Smart Summarizer: A Comparative Analysis of Abstractive and Extractive Summarization

Hemamalini Venkatesh Vanetha[1]
Dr.Sumona Mondal[2]

[1] Department of Applied Data Science, Clarkson University
[2] Co-Director of the MS Program in Applied Mathematics, Clarkson University

Abstract :

This project explores two key approaches to automatic text summarization: extractive and abstractive methods. Using benchmark datasets (XSum and CNN/DailyMail), we apply BART-Large for abstractive summarization and Lead-3 as an extractive baseline. The goal is to evaluate their performance using ROUGE metrics for content overlap. Abstractive models produce fluent, human-like summaries, while extractive methods offer simpler yet effective results. Our analysis shows that model effectiveness varies with dataset characteristics. The findings help guide practical choices for real-world summarization tasks.

Index Terms: Natural Language Processing, BART-Large, Text Summarization, Extractive Methods, Abstractive Models, ROUGE Metrics, Hugging Face Transformers, XSum, CNN/DailyMail, Comparative Study

1 Introduction

In a world overwhelmed by digital content, the quick understanding of large amounts of information has become both a necessity and a challenge. Text summarization, a fundamental task in Natural Language Processing (NLP), offers an effective solution by condensing long documents into shorter and meaningful summaries. This not only improves information accessibility, but also improves productivity in domains such as journalism, academia, and corporate communications. There are two main strategies for summarization: extractive and abstractive. Extractive summarization selects the most relevant sentences directly from the source, whereas abstractive summarization rewrites content using new phrases and sentence structures, close to resembling human-written summaries. This project investigates both methods by applying them to two widely used datasets—XSum and CNN/DailyMail—using pretrained transformer models. By evaluating performance with ROUGE metrics, the goal is to understand the strengths and limitations of each technique and determine which is more suitable for various types of articles.

2 Background

Text summarization has been a long-standing goal in NLP, with early methods relying heavily on rule-based or statistical techniques. These traditional approaches were often extractive, identifying the most important sentences using cues such as sentence position or frequency of keywords. With the advent of deep learning and large-scale datasets, abstractive methods have gained momentum. These models, particularly transformer-based architectures like BART, can generate summaries that are semantically rich and linguistically fluent. However, they require more computational resources and careful tuning compared to simpler extractive baselines.

The Lead-3 extractive method is a popular baseline that selects the first three sentences of an article. While naive, it performs surprisingly well on structured news content. On the other hand, BART (Bidirectional and Auto-Regressive Transformers) has been fine-tuned on summarization tasks and has shown strong results, especially on datasets like XSum and CNN/DailyMail. This project aims to systematically compare these two summarization styles, analyzing their effectiveness in generating coherent, concise, and accurate summaries using well-defined metrics.

3 Objective

The primary objective of this project is to perform a comparative analysis of extractive and abstractive summarization techniques on benchmark datasets. By applying both strategies to the same inputs, we aim to understand their practical effectiveness, accuracy, and use-case suitability. The specific objectives include: • To implement an extractive summarization method (Lead-3) as a baseline.

• To apply transformer-based abstractive summarization using BART-Large models.

• To evaluate and compare both methods on two benchmark datasets: XSum and CNN/DailyMail.

• To measure the summarization quality using standard ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-Lsum).

• To analyze the differences in performance and behavior of both methods across datasets with varying summary styles.

• To identify the strengths, weaknesses, and appropriate contexts for each summarization approach.

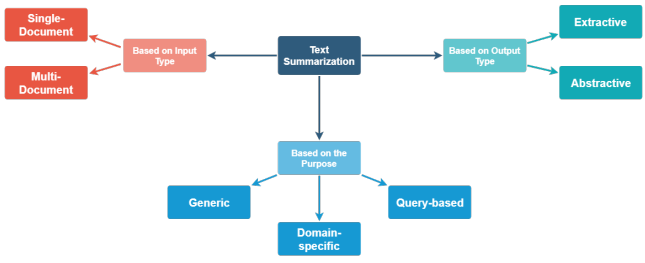

Figure 1. Text Summarization

4 Dataset Description

This project utilizes two benchmark datasets commonly used in the evaluation of text summarization systems: XSum and

CNN/DailyMail. Each dataset presents unique challenges in terms of summary length, structure, and abstraction level.

4.1 XSum Dataset

The Extreme Summarization (XSum) dataset contains BBC news articles paired with single-sentence summaries. It is known for its highly abstractive nature, as the reference summaries are not mere extractions but human-written rephrasings. This makes the dataset ideal for evaluating the capabilities of generative models like BART.

• Source: BBC News articles • Number of articles: 226,000 • Summary format: One-sentence human-written summary per article

4.2 CNN/DailyMail Dataset

The CNN/DailyMail dataset comprises news articles and bullet-point highlights, primarily used to test both extractive and abstractive summarization systems. Unlike XSum, the summaries here are longer and consist of multiple key points, making it suitable for extractive techniques like Lead-3 as well as abstractive models.

• Source: CNN and Daily Mail news websites • Number of articles: 312,000 combined • Summary format: Multiple bullet-point highlights per article

5 Methodology

Text summarization can be broadly classified into three categories based on the underlying technique:

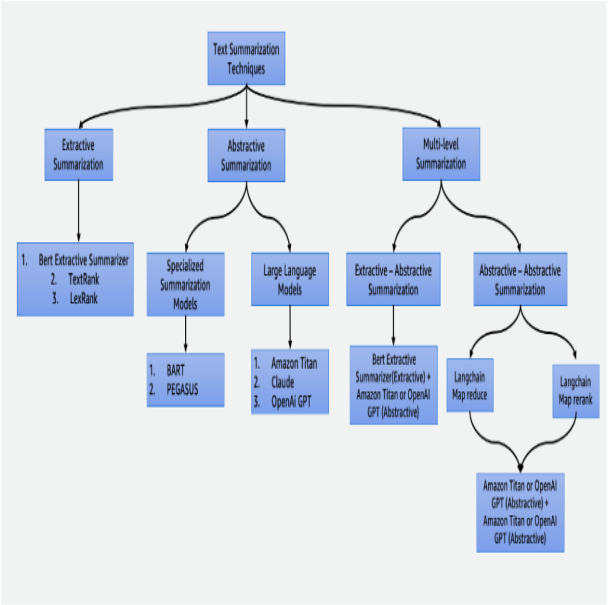

Figure 2. Types of Summarization

5.1 Extractive Summarization

Extractive summarization selects the most important sentences or phrases directly from the original text. It doesn't generate new content but relies on picking what's already present.

- Lead-3: This baseline method simply selects the first three sentences of an article. It's often effective for structured news articles like CNN/DailyMail.
- TextRank LexRank (shown in the diagram): These graph-based methods rank sentences based on their importance in the overall document using concepts similar to Google's PageRank.

Although we focused on Lead-3 for evaluation, TextRank and LexRank are other popular extractive techniques.

5.2 Abstractive Summarization

Abstractive summarization rewrites the content in a new way — much like how a human would paraphrase. It uses deep learning models trained to understand context and language generation.

- BART-Large-XSum: Fine-tuned on the XSum dataset to produce highly concise, single-sentence summaries.
- BART-Large-CNN: Trained on CNN/DailyMail to produce multi-sentence, highlight-style summaries.

These transformer-based models generate summaries not found in the source text directly.

5.3 Multi-level Summarization

Multi-level summarization combines both extractive and abstractive approaches in sequential stages, offering more nuanced summaries for complex documents.

a. Extractive → Abstractive

- Extract key parts using extractive methods like BERT or TextRank.
- Feed this shorter version into a generative model (e.g., GPT or Amazon Titan) for abstractive summarization.

b. Abstractive → Abstractive

- An initial summary is created abstractive.
- This is then passed again into another summarizer using a framework like LangChain Map-Reduce or Re-rank, improving coherence and reducing redundancy.

These are more recent research directions and offer promising results, especially for longer documents or multi-document summarization.

6 Results and Evaluation

This section provides a comprehensive evaluation of the summarization models—abstractive and extractive—across two benchmark datasets: CNN/DailyMail and XSum. The models were evaluated using standard ROUGE metrics: ROUGE-1, ROUGE-2, and ROUGE-Lsum, which quantify the overlap between model-generated summaries and human-written references.

6.1 Evaluation Metrics Used

- ROUGE-1: Measures unigram overlap between the generated and reference summaries.
- ROUGE-2: Measures bigram overlap.
- ROUGE-Lsum: Measures the longest common subsequence, capturing fluency and structure.

6.2 *Quantitative Evaluation - CNN/DailyMail Dataset*

To evaluate the quality of generated summaries, we used the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, which is widely adopted in NLP tasks for summarization. The table below summarizes the ROUGE scores for both abstractive and extractive summarization approaches on the CNN/DailyMail dataset.

Metric	Abstractive	Extractive
ROUGE-1	0.445	0.400
ROUGE-2	0.213	0.180
ROUGE-Lsum	0.375	0.331

Table 1
CNN/DailyMail ROUGE (Abstractive vs Extractive)

The ROUGE-1 score (which evaluates unigram overlap) for the abstractive model is 0.445, slightly higher than the 0.400 scored by the extractive baseline (Lead-3). This suggests that the abstractive model has a better ability to capture key terms found in the human reference summaries. The ROUGE-2 score (which measures bigram co-occurrence) shows a more noticeable improvement: 0.213 vs 0.181. This metric better reflects fluency and the ability to preserve context. The improvement here suggests that abstractive summaries maintain better phrase-level coherence. Similarly, ROUGE-Lsum, which considers the longest common subsequence of tokens and is a strong proxy for human readability, also favors the abstractive model: 0.375 over 0.332. This further demonstrates the abstractive model's strength in generating coherent, human-like summaries.

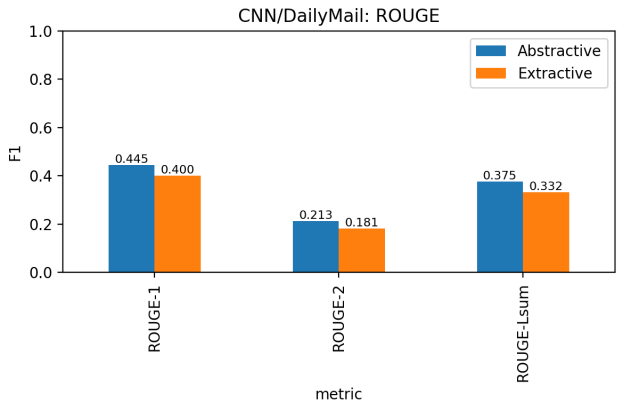

Figure 3. ROUGE Comparison on CNN/DailyMail Dataset Bar Chart

As shown in Figure 3, the abstractive model consistently outperforms the extractive baseline across all ROUGE metrics. The bars represent F1 scores, highlighting the superiority of abstractive summarization in both lexical similarity and contextual depth. The gap is especially prominent in ROUGE-2 and ROUGE-Lsum, indicating that abstractive summaries are more context-aware and fluent.

6.3 *Quantitative Evaluation - XSum Dataset*

To evaluate the quality of generated summaries, we used the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, which is widely adopted in NLP tasks for summarization. The table below summarizes the ROUGE scores for both abstractive and extractive summarization approaches on the CNN/DailyMail dataset.

Metric	Abstractive	Extractive
ROUGE-1	0.458	0.200
ROUGE-2	0.216	0.029
ROUGE-Lsum	0.363	0.126

Table 2
XSum ROUGE (Abstractive vs Extractive)

The ROUGE-1 score shows a significant lead for the abstractive model (0.458) over the extractive baseline (0.200), more than doubling the performance in unigram overlap. This clearly shows that the Lead-3 method fails to capture the critical information needed in XSum's short, highly focused summaries. Even more striking is the difference in ROUGE-2: 0.216 vs 0.029. This metric emphasizes the ability to preserve contextual meaning between consecutive words or phrases. The abstractive model's advantage here reflects its superior language generation and semantic cohesion. The ROUGE-Lsum score tells a similar story: 0.363 for the abstractive model vs 0.126 for the extractive baseline. Since ROUGE-Lsum rewards the longest matching sequences, this result supports the idea that abstractive summaries are not only more accurate but also more readable and human-like.

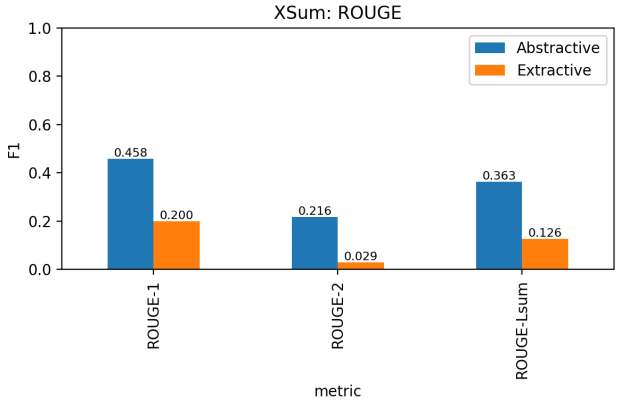

Figure 4. ROUGE Comparison on XSum Dataset Bar Chart

As shown in Figure 4, the abstractive model clearly outperforms the extractive approach across all evaluation metrics. The large gaps in ROUGE-2 and ROUGE-Lsum highlight the extractive method's inability to meet the abstraction and conciseness required by the XSum dataset.

6.4 *Dataset-wise Observations*

- CNN/DailyMail: Since the dataset contains relatively structured and factual news articles, the extractive model performs moderately well. However, the abstractive model still

leads by incorporating semantic richness and better compression.

- XSum: This dataset features a higher level of abstraction in reference summaries. The extractive method fails to generate meaningful summaries, resulting in significantly lower scores. In contrast, the abstractive model generates concise, informative outputs aligned with reference summaries.

6.5 Sample Analysis

- Abstractive summaries rephrase and reframe content—making them succinct yet informative. Extractive summaries, on the other hand, often include repetitive or redundant sentences directly copied from the original text.
- Extractive methods simply pick sentences from the document, resulting in summaries that fail to capture the core idea. Abstractive summaries generated by transformer-based models were often closer to human-written summaries, both semantically and structurally.

6.6 Visual Summary

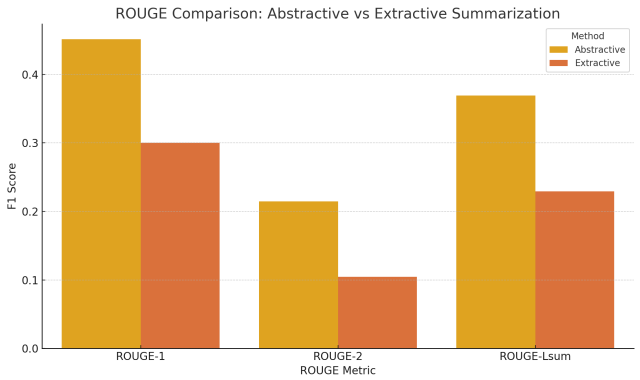

Figure 5. ROUGE Comparison: Abstractive Vs Extractive Summarization

Figure 5 shows a comparative bar chart across all metrics and datasets. It reinforces that abstractive summarization dominates extractive summarization in ROUGE metrics, especially for complex and unstructured news data.

7 Conclusion

This project set out to explore and compare two prominent approaches to text summarization — extractive and abstractive — in order to understand their effectiveness, limitations, and suitability for different types of textual data. Leveraging the Lead-3 baseline method for extractive summarization and pre-trained BART models for abstractive summarization, we implemented a complete pipeline to process, summarize, and evaluate texts from two diverse and widely-used datasets: CNN/DailyMail and XSum.

Our results reveal that abstractive summarization models, particularly those fine-tuned on specific datasets, are capable of generating more coherent, fluent, and context-aware summaries. This was evident in their higher ROUGE scores across all metrics and their ability to paraphrase and synthesize content beyond simple sentence extraction. On the other hand, extractive summarization, while computationally less demanding and easier to implement, showed limited effectiveness — especially with datasets like XSum that require highly condensed, single-sentence summaries.

Through this comparative evaluation, we gained insights into the trade-offs involved in each approach. Extractive methods maintain factual consistency by copying directly from the source, but often lack the natural flow and abstraction expected in human-like summaries. Abstractive methods, while more readable and engaging, carry a higher risk of hallucination and can require more sophisticated tuning to avoid information distortion.

Furthermore, our experiments highlighted the importance of dataset structure. The CNN/DailyMail dataset, with its multi-sentence highlights and narrative style, favors both methods reasonably well. In contrast, the XSum dataset, demanding high compression and abstraction, challenged the extractive model and brought out the strengths of the abstractive approach more clearly.

In summary, the project demonstrates that while there is no one-size-fits-all solution in summarization, abstractive methods — with appropriate model selection and dataset alignment — offer a powerful tool for condensing and rephrasing content in a human-like manner. The use of ROUGE evaluation provided objective insights, and our side-by-side comparison helped build a nuanced understanding of where each method excels or falls short. These findings set the stage for further development, especially in domain-specific applications and user-facing summarization tools.

8 Future Work

While this project provides a strong foundation for understanding the differences between extractive and abstractive summarization techniques, several avenues remain open for further exploration and enhancement.

- **Multi-Level Summarization:** In the current setup, only single-layer summaries were produced. Future iterations can explore hierarchical or multi-stage summarization approaches where longer documents are first segmented and then summarized section-wise. This would allow better handling of very large texts, such as research papers, legal judgments, and medical case reports.
- **Fine-Tuning on Domain-Specific Data:** Although pre-trained models performed well, domain adaptation can significantly improve summarization quality. Fine-tuning BART or similar models on legal, healthcare, or educational datasets could produce more context-aware and jargon-sensitive summaries.
- **Factual Consistency and Hallucination Detection:** Abstractive models often generate text that is fluent but not entirely factual. Incorporating tools to verify factual accuracy, such as QA-based post-checks or truthfulness filters, could enhance reliability, especially in high-stakes domains.
- **Deployment as a Web-Based Tool:** A practical extension of this project would be to develop a user-friendly web application that integrates both summarization modes. Users could upload documents and choose between extractive and abstractive summaries based on their needs. A future goal is to tailor this tool for summarizing healthcare records, patient discharge summaries, and legal case notes, enabling faster document review for professionals.
- **Visualization and User Feedback Loop:** Adding visual indicators of sentence importance and allowing users to give

feedback on the quality of generated summaries could help iteratively refine the summarization models and make the tool more interactive.

In essence, the work done in this project sets the stage for both deeper research and real-world application. With advancements in transformer-based models and continued interest in efficient information access, text summarization will remain a critical area in natural language processing, and this project forms a stepping stone towards building scalable, context-aware, and domain-ready solutions.

Acknowledgements

I would like to express my sincere gratitude to Prof.Sumona Mondal for her invaluable guidance, encouragement, and feedback throughout the course of this project. Her insights and support were instrumental in shaping the direction and depth of this work. I am also thankful to Prof. Michael Gilbert for his mentorship and technical suggestions, which greatly contributed to the successful completion of this project. Their expertise in the field and unwavering academic support have been a source of inspiration, and I truly appreciate their time and dedication in helping me grow both technically and intellectually.

References

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv preprint arXiv:1910.13461.

[2] See, A., Liu, P. J., & Manning, C. D. (2017). *Get To The Point: Summarization with Pointer-Generator Networks*. Proceedings of the ACL.

[3] Narayan, S., Cohen, S. B., & Lapata, M. (2018). *Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization*. Proceedings of EMNLP.

[4] Lin, C. Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. Proceedings of the ACL Workshop on Text Summarization.

[5] Hermann, K. M., et al. (2015). *Teaching Machines to Read and Comprehend*. Advances in Neural Information Processing Systems (NeurIPS).

[6] Xu, J., Li, J., Wang, M., & Sun, X. (2020). *Unsupervised Extractive Summarization by Pre-training Hierarchical Transformers*. Findings of EMNLP.

[7] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. ICML.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

[9] Grusky, M., Naaman, M., & Artzi, Y. (2018). *NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies*. Proceedings of NAACL-HLT.

[10] Hugging Face. (2023). *Transformers Library Documentation*. `https://huggingface.co/docs/transformers`