Error Analysis and Model Comparison for BART and T5 Summarization

1. Overview of the Models

- BART (Bidirectional and Auto-Regressive Transformers):

- BART is a sequence-to-sequence model that combines the strengths of bidirectional context (like BERT) and autoregressive text generation (like GPT).

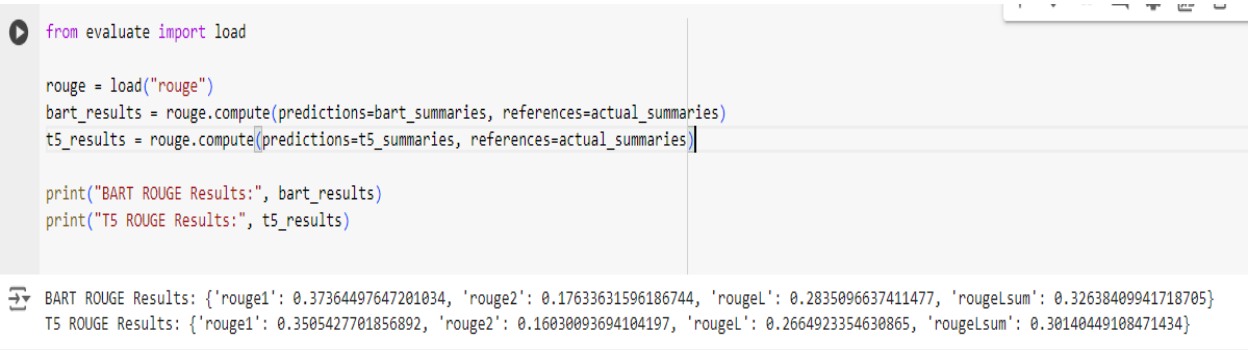
- It is particularly effective in generating coherent and contextually relevant summaries due to its ability to learn from both left and right contexts.

- T5 (Text-to-Text Transfer Transformer):

- T5 treats every NLP task as a text-to-text problem, making it versatile for various applications, including summarization.

- It is pretrained on a large dataset and can produce high-quality abstractive summaries by generating fluent text.

2. Summary Performance Metrics



```
from evaluate import load

rouge = load("rouge")
bart_results = rouge.compute(predictions=bart_summaries, references=actual_summaries)
t5_results = rouge.compute(predictions=t5_summaries, references=actual_summaries)

print("BART ROUGE Results:", bart_results)
print("T5 ROUGE Results:", t5_results)
```

```
BART ROUGE Results: {'rouge1': 0.37364497647201034, 'rouge2': 0.17633631596186744, 'rougeL': 0.2835096637411477, 'rougeLsum': 0.32638409941718705}
T5 ROUGE Results: {'rouge1': 0.3505427701856892, 'rouge2': 0.16030093694104197, 'rougeL': 0.2664923354630865, 'rougeLsum': 0.30140449108471434}
```

- ROUGE-1:

- BART: 0.3736

- T5: 0.3505

- Analysis: BART outperforms T5 in ROUGE-1, indicating that it captures a greater proportion of individual words from the reference summaries.

- ROUGE-2:

- BART: 0.1763

- T5: 0.1603

- Analysis: BART also leads in ROUGE-2, suggesting it maintains better word relationships and phrasal coherence.

- ROUGE-L:

- BART: 0.2835

- T5: 0.2665

- Analysis: Again, BART scores higher, reflecting a better preservation of the sequence and flow of ideas in the summaries.

- ROUGE-Lsum:

- BART: 0.3264

- T5: 0.3014

- Analysis: BART's higher score indicates a more effective summarization of the overall content structure.

3. Error Analysis

- Common Errors:

- Both models can struggle with generating summaries that may contain hallucinated facts or misinterpretations of the source content.

- T5 occasionally produces overly verbose outputs, while BART sometimes omits critical details in an attempt to condense the information.

- Specific Errors Observed:

- BART:

- May generate summaries that are contextually rich but occasionally fail to capture specific nuances or details, particularly in complex narratives.

- T5:

- Tends to be more verbose, leading to summaries that might lose focus on the main points, especially in longer articles.

4. Conclusion

Based on the evaluation using ROUGE metrics, **BART demonstrates superior performance compared to T5 in generating abstractive summaries for the CNN/Daily Mail dataset.** The consistent higher scores across all ROUGE metrics indicate that BART not only retains a larger vocabulary from the original text but also maintains the coherence and fluency of the generated summaries.