INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Tailored Real-Time Call Summarization System for Contact Centers

Aashraya Sachdeva, Sai Nishanth Padala, Anup Pattnaik, Varun Nathan, Cijo George, Ayush Kumar, Jithendra Vepa

Observe.AI, India

{aashraya.sachdeva, sai.nishanth, anup.pattnaik, varun.nathan cijo.george, ayush, jithendra}@observe.ai

Abstract

Contact centers are critical for delivering high-quality customer service to various businesses. Call summarization is a crucial task for contact center agents for compliance, to transfer contextual information to the next agent, or to serve as a reference for future interactions. Agents spend a substantial amount of time writing notes on or after a call, which reduces their productivity and adds to the cost per call. While there exist various pre-trained Large Language Models (LLM) for summarization, they often lack coverage of domain-specific information relevant to businesses. We propose a hybrid streaming notes generation system leveraging the generative capabilities of an LLM fine-tuned for contact center call summarization, but allowing businesses to focus notes generation around events of business interest. Our system reduces after-call work for agents by not only generating notes out-of-the-box but also allowing agents to edit them in real time due to its streaming nature.

Index Terms: real-time summarization, contact center efficiency, customizable summary

1. Introduction

Contact centers play a vital role in delivering top-quality customer service for various businesses and industries. Call summarization is often an essential task for contact center agents for compliance requirements, to pass on the context to the next agent, or simply for future reference. Agents spend a substantial amount of time writing notes on or after a call, which reduces their productivity and adds to the cost per call for businesses. For contact centers handling tens of thousands of calls a day, this has a significant impact on their overall operational cost.

Various software solutions offer call summarization through pre-trained deep-learning models. These models are trained on vast amounts of data and can generate summaries that accurately capture the overall essence of the conversation. However, they often lack coverage of domain-specific information relevant to businesses and are often not easily customizable to the needs of a business.

To address this problem, we propose a hybrid real-time streaming notes generation system leveraging the generative capabilities of a Large Language Model (LLM) fine-tuned for contact center call summarization, but also allowing businesses to focus notes generation around intents/events of business interest. These intents are often defined using key phrases and identified through automated keyword-spotting (KWS). The intent key phrases identify the most relevant parts of the call for the business while the call is happening, and this is fed into the summarization system to generate a summary snippet based on the context around the intent. Summary snippet/note generation based on only local context enables the solution to summarize in a streaming fashion.

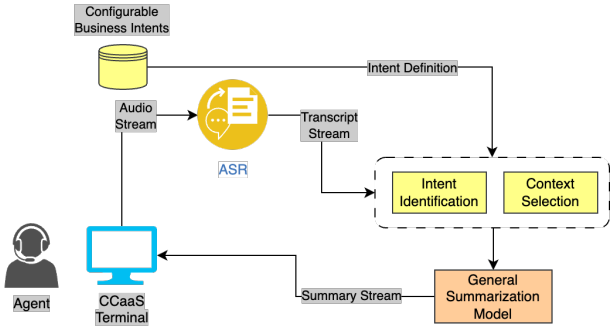

Figure 1: *System for Tailored Streaming Summarization*

Instruction: What action is taken in this dialog? ### Input: agent: okay so once this completed and you can give us a callback and we'll have your final approval in minutes ### Response: Agent suggests customer to call back for final approval after completion.

Table 1: *Sample Instruction for Training General Summarization Model. The model is trained to generate text in green with text in red as input.*

Our system reduces after-call work for agents by not only generating notes out-of-the-box but also allowing agents to edit them in real-time due to their streaming nature. By reducing the time agents spend writing notes, our solution can enhance the overall efficiency and effectiveness of call center operations. The system can also improve customer satisfaction by ensuring that agents have accurate and relevant information at their disposal. The customizable nature of our proposed solution allows businesses to tailor the summaries to their specific requirements, making it a more versatile option for contact centers.

2. Summarization System

Figure 1 shows the design of the proposed system. The key components of the system are detailed in the following sections.

2.1. General Summarization Model

NLP literature has recently witnessed a tremendous amount of activity in building models that can follow natural language instructions [1]. Following a similar methodology, we fine-tune

Agent Greeting	Customer Supervisor Escalation
have a good day	speak with owner
have great weekend	speak with supervisor
take care	speak with manager
have a nice evening	connect to one of your manager

Table 2: *Example of intents and their definitions.*

an open-source Large Language Model (LLM) on a curated instruction dataset specifically designed for contact center call summary generation and encompasses a wide range of conversation scenarios. The instruction dataset is assembled to cover multiple industry verticals to ensure that the resulting model is not only diverse but also highly adaptable to the unique needs of different business sectors. An example data point is shown in Table 1. Particularly, we fine-tune a 6.7B Cerebras-GPT [2] for 3 epochs with an instruction dataset of 10K samples. We use AdamW optimizer with a learning rate of 2e-5 and cosine decay. Generated summaries were evaluated through manual data annotation. By intentionally utilizing a smaller context for training, we enable real-time streaming predictions, significantly enhancing the system's responsiveness and efficiency in practical settings.

2.2. Custom Intent Definition and Identification

Expanding on the base summarization model's ability to create summaries that encapsulate the main activities of a call, it is crucial to recognize that not all actions may be pertinent to a specific contact center. Retraining the model for each contact center turns out to be an impractical solution. To tackle this customization issue, we enable users to specify their business intent through a collection of key phrases customized to their distinct requirements. Subsequently, we develop an in-memory keyword-spotting (KWS) algorithm [3], skilled at identifying specific business intents in real-time. This technique allows the general summarization model to produce summaries only when a relevant business intent is recognized, guaranteeing that the output corresponds to the contact center's unique requirements. Table 2 shows examples of custom business intent. To create a fluent natural language summary based on the specified business intent, we first use the KWS algorithm to identify keyphrases in the transcript stream. Once a relevant keyphrase is found, we dynamically select a context window around the intent. The General Summarization Model then generates a summary of the selected context, which is sent to the contact center agents' console, giving them real-time, relevant information tailored to the organization's specific needs.

3. Results and Application

Table 3 presents an example of a generated summary for a 10-minute voice call, effectively capturing essential details such as the call's purpose, agent greetings, and specific named entities. Importantly, the summary intentionally omits any PCI/PII-related information, such as phone numbers, due to compliance requirements. This is achieved by biasing the model during training to avoid producing such sensitive entities. The generated summary offers significant advantages in two primary scenarios. Firstly, it reduces after-call work, a process where agents are expected to write call summaries after each interaction. By automating this task, and generating it in real-time enabling agents to refine it on the go, the average call handling time is decreased, leading to increased efficiency per agent. Secondly, the summary proves invaluable when a call is transferred between agents. Presenting each agent with a concise summary of the conversation thus far ensures that the context is seamlessly passed on to the new agent, facilitating a smooth transition.

- The customer is calling to inquire about getting images for a mutual patient.
- The agent is asking for the patient's date of birth.
- The customer was inquiring about images of the lumbar spine and the agent was trying to get the images sent.
- The agent provided the customer with the fax number to send the request to and the customer placed the agent on hold to see what they could do.
- This call was to promote the services of the Memorial Medical Group, which collaborates with doctors to provide access to care, decreased wait times, and reduced paperwork.
- The group is thankful for the opportunity to serve customers and look forward to speaking with them.
- The customer was inquiring about a free colon cancer screening kit and the agent provided the customer with the contact information for the Leah M Fit Cancer Center.
- The agent also thanked the customer for their service in the healthcare industry.
- The customer provided the agent with a phone number and the agent thanked the customer for their help.
- The customer wished the agent a good day and the call ended.

Table 3: *Generated Summary for a 10 min voice call.*

4. Conclusion

Contact centers play a vital role in delivering exceptional customer service across a variety of businesses and industries. While deep-learning models have been employed to generate call summaries, they often fall short of addressing the unique needs of individual customers. Our research proposes a hybrid system that provides an out-of-box summary and allows for the integration of business-specific information through user queries in the form of keywords. This customizable solution enables businesses to align summaries with their specific goals, resulting in improved operational efficiency and customer satisfaction. The streaming nature of the solution also ensures that agents can edit/refine in real-time while on the call, thereby avoiding after-call work for summarization.

5. References

[1] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3470–3487. [Online]. Available: https://aclanthology.org/2022.acl-long.244

[2] N. Dey, G. Gosal, Zhiming, Chen, H. Khachane, W. Marshall, R. Pathria, M. Tom, and J. Hestness, "Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster," 2023.

[3] A. Białecki, R. Muir, G. Ingersoll, and L. Imagination, "Apache lucene 4," in *SIGIR 2012 workshop on open source information retrieval*, 2012, p. 17.