Text Summarization on the CNN/Daily Mail dataset using Pre-Trained BART

Matthew Murphy
University of South Florida

Abstract

This paper explores abstractive text summarization using the pre-trained BART model. We evaluate performance on the CNN/Daily Mail dataset using ROUGE metrics. Results demonstrate effective summary generation, with ROUGE-1 scores exceeding 0.40, showcasing the model's capability to generate coherent summaries while highlighting areas for improvement.

1 Introduction

In the field of Natural Language Processing (NLP) text summarization is an area that involves taking a document and condensing it into a concise summary of the document while maintaining its main concepts. With the increasing growth of digital content, thanks to many efforts like Project Gutenberg, text summarization is becoming more vital to the cataloging of digital information and understanding what the actual content of document will be pertaining to. Summarization can be classified primarily into two categories of extractive and abstractive, with extractive taking select key sentences from a source, and abstractive generates new text to capture the meaning of the source in a more coherent manner.

The use of transformer-based models has significantly increased the field of abstractive summarization. The Bidirectional and Auto-Regressive Transformer (BART) model has shown exceptional performance across various summarization tasks. BART is pre-trained using a denoising autoencoder objective, allowing it to reconstruct corrupted input sequences effectively. This makes BART well-suited for tasks like summarizing.

In this study, I evaluated the performance of the pre-trained BART model on the CNN/Daily Mail dataset. The aim was to assess the model's ability to generate coherent summaries that align with human-written headlines. The model's performance was measured using Recall-Oriented Understudy for Gisting Evaluation or ROUGE metrics. These metrics provide an insight into the overlap of words and phrases between generated and reference summaries.

Key Objectives:

1. Evaluate the effectiveness of the BART model in generating abstractive summaries.
2. Analyze the model's strengths and weaknesses based on the quantitative and qualitative results.

By utilizing the capabilities of the pre-trained model, the project demonstrates the feasibility of abstractive summarization in real-world applications.

2 Dataset

The dataset used for the project is the CNN/Daily Mail dataset, which is one of the most widely used datasets for evaluating ROUGE metrics. Its popularity stems from its extensive corpus of news articles paired with human-annotated reference summaries, making it a reliable benchmark for summarization tasks (Zhang et al., 2023). The gold-standard summaries provided by humans ensure accurate

evaluation of machine-generated summaries, allowing for meaningful comparisons across different models (Priyanka, 2020). Additionally, the structured pairing of articles and summaries facilitates straightforward calculation of ROUGE scores, further solidifying its importance in summarization research (Zhang et al., 2023; Priyanka, 2020).

The dataset is divided into three subsets:

1. **Training Set**: Contains nearly 287,000 articles, used for fine-tuning summarization models (Hermann et al., 2015).
2. **Validation Set**: Contains about 13,000 articles, used for tuning hyperparameters and intermediate evaluations (See et al., 2017).
3. **Test Set**: Comprising approximately 11,000 articles, this set is used to evaluate the model's performance (Hermann et al., 2015).

Because the BART model is pre-trained, only the test set was utilized in this project. The average article length in the dataset is about 800 words, with the highlights being 1–3 sentences long (See et al., 2017). These characteristics make the dataset suitable for both extractive and abstractive summarization tasks.

Data Preprocessing: To ensure compatibility with the BART model, the data was preprocessed by truncating articles to a maximum length of 1024 tokens to fit within the model's input constraints, while highlights were capped at a maximum of 128 tokens for efficient summarization (Lewis et al., 2020).

3 Methodology

The methodology in this project leveraged the pre-trained BART model for abstractive text summarization. The approach consisted of data preprocessing, summary generation, and evaluation using established metrics.

3.1 Model Selection

The BART model was selected for its exceptional performance in text summarization tasks (Lewis et al., 2020). BART is a sequence-to-sequence transformer model pre-trained using a denoising autoencoder objective, making it well-suited for generating coherent and meaningful summaries.

3.2 Data Preprocessing

As previously described, data preprocessing involved truncating articles and summaries:

- Articles were truncated to a maximum length of 1024 tokens to fit within the model's constraints.
- Summaries were capped at 128 tokens to ensure effective and concise outputs.

This preprocessing step ensured compatibility with the BART model while maintaining data quality for summarization tasks.

3.3 Summary Generation

Summaries were generated using the facebook/bart-large-cnn implementation from the Hugging Face library.

Key Parameters:

1. Maximum Summary Length: 130 tokens
2. Minimum Summary Length: 30 tokens
3. Sampling Strategy: do_sample=False to prioritize deterministic outputs

3.4 Evaluation

The performance of the generated summaries was evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics:

1. ROUGE-1: Measures word-level (unigram) overlap.
2. ROUGE-2: Measures two-word (bigram) sequence overlap.
3. ROUGE-L: Measures the longest common subsequence.

These metrics provided quantitative insights into the overlap and relevance of the generated summaries compared to the reference highlights (Priyanka, 2020).

3.5 Tools and Frameworks

- Hugging Face Transformers: Used to implement the BART model and manage inferences.
- Hugging Face Datasets: Used to access and preprocess the CNN/Daily Mail dataset.
- Evaluate Library: Used to calculate ROUGE scores and analyze results (Priyanka, 2020).

4 Results

The results of the summarization task were evaluated using both quantitative metrics and qualitative comparisons of the generated summaries against the reference summaries. This section details the performance of the BART model and highlights observations from sample outputs.

4.1 Quantitative Results

The performance of the BART model was measured using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics. The scores reflect the overlap between the generated and reference summaries at different granularities:

ROUGE-1: 0.4025 – Indicates a moderate overlap of individual words.

ROUGE-2: 0.2022 – Reflects a lower overlap of bigrams, indicating room for improvement in capturing contextual phrases.

ROUGE-L: 0.3155 – Highlights the model's ability to preserve fluency and structure in generated summaries.

ROUGE-Lsum: 0.3501 – Represents structural similarity between the generated and reference summaries.

These scores demonstrate that the pre-trained BART model effectively captures the essence of the articles but could improve in retaining finer details and multi-word expressions.

4.2 Qualitative Results

Sample Analysis: The table below compares the generated summaries with the reference summaries for three selected articles:

Article	Reference Summary	Generated Summary
Article 1: Covers the Palestinian Authority's accession to the ICC.	Focuses on ICC jurisdiction and opposition from Israel and the U.S., including specific temporal details.	Highlights ICC jurisdiction and opposition but lacks details like "alleged crimes committed since last June."
Article 2: Describes Theia's survival story and Sara Mellado's efforts to find her a home.	Emphasizes the emotional elements and Mellado's role in fostering Theia.	Captures Theia's injuries and recovery but omits Mellado's involvement and emotional context.
Article 3: Explores Mohammad Javad Zarif's career, including a consulate takeover and his interactions with John Kerry.	Highlights Zarif's interaction with Kerry, consulate takeover, and tweeting in English.	Focuses on Zarif's role in nuclear discussions but omits anecdotes like the consulate takeover or tweeting habit.

4.3 Observations:

Strengths:

- The generated summaries are concise, coherent, and maintain the central themes of the articles.
- The summaries perform well in aligning with key concepts from the articles.

Weaknesses:

- The summaries often omit specific details, such as timelines and emotional nuances.
- ROUGE-2 and ROUGE-L scores suggest limitations in capturing contextual relationships and maintaining sequential fluency.

5 Limitations

Despite the promising results, this project has several limitations that should be considered for future improvements:

1. Dependency on English-language Datasets:
    - The CNN/Daily Mail dataset is limited to English, which restricts the model's applicability to non-English text summarization tasks. Adapting the model for multilingual datasets could enhance its utility globally.
2. Domain-specific Summaries:
    - The dataset consists of news articles, making the model less effective in generating summaries for domain-specific content such as medical, legal, or technical documents. Fine-tuning on specialized datasets may help address this issue.
3. Resource Limitations:
    - The BART model is computationally expensive to train and fine-tune, requiring significant hardware resources such as GPUs or TPUs. These requirements may limit accessibility for researchers and developers with constrained computational resources.
4. Omission of Details:
    - While the model captures the overall meaning effectively, it occasionally omits nuanced details, as evidenced by qualitative analyses of the summaries.

Conclusion

This project evaluated the pre-trained BART model for abstractive summarization using the CNN/Daily Mail dataset. The results demonstrate that BART is effective in generating concise and coherent summaries, with moderate overlap in words and phrases as reflected by ROUGE scores (ROUGE-1: 0.4025, ROUGE-2: 0.2022, ROUGE-L: 0.3155).

The study highlighted the model's strengths, including its ability to preserve central themes, and its weaknesses, such as occasional omission of critical details. The analysis suggests that fine-tuning and parameter adjustments could further enhance performance.

In conclusion, this project illustrates the feasibility of using pre-trained transformer models for real-world text summarization applications, while also identifying areas for future exploration, including multilingual capabilities, domain adaptation, and hybrid summarization approaches.

Ethics Statement

This project adheres to ethical guidelines and ensures no biases were introduced in the data or during the processing steps. The dataset used, CNN/Daily Mail, was carefully chosen for its publicly available nature and wide acceptance in summarization research. The model was used strictly for research and educational purposes, with no intent to generate misleading or harmful summaries. Future implementations will continue to prioritize fairness and transparency in data selection and model usage.

References

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. *Advances in Neural Information Processing Systems*, 28, 1693-1701.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7871-7880.

Nallapati, R., Zhai, F., & Zhou, B. (2016). SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. *Proceedings of*

the 31st AAAI Conference on Artificial Intelligence, 3075-3081.

Priyanka. (2020). ROUGE Your NLP Results. *Medium*. Retrieved from https://medium.com/@priyankads/rouge-your-nlp-results-b2feba61053a

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the Point: Summarization with Pointer-Generator Networks. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1073-1083.

Zhang, Y., Liu, J., Wang, Y., Sun, Y., & Zhou, J. (2023). Enhancing Summarization with Latent Concepts. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 1537–1549. Retrieved from https://aclanthology.org/2023.acl-long.107.pdf