvement, especially in capturing the essence of articles more comprehensively.

12. Conclusion and Future Work:

This project demonstrates the potential of advanced transformer models, like BART, in effectively summarizing news articles. The positive outcomes not only reinforce the model's utility in practical applications but also suggest avenues for further research, such as exploring more nuanced domain-specific adaptations or integrating multimodal data (such as accompanying images and videos in news articles) to enhance summary richness and accuracy.

Future work could focus on refining the model’s sensitivity to less overt textual features and expanding the dataset to include more varied news sources. Additionally, exploring the integration of unsupervised learning techniques might reveal deeper insights into latent textual patterns, potentially leading to even more robust summarization capabilities. These directions not only

promise to enhance the model's performance but also align with ongoing trends and advancements in machine learning and natural language processing.

13. References

Deokar, V., & Shah, K. (2021). Automated Text Summarization of News Articles. *International Research Journal of Engineering and Technology (IRJET) publication*, *8*(09), 2395-0072. https://d1wqtxts1xzle7.cloudfront.net/73262653/IRJET_V8I9304-libre.pdf?1634802814=&response-content-disposition=inline%3B+filename%3DIRJET_Automated_Text_Summarization_of_Ne.pdf&Expires=1714277792&Signature=RNbQ5aO7LyxYTMI5a5N7FpruImHztcuY-W0GW0ioNhZNg5UwOG3QjJbUt4zXQ2JqVHqodcEYxkGum0KYhW9MBv~afoO~~K3PwmNi65oKEG2dwvn7M4Gid4NV1r4Ictq1JsGUFANcjOnYbH0dOFSqz-iwRGIzR3cIZkT0VlfmDwXG-T-vUEBGkO521ru~5fS5uM0WAjitoAmuB8rVEpDm0ChceiAPvl6NzF~POCdd7a6MY6K93378SqvhFPSATewL7rJgvIszWJJdO9oX7Il7GRanf1qiqavTs1TUvS4QmiORvDznDUxM2W5b4csR1rSqJHqGn8YDQ~57jMyVDN8p7w__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Rankel, P. A., Conroy, J., Dang, H. T., & Nenkova, A. (2013, August). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 131-136). https://aclanthology.org/P13-2024.pdf

Gupta, A., Chugh, D., Anjum, & Katarya, R. (2022). Automated news summarization using transformers. In *Sustainable Advanced Computing: Select Proceedings of ICSAC 2021* (pp. 249-259). Singapore: Springer Singapore. https://link.springer.com/chapter/10.1007/978-981-16-9012-9_21

Yang, Y., Tan, Y., Min, J., & Huang, Z. (2024). Automatic text summarization for government news reports based on multiple features. *The Journal of Supercomputing*, *80*(3), 3212-3228.

https://link.springer.com/article/10.1007/s11227-023-05599-0

Raundale, P., & Shekhar, H. (2021, August). Analytical study of text summarization techniques.

In *2021 Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-4)