Speaker Role Identification in Call Center Dialogues: Leveraging Opening Sentences and Large Language Models

Anonymous ACL submission

Abstract

This paper addresses the task of speaker role identification in call centre dialogues, focusing on distinguishing between the customer and the agent. We propose a text-based approach that utilises the identification of the agent's opening sentence as a key feature for role classification. The opening sentence is identified using a model trained through active learning. By combining this information with a large language model, we accurately classify the speaker roles. The proposed approach is evaluated on a dataset of call centre dialogues and achieves high accuracy. This work contributes to the field by providing an effective solution for speaker role identification in call centre settings, with potential applications in interaction analysis and information retrieval.

1 Introduction

Speaker role identification is a fundamental process that involves recognizing different speaker roles in a conversation. Its significance has grown in various settings, such as call centers, where distinguishing between agents and customers in a call transcript is critical. Speaker role identification has numerous uses, including interaction and dialogue analysis, summarisation, and information retrieval (Lavalley et al., 2010; Jahangir et al., 2021). This paper concentrates exclusively on speaker role identification in call centers. An example of the input and output of this process is demonstrated in Table 1.

In our application, speaker role identification takes place after speaker diarisation, where speaker turn information has been added to the transcripts. Within a call centre dialogue, two specific roles are present: the customer and the agent. While a typical call involves a single customer and a single agent, it is common for calls to involve more than one customers (as in Table 1), or more than one agents (such as when an agent transfers a customer to another agent). Identifying speaker roles is a challenging task due to various factors, such as transcription errors, interruptions, repetitions, multi-party conversations, and diverse topics.

Sample input dialogue:	
Person_01	hello
Person_02	hello good morning is that [NAME]
Person_01	if you hang on a sec while i just get him
Person_02	sorry
Person_01	who's calling
Person_02	it's [NAME] calling you from [ORG] it's for an application
Person_03	oh hi there hi
Person_02	hi [NAME] it's just for an application
Person_03	yes this is [NAME] yes . . .

Table 1: An example input of identifying speaker roles. The output should indicate that Person_01 is the Customer, Person_02 is the Agent, and Person_03 is the Customer.

Numerous studies have been undertaken to address the issue of speaker role identification. These efforts involve utilising text-based features (Barzilay et al., 2000; Liu, 2006; Wang et al., 2011; Sapru and Valente, 2012; Flemotomos et al., 2019), or employing multimodal approaches that integrate both text and audio features (Rouvier et al., 2015; Bellagha and Zrigui, 2020; Guo et al., 2023). In both cases, the goal is to classify each speaker in a conversation into a predefined role category. This classification task is typically accomplished using machine learning algorithms that are trained on labeled datasets, which consist of conversations where each speaker is annotated with their corresponding role category. In this paper, we focus on the text-based approach and formulate the task as a binary classification problem, with the categories being "customer" and "agent".

In call centre dialogues, distinguishing between the agent and the customer can be achieved by exploiting the language differences between them. The call centre agent typically starts the conversation by introducing themselves as a representative of their company or organization, which is referred to as the "opening sentence". We propose utilising the identification results of the opening sentence to identify the speaker roles. By combining this information with a large language model, we can accurately classify the speaker roles in call centres.

This paper makes the following two key contributions:

- We propose a model for predicting the opening sentence used by call centre agents, and provide details on how to efficiently construct the training data for this task using active learning.
- We introduce a practical approach for identifying the speaker roles in call centre dialogues by combining the opening sentence identification with a large language model.

The remainder of this paper is organised as follows. Section 2 provides a brief overview of the related work. Section 3 presents details of our methodology. Section 4 describes the experimental results and discussions. Section 5 concludes the paper and points to avenues for future work.

2 Related Work

Text-based speaker role identification often takes the form of a text classification task, aiming to categorise each speaker in a conversation into predefined role categories. Traditionally, text classification has been accomplished using machine learning algorithms trained on labeled datasets. However, with the advent of the Transformer neural network (Vaswani et al., 2017), many studies have adopted pre-trained large language models for text classification (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019). Fine-tuning these pre-trained models still requires a certain amount of labeled data. Active learning provides a means to quickly build labeled data by involving the model in the data labeling process (Settles, 2010).

On the other hand, zero-shot text classification is an approach that requires no labeled data at all (Pourpanah et al., 2022). In this method, a model is trained on a set of existing labeled examples and can subsequently classify new examples from previously unseen classes. This offers the advantage of categorising text into arbitrary categories without the requirement of data preprocessing and training. BART (Lewis et al., 2020), BLOOM (Muennighoff et al., 2022), and FLAN-T5 (Wei et al., 2021) are notable pre-trained large language models available for research purposes, offering the ability to perform zero-shot learning.

3 Method

3.1 Opening sentence Identification

Our approach for identifying the opening sentence involves using active learning methods to acquire the necessary labelled data and constructing a classifier by fine-tuning a pre-trained large language model with the labelled data.

3.1.1 Data Preparation

An active learning approach was employed to construct labeled data for opening sentence identification using a dataset of 437,135 utterances extracted from 67,719 dialogues from 10 call centres. The initial seed set of 100 samples was manually annotated using keyword searches with phrases like "calling from" and "speaking to". The identified key phrases were combined with negative examples to form a seed set. Following that, the seed set was used to train SVM classifiers, utilizing two distinct embedding strategies: BERT (Bidirectional Encoder Representations from Transformers) sentence embedding and TF-IDF. This selection was primarily made to facilitate rapid training/retraining of the classifiers during the labeling process.

The classifiers score each unlabelled sample, and based on those scores, human annotators decide which samples to label using a combination of two sampling strategies: Expected model change and Query-by-Diversity. Given the dataset's substantial class imbalance with only a few positive samples, the focus was on labeling positive samples. This approach aimed to identify the opening sentences that were most likely to have a significant impact on improving the current model. However, Query-by-Diversity sampling (Kee et al., 2018) was also employed to ensure a diverse range of opening sentences was identified. The classifiers underwent retraining either after labeling every 100 samples or when no samples had a score exceeding a threshold (0.7 in our specific case).

3.1.2 Classification

The system employed to identify the opening sentence comprises three key components: an input layer, a BERT model, and a classification layer. In this process, the input layer receives an utterance from the dialogue, and the input representation is generated by incorporating the corresponding token, segment, and position embeddings. The procedure adheres to the recommendations outlined in the work by (Devlin et al., 2019). A fully connected neural network, positioned on top of the BERT output, functions as the classification layer to determine whether the utterance is an opening sentence or not. During the training phase, the BERT layer is initialized with pre-trained parameters, and all parameters are then fine-tuned using labeled data from the data preparation step.

3.2 Speaker Role Identification

The FLAN-T5 model is used as the baseline, using a zero-shot prompting approach. Although other models could be utilized, our experiments reveal that the FLAN-T5 yields the most favorable outcomes. The prompt provided to the model is

```
Based on the utterances above,
{speaker} is
OPTIONS
- an agent from a call centre
- a customer
```

By inputting the utterances of each speaker, the model is able to assign them a role, either "customer" or "agent". This process is repeated for all speakers in the dialogue. Additionally, experiments were conducted using the entire conversation as input, and results for both approaches are reported.

To identify the role of speakers in a dialogue, we use a combination of the FLAN-T5 model and the opening sentence identification approach. First, we identify the opening sentences of the dialogue and designate their speakers as "agents". If a speaker does not have an opening sentence, they are labeled as "customers". However, in cases where there are no agents (i.e., no opening sentences detected) or no customers (i.e., all speakers have an opening sentence), we rely on the FLAN-T5 model to assign speaker roles. By combining the strengths of both approaches, we can improve the accuracy and reliability of speaker role identification.

4 Experimental Results and Discussions

4.1 Dataset

The opening sentence identification dataset consists of 67,719 dialogues, encompassing 437,135 utterances across 10 domains. These domains primarily revolve around call center scenarios, such as mobile phones, insurance, and pet food companies. A total of 867 opening utterances were labeled as positive examples, indicating they were opening sentences, while 1,982 utterances were labeled as negative examples, representing non-opening sentences, using the active learning approach. A subset of 321 dialogues from seven domains was selected for speaker role identification, which includes speaker diarisation information.

4.2 Opening sentence identification

To ensure balanced representation of positive and negative samples, we divided the opening sentence identification data into a train set and a test set, following an 80-20 split while maintaining an equal ratio of positive and negative samples between the two sets. Since the data was generated through active learning, there is a potential bias due to the deliberate selection of samples for labelling. To address this, we generated an additional test set by randomly selecting 100 dialogues and manually assigning labels to them. We presented the results obtained from the SVM classifiers as well as the classification performance using BERT (bert-base-uncased and bert-large-uncased). We trained the BERT classifier model for 3 epochs.

Table 2: The accuracy, precision, recall and F1 scores of different classifiers on opening sentence identification

Method	Acc	Pre	Rec	F1
Test Set				
SVM-TF-IDF	91.78	86.88	80.35	83.48
SVM-BERT	94.92	**92.12**	87.86	89.94
BERT base	**96.41**	92.09	94.22	**93.14**
BERT large	95.37	86.60	**97.11**	91.55
Additional Test Set				
SVM-TF-IDF	99.88	**98.91**	86.67	92.39
SVM-BERT	99.85	91.43	91.43	91.43
BERT base	99.05	94.23	93.33	**93.78**
BERT large	**99.89**	90.27	**97.14**	93.58

The SVM-TF-IDF method achieved an accuracy of 91.78%, highlighting its proficiency in accurately identifying opening sentences. In contrast,

the SVM-BERT approach outperformed the SVM-TF-IDF method with an accuracy of 94.92%. This improvement can be attributed to the utilization of BERT embeddings, which incorporate the semantic meaning and contextual information of words. However, the SVM-BERT approach only utilizes the last layer of BERT for embedding, resulting in slightly lower performance compared to other BERT models.

Among the evaluated methods, the BERT base model achieved the highest accuracy of 96.41%. This demonstrates the effectiveness of leveraging pre-trained language models like BERT for opening sentence identification. Although the BERT large model achieved a slightly lower accuracy of 95.37% compared to BERT base, it excelled in recall with a score of 97.11%. This indicates its strength in correctly identifying positive samples, albeit with a slightly lower precision compared to BERT base. Furthermore, the precision, recall, and F1 scores are notably high, highlighting a well-balanced trade-off in accurately identifying both positive and negative samples. The results obtained from the additional test set further validate this observation.

4.3 Speaker Role Identification

For speaker role identification, a subset of 321 dialogues from seven domains was selected. The evaluation focused on measuring the accuracy of two approaches: FLAN-T5 and the combined use of opening sentence identification and FLAN-T5. The results obtained from these evaluations are presented in Table 3.

Table 3: Accuracy of different approaches on speaker role identification

Method	Acc
FLAN-T5-Large dialogue as context	70.17
FLAN-T5-Large utterances	81.25
FLAN-T5-XL utterances	86.36
Using Opening Sentence	89.49
Opening Sentence + FLAN-T5-XL	**93.61**

FLAN-T5-Large, when considering the whole dialogue as context, achieved an accuracy of 70.17%. However, when using utterances from a specific speaker as context, FLAN-T5-Large demonstrated improved performance with an accuracy of 81.25%. This approach outperformed the dialogue-level context approach, highlighting the benefits of considering individual utterances. The FLAN-T5-XL variant achieved an accuracy of 86.36%, surpassing the previous approaches. This improvement can be attributed to its larger model configuration, which enhances its ability to capture complex patterns and representations. While FLAN-T5-XXL has the potential to produce superior results, it was not feasible to incorporate it into our current infrastructure.

The utilisation of the opening sentence identification approach resulted in an accuracy of 89.49%. This method leverages labelled data, providing an advantage over FLAN-T5, which is a zero-shot approach. Combining the opening sentence identification with FLAN-T5-XL yielded the highest accuracy of 93.61%. This combination proves to be the most effective for accurate identification. Upon analysing the incorrect cases, we identified several primary causes: (1) Inaccurate speaker identifiers in the input data, particularly due to speech diarisation errors. (2) Complex contextual scenarios that pose challenges even for human understanding. (3) Instances where agents engage in conversations with each other, making it difficult to distinguish their roles. (4) Situations involving business numbers being contacted, which often share the same opening sentence pattern and are prone to misidentification as agents.

5 Conclusion

This paper proposes a text-based approach for speaker role identification in call centre dialogues. By combining the identification of the agent's opening sentence with a large language model, our approach achieves high accuracy in classifying speaker roles. This has practical implications for call centre applications, enabling improved customer-agent interaction analysis and call pattern analysis.

The use of active learning allows for efficient construction of the training dataset for opening sentence identification. Integrating this information into the classification process significantly improves the accuracy of speaker role identification.

Future work can explore enhancements to the system, such as incorporating additional contextual features and exploring multimodal approaches. Evaluating the approach on larger and more diverse datasets would also provide a better understanding of its generalisability.

References

Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *AAAI/IAAI*, pages 679–684.

Mohamed Lazhar Bellagha and Mounir Zrigui. 2020. Speaker naming in tv programs based on speaker role recognition. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Linguistically aided speaker diarization using speaker role information. *arXiv preprint arXiv:1911.07994*.

Dongyue Guo, Jianwei Zhang, Bo Yang, and Yi Lin. 2023. A comparative study of speaker role identification in air traffic communication using deep learning approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–17.

Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, and Ihsan Ali. 2021. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, 171:114591.

Seho Kee, Enrique Del Castillo, and George Runger. 2018. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454:401–418.

Rémi Lavalley, Chloé Clavel, Patrice Bellot, and Marc El-Beze. 2010. Combining text categorization and dialog modeling for speaker role identification on call center conversations. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yang Liu. 2006. Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 81–84.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*.

Michael Rouvier, Sebastien Delecraz, Benoit Favre, Meriem Bendris, and Frederic Bechet. 2015. Multimodal embedding fusion for robust speaker role recognition in video broadcast. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 383–389. IEEE.

Ashtosh Sapru and Fabio Valente. 2012. Automatic speaker role labeling in ami meetings: recognition of formal and social roles. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5057–5060. IEEE.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey. 2011. Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5556–5559. IEEE.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.