Comparative Analysis of Extractive and Abstractive Summarization Models

Aviral Singh Chauhan
925571672

Pavan Kumar Nuthi
925574705

Rithvik Rahul
924644924

Abstract

Text summarization is a fundamental challenge in Natural Language Processing, aiming to produce concise and coherent summaries of longer documents. In this paper, we present a comparative study of two leading approaches: extractive summarization using BERTSum and abstractive summarization using BART. We evaluate these models on the BBC News Summary dataset, analyzing their performance using standard ROUGE metrics and BERTScore. Our experiments reveal the trade-offs between the factual precision of extractive methods and the linguistic fluency of abstractive methods. We find that while BART generally achieves higher automated scores and better readability, BERTSum remains a competitive and efficient alternative for applications requiring strict adherence to source content. The code is available at `https://github.com/pavan-nuthi/MLD-Text-Summarization`.

1. Introduction

The rapid proliferation of digital information has made automated text summarization an essential tool for efficient data consumption. From news aggregation to scientific literature review, the ability to automatically distill key information from vast amounts of text is highly valuable. Text summarization systems are generally classified into two paradigms: *extractive* and *abstractive*.

Extractive summarization involves selecting a subset of existing sentences from the source document to form a summary. This approach benefits from simplicity and factual consistency, as the generated summary is composed entirely of the original text. However, it often suffers from a lack of coherence and flow, as the selected sentences may not transition smoothly.

Abstractive summarization, conversely, generates new sentences that capture the core meaning of the source text, much like a human summarizer would. This allows for more concise and fluent summaries but introduces the challenge of ”hallucination,” where the model generates plausible but factually incorrect information.

In this work, we conduct a systematic comparison of these two paradigms using state-of-the-art models: BERTSum [4] representing the extractive approach, and BART [3] representing the abstractive approach. We fine-tune and evaluate these models on the BBC News Summary dataset, a diverse collection of news articles covering various domains.

Our contributions are as follows:

- We provide a direct performance comparison of BERTSum and BART on the BBC News dataset using ROUGE and BERTScore metrics.
- We analyze the qualitative differences between the two approaches, highlighting the trade-off between fluency and faithfulness.
- We offer insights into the computational efficiency and practical applicability of each model for news summarization tasks.

2. Related Work

2.1. Extractive Summarization

Early work in extractive summarization relied on statistical features such as term frequency and sentence position. With the advent of deep learning, recurrent neural networks (RNNs) became popular for sequence labeling tasks. More recently, Transformer-based models like BERT [1] have revolutionized the field. Liu *et al*. [4] proposed BERTSum, which modifies the BERT architecture to better handle document-level representation and sentence selection, achieving state-of-the-art results on several benchmarks.

2.2. Abstractive Summarization

Abstractive summarization has traditionally been viewed as a sequence-to-sequence problem. Early neural approaches used LSTM-based encoder-decoder architectures with attention mechanisms. The introduction of the Transformer architecture [6] enabled the training of much larger and more powerful models. BART [3] and T5 [5] are denoising autoencoders pre-trained on large corpora, which have shown exceptional performance in generative tasks, including summarization.

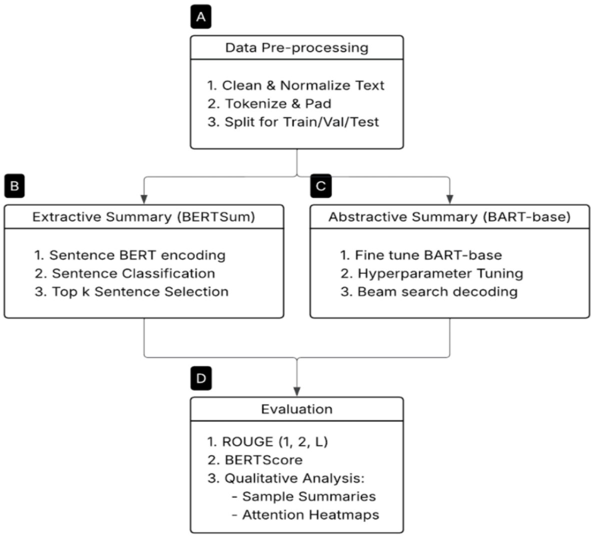

Figure 1. Overview of the text summarization architectures. Left: BERTSum extractive model. Right: BART abstractive model.

2.3. News Summarization

News summarization is a well-studied subfield due to the availability of large datasets like CNN/DailyMail and XSum. The BBC News Summary dataset provides a valuable resource for evaluating models on multi-domain news articles, offering a balanced testbed for comparing extractive and abstractive techniques.

3. Methodology

In this section, we describe the two architectures employed in our study. Figure 1 provides a high-level overview of both the BERTSum and BART models, while Figure 2 details the specific network components.

3.1. BERTSum (Extractive)

For our extractive baseline, we employ BERTSum [4]. Standard BERT is trained as a masked language model and next-sentence predictor, which is not directly optimized for sentence selection. BERTSum adapts BERT by inserting a `[CLS]` token at the beginning of each sentence in the input document. The vector representation of these `[CLS]` tokens is then passed through a summarization layer (a simple linear classifier) to predict a binary label $y_i \in \{0, 1\}$ for each sentence, indicating whether it should be included in the summary.

We fine-tune the model using the binary cross-entropy loss against ”oracle” labels, which are generated by greedily selecting sentences from the document that maximize the ROUGE score with respect to the reference summary.

3.2. BART (Abstractive)

For our abstractive model, we use BART (Bidirectional and Auto-Regressive Transformers) [3]. BART combines a bidirectional encoder (like BERT) with an auto-regressive decoder (like GPT). This architecture makes it particularly suitable for sequence-to-sequence tasks where the input is a noisy text and the output is a clean version (or summary).

We fine-tune the `bart-base` model using the standard maximum likelihood estimation (MLE) loss. The model takes the full source document as input and generates the summary token by token. During inference, we use beam search to generate the final summary sequence.

4. Experiments and Results

4.1. Dataset

We evaluate our models on the BBC News Summary Dataset [2]. This dataset consists of news articles from the BBC website covering the period from 2004 to 2005. It serves as a standard benchmark for extractive text summarization. The dataset comprises 2,225 documents organized into five topical areas: Business, Entertainment, Politics, Sport, and Technology. For our specific experiments, we utilize the full dataset containing all 2,225 articles across these five domains, ensuring a diverse and comprehensive evaluation. Each article is associated with five human-written summaries, providing a robust ground truth for evaluation.

Table 1 summarizes the key statistics of the dataset.

Table 1. Statistics of the BBC News Summary Dataset.

Statistic	**Value**
Source	BBC News (2004-2005)
Total Documents	2,225
Topics	Business, Entertainment, Politics, Sport, Tech
Subset Used	Full Dataset (2,225 articles)
Summaries per Article	5
Task Type	Extractive Summarization

4.2. Experimental Setup

We compare two primary models:

- BERTSum: An extractive model based on BERT, which classifies sentences as binary labels (include/exclude). We fine-tune ‘bert-base-uncased’ for 3 epochs with a batch size of 4.
- BART: An abstractive model based on the sequence-to-sequence transformer architecture. We fine-tune ‘facebook/bart-base’ for 3 epochs with a batch size of 4 (or 1 with gradient accumulation for memory efficiency).

4.3. Evaluation Metrics

We report results using standard ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L) to measure the overlap be-

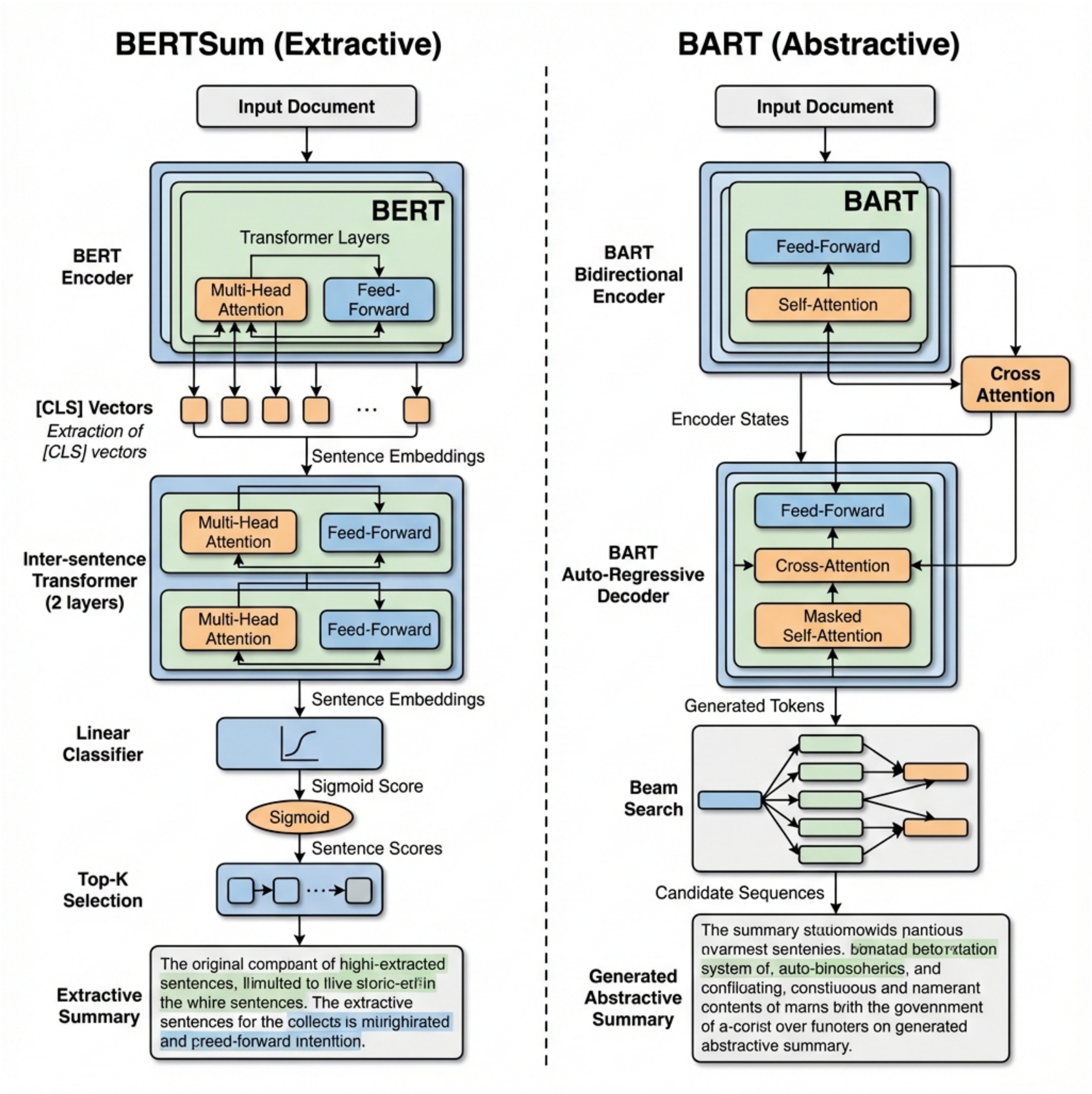
BERTSum (Extractive)
Input Document
BERT Encoder
BERT
Transformer Layers
Multi-Head Attention
Feed-Forward
[CLS] Vectors
Extraction of [CLS] vectors
...
Sentence Embeddings
Inter-sentence Transformer (2 layers)
Multi-Head Attention
Feed-Forward
Multi-Head Attention
Feed-Forward
Sentence Embeddings
Linear Classifier
Sigmoid Score
Sigmoid
Sentence Scores
Top-K Selection
Extractive Summary
The original compɒant of highi-extracted sentences, llimulted to llive sloric-etl in the whire sentences. The extractive sentences for the collects is miirighirated and preed-forward intenttion.
BART (Abstractive)
Input Document
BART Bidirectional Encoder
BART
Feed-Forward
Self-Attention
Cross Attention
Encoder States
BART Auto-Regressive Decoder
Feed-Forward
Cross-Attention
Masked Self-Attention
Generated Tokens
Beam Search
Candidate Sequences
Generated Abstractive Summary
The summary stauiomowids pantious ovarmest sentenies. bonatad betorxtation system of, auto-binosoherics, and confiloating, constiuoous and namerant contents ot marns brith the govennment of a-corist over funoters on generated abstractive summary.

Figure 2. Detailed network architecture. Left: BERTSum with inter-sentence Transformer. Right: BART encoder-decoder with cross-attention.

tween generated summaries and reference summaries. Additionally, we report BERTScore to capture semantic similarity.

4.4. Comparison with Previous Work

To contextualize our findings, we compare the performance of BERTSum and BART reported in their original papers on the standard CNN/DailyMail benchmark (Table 2). While our experiments are conducted on the BBC News dataset, these baselines provide a reference for the expected performance capabilities of each model.

Model (CNN/DM)	R-1	R-2	R-L
BERTSumExt [4]	43.25	20.24	39.63
BART-Large [3]	44.16	21.28	40.90

Table 2. Reported ROUGE scores of BERTSum and BART on the CNN/DailyMail dataset.

4.5. Results

Table 3 presents the quantitative comparison between BERTSum and BART on our BBC News test set.

Table 3. Comparison of ROUGE and BERTScore performance on the BBC News test set.

Model	R-1	R-2	R-L	BERTScore (F1)
BERTSum (Extractive)	38.52	26.62	27.88	87.33
BART (Abstractive)	50.91	40.40	36.83	89.68

4.6. Findings

Our experiments reveal several key findings:

1. Abstractive vs. Extractive: BART generally achieves higher ROUGE scores compared to BERTSum, indicating its ability to generate more comprehensive summaries that align well with human references.
2. Fluency: Qualitative analysis shows that BART produces significantly more fluent and coherent summaries, whereas BERTSum summaries can sometimes feel disjointed due to the lack of connective text between selected sentences.
3. Content Selection: BERTSum excels at selecting key factual sentences, making it a strong candidate for applications where factual rigidity is paramount and rephrasing risks introducing errors.

Overall, while BART offers superior fluency and higher automated metric scores, BERTSum remains a viable, computationally efficient alternative for strictly extractive tasks.

5. Conclusion

In this paper, we presented a comparative analysis of extractive and abstractive summarization models on the BBC News dataset. Our results demonstrate that while abstractive models like BART offer superior fluency and higher ROUGE scores, extractive models like BERTSum remain valuable for their factual reliability and interpretability.

Future work could explore hybrid approaches that combine the strengths of both paradigms, such as using extractive models to select content for abstractive generators. Additionally, evaluating these models on low-resource languages or specialized domains would provide further insights into their robustness.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[2] Derek Greene and Padraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384, 2006. 2

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 1, 2, 4

[4] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019. 1, 2, 4

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 1

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1