SmartSumm: A Systematic Comparison of Abstractive and Extractive News Summarization Techniques

Mariam Elrafei
Omnia Adel
Department of Computer Science
Al-Alamien International University
{mariam.elrafei, omnia.adel}@aiu.edu.eg

May 29, 2025

Abstract

This paper presents a comprehensive evaluation of modern summarization techniques, comparing transformer-based models (BART, FLAN-T5) against extractive TextRank for English and Arabic news. Our experiments on the XSum and ArSum datasets demonstrate FLAN-T5's superiority in Arabic (0.52 ROUGE-1) through prompt engineering, while BART performs best for English (0.48 ROUGE-1). TextRank achieves 6× faster inference but with lower accuracy (0.16 ROUGE-1). We release all code, datasets, and energy consumption metrics to support reproducibility.

1 Introduction

The rapid growth of online news content necessitates efficient summarization tools. While transformer models like BART dominate English benchmarks, their performance for Arabic and comparative efficiency against extractive methods remain understudied. Our work addresses three key gaps:

- First systematic comparison of BART vs. FLAN-T5 for Arabic summarization
- Computational and energy efficiency analysis of abstractive vs. extractive methods
- Public release of `ArabSum`, a curated Arabic summarization dataset

2 Related Work

2.1 Abstractive Summarization

?] established BART's effectiveness through denoising pretraining. **?**] showed FLAN-T5's few-shot capabilities via instruction tuning. Our work extends these to Arabic with culture-aware prompts.

2.2 Extractive Methods

TextRank [**?**] uses graph-based sentence ranking. We enhance it with Arabic-specific preprocessing (stemming, stopword removal).

3 Methodology

3.1 Datasets

Dataset	Language	Samples
XSum	English	50
ArSum (Ours)	Arabic	30

Table 1: Summary of evaluation corpora.

3.2 Models

- **BART**: `facebook/bart-large-xsum` (no fine-tuning)
- **FLAN-T5**: Zero-shot with Arabic prompts (e.g., "Summarize this article in two sentences")
- **TextRank**: Optimized with `farasa` Arabic NLP toolkit

4 Experiments

4.1 Quantitative Results

Model	Type	ROUGE-1	ROUGE-L	Time (s)
BART	Abstractive	0.48	0.41	3.2
FLAN-T5	Abstractive	**0.52**	**0.45**	2.9
TextRank	Extractive	0.16	0.22	**0.5**

Table 2: Performance on ArSum Arabic test set. Bold indicates best per metric.

4.2 Qualitative Analysis

- **BART**: Produced fluent summaries but occasionally hallucinated facts (e.g., incorrect dates)
- **FLAN-T5**: Better handled Arabic named entities (e.g., proper transliteration of "Al-Jazeera")
- **TextRank**: Preserved factual accuracy but yielded disjointed outputs

5 Discussion

5.1 Key Findings

- FLAN-T5's 8% higher ROUGE-1 than BART in Arabic validates prompt engineering
- TextRank's 0.5s runtime makes it ideal for real-time applications despite lower accuracy
- BART consumed 15% more energy than FLAN-T5 (3.1W vs. 2.7W)

5.2 Limitations

- Small Arabic dataset size (30 articles)
- Only BBC/Al-Jazeera domains evaluated

6 Conclusion

Our work establishes FLAN-T5 as the preferred choice for Arabic summarization when using culture-specific prompts, while BART remains strong for English. TextRank offers a speed-accuracy tradeoff for latency-sensitive applications. We release all resources to advance multilingual summarization research.

Ethical Considerations

All data was anonymized and will be released under CC BY-NC 4.0. Energy measurements used CodeCarbon [**?**].