Text Summarization for Call Center Transcripts

Ishrat Ahmed[1(✉)], Yu Zhou[2], Nikhita Sharma[2], and Jordan Hosier[2]

[1] University of Pittsburgh, Pittsburgh, USA
isa14@pitt.edu
[2] Vail Systems, Inc., Chicago, IL, USA
{yzhou,nsharma,jhosier}@vailsys.com

Abstract. While text summarization of transcripts in call centers is needed for detailed analysis, it presents challenges stemming from the call itself (context switching among speakers, cross talk, etc.) and from the resulting transcript (ASR transcription errors). This work aims to develop a summarization model suitable for on-premise deployment at call centers by fine-tuning pre-trained open-source large language models, assisted with reference summaries generated by GPT-3. The results are analyzed using ROUGE and human evaluation scores, and the correlation of these two metrics is examined. A fine-tuned BART model outputs satisfactory summaries with a human evaluation score of 6.95, approaching the GPT-3 score of 7.69.

Keywords: Summarization · Call Transcript · Large Language Model · GPT-3

1 Introduction

Call centers process a large volume of calls each day, of which only a small portion are selected later for manual review [1]. To analyze these calls (i.e., categorizing issues reported by the customers and identifying gaps and opportunities in provided services), it is necessary to automatically generate a text summary for each call. In recent years, transformer-based large language models (LLM) have shown promise in text summarization. These generative language models are especially skilled in extracting key contents from long documents and producing abstractive summaries. These summaries are well suited for transcripts of conversations between customers and call center agents, as these calls often last many minutes and cover a wide range of topics.

While pre-trained LLMs effectively summarize various text documents, call center transcripts present some unique challenges [2]. The lower-quality audio, often recorded at 8K sampling rate, and noisy environments result in a high word error rate (WER) when transcribed by an automatic speech recognition (ASR) model, as ASR models are commonly trained using higher-quality clean audio datasets such as LibriSpeech [3]. Cross-talk can further confuse both acoustic

K. Arai (Ed.): IntelliSys 2023, LNNS 822, pp. 542–551, 2024.
https://doi.org/10.1007/978-3-031-47721-8_36

and language models within the ASR engine. In addition, multiple topics can be scattered across utterances, so producing a concise summary is non-trivial.

The common LLMs used for text summarization includes BART [5], Pegasus [7], T5 [6], and more recently GPT-3 [4]. While GPT-3 produces a satisfactory summary, it can only be accessed via external APIs, which is unsuitable for call center services due to privacy concerns over personal identifiable information (PII). In this work, we experimented with open-source LLMs to summarize call transcripts that are suitable for deploying on-premise. However, these pre-trained open-source LLMs do not perform well on call transcripts without fine-tuning using domain-specific data. Fine-tuning requires reference summaries of a large number of call transcripts to accommodate the vast variation in call center services. This is an expensive task when using human annotators. As a result, we adopt an approach that leverages the capability of GPT-3 to generate reference summaries for properly redacted call transcripts. These summaries are then used as training samples to fine-tune open-source LLMs so that their output can resemble the quality of GPT-3 on our domain-specific call transcripts. The models are evaluated using ROUGE scores and human evaluation scores, and the correlation between these two metrics is examined.

The main contributions of this work are: (1) fine-tuning open source LLMs for summarization tasks and using GPT-3 to generate ground truth of training samples; (2) demonstrating LLMs can generate summaries for imperfect texts, and (3) analyzing the correlation between ROUGE scores and human evaluation scores across the studied LLMs.

2 Related Work

Automatic text summarization has been extensively studied in the Natural Language Processing (NLP) domain. Traditional summarization methods can be categorized into two classes: extractive and abstractive. The extractive approach selects the most important words and sentences within the original document to generate a summary. In contrast, the abstractive approach generates a whole new summary based on the original text, often including text that doesn't appear in the original document. Early methods such as the TextRank [8] algorithm and Latent Semantic Analysis [9] focused on extractive summarization. More recently, transformer-based LLMs such as BART [5], Pegasus [7], and T5 [6] have been utilized for both types of summarization methods. LLMs are particularly utilized in spoken dialogue summarization. Analysis of such dialogues (e.g., online meetings, customer service calls, etc.) combines speech recognition efforts with text summarization [10].

Much work in summarization has been specifically aimed at call transcripts, as it presents unique challenges (i.e., noisy environments, cross talk, etc.). Chandramouli et al. [11] presented an unsupervised approach to extract meta-data from call transcripts using BERT [12], including key topics and intents to classify transcripts into pre-defined categories. They used an unsupervised method due to the expense of tagging call transcripts. Biswas et al. [13] developed a method

combining topic modeling and sentence selection with punctuation restoration to condense ill-punctuated call transcripts to produce readable extractive summaries. Uma and Sityaev [20] evaluated several extractive text summarization techniques (e.g., Text Rank, BERTSum, etc.) to produce summaries for call center transcripts, focusing in particular on abstractive summaries for call transcripts. Extractive summarization of the call transcripts may be inappropriate due to a high rate of ASR transcription error and multiple topics scattered across utterances from multiple speakers in the transcripts. Stepanov et al. [1] describe an abstractive summarization technique where hand-written templates are filled with entities detected in the transcript using Named Entity Recognition (NER), PoS-tagging, chunking, and dependency parsing.

Our work uses pre-trained and fine-tuned LLMs to generate abstractive summaries of call transcripts directly. Because the call transcripts may cover a wide range of topics, fine-tuning LLMSs will likely generate summaries that provide a wider conversational perspective. To evaluate the model-generated summaries, we compare them against ground truth or reference summaries. Ground truth or reference summaries can be derived manually by human readers, [13], by using the title or heading text, and topic descriptors [14]. Generating human summaries is expensive and non-scalable, while using topic descriptors as ground truth summaries can be vague and may lack details about the call.

Recently, GPT-3 has been used as a source of reference summaries [2,15]. Asi et al. [2] used GPT-3 generated pseudo-labels per call segment, combined with human labels as summaries to fine-tune their model on conversational text. Similarly, Wang et al. [15] leveraged GPT-3 as a reference summary generator. In this vein, we use GPT-3 to generate short summaries for the call transcripts, which are used as the reference for fine-tuning and evaluation.

3 Methodology

3.1 Dataset

The data comes from two PII redacted sources: a financial service and a food ordering service. It consists of 5,452 call transcripts between callers and customer service agents. The separation between customer and agent text is removed in the transcripts, and the dialogue is combined to form a single, long-form paragraph per call. Among these call transcripts, 5,000 are used for training, 389 are set aside as validation datasets for hyper-parameter tuning, and the remaining 63 transcripts are used as the test set. The test set is kept small to facilitate the human evaluation of the model-generated summaries.

3.2 Experiment Details

Pre-Trained Models We use pre-trained BART [5], Pegasus [7], and T5 [6] models for summarization. Both BART and Pegasus models are trained on the Extreme Summarization (XSum) [16] dataset. The data used to pre-train the

Table 1. Pre-Trained Models

Models	Description	Parameters
bart.large.xsum	bart.large fine-tuned on Xsum	400 M
pegasus.xsum	pegasus fine-tuned on Xsum	568 M
T5-small	T5 pre-trained on C4	60 M

T5 model is known as C4 (Colossal Clean Crawled Corpus (700GB)) [17]. These models are summarized in Table 1.

Fine-Tuned Models We fine-tune both BART and T5 models for summarization and compare their summaries against GPT-3 generated summaries. To do this, GPT-3 is tasked with generating a short summary for each call. We use this summary as the ground truth. For each transcript, the following question-based prompt is concatenated with the transcript's content: "what is a one-sentence Tl;dr of this call:". GPT-3 parameters *temperature*, *top-p*, *frequency* and *presence* are adjusted to obtain optimal results. Huggingface transformers library [18], along with FastAI and Blurr packages, are used to fine-tune the models based on the call transcripts and corresponding reference summaries. The best results on the validation set are seen after fine-tuning three epochs for BART and eight epochs for T5. After training, the model inference is conducted on the test dataset.

Evaluation Most studies in text summarization use Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [19] score as the primary evaluation metric. ROUGE score is used to approximate the similarity of the model-generated summaries with the reference summaries. It consists of F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L that measure the word overlap, bi-gram overlap, and longest common sequence between the ground truth and the generated summary, respectively. ROUGE score has limited capability of capturing semantic similarities such as paraphrasing, which is common in abstractive summarization. In our work, the ROUGE score and human evaluation are both used to investigate model performance. Human evaluation is important to evaluate the quality of the summaries as models are usually optimized for metrics like ROUGE. During the human evaluation, the model-generated summaries were manually verified for content and readability by seven different evaluators.

4 Results

In this section, BART and T5 stand for BART-large-xsum and T5-small models, respectively. Table 2 shows a sample call transcript along with summaries generated by various pre-trained and fine-tuned models, in each case demonstrating apparent ASR transcription errors. In this particular example, pre-trained Pegasus and BART models do not produce summaries that are relevant to the transcript, while both pre-trained and fine-tuned T5 models show tendencies of yielding extractive summaries. Among all models, GPT-3 generates the highest quality summary, as expected.

Table 2. Example Summaries Generated by the Fine-Tuned Models

Original transcript (abbreviated)	...if you would prefer to prepay please have your credit card ready at the end of the call...what can I get for you today. Can I get a kids low mane with chicken. No vegetables. And then a kids low Maine with steak no vegetables. Into Santa tie Donna or fudge brownie today...that bring your total to eleven fifty two and that'll be ready for pickup and about twelve minutes.
Pre-trained Pegasus	If you have any questions about this or any of our other shows please call us at 800-273-3217 800-273-3217 800-273-3217 800-273-3217 or email us at 800-273-3217 800-273-3217 800-273-3217 800-273-3217.
Pre-trained BART	For the first time in more than a decade we have an online ordering option for you to play the game Pokemon Go on your mobile phone or tablet.
Fine-tuned BART	The message is asking if the person would like to prepay for their order, and if so, to have their credit card ready. The total cost of the order is $11.52.
Pre-trained T5	I have one order of the kids old man with chicken no vegetables and then a kids low Maine with steak no veggies, and another order the children bowl made from steak.
Fine-tuned T5	If you would like to prepay for prepay, please have your credit card ready at the end of the call.
GPT-3	The caller is ordering two kids meals, one with chicken and one with steak, and requests no vegetables. The total cost will be $11.52 and the food will be ready for pickup.

4.1 Quantitative Results

Since GPT-3 output is used as the reference in this work, we evaluate the summaries generated by all other models by comparing them with the GPT-3 result for a given call transcript. The average ROUGE-1, ROUGE-2, and ROUGE-L F1-scores for the test set across four models, pre-trained and fine-tuned BART-large and T5-small, are presented in Table 3. Note we also experimented briefly with the pre-trained Pegasus summarization model, however, it does not gen-

Table 3. Rouge Scores Across the Models

Models	Rouge-1	Rouge-2	Rouge-L
Pre-trained BART	24.5	5.7	16.6
Fine-tuned BART	**37.8**	**16.8**	**31.0**
Pre-trained T5	21.9	5.1	15.2
Fine-tuned T5	30.6	10.3	24.1

eralize well to the call center dataset used in this study. Therefore, it is not fine-tuned nor included in subsequent analysis.

As seen in Table 3, for both BART and T5, a model fine-tuned with our domain-specific dataset shows significant improvement over the pre-trained model. This is not surprising because call transcripts have unique characteristics, such as ASR transcription errors and various topics scattered over short utterances by multiple speakers, which are not represented in typical LLM training data. Overall, fine-tuned BART-large-xsum model exhibits the highest ROUGE scores. It is worth noting that a high ROUGE score only indicates close resemblance to the reference text generated by GPT-3, which does not necessarily ensure a high-quality summary. For that purpose, human evaluation is needed.

4.2 Qualitative Results

To better estimate the model efficacy, we employed domain experts to conduct a human evaluation of the summary quality. Each generated summary, with the model name anonymized, is read by multiple reviewers and receives a score in the range of [1, 11] from each reviewer, where higher values reflect more satisfactory summaries.

Figures 1, 2 and 3 present the histograms of human evaluation scores for the test set samples for each model. Scores for GPT-3 outputs are predominantly in the range of [9, 11]. In contrast, most of the scores produced by pre-trained BART and T5 are found at the lower end of the range. After fine-tuning using domain-specific data, both BART and T5 exhibit notable improvements. In particular, fine-tuned BART in Fig. 2 demonstrates a score distribution similar to that of GPT-3 in Fig. 1, which is further supported by Table 4. These results indicate that, per human perception, the quality of call transcript summaries generated by this fine-tuned BART model approaches that of GPT-3. Thus, it is a candidate suitable for on-premise deployment in call center applications.

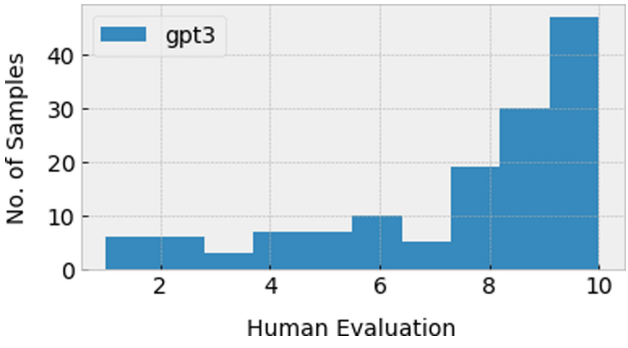

Fig. 1. Human Evaluation Scores for Summaries Generated by GPT-3

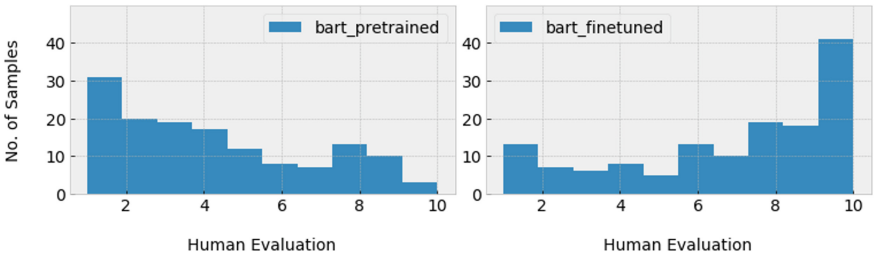

Fig. 2. Human Evaluation Scores for Pre-Trained and Fine-Tuned BART

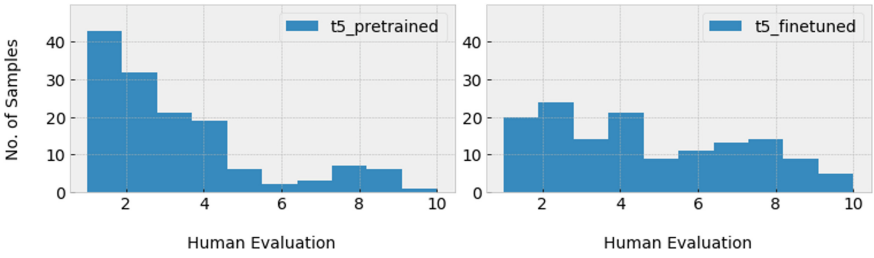

Fig. 3. Human Evaluation Scores for Pre-Trained and Fine-Tuned T-5

Table 4. Mean Human Evaluation Scores Across the Models

Models	Avg score
GPT-3	**7.69**
Pre-trained BART	4.12
Fine-tuned BART	**6.95**
Pre-trained T5	3.06
Fine-tuned T5	4.56

4.3 Comparison of ROUGE Score and Human Evaluation Score

As discussed earlier, in this study, ROUGE score is not a direct measure of the summary quality, rather, it assesses how much the model-generated text matches the GPT-3 output. With recent adoptions of leveraging the output from a state-of-the-art LLM such as GPT-3 as ground truth to fine-tune or domain-adapt a smaller model for application deployment [2,15], it is worthwhile to investigate whether an evaluation metric such as ROUGE based on the model-generated ground truth summaries can still reflect the benchmark it is intended to measure.

Figure 4 displays ROUGE-2 (ROUGE-1/L omitted to reduce clutter) and the human evaluation scores of all models, using the data from Tables 3 and 4. It shows qualitative agreement between ROUGE and human evaluation, namely, the order remains the same when ranking the models by ROUGE score and by human evaluation score.

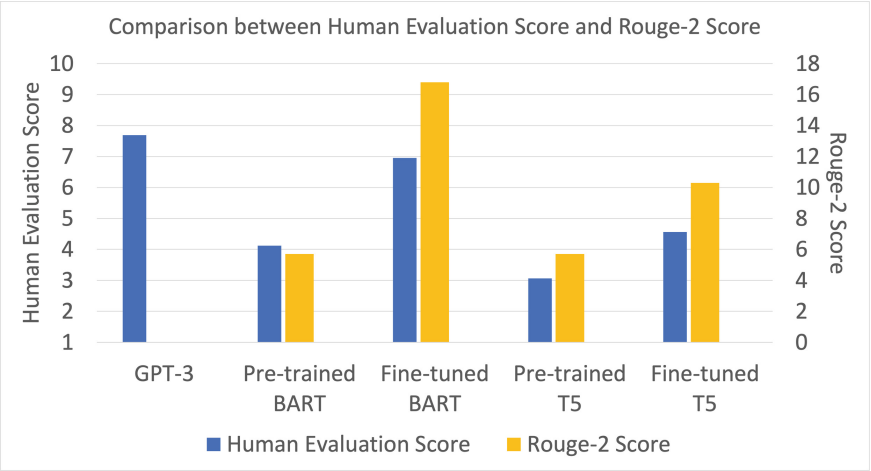

Fig. 4. ROUGE Score and Human Evaluation Score Across the Models

For quantitative comparison, the correlation coefficient between ROUGE scores and human evaluation scores of all samples in the test set is computed for each model, with one consideration: for a specific test sample, if the reference summary generated by GPT-3 is deficient, then there is no reason to expect a close match (i.e., a high ROUGE score) from the summary of another model regardless of its actual quality. Therefore, to understand how ROUGE scores correlate with human evaluation scores, a more meaningful result can be obtained by filtering out test samples with low-quality reference summaries from GPT-3.

Table 5. Correlation between ROUGE Scores and Human Evaluation Scores

Models	Pre-trained BART	Fine-tuned BART	Pre-trained T5	Fine-tuned T5
ROUGE-1	0.48	0.58	0.19	0.72
ROUGE-2	0.35	0.43	0.12	0.69
ROUGE-L	0.50	0.50	0.26	0.59

Table 5 presents the correlation coefficient for samples in the test set for which GPT-3 produces summaries with human score >8.0. Fine-tuned models result in higher correlation between ROUGE and human scores than pre-trained models. Interestingly, the highest correlation is observed for the fine-tuned T5 model, even though its efficacy is lower than that of fine-tuned BART (Fig. 4). This is a consequence of T5 being more extractive, i.e. key phrases in the original text are selected as summary directly, which impacts both the ROUGE score and human evaluation score. On the other hand, the correlation is weaker for a more abstractive model such as BART, where a concise summary yielding a high human evaluation score doesn't necessarily contain the exact phrases found

in GPT-3 output. Therefore, the ROUGE score is better suited to evaluate extractive summaries than abstractive summaries. However, when a model is ineffective, the ROUGE score computation is dominated by the random matches between irrelevant phrases in its output and the reference summary, resulting in a low correlation between ROUGE and human evaluation, as observed in Table 5 for the pre-trained T5 model.

5 Conclusion

This work aims to develop a text summarization model for call transcripts that can be deployed on-premise at call centers, and demonstrate that LLMs can generate satisfactory summaries for deficient transcription text via fine-tuning. To overcome the unique challenges presented by call transcripts (e.g., high ASR transcription errors and multiple topics scattered across utterances from multiple speakers), pre-trained text summarization language models are fine-tuned using call center transcripts, assisted with GPT-3-generated reference summaries. This approach suggests that LLMs have the potential to generate a reasonably good summary of such imperfect texts, however, it does require fine-tuning. Fine-tuned BART-large-xsum model is found to output summaries with high ROUGE scores as well as satisfactory human evaluation results approaching that of GPT-3. In addition, we examine how ROUGE scores based on reference text generated by GPT-3 compare with human evaluations of the quality of text summaries and find qualitative agreement in model rankings using these two evaluation metrics. Moreover, their correlation exhibits a larger variation with model efficacy when the model summary is more extractive than abstractive. Additional evaluation metrics, such as relevance and factuality will be examined in a future study.

References

1. Stepanov, E., Favre, B., Alam, F., Chowdhury, S., Singla, K., Trione, J., Bechet, F., Riccardi, G.: Automatic summarization of call-center conversations. In: Conference: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015) (2015)
2. Asi, A., Wang, S., Eisenstadt, R., Geckt, D., Kuper, Y., Mao, Y., Ronen, R.: An End-to-End Dialogue Summarization System for Sales Calls (2022). arXiv:2204.12951
3. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. IEEE (2015)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Amodei, D.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
5. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Zettlemoyer, L.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019). arXiv:1910.13461

6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2019)
7. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning. PMLR (2020)
8. Mihalcea, R., Tarau, P.: Textrank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
9. Steinberger, J., Jezek, K.: Using latent semantic analysis in text summarization and summary evaluation. Proc. ISIM **4**(8), 93–100 (2004)
10. El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K.: Automatic text summarization: a comprehensive survey. Expert Syst. Appl. **165**, 113679 (2021)
11. Chandramouli, A., Shukla, S., Nair, N., Purohit, S., Pandey, S., Dandu, M.M.K.: Unsupervised paradigm for information extraction from transcripts using BERT (2021). arXiv:2110.00949
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). arXiv:1810.04805
13. Biswas, P.K., Iakubovich, A.: Extractive summarization of call transcripts (2021). arXiv:2103.10599
14. Goo, C.W., Chen, Y.N.: Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 735–742. IEEE (2018)
15. Wang, S., Liu, Y., Xu, Y., Zhu, C., Zeng, M.: Want to reduce labeling cost? GPT-3 can help (2021). arXiv:2108.13487
16. Narayan, S., Cohen, S., Lapata, M.: Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1797–1807. Association for Computational Linguistics (2018)
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 1–67 (2020)
18. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Rush, A.M.: Huggingface's transformers: state-of-the-art natural language processing (2019). arXiv:1910.03771
19. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
20. Uma, A.N., Sityaev, D.: Comparing Methods for Extractive Summarization of Call Centre Dialogue (2022). arXiv:2209.02472