Integrating Extractive and Abstractive Summarization: A Hybrid Approach

P. Yeshwanth Chowdary, K.Vishruth Solomon Kumar, B. Shashi Kiran and S. Aswani
Department of CSE (AI&ML), Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India

Keywords: Text Summarization, Extractive Summarization, Abstractive Summarization, KL Divergence, BART, Key Sentences, Hybrid Summarization.

Abstract: This project presents a comprehensive methodology for text summarization that integrates both extractive and abstractive techniques to enhance the quality of generated summaries. In the extractive summarization phase, KL divergence is utilized to identify and select key sentences or phrases from the source text, effectively capturing the most relevant information. These selected segments are then passed as input to an abstractive summarization model, specifically BART (Bidirectional and Auto-Regressive Transformers), which processes and refines the extracted information to produce a coherent and fluent summary. By combining the precision of extractive summarization with the fluency and coherence of abstractive approaches, the proposed methodology aims to generate high-quality summaries that offer improved coverage of the source text, enhanced fluency, and a significant reduction in redundancy.

1 INTRODUCTION

In today's data-driven world, the volume of text generated across various industries, such as journalism, research, and business, is overwhelming. The challenge lies in efficiently extracting relevant information from these large text corpora while preserving the core meaning and essential details. Text summarization is a vital tool for addressing this issue by condensing extensive documents into shorter, more manageable summaries. However, traditional methods, such as frequency-based techniques, often fail to capture the deeper semantic meaning and context of the text, resulting in suboptimal summaries that may lack coherence or relevance.

Our project aims to develop a more sophisticated text summarization system that leverages advanced techniques like KL-Divergence for extractive summarization, identifying key sentences based on their divergence from the overall text distribution. Additionally, we use BART, a state-of-the-art transformer model, for abstractive summarization, which generates fluent, coherent summaries by rephrasing or creating new sentences based on the original content. This hybrid approach ensures that the final summaries are both informative and concise, making them useful across a wide range of industries and domains that require efficient text processing and analysis

2 LITERATURE SURVEY

Numerous studies have explored text summarization across various domains, applying both traditional and modern techniques. Machine learning offers various potential methods, but its effectiveness hinges on selecting an appropriate algorithm tailored to the specific domain.

(B Rajesh, K Nimai Chaitanya, P Tejesh Govardhan, K Krishna Mahesh, & B Sudarshan, 2024) Proposed a systematic approach to extractive text summarization, focusing on text preprocessing, sentence scoring, selection, and post-processing. Utilizing Python libraries like SpaCy and NLTK, they score sentences based on word frequencies and merge similar sentences for coherent summaries. Challenges include potential redundancy in extracted sentences and limitations in handling diverse text formats.

(J.N. Madhuri & R. Ganesh Kumar, 2019) This work presents a statistical method for single-document extractive summarization by ranking sentences based on assigned weights. Their approach aims to condense text into concise summaries,

Chowdary, P. Y., Kumar, K. V. S., Kiran, B. S. and Aswani, S.
Integrating Extractive and Abstractive Summarization: A Hybrid Approach.
DOI: 10.5220/0013602400004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 784-788
ISBN: 978-989-758-763-4

evaluated through weights prioritizing important sentences. They identify challenges such as context loss when extracting sentences in isolation and difficulties in accurate weight assignment.

(Asha Rani Mishra, V.K Panchal, & Pawan Kumar, 2019) Address the challenge of extracting insights from large textual datasets using topic modelling and key phrase extraction. Their multifaceted approach employs techniques like LSI and TF-IDF to identify key topics and phrases, while summarization methods like LSA are applied for concise outputs. Challenges include combining techniques effectively and handling diverse text structures.

(Sanchit Agarwal, Nikhil Kumar Singh, & Priyanka Meel, 2018) Introduced a method for extractive summarization that combines K-Means clustering with sentence embeddings. By clustering sentences based on semantic similarity, they select the most relevant sentences for summaries, evaluated using ROUGE scores on the DUC 2001 dataset. Challenges include sensitivity to cluster numbers and the need for high-quality embeddings for accurate results.

(Wen Xiao & Giuseppe Carenini, 2019) Presented a novel extractive summarization model that combines global and local context to identify key information from long documents. Their methodology evaluates content using both contexts, achieving superior ROUGE-1 and ROUGE-2 scores on Pubmed and arXiv datasets compared to traditional models. However, the authors note challenges in effectively integrating these contexts, the risk of redundancy in summaries, and the need for robust evaluations across diverse document types.

(L. Lebanoff, K. Song, & F. Liu, 2018) The authors address the challenge of generating text abstracts from multiple documents, utilizing a neural encoder-decoder framework traditionally designed for single-document summarization. Their approach incorporates the Maximal Marginal Relevance (MMR) method to select representative sentences from various documents, subsequently fusing these sentences into an abstractive summary. This method does not require additional training data, demonstrating its robustness in Multi document contexts.

(Glorian Yapinus, Alva Erwin, Maulhikmah Galinium, & Wahyu Muliady, 2014) This study introduces a hybrid approach to multi-document summarization specifically designed for Indonesian documents. The authors aim to effectively condense information from multiple sources into a coherent summary by combining WordNet-based text summarization (abstractive) with title word-based summarization (extractive). This method is evaluated against Latent Semantic Analysis (LSA), highlighting its ability to generate well compressed and readable summaries.

(A. Ghadimi & H. Beigy, 2022) This research presents HMSumm, a hybrid approach to multi document summarization which integrates extractive and abstractive techniques by utilizing pre-trained language models. The methodology involves generating an extractive summary by selecting key sentences from the documents while employing a determinantal point process (DPP) to minimize redundancy. Subsequently, the extractive summary is passed to BART and T5 models for abstractive summarization, with the final output chosen based on sentence diversity. The study highlights the effectiveness of combining multiple models to enhance summarization quality.

(Christian, H. Agus, M.P., & Suhartono, 2016) This study investigates the application of the TF-IDF algorithm for single-document automatic text summarization, aiming to enhance information retrieval amidst the abundance of online content. The methodology involves ranking sentences based on the frequency of important terms while minimizing the impact of common words. The performance of the TF-IDF summarizer is assessed against other summarization tools using the F-Measure for comparison. While effective, the paper notes limitations such as the algorithm's reliance on term frequency, which may overlook critical contextual information.

(G. Di Fabbrizio, A. Stent, & R. Gaizauskas, 2014) This study presents STARLET-H, a hybrid summarization system designed for synthesizing reviews of products and services. The methodology involves a dual approach, utilizing extractive methods to select key quotes from reviews, which are then blended into an abstractive summary. However, the paper highlights challenges such as potential inconsistencies when merging extracted quotes with generated text, which can disrupt the narrative flow and affect overall coherence."

3 DESIGN AND PRINCIPLE OF MODEL

3.1 Methodology

In this study, we developed a hybrid document and text summarization system that integrates both

extractive and abstractive techniques using the BART pre-trained model. The primary objective was to generate comprehensive, informative, and coherent summaries that effectively convey the main ideas of the source documents.

3.1.1 Pre-trained Model Selection and Pre-Processing:

For abstractive summarization, we employed the BART model. The model's architecture allows it to generate fluent and contextually rich summaries based on the input text.

The preprocessing step cleans and normalizes the input text to improve summarization. It removes emojis, emails, URLs, phone numbers, and HTML tags while normalizing hyphenated words, extra spaces, Unicode characters, quotation marks, and bullet points. These steps ensure the text is clean and ready for summarization

3.1.2 Extractive Summarization

The extractive summarization process begins with the implementation of the KL Divergence algorithm to identify key sentences from the source text. The algorithm calculates the divergence between the probability distributions of words across the entire text and the candidate summary sentences. Sentences with the lowest KL Divergence scores are selected, ensuring that the extractive summary retains the most relevant and informative parts of the text.

The extractive summarization function processes the input text by first splitting it into sentences. The algorithm then computes the importance of each sentence by analysing word frequencies, ultimately selecting the top sentences that best represent the original content.

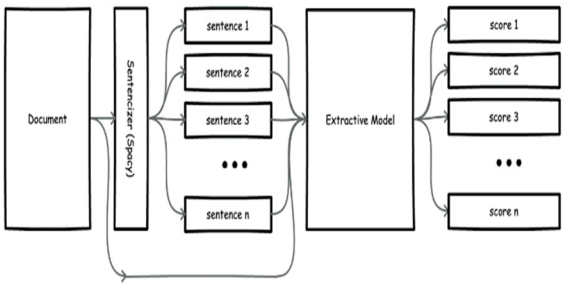

Figure 1: Extractive Summarization Approach.

3.1.3 Abstractive Summarization

The key sentences identified in the extractive phase serve as input for the BART model, which generates an abstractive summary. This step enhances the relevance and coherence of the final output by allowing the model to focus on the most critical information from the extracted sentences.

The abstractive summarization function tokenizes the input text and generates a summary using the BART model's capabilities. The model's ability to understand and rephrase content ensures that the summaries produced are succinct while retaining the original meaning.

3.1.4 Hybrid Summarization

The hybrid summarization methodology leverages the extracted key information to guide the generation of the abstractive summary. By performing summarization on the key sentences obtained from the KL Divergence algorithm, the system combines the strengths of both extractive and abstractive approaches. This results in high-quality summaries that are coherent and rich in content, effectively addressing the limitations of purely extractive or purely abstractive methods.

Overall, the proposed system exemplifies a robust approach to document summarization by effectively integrating a variety of advanced techniques. This integration allows the system to produce summaries that are not only informative and concise but also contextually relevant to the source material.

4 RESULTS

The performance of the summarization model is evaluated using the ROUGE metric suite. The summarization model was evaluated using a dataset of news articles to measure its performance. The model's performance was assessed on a representative subset of a news article dataset, which provided an initial indication of the model's capability. The results are summarized through the average ROUGE scores as follows:

ROUGE-1 Score (Unigram Overlap): The ROUGE-1 score evaluates the overlap of unigrams. A recall of 0.6466 indicates that 64.66% of the unigrams in the reference summary were captured in the generated summary. The precision of 0.5636 suggests that 56.36% of the unigrams in the generated summary are relevant to the reference summary. The F1-score of 0.5783, which balances recall and precision, highlights a reasonable level of summarization accuracy.

ROUGE-2 Score (Bigram Overlap): The ROUGE-2 score measures the overlap of bigrams. A recall of 0.5359 indicates that 53.59% of the bigrams

in the reference summary were matched in the generated summary. The precision of 0.4316 shows that 43.16% of the bigrams in the generated summary are relevant to the reference summary. The F1-score of 0.4521 reflects a balanced measure of bigram overlap.

ROUGE-L Score (Longest Common Subsequence): The ROUGE-L score assesses the longest common subsequence between the generated and reference summaries. A recall of 0.6247 suggests that 62.47% of the longest common subsequence in the reference summary was captured in the generated summary. The precision of 0.5438 indicates that 54.38% of the longest common subsequence in the generated summary is relevant. The F1-score of 0.5585 provides a balanced evaluation of recall and precision for LCS overlap.

Table 1: ROUGE Scores.

Model	Recall	Precision	F-measure
Rouge-1	0.6466	0.5636	0.5783
Rouge-2	0.5359	0.4316	0.4521
Rouge-L	0.6247	0.5438	0.5585

Table 1 presents the ROUGE scores, highlighting the model's accuracy in generating coherent and informative summaries.

The performance of the summarization model is further evaluated using BERT Score, a metric that leverages pre-trained language models to assess the semantic similarity between the generated and reference summaries. BERT Score computes precision, recall, and F1 scores based on contextualized embeddings, providing a more nuanced evaluation compared to traditional overlap-based metrics.

The summarization model's performance was assessed on a representative subset of a dataset containing summaries, providing a robust indication of its effectiveness in generating semantically relevant summaries. The results, as summarized by the average BERT Score metrics, are as follows:

BERT Precision: The precision score of 0.8791 indicates that 87.91% of the words in the generated summary were relevant, capturing the key information from the reference summary.

BERT Recall: The recall score of 0.8887 shows that 88.87% of the key words in the reference summary were captured in the generated summary. This high recall demonstrates the model's ability to capture a large proportion of the essential information from the reference summary, ensuring completeness.

BERT F1-Score: The F1-score of 0.8836 balances both precision and recall, indicating that the generated summaries are both highly relevant and comprehensive.

Table 2: BERT Score.

Metric	Score
BERT Precision	0.8791
BERT Recall	0.8887
BERT F1-Score	0.8836

Table 2 presents the BERT Score metrics, highlighting the model's ability to generate semantically relevant and accurate summaries based on contextualized embeddings.

In addition to the ROUGE, BERT Score evaluations, the text summarization application developed using Streamlit provides users with an intuitive interface to generate summaries from various text inputs. The application consists of a user-friendly layout where users can upload documents or paste text directly into a designated input area. Upon submission, the application processes the input and displays both the summarized and original texts, alongside their respective word counts

Key Features and Functionalities.

Input Interface: Users can easily upload documents in PDF or Word format or input text directly.

Text Summarization: The application utilizes a robust BART model for abstractive summarization and a KL Divergence approach for extractive summarization.

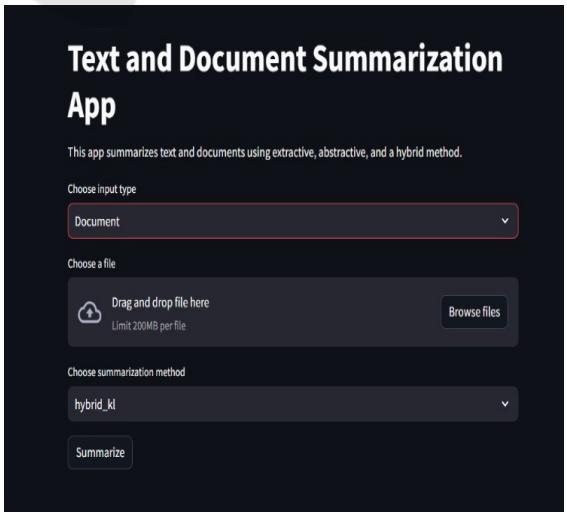

Figure 2: Web App for Text and Document Summarization.

5 FUTURE WORK

In future work, several avenues can be explored to enhance the capabilities of the text summarization system. First, efforts will focus on improving model performance by integrating advanced speech-to-text (STT) technologies, allowing the system to generate accurate transcripts from video audio, thereby broadening the range of input sources.

Another significant direction involves extending the hybrid summarization approach to handle multiple documents simultaneously. This enhancement would enable more comprehensive content synthesis by integrating information from various sources and identifying common themes or patterns. It would also facilitate cross-document analysis, enabling users to draw richer and more insightful conclusions from diverse datasets, thus broadening the scope and applicability of the summarization system.

Additionally, deploying the summarization pipeline in real-time applications represents a promising opportunity. By adapting the system for platforms like news aggregators or chatbot interfaces, users could receive timely and relevant information summaries, improving the overall user experience.

Lastly, addressing multilingual summarization is crucial for expanding the system's reach. By leveraging transformer models adept at handling diverse languages, the methodology could support a wider audience and cater to the global demand for effective text summarization.

By pursuing these future directions, the project aims to significantly advance the effectiveness and applicability of text summarization technologies.

6 CONCLUSION

This research project focused on hybrid text summarization using KL Divergence and BART, demonstrating significant potential in generating concise and informative summaries from textual data. By integrating both extractive and abstractive techniques, the project effectively leveraged the strengths of KL Divergence for content relevance in sentence selection and BART for producing fluent and coherent summaries.

The findings highlight the efficacy of this hybrid approach, showcasing its ability to create more effective and contextually aware summarization solutions. As natural language processing technology evolves, the integration of diverse methods becomes increasingly essential for addressing the growing demand for efficient information extraction and synthesis across various applications.

The project lays a solid foundation for further advancements in the field of text summarization. The combination of statistical methods and deep learning techniques presents a robust framework for developing innovative solutions. Future enhancements could involve refining the model through feature engineering, real-time data integration, and exploring additional transformer architectures, ultimately contributing to the ongoing evolution of text summarization methodologies and their applications in a wide range of domains.

REFERENCES

A. Ghadimi, & H. Beigy. (2022). Hybrid Multi Document Summarization Using Pre-Trained Language Models. *Expert Systems with Applications*.

Asha Rani Mishra, V.K Panchal, & Pawan Kumar. (2019). Extractive Text Summarization- An effective approach to extract infromation from text. *International Conference on Contemporary Computing and Informatics (IC3I).*

B Rajesh, K Nimai Chaitanya, P Tejesh Govardhan, K Krishna Mahesh, & B Sudarshan. (2024). Text Summarization Using NLP. *International Research Journal of Engineering and Technology(IRJET).*

Christian, H. Agus, M.P., & Suhartono. (2016). Single Document Automatic Text Summarization Using Term Frequency Inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, 285-294.

G. Di Fabbrizio, A. Stent, & R. Gaizauskas. (2014). A Hybrid Approach to Multi-Document Summarization of Opinions in Reviews. *INLG*, 54-63.

Glorian Yapinus, Alva Erwin, Maulhikmah Galinium, & Wahyu Muliady. (2014). Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive Extractive Sumarization Technique. *International Conference on Information Technology and Electrical Engineering (ICITEE).*

J.N. Madhuri, & R. Ganesh Kumar. (2019). Extractive Text Summarization Using Sentence Ranking. *International Conference on Data Science and Communication (IconDSC).*

L. Lebanoff, K. Song, & F. Liu. (2018). Adapting the Neural Encoder Decoder Framework from Single to Multi Document Summarization. *ACLEMNLP.*

Sanchit Agarwal, Nikhil Kumar Singh, & Priyanka Meel. (2018). Single Document Summarization Using Sentence Embeddings and K-Means Clustering. *International Conference on Advances in Computing, Communication Contral and Networking (ICACCCN).*

Wen Xiao, & Giuseppe Carenini. (2019). Extractive Summarization of Long Documents by Combining Global and Local Context. *EMNLPIJCNLP.*