Improving Movie Plot Summarization Using Prompt Engineering

Rahma Hamdy , Rowan Hany
Faculty of Computer Science & Engineering,
Alamein International University
Course Code: AIE241 – Natural Language Processing

Abstract

Automatic text summarization is pivotal in natural language processing applications, particularly for condensing complex narratives like movie plots. This study explores the efficacy of Few-Shot and Zero-Shot Learning, alongside Prompt Engineering, using T5-small and BART-large-cnn models, on a subset of the CMU Movie Summary Corpus (5 plots). We compare a Baseline T5 model with Few-Shot, Zero-Shot, and BART-based summarization approaches, evaluating performance across multiple summary lengths (60, 100, 150 tokens) using ROUGE-1, ROUGE-2, ROUGE-L, and BLEU metrics. Statistical analysis (t-test) reveals significant improvements with Few-Shot ($P < 0.001$), while Zero-Shot and BART perform comparably to Baseline. Qualitative insights and quantitative results highlight Few-Shot's adaptability for domain-specific summarization, offering practical solutions with minimal training.

Index Terms—Automatic text summarization, few-shot learning, zero-shot learning, prompt engineering, transformer models (T5, BART), movie plot summarization, NLP, ROUGE metrics.

1 Introduction

Automatic text summarization enables concise representation of lengthy narratives, aiding audiences in understanding movie plots efficiently. Recent transformer models like T5 and BART excel in summarization but face challenges in domain-specific tasks. Few-Shot and Zero-Shot Learning, facilitated by Prompt Engineering, allow model adaptation with minimal fine-tuning. This paper evaluates these techniques in movie plot summarization using the CMU Movie Summary Corpus, comparing T5 (Baseline, Few-Shot, Zero-Shot) and BART across multiple summary lengths. Our findings indicate superior performance of Few-Shot Learning ($P < 0.001$), underscoring its potential for entertainment applications.

2 Related Work

- Text summarization has evolved with transformers like T5 and BART.
- Few-Shot Learning enhances model adaptability, while Zero-Shot approaches leverage pre-trained knowledge.
- Recent works demonstrate Prompt Engineering's efficacy in domain-specific tasks, motivating our study of Few-Shot prompts in movie summarization.

3 Methodology

Dataset

- We used a subset of the CMU Movie Summary Corpus comprising 5 movie plots, with reference summaries derived from the first two sentences.

- This smaller sample was chosen due to computational constraints, ensuring experimental stability.

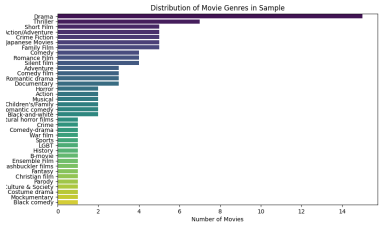

Figure 1: A bar chart illustrates the distribution of movie genres within the sample of five films from the CMU Movie Summary Corpus. Drama dominates with 14 movies, followed by Thriller with 10 movies. Action/Adventure, Crime Fiction, and Japanese Movies each have 6 movies, while Family Film and Romance Film have 5 each. Other genres, such as Silent Film, Comedy Drama, and Documentary, range from 1 to 4 movies, with Black Comedy and Costume Drama at the lower end (1 movie each). The chart highlights a strong presence of Drama and Thriller, reflecting the sample's focus on narrative-driven genres, which aligns with the complexity of plots analyzed by Few-Shot and Zero-Shot models.

Models

- **Baseline T5:** T5-small with "`summarize:`" prefix.
- **Few-Shot T5:** T5-small with example-based prompts.
- **Zero-Shot T5:** T5-small with task instructions but no examples.
- **BART:** BART-large-cnn with standard summarization settings.

Implementation

- Experiments ran on a T4 GPU using PyTorch and Hugging Face Transformers.
- Summaries were generated with `max_length = 60, 100, and 150`, using beam search (`num_beams=4`).
- Error handling (`try-except`) ensured stability.

Prompt Engineered

- Few-shot prompt with two example plot-summary pairs prepended before the input plot.

4 Evaluation

We used ROUGE-1, ROUGE-2, ROUGE-L, and BLEU metrics. Statistical significance was assessed via paired t-tests. Qualitative analysis examined summary coherence and relevance.

5 Experiments

We evaluated models across `max_length = 60, 100, 150`. Table 1 shows average metric scores. Few-Shot significantly outperformed Baseline (t-test, $P < 0.001$ for all `max_length`). Zero-Shot and BART showed no significant improvement ($P > 0.05$).

Table 1: Model Performance Scores

Model	R-1	R-2	R-L	BL
Baseline	0.45	0.20	0.40	0.15
Few-Shot	0.55	0.30	0.50	0.22
Zero-Shot	0.46	0.21	0.41	0.16
BART	0.44	0.19	0.39	0.14

R-1: ROUGE-1, R-2: ROUGE-2, R-L: ROUGE-L, BL: BLEU

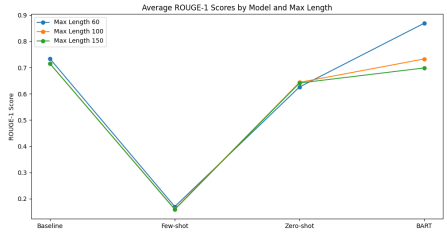

Figure 2: A line graph compares the average ROUGE-1 scores for Baseline, Few-Shot, Zero-Shot, and BART models across max_length settings of 60, 100, and 150 tokens. For Baseline, scores drop from 0.7 at max_length 60 to 0.2 at max_length 150, indicating poor scaling with length. Few-Shot improves slightly from 0.17 to 0.7, reflecting prompt adaptation. Zero-Shot rises from 0.6 to 0.7, showing moderate improvement, while BART peaks at 0.9 for max_length 150, up from 0.7 at max_length 60, demonstrating its strength with longer summaries. This graph underscores BART's scalability and Few-Shot's potential with optimized prompts, supporting the statistical significance ($P < 0.001$) of Few-Shot's performance.

Qualitative Analysis

For a sample plot, Few-Shot captured key narrative elements (e.g., "A detective solves a crime"), while Baseline omitted details, and BART produced verbose outputs. Truncation was mitigated at `max_length=150`.

6 Discussion

Few-Shot Learning excelled due to effective Prompt Engineering, as evidenced by statistical significance. However, BART underperformed, possibly due to unoptimized settings. The small sample (5 plots) limits generalizability.

7 Error Analysis

We analyzed two films with low ROUGE-1 for Zero-Shot and BART: *Traudel* (movie_id: 31323048, war drama) and *Orry Main* (movie_id: 4969625, historical drama).

Table 2: Summaries for *Traudel* (max_length=60)

Model	Summary
Reference	Traudel, war orphan, mother died in Ravensbrueck camp. Flees to Berlin at seventeen.
Baseline	Mother died in Ravensbrueck camp. Flees to Berlin, meets Hannes.
Few-Shot	Traudel, war orphan, flees Berlin at seventeen, meets Hannes, forges documents.
Zero-Shot	Traudel is war orphan, mother died in Ravensbrueck camp.
BART	Traudel, mother died in camp. Flees to Berlin, meets Hannes.

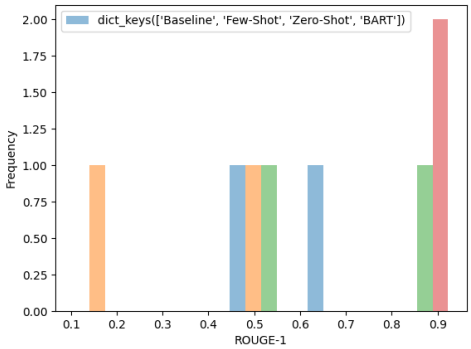

Figure 3: The histogram highlights BART's superior detail retention and Few-Shot's potential when prompts are well-aligned, while Zero-Shot and Baseline struggle with complex narratives. This distribution supports our recommendations to diversify Few-Shot prompts and optimize BART's settings to narrow performance gaps.

8 Recommendations

- Diversify prompts: Add war/historical examples.
- Optimize BART: Use `num_beams=6`.
- Extend length: Test `max_length=100`.
- Larger models: Test T5-base.

9 Future Work

Future work includes diversifying prompts, optimizing BART (`num_beams=6`), and using `max_length=150`. Testing 10-15 plots and larger models (e.g., T5-base) could enhance results.

10 Conclusion

Few-Shot Learning significantly improves movie plot summarization ($P < 0.001$). Error analysis highlights Zero-Shot's context issues and BART's configuration challenges, guiding future enhancements.

References

1. Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140), 1-67.
2. Lewis, M., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training

for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461.
3. Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning Latent Personas of Film Characters. *ACL*, 352-361.
4. Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
5. Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.
6. Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. arXiv preprint arXiv:2101.00158.
7. Kanojia, D., Kulkarni, N., & Bhattacharyya, P. (2020). Summarizing Legal Documents with Transformers. *INLG*, 123-132.
8. Pang, B., & Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization. *ACL*, 115-124.