Text Summarization

Saurav Kumar Singh[1], Shreya Surendra Choure[2], Upasana Chaudhari[1], and Pranav Dharamthok[2]

[1]saurav2@pdx.edu, upasana@pdx.edu
[2]schoure@pdx.edu, pran4@pdx.edu

1 Project Overview

Our NLP project focuses on developing conversation summarization models utilizing techniques such as TF-IDF, TextRank, and advanced Large Language Models including T5 and BART. We perform comprehensive quantitative analyses to evaluate their performance. Additionally, the models are integrated into a Streamlit-based interface for enhanced user interaction.

2 Main Goals

Our NLP Project develops conversation summarization models using TF-IDF, TextRank, and Large Language Models like T5, BART and evaluates its performance with quantitative analyses. Further It is deployed to "streamlit" for User Interaction.

Goals:

1. How does Transformers/ LLM perform over the traditional text summarization approaches?
2. How much accuracy improvements fine tuning the model on a particular domain brings?
3. Can we detect issues of a customer from a call center Transcript ?

Industry use cases:

- Automated Call Categorization: identify high-frequency call categories and areas for improvement, contributing to enhanced services.
- AI-Driven Issue Extraction: provide a quick and accurate summary for efficient issue resolution.

3 What NLP task(s) do you address?

Data Preprocessing: Removed special characters, punctuation, symbols, and stop words to clean the text. Additionally, performed tokenization and addressed noisy data for improved data quality.

Traditional TF-IDF and Text Rank for Summarization: We have chosen TF-IDF (Term Frequency-Inverse Document Frequency) and the Text Rank algorithm as our baseline models for summarization. These models employ an extractive summarization approach, which involves selecting important sentences from a text to create a summary.

Untrained BART and T5: We are incorporating untrained versions of BART (Bidirectional and Auto-Regressive Transformers) and T5 (Text-To-Text Transfer Transformer) These models, in their original state, possess the capacity for abstractive summarization, generating novel sentences to encapsulate the core content of the conversation.

Trained and Fine-Tuned LLMs: The project involves the training and fine tuning of the T5 and BART. This process enhances the models' summarization capabilities by adapting them to specific tasks and domains, ensuring more precise and contextually relevant generation of summaries for conversations.

GPT 4: The task involves efficiently employing prompt engineering techniques to ensure accurate and effective summarization. This model excels in generating summaries that capture the essence of the conversation

Performance Evaluation: Quantitative analyses are conducted to evaluate the performance of the summarization models. This involves using metrics ROUGE scores to assess the similarity between the generated summaries and reference summaries.

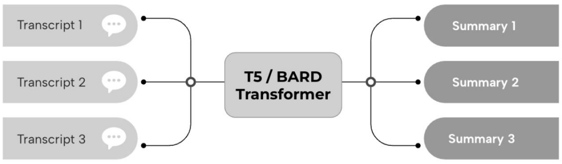

Figure 1: Model Flow

4 Data Set Description

The DialogSum Corpus is a significant dataset specifically tailored for dialogue summarization. It contains 13,460 dialogues and an extra set of 1,000 dialogues specifically for testing purposes, effectively categorized into training, testing, and validation groups. Each data instance in this dataset includes a unique identifier, the dialogue text, a human-written summary, and a topic. The dataset is meticulously organized, with 12,460 instances dedicated to training, 500 for validation, and 1,500 for testing. Additionally, there are 100 holdout instances that feature three key elements: ID, dialogue, and topic.

DialogSum's dialogues are characterized by their real-life scenario relevance, diversity in content, and clear communication patterns and intents. The annotation process for creating summaries was thorough, focusing on elements like salience, brevity, preservation of named entities, observer perspective, and formal language. The dataset was developed and annotated by linguistics experts and language professionals, ensuring high-quality and realistic dialogue scenarios. https://huggingface.co/datasets/knkarthick/dialogsum/viewer

5 Methodology

Text summarization Approaches :

Extractive Summarization: Imagine you have a big article, and you want to make a summary of it. Extractive summarization is like using a highlighter. You go through the article and highlight the most important sentences. Then, you just take those sentences and put them together to make a summary. You're not changing the sentences or writing anything new; you're just pulling out the key parts as they are.

5.1 TF-IDF: Term Frequency-Inverse Document Frequency

TF-IDF, which stands for Term Frequency-Inverse Document Frequency, is a popular technique used for text summarization.

TF-IDF helps in figuring out which words are most important in a piece of text. It does this by looking at how often a word appears in a specific document (Term Frequency) and comparing it with how rare the word is across all documents (Inverse Document Frequency).

Some words appear a lot but aren't very meaningful (like 'the' or 'is'). TF-IDF is great because it gives higher scores to words that are frequent in a particular document but not so common in others. This balance helps in pinpointing words that are crucial in understanding the main ideas of the text.

Once you know which words are most significant, you can use them to find the most important sentences in a document. Typically, sentences containing high TF-IDF scoring words are more likely to be key parts of the summary.

TF-IDF is a straightforward, automated method, making it efficient for handling large volumes of text. This is especially handy for summarizing long documents or processing many documents quickly.

5.2 TextRank

TextRank is an extractive and unsupervised text summarization technique that also excels in keyword extraction. It draws inspiration from the PageRank algorithm, famously used in Google's search engine for ranking websites. This innovative algorithm treats sentences in the text as nodes in a graph, similar to how PageRank views web pages.

The relationships, or edges, between these sentence nodes are determined based on their similarity, creating a framework that mirrors the connectivity used in PageRank for web pages. By leveraging this structure, TextRank effectively identifies and highlights key sentences and phrases, making it an invaluable tool for distilling the essence of large volumes of text in a manner that's automated and efficient.

In TextRank, sentences from the text are treated as nodes in a graph, with the edges representing the similarity between these sentences. This similarity is quantified and stored in a matrix, similar to the approach of PageRank's matrix. The algorithm begins by combining all the text from the source documents and then dividing this content into individual sentences. Each sentence is then transformed into a vector using word embeddings, capturing the semantic essence of the words.

These vectors facilitate the calculation of similarities between sentences, which are used to construct a graph. The sentences are then ranked based on their connectivity and importance in this graph. The top-ranked sentences are selected to form the final summary, offering a concise representation of the original text's key points. This method is particularly effective for summarizing large volumes of text in an unsupervised manner, making it a valuable tool in the field of natural language processing.

Abstractive Summarization Abstractive summarization is a bit like telling a friend about that same article in your own words. Instead of just picking out certain sentences, you read the whole thing and then write a summary that captures the main ideas, but in a new and shorter way. You're not just copying and pasting; you're kind of like a storyteller, retelling the article in a brief, fresh manner, often using different words and maybe even combining ideas from different parts of the article.

5.3 T5 Text-To-Text Transfer Transformer

Untrained T5 Model:

Selection Rationale: T5 is chosen for its effective grasp of contextual information. It leverages pre-training knowledge from a vast text corpus, enhancing its understanding of nuances within sentences.

Pre-training Knowledge: T5 benefits from extensive pre-training, allowing it to capture diverse linguistic patterns and semantic relationships.

Sentence Generation: T5 employs a text-to-text approach, predicting relevant information during fine-tuning. This enables it to generate new and original sentences in the summarization process.

T5 Fine-Tuned Model:

Training Data: The model is fine-tuned using the curated DialogSum dataset, which includes dialogues, a referential human-written summaries, and topics.

Optimization Objective: The fine-tuning process optimizes the model to generate coherent and contextually relevant summaries. Training involves minimizing a loss function that measures the disparity between the generated summaries and the human-written references.

5.4 BART (Bidirectional and Auto-Regressive Transformers)

Untrained BART Model:

Selection Rationale: BART is chosen for its bidirectional and auto-regressive transformer design. This architecture is well-suited for various NLP tasks, including text summarization.

Bidirectional Design: BART's bidirectional architecture allows it to consider both past and future context when generating summaries. This contributes to the production of coherent and contextually relevant summaries with original sentences.

BART Fine-Tuned Model:

Training Data: Similar to T5, BART undergoes a fine-tuning process using the DialogSum dataset to adapt its capabilities to the dialogue summarization task.

Training Objective: The model is trained to minimize the difference between its generated summaries and the human-written summaries in the dataset. This involves adjusting the model's parameters to enhance its summarization performance. **Comparison of T5 and BART:**

Both T5 and BART, in their fine-tuned states, aim to generate abstractive summaries that capture the essential information from dialogues.

The comparison involves evaluating their performance based on metrics such as ROUGE scores to measure the similarity between the generated summaries and reference summaries in the DialogSum dataset.

5.5 GPT-4

GPT 4 has a better understanding of context and nuances in language. This improved contextual understanding can lead to more coherent and contextually relevant summaries.

GPT-4 demonstrates better abilities to abstract information and generalize across different domains. This can result in more concise and informative summaries that capture the key points of a text.

It is considered better equipped to handle ambiguous language and complex texts. This is particularly important for summarization tasks where the source material may be intricate or contain multiple layers of meaning.

Prompt engineering allows for an iterative process which allows experiment with different prompts, observe the model's responses, and refine the approach based on the results. This iterative refinement is crucial for achieving optimal performance. By crafting a well-designed prompt, the model can be guided to generate a summary that aligns with your specific requirements. This makes summaries more explicit about the format, length, or style of the desired summary. Effective prompt engineering helps you control the quality of the summarization output.

ChatGPT-4 being the most advanced multimodal AI from OpenAI that outperforms T5 and BART untrained models with superior text understanding, image processing, and contextually aware responses.

6 Result

- TFIDF has moderate ROUGE-1 and ROUGE-L scores but low ROUGE-2, indicating relevance to the reference summary but missing detailed phrasing.

- Text Rank captures some key concepts with a ROUGE-1 score of 0.17 but struggles with detailed elements and overall structure.

- T5 RAW presents an F1-Score of 0.19 in ROUGE-1, with moderate improvements noted upon fine-tuning, reaching a 0.39 score.

- BART RAW shows high recall but low precision, suggesting summaries are exhaustive but possibly include irrelevant information. Fine-tuning raises its ROUGE-1 F1-Score to 0.40.

- GPT-4 outperforms BART RAW and T5 RAW in ROUGE-1 and ROUGE-L, indicating more accurate summary content matching with reference summaries. However, its recall scores suggest moderate enhancement in capturing the entirety of reference content compared to BART RAW.

Overall, fine-tuning significantly improves the accuracy of T5 and BART models, while ChatGPT-4 shows superior performance in text understanding and contextual responses.

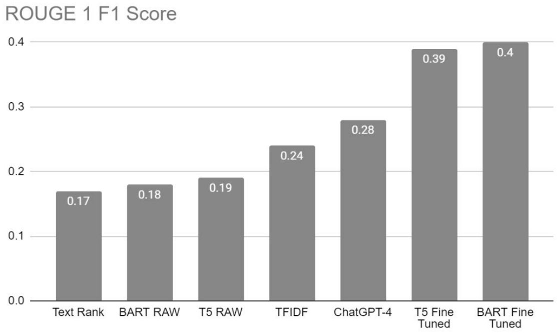

Figure 2: Rouge 1 F1 Score for all the models

Model	R-1 F1	R-2 F1	R-L F1
TFIDF	0.24	0.07	0.19
Text Rank	0.17	0.05	0.16
T5 RAW	0.19	0.07	0.19
T5 Fine Tuned	0.39	0.14	0.37
BART RAW	0.18	0.05	0.17
BART Fine Tuned	0.40	0.15	0.36
GPT-4	0.28	0.08	0.25

Table 1: Model Performance Metrics

7 Interesting insights

1. Many conversations range from 200 to 500 characters, with human-generated summaries typically spanning 50 to 130 characters sometimes missing out the important gist. However, AI-generated summaries, spanning 200 to 350 characters, consistently capture the overall essence of the conversation.

2. Fine-tuning BART and T5 models significantly improved the ROUGE score from 0.18 to 0.4, underscoring the impactful role of training in enhancing the models' ability to generate summaries that align more closely with human-written references.

3. GPT-4 demonstrates superior performance compared to traditional and untrained mod-

els, showcasing the effectiveness of extensive pre-training. However, in head-to-head comparisons with fine-tuned models, pretrained models consistently outperform ChatGPT. This suggests that while general pretraining is powerful, task-specific fine-tuning remains crucial for optimizing models for specialized applications.

8 Ethical Considerations of Your Research Project

1. **Removal of PII from the Dataset:** Maintaining compliance with data protection laws like GDPR by removing personally identifiable information (PII) from the dataset.

2. **Confidentiality in Health Center or Similar Transcripts:** Ensuring the confidentiality of information contained in transcripts from health centers or similar settings.

3. **Consent and Authorized Use of Training Dataset:** Securing consent and ensuring authorized use of the training dataset, especially considering its sensitive nature and the potential presence of personal information.

9 Other Additional Contributions

Deployment and Demonstration

- The untrained T5 and BART models were saved in an offline format.

- The versions trained on our dataset were also saved for offline use.

- The models were uploaded to GitHub for version control and were deployed using Streamlit for hosting.

- These models are invoked by a separate function designed for summarization using the trained datasets.

- To test the summarization capabilities interactively, we utilized Streamlit to create a user interface.

10 Future Work

1. Enhance the model capability to process and summarize information from multimedia sources.

2. Improve the user experience by providing comprehensive summaries across different data types.

3. Implement parallel processing and distributed computing techniques to improve summarization speed for large datasets.

11 Group Contribution

Option 1: We agree that all group members made a valuable contribution and therefore believe it is fair that each member receive the same grade for the discussion.

12 References

1. Papers with Code. ”DialogSum Dataset.” https://paperswithcode.com/dataset/dialogsum.

2. Hugging Face Blog. ”Audio Datasets.” https://huggingface.co/blog/audio-datasets.

3. Voice Tech Podcast. ”Automatic Extractive Text Summarization using TF-IDF.” https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf-3fc9a7b26f5.

4. Google AI Blog. ”Exploring Transfer Learning with T5.” https://blog.research.google/2020/02/exploring-transfer-learning-with-t5.html.

5. Hugging Face. ”facebook/bart-large-cnn.” https://huggingface.co/facebook/bart-large-cnn.

6. F. Falcão. ”Metrics for Evaluating Summarization of Texts Performed by Transformers: How to Evaluate.” https://fabianofalcao.medium.com/metrics-for-evaluating-summarization-of-texts-performed-by-transformers-how-to-evaluate-the-b3ce68a309c3.

7. Towards Data Science. ”Introduction to Text Summarization with ROUGE Scores.” https://towardsdatascience.com/introduction-to-text-summarization-with-rouge-scores-84140c64b471.

8. Streamlit Documentation: https://docs.streamlit.io/.

9. Text Summarization TF-IDF: https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf-3fc9a7b26f5

10. Medium Text Summarization: https://medium.com/@ashins1997/text-summarization-f2542bc6a167

11. Towards Data Science Text Summarization TF-IDF: https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3

12. Analytics Vidhya Text Summarization Text Rank: https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

13. Medium Text Summarization Text Rank: https://medium.com/data-science-in-your-pocket/text-summarization-using-textrank-in-nlp-4bce52c5b390