Text Summarization with Extractive and Transformer Models

Sriram Theerdh Manikyala
manikyala@arizona.edu

GitHub Repository

Abstract

The goal of this research is to summarize text using both cutting-edge transformer models and conventional NLP techniques. The main objective is to provide succinct summaries from lengthy research articles while maintaining important context and information. Sentence grading based on token frequency follows data pretreatment, which includes text normalization, tokenization, and stopword removal. After that, we use a transformer-based model (BART) for ab- stractive summarization and evaluate its effec- tiveness against conventional extractive meth- ods. The findings show that transformer models provide more natural and fluid summaries, even while extractive techniques are good at keeping important features intact. Along with highlight- ing the advantages and disadvantages of each strategy, this paper offers a useful framework for automatic text summarizing.

1 Introduction

In Natural Language Processing (NLP), text summarization is a crucial activity that aims to reduce lengthy texts while keeping important information. Both extractive and abstractive summarizing techniques are investigated in this study. Tokenization, stopword elimination, and sentence grading based on token frequency are the first preprocessing techniques we use on the incoming text. The BART transformer model is then used for abstractive summarization. We assess how well these methods perform in producing succinct and insightful sum- maries of research articles by fusing classic and contemporary methodologies.

2 Related Work

Over time, automatic text summarization has undergone substantial development. Initially, the emphasis was on extractive summarization, in which individual sentences or segments were chosen from the original text using statistical metrics such as Latent Semantic Analysis (LSA) and TF-IDF (Term Frequency-Inverse Document Frequency). Although these methods were successful in finding pertinent content, they frequently resulted in summaries that were inconsistent and unnatural since they just extracted individual sentences without considering the larger context.

Abstractive summarization techniques became more popular as deep learning gained traction. Models were able to produce summaries by rephrasing the original text thanks to techniques utilizing Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which produced more coherent outputs. This has been extended by transformer models like BERT, GPT-3, and BART. BERT's bidirectional attention mechanism allows it to do exceptionally well in extractive sum- marization, while GPT-3 and BART achieve impres- sive results in abstractive summarization by pro- ducing text that flows and sounds human. Hybrid approaches that combine extractive and abstractive techniques to improve summarization quality are being researched, although problems such as pro- ducing irrelevant or factually incorrect summaries continue to exist.

3 Methodology

The processes involved in text summarization are described in detail in this part, including sentence scoring, model selection, data preprocessing, and the application of transformer models. The approach creates excellent summaries by fusing cutting-edge neural network models with conventional NLP techniques.

3.1 Dataset

The dataset is a compilation of research papers that were given from 1987 to 2016 at the Neural Information Processing Systems (NIPS) conference. It

provides a thorough historical account of the development of machine learning research.

3.2 Data Preprocessing

Cleaning and standardizing the text in preparation for summary is part of data preparation:

Text Cleaning: HTML tags, numbers, URLs, and parentheses are eliminated, and the text is changed to lowercase. Tokenization: NLTK's tokenization is used to divide the cleaned text into distinct words. Stopword Removal: To concentrate on important keywords, common stopwords like "and," "the," and "of" are eliminated using the NLTK stopwords list. A cleaned document and a set of tokens prepared for examination are the end products of this operation.

3.2.1 Sentence Segmentation

Sentence segmentation comes next after preprocessing. SpaCy's sentence segmentation feature divides the processed document into separate sentences. This makes it possible for the summarization algorithm to assess each sentence's applicability to the final summary on its own.

3.2.2 Sentence scoring using token frequencies

We rate each phrase according to the frequency of words it contains in order to assess its value. This is how this procedure operates:

Word Frequency Analysis: Over the course of the document, the frequency of each token—noun, verb, adjective, or interjection—is determined. Sentence Scoring: The total frequency of all the words in a sentence is added up to determine its score. Higher word frequency sentences are given a higher score since they are seen more significant. This stage produces a dictionary of sentences with scores associated with them; sentences with higher scores are considered more significant.

3.2.3 Extractive summary generation

The extractive summary is formed by selecting the top N highest-scoring sentences based on the sentence scores obtained in the preceding stage. Selecting the most pertinent information straight out of the document is the main goal of this approach.

3.3 Transformer based summarisation

An abstractive summary is produced by applying a transformer-based model in addition to the extractive approach:

Model Selection: For this work, the Hugging Face BART transformer model is utilized. Large text datasets are used to pre-train the model, which is then optimized to handle text summarization. Contribution to the Model: The BART model receives the tokenized and cleaned text, processes it, and produces a summary using its prior knowledge. Compared to the extractive method, the output of the transformer model is a more natural and fluid abstractive summary that may contain rephrased content.

4 Evaluation Metrics

The ROUGE measure, which analyzes the overlap of n-grams (bigrams, unigrams, etc.) between the generated summary and the reference summary, is used to assess the quality of the created summaries. Greater ROUGE scores show that more of the significance and important details of the original material are preserved in the generated summary.

5 Results and Discussion

Although it lacked fluency, the extractive approach successfully stored important information. Although it occasionally included extraneous features, the BART transformer model produced summaries that were more fluid. While the extractive method main- tained factual accuracy, BART was overall more coher- ent, indicating that a hybrid approach could com- bine the best features of both.

References

[1] Vaswani, A., et al. (2017). *Attention Is All You Need*. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 2017. https://arxiv.org/abs/1706.03762

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of NAACL-HLT 2019*. https://arxiv.org/abs/1810.04805

[3] Radford, A., et al. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[4] Raffel, C., Shinn, K., Lee, S., & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. https://arxiv.org/abs/1910. 10683