Using ROGUE-N to Summarise a Passage - Assignment 1/Task 2

David Ewing

2024.09.07

Abstract

In this report, we evaluate the performance of two large language models (LLMs), ChatGPT 4.0 and Google Bard, in summarising a passage about the history of the Olympic Games using the ROGUE-N metric. The reference summary was generated by Claude. We calculated ROGUE-1, ROGUE-2, ROGUE-3, and ROGUE-4 scores, as well as the overall ROGUE-N score for each model. The results are discussed, and conclusions are drawn regarding the relative performance of the models.

1 Introduction

The ROGUE-N metric evaluates the quality of a summary by comparing it with a reference summary using precision and recall for different n-grams. In this task, we used two LLMs, ChatGPT 4.0 and Google Bard, to generate 40-word summaries of a passage on the Olympics, using a summary generated by Claude as the reference.

2 Methodology

The two LLMs were used to generate summaries of a 400-word passage. The ROGUE-N scores (for n-grams of 1 to 4) were calculated for each summary using the following formula:

$$\text{ROGUE-N} = \frac{\text{Matches of N-grams in both candidate and reference}}{\text{Total N-grams in the reference summary}} \tag{1}$$

The overall ROGUE-N score was computed by calculating the geometric mean of the individual ROGUE-n scores as follows:

$$\text{ROGUE-N} = \sqrt[4]{(\text{ROGUE-1}) \times (\text{ROGUE-2}) \times (\text{ROGUE-3}) \times (\text{ROGUE-4})} \tag{2}$$

The calculated results for ChatGPT 4.0 and Google Bard are shown in the following table.

3 Results

Model	Overall Score	ROGUE-1	ROGUE-2	ROGUE-3	ROGUE-4
ChatGPT 4.0	0.1559	0.5278	0.3143	0.1176	0.0303
Google Bard	0.1137	0.4000	0.1471	0.0909	0.0312

Table 1: ROGUE-N Scores for ChatGPT 4.0 and Google Bard

4 Discussion

4.1 ChatGPT 4.0 Summary

ChatGPT 4.0 provided a summary with a higher overall ROGUE-N score, suggesting better alignment with the reference summary generated by Claude. Its high ROGUE-1 score indicates strong unigram precision, while lower n-gram precision scores reflect some loss of coherence across longer phrases.

4.2 Google Bard Summary

Google Bard, while still providing an accurate summary, had a lower overall ROGUE-N score. It exhibited lower performance across n-grams, particularly in bigrams and trigrams, suggesting some issues in capturing more complex phrase-level patterns when compared to the reference summary.

4.3 Claude as the Reference

Claude, used as the reference summary, performed the task of summarisation efficiently by capturing the key elements of the passage in a coherent manner. Its summary provided a balanced structure, containing details about both the historical origins and modern evolution of the Olympics. Claude's summary effectively set the standard for evaluating the other two models. Claude's tendency to capture both broader themes and specifics made it an ideal reference for this evaluation.

5 Conclusion

This comparative analysis using ROGUE-N metric demonstrates that ChatGPT 4.o outperformed Google Bard in summarising the provided passage about the Olympic Games. ChatGPT's higher ROGUE-1 score indicates stronger performance in capturing unigrams, while its slightly better performance in higher n-grams suggests that coherence is preserved over longer phrases. Its performance dropped significantly as n-gram length increased, indicating that it may struggle to maintain 'contextual flow' in longer sequences of words.

Google Bard, performed less effectively across all n-grams. This may indicate Bard was less effective at capturing relationships between words and phrases within the passage. Its lower ROGUE-1 score may also suggest that it missed some details found in the reference summary.

Claude, used as the reference summary, proved to be a balanced and well-structured reference. It successfully captured both historical details and modern developments of the Olympics, which enabled effective evaluation of the other models. Claude's ability to blend specific and general information made it an ideal baseline.

In conclusion, both ChatGPT 4.0 and Google Bard demonstrated reasonable summarisation capabilities, however, the author's experience in these matters has been limited to this Assignment. ChatGPT 4.0 exhibited a better balance between precision and coherence. Future research could explore these models in different contexts, such as summarising technical or creative texts, to better understand their limitations and strengths across various domains.

It is the author's gut feeling that diagramming the sentence structure and providing more structure to the comparison is an area of research that has not yet been explored.