Abstractive Dialogue Summarization: A Comparative Evaluation of Fine-Tuned BART and Gemini

Erica Solinas

Introduction

In today's digital age, we are constantly generating and consuming vast amounts of conversational data, especially through group chats and messaging platforms. Navigating through long threads can be time-consuming and overwhelming, often making it difficult to identify key information. In such contexts, having access to concise summaries would greatly enhance user efficiency and comprehension. This project tackles the task of **Text Summarization**, with a particular focus on the creation of a Conversation Summarizer.

Summarization Task

There are two main strategies to approach summarization tasks:

1. **Extractive Summarization**: this method selects a subset of sentences directly from the original text to build the summary. It is typically framed as a binary classification problem, where each sentence is evaluated on whether it should be included in the final summary. Extractive methods are relatively simple to implement and evaluate, often using accuracy-based metrics.

2. **Abstractive Summarization**: these models typically rely on text generation architectures, making evaluation more complex. However, they tend to produce more fluent summaries and are better suited to tasks requiring semantic abstraction.

For this project, we adopted the abstractive summarization approach, comparing two different methodologies for conversation summarization. First, we fine-tuned Facebook's BART model on the SAMSum dataset, then we compared its performance with Gemini, a state-of-the-art large language model (LLM) developed by Google AI. Gemini was accessed via API and employed in a few-shot learning setting. The comparison was carried out using both **ROUGE** and **BERTScore** metrics, allowing us to evaluate the generated summaries in terms of content overlap and semantic similarity with the reference summaries.

Dataset

We used the **SAMSum dataset**, developed by Samsung RD, which consists of multi-turn chat dialogues paired with human-written abstractive summaries. The conversations were written by linguists who were asked to create dialogues similar to their real-life messenger conversations, covering a wide range of topics such as casual small talk, gossip, planning meetings, political discussions, and academic consultations. The dialogues reflect varying degrees of formality and include informal expressions, slang, emoticons, and occasional typos, making the dataset a realistic representation of everyday messaging. The summaries follow specific guidelines: they had to be concise, extract key pieces of information, mention the participants by name, and be written in the third person. Each dialogue has exactly one reference summary. The dataset contains 14,732 dialogues for training, 818 for validation, and 819 for testing. Approximately 75% of the dialogues involve two speakers, with the remainder featuring conversations between three or more participants.

Model and Fine-Tuning

Due to computational constraints, we selected a model that could be efficiently fine-tuned using Google Colab. We chose **Facebook's BART** (Bidirectional and Auto-Regressive Transformers), a transformer-based sequence-to-sequence model that combines the strengths of BERT (as encoder) and GPT (as decoder). The encoder is bidirectional which takes inspiration from BERT, while the decoder is autoregressive from left to right which takes inspiration from GPT-3 (Figure 1).

BART is pre-trained specifically for summarization tasks and has demonstrated strong performance across various text generation tasks, including abstractive summarization. However, BART was originally pre-trained on well-structured textual corpora such as news articles and stories, which

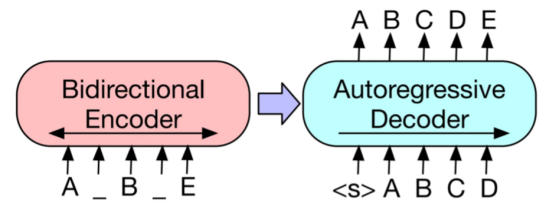

Figure 1: BART architecture

differ significantly from the dialogues found in messaging applications. Conversations are characterized by short, informal, and often fragmented utterances involving multiple speakers and lacking the clear paragraph structure typical of formal texts. This structural discrepancy presents a challenge, as many summarization models are not inherently designed to handle the dynamic and interactive nature of dialogues. To bridge this gap, we fine-tuned the 'facebook/bart-large-xsum' variant, which was initially trained on the XSum dataset, containing over 220,000 BBC news articles paired with one-sentence abstractive summaries. Fine-tuning was performed using the Hugging Face Transformers library with GPU acceleration on Google Colab. Before launching full-scale training, we conducted an initial experiment on a subset of 1,500 samples for 3 epochs. During this test, training loss decreased steadily, while validation loss initially improved but then worsened in the third epoch, rising from 0.3811 to 0.5019. This trend, along with a plateau in ROUGE scores, suggested the onset of overfitting. Based on these results, we decided to train the full model for only 2 epochs (Figure 2). Figure 3 shows the results

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	0.308600	0.303632	53.331100	28.748700	44.169500	44.114600	27.515900
2	0.215700	0.316653	54.527800	30.179400	44.986200	44.990100	29.183400

Figure 2: Results of the first fine-tuning

of the final fine-tuning on the full dataset SAMsum.

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	0.198500	0.381135	49.921400	24.790700	40.674700	40.610900	27.110000
2	0.143100	0.405574	50.698700	25.209500	41.161700	41.162100	27.420000
3	0.079100	0.501908	50.621000	24.752500	40.462000	40.440200	29.846000

Figure 3: Results of the final fine-tuning

Comparison with Gemini

To evaluate the effectiveness of our fine-tuned model, we compared its performance with a state-of-the-art model: **Gemini Flash 2.0**, accessed via API using prompting techniques. We initially tested the model in a zero-shot setting by asking it to summarize conversations directly. However, we observed better results using a few-shot approach. In this setup, we included three example dialogues and their summaries taken from the SAMSum training set directly in the prompt, followed by the new conversation to summarize. Additionally, we explicitly defined the model's role and clearly stated the task to guide its output generation more effectively.

Evaluation Metrics and Results

After training, we evaluated the fine-tuned BART model on the SAMSum test set. For each dialogue, the model generated a summary which was stored in a CSV file. To assess the quality of the generated summaries, we compared them to the reference (gold standard) summaries using both ROUGE and BERTScore metrics:

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap between the generated summary and the reference summary in terms of n-grams. In particular, ROUGE-1 and ROUGE-2 evaluate unigram and bigram matches, while ROUGE-L focuses on the longest common subsequence, capturing fluency and sentence-level coherence.
- **BERTScore**, on the other hand, goes beyond surface-level overlap by comparing contextual embeddings of words using a pre-trained BERT model. It evaluates the semantic similarity between the generated and reference summaries, providing a more nuanced understanding of content preservation and meaning.

The finetuned model achieved solid results. However, to establish a meaningful benchmark, we also evaluated summaries generated by Gemini Flash 2.0 on the same test set, using a few-shot prompting strategy. These summaries were saved and evaluated in the same way. We can see the results for ROUGE Score in Table 1 and BERTScore in Table 2.

Metric	BART	Gemini
ROUGE-1	0.5165	0.4255
ROUGE-2	0.2726	0.1744
ROUGE-L	0.4313	0.3338
ROUGE-Lsum	0.4310	0.3338

Table 1: ROUGE scores for BART and Gemini

Discussion

These results show that the fine-tuned BART model not only outperformed Gemini in both lexical overlap (ROUGE) and semantic similarity

Metric	BART	Gemini
Precision	0.9210	0.8944
Recall	0.9207	0.9186
F1 Score	0.9207	0.9062

Table 2: BERTScore for BART and Gemini

(BERTScore) but also proved to be more effective in adapting to the structure and style of chat-based dialogues. This suggests that task-specific fine-tuning, even with limited resources, can achieve superior performance compared to few-shot prompting with a general-purpose LLM like Gemini. The gap between the two models is particularly noticeable in the ROUGE-2 and ROUGE-L scores, which reflect the ability to generate coherent, fluent, and contextually appropriate summaries. This suggests that the BART model, after being fine-tuned on a domain-specific dataset such as SAMSum, is better equipped to handle the informal, fragmented nature of conversational data. The strong BERTScore also confirms that the model is not only accurate at the surface level but also captures the underlying meaning of the dialogues effectively.

Interestingly, the Gemini model showed a high recall in BERTScore (0.9186), which indicates that it tends to include semantically relevant content from the input, but this comes at the cost of precision. In practice, this may result in slightly longer or more generic summaries that do not match the gold standard as closely in form or structure. This observation aligns with the nature of few-shot prompting: while flexible, it lacks the domain adaptation that fine-tuning provides.

Conclusion and Future Work

This project demonstrates the effectiveness of fine-tuning a pre-trained summarization model, such as BART, on a conversation-specific dataset like SAMSum. Despite hardware limitations, it was possible to adapt a large language model using Google Colab and achieve performance that outperforms a cutting-edge model like Gemini when used in a few-shot setting. The results confirm that domain adaptation via fine-tuning can achieve significant improvements in both lexical and semantic summary quality, even with a relatively small dataset and modest resources. The comparison with Gemini also highlights the limitations of prompting-based methods for tasks involving complex discourse structures such as dialogue. While prompting remains a powerful and flexible tool, especially for rapid prototyping or low-resource settings, fine-tuning remains a more robust approach when high-quality, domain-specific performance is required. For future work, several directions could be explored. First of all, experimenting with other conversation-oriented datasets could help assess the model's generalizability. Finally, integrating newer evaluation metrics such as DeepEval could provide a more comprehensive picture of summary quality, especially in terms of coherence and factual consistency.

References

- Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019). SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pp. 70–79.
- Google AI. Gemini API documentation. `https://ai.google.dev/gemini-api/docs`
- Hugging Face Transformers library. `https://huggingface.co/docs/transformers/en/index`
- Zhu, C. (2021). Chapter 8 - Applications and future of machine reading comprehension. In Machine Reading Comprehension (pp. 185-207). Elsevier.
- Lewis, M. et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation. `https://arxiv.org/abs/1910.13461`
- Lin, C.-Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004).
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. `http://arxiv.org/abs/1904.09675`