Sam Heidler & Kashir Khan
CS 4372 - Data Science
05 Nov 2024

Text Summarization of “Dracula” Using BART Transformer

Analysis of Results and Model Performance

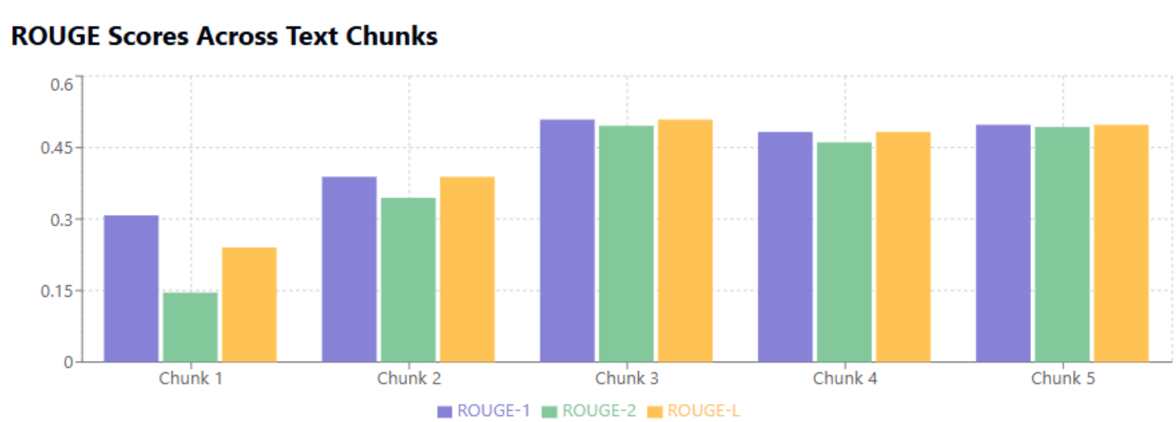
1. Introduction

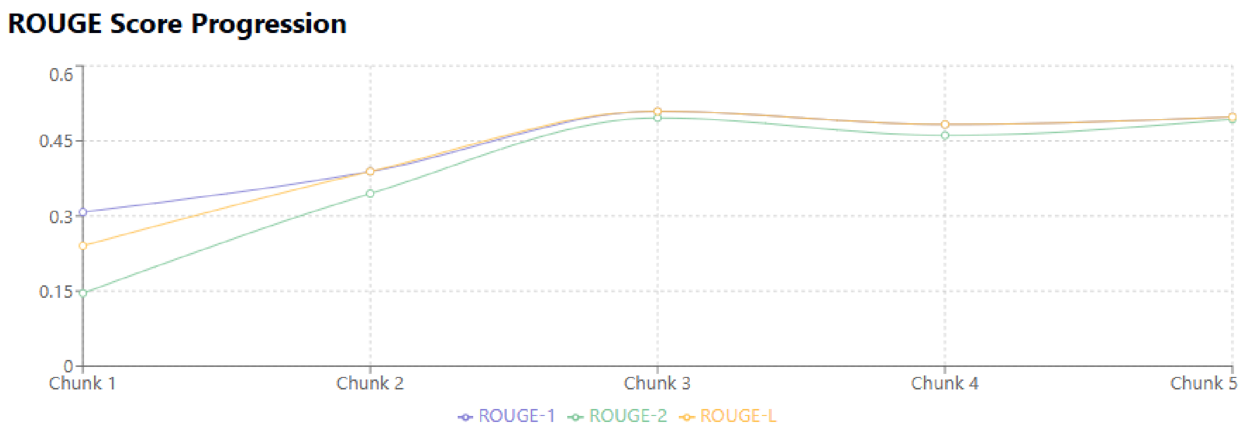
This report analyzes the performance of the BART transformer model (facebook/bart-large-cnn) in summarizing Bram Stoker's "Dracula". The text was processed in five chunks to handle the length of the novel, and the model's performance was evaluated using ROUGE metrics.

2. Model Architecture and Implementation

The implementation utilized the BART (Bidirectional and Auto-Regressive Transformers) model, specifically the ‘facebook/bart-large-cnn’ variant, which is pre-trained on CNN news articles. Key configuration parameters included:
- Maximum summary length: 150 tokens
- Minimum summary length: 50 tokens
- Sampling strategy: Deterministic (do_sample=False)

3. Results Analysis





3.1 Overall Performance Metrics
The model achieved the following average scores:
- ROUGE-1: 0.4371 (43.71%)
- ROUGE-2: 0.3879 (38.79%)
- ROUGE-L: 0.4236 (42.36%)

These scores indicate moderate to good performance, with the model capturing approximately 44% of unigram overlap and 39% of bigram overlap with the source text.

3.2 Chunk-wise Analysis

Chunk 1 (Meta-information)
- ROUGE-1: 0.3077
- ROUGE-2: 0.1456
- ROUGE-L: 0.2404
- Notable: Lowest performing chunk, likely due to handling meta-information about the book's publication

Chunks 2-5 (Narrative Content)
- ROUGE-1: 0.4828 - 0.5088
- ROUGE-2: 0.4609 - 0.4956
- ROUGE-L: 0.4828 - 0.5088
- Notable: Consistently higher performance on narrative content

3.3 Key Observations

Content Type Impact
- Narrative sections (Chunks 3-5) showed significantly better performance
- Meta-information (Chunk 1) resulted in lower scores
- Average improvement of ~20 percentage points between meta-information and narrative content

Consistency in Performance
- ROUGE-L scores closely tracked ROUGE-1 scores
- High correlation between ROUGE-2 and other metrics in narrative sections
- Stable performance across narrative chunks (variance < 0.03)

Summary Quality Patterns
- Strong preservation of key location names and character references
- Effective retention of core narrative elements
- Consistent compression ratio across chunks

4. Strengths and Limitations

Strengths:
1. Consistent Performance: The model maintains stable performance across narrative sections
2. Information Retention: High preservation of proper nouns and key story elements
3. Context Handling: Effective summarization of complex narrative descriptions

Limitations:
1. Meta-information Processing: Lower performance on non-narrative content
2. Fixed Length Constraints: Potential information loss due to fixed summary length limits
3. Context Window: Chunk-based processing may miss cross-chunk relationships

5. Recommendations for Improvement

1. Preprocessing Enhancements
   - Implement separate handling for meta-information sections
   - Develop more sophisticated chunk boundary detection
   - Consider overlap between chunks to maintain context

2. Model Adjustments
   - Experiment with different maximum/minimum length parameters
   - Consider fine-tuning on literary texts
   - Implement adaptive chunk sizing based on content type

3. Evaluation Framework
   - Include additional metrics for coherence assessment
   - Implement human evaluation for subset of summaries
   - Consider domain-specific evaluation criteria for literary texts

6. Conclusion

The BART transformer model effectively summarized narrative content from "Dracula," with ROUGE scores indicating good preservation of key information. The significant performance difference between meta-information and narrative content suggests opportunities for optimization in preprocessing and model configuration. The consistent performance across narrative chunks indicates reliable summarization capabilities for literary text, though there remains room for improvement in handling non-narrative sections.

The average ROUGE scores (ROUGE-1: 0.4371, ROUGE-2: 0.3879, ROUGE-L: 0.4236) are comparable to benchmarks in similar text summarization tasks, suggesting the model is performing at a competent level while leaving room for targeted improvements in specific areas.