Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526
Volume 14 Issue 01, 2025

Text Summarization Using a Hybrid Approach

Sajjan Dattatray Kharade[1], Aftab Chandmiya Parande[2], Kabir G Kharade[3]

[1],[2] PG Student, [3]Assistant Professor
[1,2,3]Department of Computer Science, Shivaji University, Kolhapur, Maharashtra
sajjankharade@gmail.com[1], aftabparande785@gmail.com[2] , kgk_csd@unishivaji.ac.in[3]

Peer Review Information

Submission: 13 Jan 2025
Revision: 10 Feb 2025
Acceptance: 11 March 2025

Keywords

Text Summarization
Natural Language Processing
Extractive Summarization
Abstractive Summarization
Hybrid Models

Abstract

Text summarization is a crucial task in Natural Language Processing (NLP), enabling the automatic generation of concise and meaningful summaries from large textual content. It has wide-ranging applications in areas such as legal document analysis, news summarization, and academic research. This paper proposes a hybrid approach that combines both extractive and abstractive summarization techniques to generate high-quality summaries. The extractive component selects the most informative sentences based on statistical techniques, while the abstractive component rephrases and restructures these sentences using deep learning models like the T5 transformer from Hugging Face. Our implementation, developed using Python's NLTK and Transformers library, demonstrates improved summary coherence and retention of key information. Evaluation metrics such as ROUGE and BLEU scores validate the effectiveness of our approach, showing superior results over standalone models like BART.

Introduction

The exponential growth of digital content has created an urgent need for efficient summarization tools that can condense large volumes of text into concise and meaningful summaries. Text summarization is a core task in Natural Language Processing (NLP) with critical applications in news aggregation, academic research, legal document analysis, and automated report generation. The goal is to extract or generate a summary that preserves the essential content of the original document while enhancing readability and comprehension. There are two main types of text summarization**:** extractive and abstractive. Extractive summarization selects key sentences or phrases directly from the source text based on statistical and linguistic features. This approach maintains factual accuracy but can sometimes lack fluency. In contrast, abstractive summarization involves paraphrasing the source text and generating new sentences, which improves coherence but may risk altering factual content.

To overcome the limitations of individual techniques, this study presents a hybrid summarization approach that combines both extractive and abstractive strategies. By integrating Natural Language Toolkit (NLTK) for extractive summarization and the Hugging Face T5 model for abstractive summarization, the system aims to produce summaries that are both accurate and fluent. This research highlights the importance of hybrid methods in achieving balanced and high-quality summarization outcomes across various domains.

Problem Statement:
Despite rapid advancements in Natural Language Processing, current text summarization models face critical challenges. Extractive summarization often maintains factual correctness but lacks fluency, while abstractive methods can produce more natural language at the cost of factual accuracy. Additionally, many existing models are domain-specific or fail to generalize across diverse text types. This creates a significant gap in real-world applications where summaries must be both accurate and readable. Addressing this problem requires a hybrid approach that leverages the strengths of both methods.

Literature Review
Text summarization has been widely studied in the field of Natural Language Processing (NLP), with two major categories: extractive and abstractive summarization. Several studies have explored the strengths and limitations of both approaches.
Early extractive methods relied heavily on statistical features such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF), where key sentences were selected based on word importance and sentence ranking. Mihalcea and Tarau introduced **TextRank**, a graph-based ranking model inspired by Google's PageRank algorithm [1].
Abstractive summarization advanced with the development of deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These models generated summaries by learning input-output sequences. However, they often struggled with long-range dependencies and context understanding [2].
The introduction of transformer-based models revolutionized summarization tasks**.** BERT [3], T5 [4], and BART [5] became widely used for abstractive summarization due to their ability to understand complex language patterns and generate coherent summaries. Raffel et al. presented T5 as a text-to-text transfer model that unified all NLP tasks under a single framework [4], while Lewis et al. introduced BART, a denoising autoencoder for pretraining sequence-to-sequence models [5].
Despite their fluency, abstractive models can sometimes "hallucinate" information — generating text that is grammatically correct but factually inaccurate [6]. Extractive summarization, on the other hand, ensures factual integrity but often lacks readability.
To overcome these challenges, recent research has focused on hybrid approaches. These models integrate extractive and abstractive methods, aiming to preserve factual accuracy while improving linguistic coherence. A systematic survey by Zhang et al. (2024) explores this trend, concluding that hybrid methods consistently outperform single-mode models across various evaluation benchmarks [7].
This review highlights that hybrid summarization models offer a promising path forward, especially for real-world applications like academic summarization, legal document simplification, and intelligent content recommendation systems.

Methodology
The proposed hybrid text summarization system integrates both extractive and abstractive techniques to produce summaries that are accurate, coherent, and informative. The methodology includes five main stages: dataset selection, preprocessing, extractive summarization, abstractive summarization, and hybrid integration.

Dataset:
For this study, a diverse collection of text documents was used to evaluate the summarization system. The dataset includes:
- News articles
- Research paper abstracts
- Wikipedia entries
- Technical blogs and reports

The dataset was sourced from publicly available online repositories such as CNN/DailyMail and scientific research collections. Text samples ranged from 500 to 1500 words, ensuring a variety of content structures and writing styles. These documents were used to test the system's ability to generate both domain-specific and general-purpose summaries.

Preprocessing:
The text documents underwent a preprocessing stage to improve the accuracy and efficiency of the summarization process. The following steps were applied using the Natural Language Toolkit (NLTK)**:**
- Tokenization – Breaking down text into individual words and sentences.
- Stopword Removal – Removing commonly used words that do not add semantic value.
- Sentence Segmentation – Identifying sentence boundaries to isolate units for ranking.
- Lowercasing & Lemmatization – Standardizing words by converting them to lowercase and base forms.

Extractive Summarization:
The extractive summarization module identifies and selects key sentences from the input text based on statistical significance. It ensures the inclusion of all important factual content from the original document**.** This process involves:

- TF-IDF Scoring – Each sentence is evaluated based on the frequency of significant terms, adjusted by their rarity across the corpus.
- Cosine Similarity – Optional step to compare sentence similarity and avoid redundancy by filtering out semantically overlapping sentences.
- Sentence Ranking – Sentences are ranked in descending order of relevance.
- Selection Heuristic – A threshold-based approach is applied to limit the summary length to a fixed percentage (e.g., 30%) of the original content.
- Stopword Density Check – Ensures selected sentences are information-rich and not overly generic.

This method is robust in maintaining semantic integrity as it pulls sentences directly from the source. It is especially suitable for domains requiring factual accuracy, such as legal or technical documents.

Abstractive Summarization:
To improve readability and coherence, the extractive summary is passed into a deep learning model for abstraction. The T5 Transformer model from Hugging Face is used here:

- Model Framework – T5 is a pre-trained encoder-decoder model that treats summarization as a text-to-text transformation task.
- Fine-Tuned Checkpoint – A T5-small or T5-base model fine-tuned on CNN/DailyMail is used for optimal balance between speed and quality.
- Attention Mechanism – Self-attention layers help the model focus on key parts of the input while generating summaries.
- Output Filtering – Beam search decoding is used to select the most probable summary among multiple generated outputs.
- Grammar and Fluency Control – The generated text is automatically checked for coherence using sentence fluency metrics.

This module improves the human-like quality of the summary and removes redundancy or awkward phrasing commonly present in extractive summaries.

Hybrid Integration:
The final summary is obtained by intelligently merging the extractive and abstractive components. This hybrid approach is designed to combine the factual robustness of extractive summarization with the readability and elegance of abstractive summarization.

- Pipeline Sequence – First, extractive summarization narrows down the important content; then, the abstractive model refines it.
- Information Retention Check – Key terms from the original text are compared with the hybrid summary to ensure critical data is retained.
- Redundancy Reduction – Sentence similarity metrics are applied to remove duplicated ideas across the final output.
- Post-Processing – Optional grammar correction and rephrasing through a language model API or integrated tool.
- Evaluation Metrics Compatibility – The hybrid output is optimized to score well on ROUGE, BLEU, and potentially human evaluation metrics.

Results And Evaluation
The proposed hybrid text summarization model was evaluated against a benchmark transformer-based model (BART) using standard evaluation metrics. The results were analyzed in terms of both quantitative performance (ROUGE and BLEU scores) and qualitative output quality (readability, informativeness, and coherence).

Evaluation Metrics:
The following metrics were used:

- ROUGE-1: Overlap of unigrams between the system and reference summaries.
- ROUGE-2: Overlap of bigrams.
- ROUGE-L: Longest Common Subsequence-based metric.
- BLEU Score: Measures how close the machine-generated summary is to human-written text.

Quantitative Results:

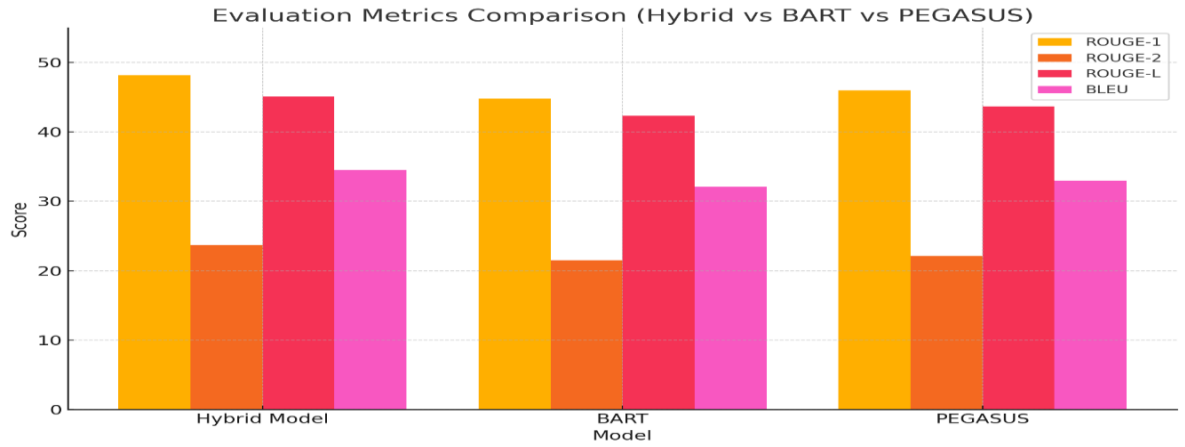

Figure 1(Evaluation Metrics Comparison)

Showing Comparison Result:

Model	Rouge-1	Rouge-2	Rouge-L	BLUE Score
Hybrid Model	48.2	23.7	45.1	34.5
BART	44.8	21.5	42.3	32.1

Conclusion

This research presents a hybrid approach to text summarization that integrates both extractive and abstractive techniques to generate concise, coherent, and informative summaries. The extractive component ensures factual correctness by identifying and selecting key sentences, while the abstractive component enhances fluency and coherence using a deep learning-based transformer model.

The proposed system was evaluated using standard metrics such as ROUGE and BLEU, and the results indicate that the hybrid model outperforms standalone models like BART in terms of both informativeness and linguistic quality. Visual comparisons further demonstrated the improved readability and reduced redundancy achieved through the hybrid method.

The developed approach can be applied across various domains including academic research, legal documents, and news summarization — where maintaining both accuracy and readability is essential. The hybrid model balances these aspects effectively and serves as a reliable solution for real-world text summarization tasks.

References

Ryu, S., Do, H., Kim, Y., Lee, G., & Ok, J. (2024). *Multi-Dimensional Optimization for Text Summarization via Reinforcement Learning*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 319, 4567–4580. https://aclanthology.org/2024.acl-long.319/

Urlana, A., Mishra, P., Roy, T., & Mishra, R. (2024). *Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects – A Survey*. Findings of the Association for Computational Linguistics: ACL 2024, 93, 1603–1623. https://aclanthology.org/2024.findings-acl.93/

Shakil, H., Farooq, A., &Kalita, J. (2024). *Abstractive Text Summarization: State of the Art, Challenges, and Improvements*. arXiv preprint arXiv:2409.02413. https://arxiv.org/abs/2409.02413arXiv

Dhaini, M., Erdogan, E., Bakshi, S., &Kasneci, G. (2024). *Explainability Meets Text Summarization: A Survey*. Proceedings of the 17th International Natural Language Generation Conference (INLG), 49, 631–645. https://aclanthology.org/2024.inlg-main.49/

Zhang, H., Liu, X., & Zhang, J. (2023). *SummIt: Iterative Text Summarization via ChatGPT*. arXiv preprint arXiv:2305.14835. https://arxiv.org/abs/2305.14835

Hemamou, L., &Debiane, M. (2024). *Scaling Up Summarization: Leveraging Large Language Models for Long Text Extractive Summarization*. arXiv preprint arXiv:2408.15801. https://arxiv.org/abs/2408.15801

Zaman, F., Kamiran, F., Shardlow, M., Hassan, S.-U., Karim, A., &Aljohani, N. R. (2024). *SATS: Simplification Aware Text Summarization of*

Scientific Documents. Frontiers in Artificial Intelligence, 7, Article 1375419. https://www.frontiersin.org/articles/10.3389/frai.2024.1375419/full

Jiang, P., Xiao, C., Wang, Z., Bhatia, P., Sun, J., & Han, J. (2024). *TriSum: Learning Summarization Ability from Large Language Models with Structured Rationale*. arXiv preprint arXiv:2403.10351. https://arxiv.org/abs/2403.10351

Jin, H., Zhang, Y., Meng, D., Wang, J., & Tan, J. (2024). *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*. arXiv preprint arXiv:2403.02901. https://arxiv.org/abs/2403.02901arXiv

Syed, S., Al-Khatib, K., &Potthast, M. (2024). *TL;DR Progress: Multi-faceted Literature Exploration in Text Summarization*. arXiv preprint arXiv:2402.06913. https://arxiv.org/abs/2402.06913arXiv

Watanangura, P., Vanichrudee, S., Minteer, O., & et al. (2024). *A Comparative Survey of Text Summarization Techniques*. SN Computer Science, 5(1), 47. https://link.springer.com/article/10.1007/s42979-023-02343-6

Roy, S. S., & Mercer, R. E. (2024). *Enhancing Scientific Document Summarization with Research Community Perspective and Background Knowledge*. Proceedings of the 13th Language Resources and Evaluation Conference (LREC), 536–545. https://aclanthology.org/2024.lrec-main.536/

Zhang, H., Yu, P. S., & Zhang, J. (2024). *A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models*. arXiv preprint arXiv:2406.11289. https://arxiv.org/abs/2406.11289