# Improving Role-Oriented Dialogue Summarization with Interaction-Aware Contrastive Learning

**Weihong Guan**[1], **Shi Feng**[1*], **Daling Wang**[1],
**Faliang Huang**[2], **Yifei Zhang**[1], **Yuan Cui**[3]
[1]Northeastern University, Shenyang, China
guanweihong@stumail.neu.edu.cn, {fengshi, wangdaling, zhangyifei}@cse.neu.edu.cn
[2]Guangxi Key Lab of Human-machine Interaction and Intelligent Decision, Nanning Normal University
Nanning, China
faliang.huang@gmail.com
[3]Shenyang Polytechnic College, Shenyang, China
cyuan401@163.com

**Abstract**

Role-oriented dialogue summarization aims at generating summaries for different roles in dialogue, e.g., user and agent. Interaction between different roles is vital for the task. Existing methods could not fully capture interaction patterns between roles when encoding dialogue, thus are prone to ignore the interaction-related key information. In this paper, we propose a contrastive learning based interaction-aware model for the role-oriented dialogue summarization namely CIAM. An interaction-aware contrastive objective is constructed to guide the encoded dialogue representation to learn role-level interaction. The representation is then used by the decoder to generate role-oriented summaries. The contrastive objective is trained jointly with the primary dialogue summarization task. Additionally, we innovatively utilize different decoder start tokens to control what kind of summary to generate, thus could generate different role-oriented summaries with a unified model. Experimental results show that our method achieves new state-of-the-art results on two public datasets. Extensive analyses further demonstrate that our method excels at capturing interaction information between different roles and producing informative summaries.

**Keywords:** Dialogue Summarization, Role Interation, Contrastive Learning

## 1. Introduction

Dialogue summariztaion aims at condensing a long dialogue into a short summary (Feng et al., 2021). In the era of information explosion, dialogue summarization is valuable as it can quickly capture the key information in dialogue which promotes productivity. In many real life dialogue scenarios, each speaker has an official role (e.g. user or agent) acting for corresponding responsibility (e.g. raise questions or give suggestions). Different roles interact with each other to achieve goal. It is equally important to summarize the main content for each role in addition to the whole dialogue. Lin et al. (2021) propose the role-oriented dialogue summarization task and construct a related dataset CSDS based on customer service dialogues. Figure 1 shows an example from CSDS. Apart from an overall summarization for the whole dialogue (Final Summary), summaries for both user and agent are also required, which could reflect the user's demand and instruct the agent to handle similar issues.

As discussed in Lin et al. (2022), interaction between different roles is crucial for the role-oriented dialogue summarization. First, interaction can help track the key information scattered across different roles. As shown In Figure 1, to generate the first sentence in the agent summary, the model not only needs to focus on the key utterance of the

| Dialogue | |
|---|---|
| 1 Q: 我**购买的商品可以更换地址**吗? (Can I **change the address of the goods I purchased**?)<br>2 A: **下单之后是改不了**的呢。(It **cannot be changed after placing the order**.)<br>3 Q: 那我取消吧。(Then I'll cancel.)<br>4 A: 是哪个订单的呢。(Which order is it?)<br>5 Q: 我已经取消了。(I have canceled.)<br>6A: **全部取消**？您只**取消**了一部分呢(**Cancel all**? You only **canceled** a part of it.)<br>7 Q: **请全部取消**，其他的显示关联订单**无法取消**。(Please **cancel all of it**. Other parts **cannot be canceled** because of associated orders.)<br>8 A: 好的，您稍等。这个订单是三方商家的，**需要商家审核一下**。(Okay, just a moment. This order is from a third-party merchant and needs to **be reviewed by the merchant**.)<br>9 Q：好的 。(Ok.)<br>10 A：请问还有其他需要帮忙吗。(Is there anything else you need help with?)<br>11 Q: 没有了,谢谢。(That is all, thanks.) | |
| **User Summary** | 用户**询问购买的商品能否更换地址**。用户希望**客服帮助取消全部订单**。(The user asks **whether the address of the purchased product could be changed**. The user **wants the agent to help cancel all orders**.) (1,6,7) |
| **Agent Summary** | 客服表示**下单后不可以更换地址**。客服表示**需要商家审核**。(The agent states that **the address cannot be changed after placing the order**. The agent says it **needs to be reviewed by the merchant.**) (1,2,8) |
| **Final Summary** | 用户询问**购买的商品能否更换地址**。客服表示**下单后不可以**。用户希望**客服帮助取消全部订单**。客服表示**需要商家审核**。(The user asks **whether the address of the purchased product could be changed**. The agent says it **was not possible after placing the order**. The user **wants the agent to help cancel all orders**. The agent says it **needs to be reviewed by the merchant.**) (1,2,6,7,8) |

Figure 1: An example of the role-oriented dialogue summarization. Red numbers represent key utterance indexes, and blue texts represent key information.

*Corresponding author.



*LREC-COLING 2024*, pages 8913–8924

agent (utterance 2), but also needs to integrate the key information "address" from the interaction-related utterance of the user (utterance 1). Second, interaction can help grasp the role's main content. In Figure 1, the first sentence of the user summary describes user's question about "change address". Thus the first sentence of the agent summary should describe the corresponding answer from the agent. Therefore, generating summary for a given role could benefit from referring to the other role's summary. The overall summary is also important for the generation of role-oriented summaries as it could indicate sentence-level logical relationship between different role's summaries.

Several methods have been proposed for the role-oriented dialogue summariztaion task. Lin et al. (2021) train different models for different role-oriented summaries. But they completely ignore interaction between different roles. Lin et al. (2022) employ two novel attention in decoder to capture interaction between roles. However, their method needs to assign each role a decoder which could introduce a lot of extra parameters. Besides, their method could not generate the overall summary directly and fails to utilize the overall summary for the generation of role-oriented summaries. Liang et al. (2023) use discrete role prompts attached to dialogue to control model to generate different role-oriented summaries and use a global-local centrality model to capture key utterances and topics. Nevertheless, the discrete prompt is difficult to design and the embedding of the prompt will affect the embedding of the dialogue. Moreover, existing methods do not build mechanisms for the encoder to understand interaction patterns between roles. The encoder can not fully comprehend the role-level interaction from the flat-concatenated utterances, thus the summarization model are prone to ignore vital interaction clues in dialogue and generate inaccurate summaries.

In this paper, we propose a **C**ontrastive Learning based **I**nteraction-**A**ware **M**odel (CIAM) for the role-oriented dialogue summarization task. An interaction-aware contrastive learning method (ICAL) is proposed to capture interaction patterns between different roles. Specifically, we first divide the representation derived from the encoder into self-representations for different roles. Then we calculate the masked dot-pot attention between self-representations to extract interaction information and finally obtain an interaction-aware representation for each role. The interaction-aware representation could guide the self-representation to learn interaction patterns by optimizing a contrastive loss. The negative samples for the contrastive learning are constructed by destroying interaction between roles in the original dialogue. With the guidance of the contrative objective, the encoder could generate informative dialogue representations that contain sufficient interaction information. In addition, we use different decoder start tokens to control the generation of different role-oriented summaries instead of using discrete prompt attached to the start of dialogue like Liang et al. (2022). The decoder start token is a randomly initialized special token without manual design. During training, the contrasting objective acts as an auxiliary task for the dialogue summarization task. At inference time, we control the model to generate summaries for different roles by decoding from different start tokens. In this way, our method could fully capture the interaction between different roles and generate different summaries with a unified seq2seq framework.

To evaluate our methods, we conduct experiments on two public datasets (Lin et al., 2021; Song et al., 2020) in different domains (customer service, medical inquiry). We apply our method on two widely-used summarization frameworks: BERTabs and BART. Experimental results show that our method could improve both performances of the two frameworks, and our model based on BART achieves new state-of-the-art performance on the two datasets.

The main contributions of this paper include: (1) We propose an interaction aware contrastive learning for the role-oriented dialogue summarization, which could help the model capture interaction patterns between different roles and understand the structure of the dialogue. (2) We employ decoder start tokens to control what kind of summary to generate which could generate different summaries with a unified model. (3) Our method can be applied to different seq2seq models and can outperform previous SOTA models and powerful LLMs on two public datasets.[1]

## 2. Related Work

### 2.1. Dialogue Summarization

Recently, dialogue summarization has drawn much attention in many domains, e.g., meeting (Carletta et al., 2005), chit-chat (Gliwa et al., 2019), customer service (Lin et al., 2021) and health care (Krishna et al., 2020). Considering the unique structure of dialogue, many existing studies pay attention to utilizing dialogue-related features. Liu et al. (2019) introduce auxiliary key point sequences to ensure the logic and correctness of the summary. Zou et al. (2021) employ a novel saliency-aware topic model to incorporate topic information for the dialogue summarization model. Chen and Yang (2021) explicitly model conversation structure by incorporating discourse relations and dialogue act triplets

[1] our code is available at https://github.com/neugwh/CIAM

through graph. Kim et al. (2022) inject commonse knowledge into the dialogue summarization model by utilizing an external knowledge model.

All the above studies focus on summarizing the whole dialogue. A few studies have paid attention to the role-oriented dialogue summarization task. Zhang et al. (2021) employ a variational auto-encoder based framework to generate the user summary and the agent summary separately in a unsupervised way. Lin et al. (2021) construct a dataset for the role-oriented dialogue summarization namely CSDS. Lin et al. (2022) first consider the information from other roles and propose two novel role-interaction attention in the decoder. Liang et al. (2022) control the generation of different summaries with a discrete role prompt and capture salient content with a role-level centrality model. Different from previous methods, the idea of our method is to help the representation derived from the encoder learn role-level interaction information by an auxiliary task.

### 2.2. Contrastive Learning

Contrastive learning has been widely used in many NLP tasks, such as text representation learning (Gao et al., 2021; Liu et al., 2022; Li et al., 2023), knowledge graph (Wang et al., 2023), machine translation (Pan et al., 2021) and text summarization (Liu and Liu, 2021; Xu et al., 2022). For dialogue summarization, contrastive learning is often utilized to learn dialogue-specific features. Liu et al. (2021) use two topic-aware contrastive objective to implicitly capture the topic structure of a dialogue. Geng et al. (2022) use speaker-aware supervised contrastive learning tasks to identify speakers. They only focus on identifying different speakers but ignore interaction information.

In this paper, we propose an interaction-aware contrastive objective to capture interaction patterns between different roles. Our idea is similar to Liu et al. (2022) who utilize contrastive learning to learn interaction-aware dialogue embedding. However, they force the encoder to only focus on interaction-aware information and eliminate interaction-free information, thus losing each role's own important information. Our method aims to acquire interaction-related key information from other roles while retaining each role's own key information. Besides, they focus on learning unsupervised dialogue embedding, while we focus on improving role-oriented dialogue summarization with the assistance of the contrastive learning objective.

## 3. Methodology

The main structure of our proposed model CIAM is shown in the part (b) of the Figure 2. Our model is built on the seq2seq framework, consisting of an encoder, an interaction-aware contrastive learning and a role-oriented decoder.

### 3.1. Task Formulation

Given a dialogue $D$ with $m$ utterances $\{u_1, \ldots, u_m\}$ and $p$ roles $\{r_1, \ldots, r_p\}$. Each utterance consists of a sentence $s_i$ and a role $r_k$. We concatenate all utterances and roles together to get the final input sequence $\{x_1, \ldots, x_n\}$. The role oriented dialogue summarization task aims at generating a summary for each role in addition to an overall summary. All public datasets for this task have two roles, one asking questions and one answering questions. Following Liang et al. (2023), we use $y^{user}$ and $y^{agent}$ to represent summaries of the two roles and $y^{final}$ to represent the overall summary. It is worth mentioning that our model could be extended to multi-role dialogue scenarios through the OVR (one vs rest) strategy with a few modification.

### 3.2. Encoder

We adopt a pretrained encoder (e.g. BERT or BART) to encode the input sequence and obtain the encoder hidden states $H$ as :

$$\{h_1, \ldots, h_n\} = \text{Encoder}\left(\{x_1, \ldots, x_n\}\right) \quad (1)$$

where $h_i \in R^d$ and $d$ is hidden size.

### 3.3. Contrastive Learning

The encoder simply treats the dialogue as a flat sequence, thus failing to understand interaction patterns between different roles. Therefore, we propose an interaction-aware contrastive learning method (IACL) based on the encoded hidden states, which could assist the encoded dialogue representation in grasping vital interaction information between different roles. The detailed structure of the IACL is shown in the part (a) of the Figure 2.

**Self-Representation** The first step is to divide the encoded representation into self-representations for different roles. Similar to Liu et al. (2022), We create two binary role masks $m^u \in R^n$ and $m^a \in R^n$ for user and agent respectively. The $i$-th element in $m^u$ is set to 1 if the token $x_i$ is spoken by user, otherwise it is 0. Similarly, we can generate $m^a$. Then we can get user's self-representation $H^u$ and agent's self-representation $H^a$ by:

$$\begin{aligned} H^u &= H \odot m^u, \\ H^a &= H \odot m^a, \end{aligned} \quad (2)$$

where $\odot$ denotes the broadcast element-wise multiplication.

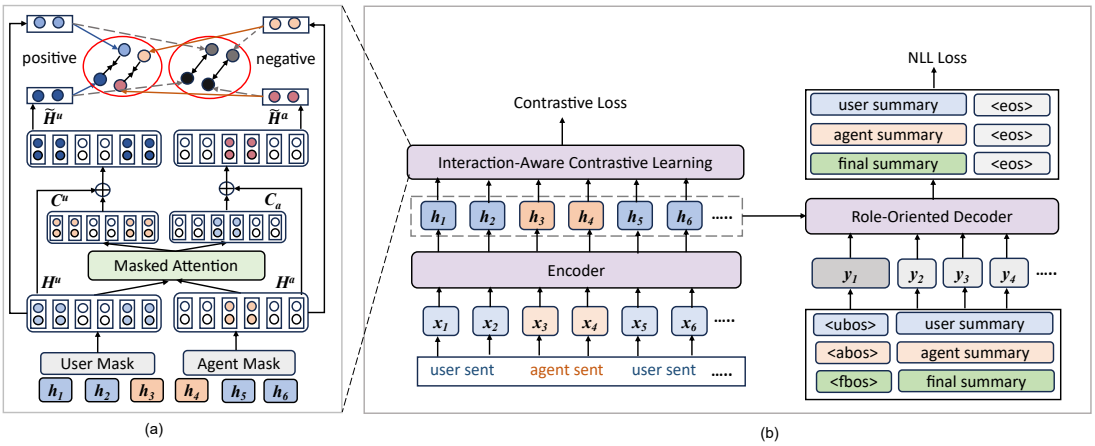


Figure 2: The overview of CIAM. Part (a) is the detailed process of interaction-aware contrastive learning. Part (b) is the overall structure of the role-oriented dialogue summarization model.

**Interaction-Aware-Representation** To acquire the interaction information from the other role, we perform the masked dot-product attention between $H^u$ and $H^a$. First, we compute two attention score matrices by:

$$A_{u2a} = \text{softmax}\left(\frac{H^u\left(H^a\right)^T}{\sqrt{d}} + M_{u2a}\right),$$
$$A_{a2u} = \text{softmax}\left(\frac{H^a\left(H^u\right)^T}{\sqrt{d}} + M_{a2u}\right), \quad (3)$$

where $d$ is hidden size, $A_{u2a}$ and $A_{a2u}$ are both $n * n$ square matrices. We do not use linear layers to map $H^u$ and $H^a$ before the attention calculation, as we found in experiments that using the original representation could result in better results. We use two attention masks $M_{u2a} \in R^{n\times n}$ and $M_{a2u} \in R^{n\times n}$ defined as:

$$M_{u2a}[i,j] = \begin{cases} 0, & m_j^a = 1, |I_i - I_j| \le w \\ -\infty, & \text{otherwise} \end{cases}$$
$$M_{a2u}[i,j] = \begin{cases} 0, & m_j^u = 1, |I_i - I_j| \le w \\ -\infty, & \text{otherwise} \end{cases}$$

where $m_i^a$ and $m_i^u$ are the $i$th elements of $m^a$ and $m^u$, $I_i$ and $I_j$ are indexes of the utterances that the $i$-th and $j$-th token belong to. Note that we introduce a hype-parameter $w$ to mask the long-range attentions among utterances as distant utterances are usually irrelevant. Then we can obtain a cross representation for each role by:

$$C^u = \left(A_{u2a} \odot m^u\right) H^a,$$
$$C^a = \left(A_{a2u} \odot m^a\right) H^u, \quad (4)$$

$C$ represents the interaction information from the other role. Finally we can obtain an interaction-aware representation for each role by fusing the self-representation and the cross-representation:

$$\tilde{H}^u = \left(H^u \oplus C^u\right) W^u,$$
$$\tilde{H}^a = \left(H^a \oplus C^a\right) W^a, \quad (5)$$

where $\oplus$ denotes the concatenation operator, $W_u \in R^{2d\times d}$ and $W_a \in R^{2d\times d}$ are trainable parameters. The interaction-aware representation can guide the self-representation to learn role-level interaction information by optimizing a contrastive loss.

**Training Samples Construction** To train the interaction-aware objective, we need to construct a positive sample and multiple negative samples for a given dialogue $D$. We treat the original dialogue $D$ as the positive sample and construct negative samples by destroying interactions between different roles in dialogue. Specifically, we keep all utterances of the user and mask all utterances of the agent. Then we randomly select a different dialogue $\widetilde{D}$ and sample a consecutive of utterances of the agent in $\widetilde{D}$ to fill the masked utterances of $D$. We repeat the process until all the masked utterances in $D$ has been filled. We repeat this operation multiple times and generate $K$ negative samples where $K$ is a hyper-parameter. Similarly, we keep all utterances of the agent and replace all utterances of the user, obtaining another $K$ negative samples. Thus, for each dialogue, we can obtain $2K + 1$ training samples in total, where the first sample is the positive sample and remaining samples are negative samples. We generate $H_i^u$, $H_i^a$, $\tilde{H}_i^u$ and $\tilde{H}_i^a$ for the $i$-th sample through steps in above sections.

**Contrastive Objective** We adopt the InfoNCE loss (Oord et al., 2018) to train our interaction-

aware objective. The loss is defined as:

$$\mathcal{L}^{user} = -\log \frac{e^{sim\left(H_1^u, \tilde{H}_1^u\right)/\tau}}{\sum_{j=1}^{(2K+1)} e^{sim\left(H_j^u, \tilde{H}_j^u\right)/\tau}},$$
$$\mathcal{L}^{agent} = -\log \frac{e^{sim\left(H_1^a, \tilde{H}_1^a\right)/\tau}}{\sum_{j=1}^{(2K+1)} e^{sim\left(H_j^a, \tilde{H}_j^a\right)/\tau}}, \quad (6)$$
$$\mathcal{L}_{con} = \mathcal{L}_{user} + \mathcal{L}_{agent}.$$

where $\tau$ is the hyper-parameter of temperature. $sim(,)$ represents a pooling operation followed by a cosine similarity calculation. The loss pulls close the self-representation and the interaction-aware representation if the dialogue has correct interaction, and pushes away otherwise. It guides the dialogue representation output by the encoder to correctly capture interaction information between different roles.

### 3.4. Role-Oriented Decoder

In this paper, we propose a role-oriented decoder which could generate different role-oriented summaries with a single unified model. Specifically, we employ different decoder start tokens to control what kind of summary to generate. The decoder start token is a specific token attached to the start of the decoder's input to indicate the start point of decoding, which is often represented by "<bos>". As shown in Figure 2, we use three different start tokens, "<ubos>", "<abos>", "<fbos>" to guide the decoder to generate the user summary, the agent summary and the final summary, respectively. We do not use discrete role prompts as in Liang et al. (2022) for two reasons. First, the discrete prompt needs to be carefully designed, while the decoder start token is randomly initialized without any manual annotation. Second, the embedding of the prompt will affect the embedding of the dialogue, making it difficult to obtain the self-representation for each role. Our method only use the decoder to control the generation of different summaries, which is compatible with our proposed contrastive objective.

During training, we attach the role-oriented decoder start token to the start of the corresponding summary and get the input sequence for the decoder$\{y_1, \ldots, y_T\}$, where $T$ is the length of the summary. Then we optimize the summarization model with a Negative Log-Likelihood loss:

$$\mathcal{L}_{nll} = -\sum_{i=1}^{T} \log P\left(y_i \mid y_{<i}, X\right) \quad (7)$$

At the inference time, we could generate summaries for different roles by decoding from different decoder start tokens.

### 3.5. Multi-Task Learning

During training, the interaction-aware contrastive learing acts as an auxiliary task for the dialogue summarization task. We combine the contastive loss and the nll loss as:

$$\mathcal{L} = \mathcal{L}_{nll} + \gamma \mathcal{L}_{con} \quad (8)$$

where $\gamma$ is the weight of the contrastive loss. The two tasks can effectively supplement each other.

## 4. Experiments

### 4.1. Datasets and Metrics

We evaluate our method on two public role-oriented dialogue summarization datasets: **CSDS**[2] (Lin et al., 2021) and **MC**[3] (Song et al., 2020). The statistics of the two datasets are shown in the Appendix A.

CSDS is a Chinese customer service dialogue summarization dataset which provides a user summary, an agent summary and an overall summary for each dialogue. MC is a Chinese medical inquiry summarization dataset, which provides a summary of patient's question and a summary of doctor's suggestion for each dialogue. We note them as the user summary and the agent summary respectively. CSDS is more abstractive and more specific for the role-oriented summarization, thus is more challenging than MC.

To evaluate models, we employ lexical-level metrics **ROUGE-1/2/L**[4] (Lin, 2004) and semantic-level metric **BERT-score**[5] (Zhang et al., 2019) to measure the similarity of references and generated summaries. For ROUGE, we follow the setting in Lin et al. (2022) and convert all the Chinese characters into number ids for calculation. We calculate the F1 score of all metrics.

### 4.2. Baselines

We apply our CIAM on two widely used seq2seq summarization models: **BERTAbs** (Liu and Lapata, 2019) and **BART** (Lewis et al., 2020). BERTAbs employs a pretrained encoder BERT to encode dialogue and employs a non-pretrained transformer decoder to generate summaries. BART is a pretrained seq2seq model, which achieves great success on summarization datasets. We add our proposed CIAM method on the two models and note them as **BERT-CIAM** and **BART-CIAM**. We also compare

[2] https://github.com/xiaolinAndy/CSDS

[3] https://github.com/cuhksz-nlp/HET-MC. We use the official crawling script to obtain the dataset.

[4] https://github.com/pltrdy/files2rouge

[5] https://github.com/Tiiiger/bert_score

| CSDS | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore |
|---|---|---|---|---|
| **BERT** | 55.41/52.71/49.61 | 39.42/36.39/33.88 | 53.41/50.45/46.88 | 78.52/79.23/76.39 |
| **BERT-Both** | 57.24/57.36/51.92 | 40.12/40.70/36.37 | 54.87/55.17/49.52 | 79.85/80.70/77.23 |
| **BERT-RAC** | 57.35/57.75/52.23 | 40.34/41.05/36.75 | 55.12/55.53/49.89 | 79.89/80.69/77.27 |
| **BERT-GLC** | 57.59/58.14/52.34 | **41.28**/41.84/36.48 | **55.74**/55.86/50.16 | 79.89/80.71/77.28 |
| **BERT-CIAM** | **57.66**/**58.73**/**52.55** | 41.12/**42.01**/**36.92** | 55.51/**56.72**/**50.20** | **79.90**/**81.25**/**77.39** |
| **w/o IACL** | 56.24/57.20/51.21 | 40.28/40.92/36.09 | 54.38/55.29/49.16 | 79.61/80.89/76.79 |
| **BART** | 58.66/60.35/54.13 | 43.35/45.09/39.37 | 56.60/58.13/51.18 | 79.54/81.14/77.31 |
| **BART-Both** | 59.21/60.53/54.22 | 43.88/45.39/39.96 | 57.32/58.28/51.90 | 79.74/81.37/77.41 |
| **BART-RAC** | 59.86/61.67/54.83 | 44.42/46.14/40.29 | 57.86/59.45/52.43 | 79.97/81.92/77.60 |
| **BART-GLC** | 60.07/61.72/54.82 | 44.55/46.21/40.11 | 58.06/59.51/52.46 | **80.10**/81.90/77.61 |
| **BART-CIAM** | **60.27**/**62.21**/**55.04** | **44.63**/**46.35**/**40.46** | **58.20**/**59.88**/**52.69** | 80.01/**82.03**/**77.63** |
| **w/o IACL** | 59.39/61.69/54.68 | 43.85/46.13/40.12 | 57.34/59.37/52.39 | 79.77/81.88/77.58 |

Table 1: Results on the CSDS dataset test set. Each block has three values, representing the final summary/user summary/agent summary from left to right.

| MC | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore |
|---|---|---|---|---|
| **BERT** | 84.07/95.13/81.66 | 79.90/94.50/76.73 | 83.04/95.08/80.42 | 92.68/97.86/91.71 |
| **BERT-Both** | 84.69/95.19/82.11 | 80.76/94.63/77.49 | 83.68/95.14/80.92 | 93.02/97.90/91.91 |
| **BERT-RAC** | 85.12/95.50/82.62 | 81.30/94.80/77.91 | 84.07/95.72/81.36 | 93.11/97.89/92.29 |
| **BERT-GLC** | 85.64/95.49/82.87 | 81.44/94.97/78.05 | 84.16/96.10/81.57 | 93.15/97.92/**92.36** |
| **BERT-CIAM** | **85.87**/**95.96**/**83.04** | **81.82**/**94.98**/**78.53** | **84.62**/**96.11**/**81.69** | **93.39**/**98.16**/92.14 |
| **w/o IACL** | 84.83/95.20/82.25 | 80.93/94.59/77.63 | 84.02/95.21/80.96 | 93.05/97.88/91.95 |
| **BART** | 88.37/95.42/86.33 | 84.75/94.99/82.33 | 87.38/95.37/85.30 | 93.65/97.94/92.63 |
| **BART-Both** | 88.52/95.63/87.06 | 85.22/95.42/82.89 | 87.55/95.96/85.79 | 93.72/97.89/92.67 |
| **BART-RAC** | 89.43/96.78/88.21 | 86.29/95.86/84.58 | 88.47/96.12/86.56 | 94.01/98.13/92.84 |
| **BART-GLC** | 89.55/96.84/88.47 | 86.47/96.14/**84.62** | 88.56/96.23/86.77 | 94.17/98.25/**92.96** |
| **BART-CIAM** | **89.85**/**96.86**/**88.73** | **86.93**/**96.31**/84.56 | **88.83**/**96.74**/**86.84** | **94.26**/**98.55**/92.90 |
| **w/o IACL** | 88.87/96.23/87.95 | 85.78/95.84/83.71 | 87.96/96.10/86.14 | 93.89/98.06/92.78 |

Table 2: Results on the MC dataset test set. The values in each block represent the same as in Table 1.

our methods with following previous SOTA methods based on the two models: **BERT-both** and **BART-both** from Lin et al. (2022) which employ two novel role attentions to model interaction between different roles. **BERT-RAC** and **BART-RAC** from Liang et al. (2022) which employ a discrete role prompt to control model for different summaries, and employ a centrality model to capture salient utterances. **BERT-GLC** and **BART-GLC** from Liang et al. (2023) which employ global-to-local centrality scores to capture sub topics on the basis of RAC.

### 4.3. Implementation Details

We use chinese-bert-wwm[6] and bart-base-chinese[7] to initialize our BERT-based models and BART-based models respectively. We train all models on a single RTX-3090 GPU. The best checkpoint is chosen based on the performance on validation set. For all models, the maximum input length is 512. For CIAM, the window size $w$ is 8 and the number of negative samples $K$ is 3. The temperature $\tau$ and the weight of the contrastive loss $\gamma$ is selected based on the performance on the validation set. More implementation details are shown in the Appendix B.

[6] https://huggingface.co/hfl/chinese-bert-wwm

[7] https://huggingface.co/uer/bart-base-chinese-cluecorpussmall

## 5. Results and Analysis

### 5.1. Main Results

The results of automatic metrics are shown in Table 1 and Table 2. The results on the two datasets are similar. We can see that our proposed CIAM could bring remarkable improvement on both BERTAbs and BART on the two datasets. And the improvement on the BERTAbs is more remarkable than on the BART. We guess the reason is that the BART could better capture the interaction between different roles than BERT. For most of metrics, our proposed CIAM could outperform previous methods based on the same backbone model. BART-

| CSDS | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore |
|---|---|---|---|---|
| **BART-CIAM** | **60.27/62.21/55.04** | **44.63/46.35/40.46** | **58.20/59.88/52.69** | **80.01/82.03/77.63** |
| **ChatGPT** | 51.96/48.85/46.21 | 33.89/31.76/28.56 | 48.25/44.98/42.25 | 75.46/76.74/73.90 |
| **GPT4** | 53.04/49.78/47.97 | 35.10/32.98/30.97 | 49.72/46.32/43.98 | 77.12/77.15/75.10 |

Table 3: The results compared with ChatGPT and GPT4. The values in each block represent the same as in Table 1.

| CSDS | Informativeness | Conciseness | Human Preference |
|---|---|---|---|
| **GPT4** | 1.45/1.38/1.34 | 1.29/1.16/1.19 | 0.40/0.30/0.29 |
| **BART** | 1.52/1.48/1.42 | 1.10/1.18/1.16 | 0.19/0.38/0.31 |
| **BART-GLC** | 1.48/1.44/1.45 | 1.43/1.46/1.53 | 0.45/0.45/0.40 |
| **BART-CIAM** | **1.59/1.51/1.53** | **1.54/1.58/1.56** | **0.57/0.56/0.53** |

Table 4: The results of human evaluation. The values in each block represent the same as in Table 1.

CIAM achieves new state-of-the-art results on the two datasets. Overall, our proposed CIAM is effective on different backbone structures and different datasets.

We also conduct ablation study for the interaction-aware contrastive learning (IACL). We can see that all metrics drop substantially after removing the IACL, which indicates that our proposed interaction-aware objective could help generate better role-oriented summaries. The results after removing the IACL are still much higher than the results of BERTAbs and BART, demonstrating the effectiveness of generating different role-oriented summaries with the control of decoder start tokens. With the guidance of different decoder start tokens, we can generate summaries for different roles by only training a single model and can fully utilize the strong relatedness between different role-oriented summaries. We further verify the effectiveness of our decoder-start token based control strategy in Section 5.4.

## 5.2. Comparasion with LLMs

Recently, large language models (LLMs), such as ChatGPT (Zhong et al., 2023) and GPT-4, have demonstrated promising performances in various natural language applications. Pu et al. (2023) find summaries generated by LLMs could outperform summaries generated by fine-tuned methods. But they do not conduct experiments on the role-oriented dialogue summarization datasets. The role-oriented dialogue summarization task might be more difficult for LLMs compared with other summarization tasks. First, datasets for the task usually focus on specific domains such as customer service. LLMs might lack enough prior knowledge for these domains. Second, each role has a specific goal and the summary of each role should reflect the corresponding goal, which requires the LLMs to fully understand the responsibility of each role. We compare our proposed BART-CIAM with two powerful LLMs: **ChatGPT** (gpt-3.5-turbo-0613) and **GPT4** (gpt-4-0613) on the CSDS dataset. The detail of the prompt used is provided in Appendix C. We provide four examples to instruct LLMs as we find the two models fail to generate role-oriented summaries in zero-shot setting.

Table 3 shows the results. We can see that the LLMs can obtain satisfactory results with only a few examples, which highlights the potential of LLMs, but the results of automatic metrics are still lower than the fine-tuned methods. And the results of the GPT4 are slightly higher than the results of the ChatGPT. Besides, we find LLMs are obviously better at generating the overall summary than generating the role-oriented summaries, possibly because the LLMs can not fully understand the goal of each role in few-shot setting. The automatic results might not reflect the quality of the summary, thus we also conduct human evaluation experiments for the GPT4 in the next section.

## 5.3. Human Evaluation

We also conduct human evaluation for our proposed BART-CIAM and three other baselines. Specifically, we randomly sample 100 samples from the CSDS dataset with corresponding generated summaries. We recruited 4 graduate students with Chinese proficiency to rate summaries according to the following two aspects: (1) **Informativeness**: whether the generated summary could correctly contain the key information (2) **Conciseness**: whether the generated summary could avoid redundant and unnecessary information. The summaries are rated with a score ranged from 0 to 2, with 2 being the best. We also conduct a **Human Preference** test where evaluators are asked to select a best summary or several best summaries from the generated summaries of different models and we calculate the percentage being selected

for each model. We train the evaluators with the evaluation rules and require them to evaluate summaries according to both the reference summary and the original dialogue. Following Wang et al. (2022), we run an inter-annotator agreement study and the average kappa score is 0.46.

As shown in Table 4, our method can generate more informative and more concise summaries compared with baselines. And the summaries generated by our model are most prefered by human evaluators. Besides, we can see that the performance of the GPT4 is comparable to the fine-tuned BART, but still lags behind our method, especially in conciseness. We found the summaries generated by LLMs are prone to contain unnecessary information such as greeting, thanks, or even hallucinated content. Therefore, a fine-tuning method specific for the role-oriented dialogue summarization is still meaningful.

## 5.4. Further Analysis

**Decoder Start Token vs Discrete Prompt** Our proposed CIAM employs decoder start tokens to control the generation of different role-oriented summaries instead of employing discrete role prompts as previous studies (Liang et al., 2022, 2023). To understand the impact of different control strategies, we introduce two variants of our proposed CIAM, **CIAM-prefix** and **CIAM-suffix**. CIAM-prefix employs a prompt attached to dialogue to control the model as in Liang et al. (2022) while CIAM-suffix employs a prompt attached to the input of the decoder to control the model. The interaction-aware contrastive objective is also employed on the two variants. Following Liang et al. (2023), we use "[user summary]", "[agent summary]", "[final summary]" as prompts for required summaries. We use BART as the backbone for all models and conduct experiments on the CSDS dataset.

| CSDS | ROUGE-2 | ROUGE-L |
|---|---|---|
| **CIAM** | **44.63/46.35/40.46** | **58.20/59.88/52.69** |
| **CIAM-prefix** | 44.20/46.12/40.08 | 57.54/59.39/52.25 |
| **CIAM-suffix** | 44.48/46.27/40.32 | 57.89/59.64/52.57 |

Table 5: Results of different control strategies on the CSDS dataset. The values in each block represent the same as in Table 1.

The results are shown in Table 5. We can see that the results of the CIAM-prefix are worse than other models significantly. The reason is that the representation of the prompt affects the representation of the dialogue. And the results of the CIAM-suffix are slightly worse than CIAM. One possible reason is that the prompt we use is not optimal as the discrete prompt is hard to design. Besides, employing prompt on the decoder is difficult to implement, especially for BERTAbs. The decoder start token can be seen as a specific prompt for the decoder, which requires no annotation and can be easily implemented for any generation models.

**Coefficient of the Multi-Task Loss $\tau$** The coefficient $\tau$ controls the weight of the contrastive loss in the multi-task loss. To understand the impact of $\tau$, we conduct experiments with BART-CIAM on the CSDS dataset. As shown in Table 6, the weight

| $\tau$ | ROUGE-2 | ROUGE-L |
|---|---|---|
| **0** | 43.85/46.13/40.12 | 57.34/59.37/52.39 |
| **0.2** | 44.02/46.23/40.29 | 57.39/59.58/52.65 |
| **0.4** | 44.27/46.25/40.38 | 57.69/59.64/52.53 |
| **0.6** | **44.63/46.35/40.46** | **58.20/59.88/52.69** |
| **0.8** | 44.51/46.27/40.34 | 57.90/59.72/52.47 |
| **1.0** | 44.49/46.30/40.41 | 57.84/59.69/52.43 |
| **2.0** | 44.07/45.96/40.17 | 57.42/59.36/52.24 |
| **10.0** | 43.36/45.77/39.85 | 56.87/59.04/51.89 |

Table 6: The results of the BART-CIAM with different $\tau$ coefficients. The values in each block represent the same as in Table 1.

of the contrastive loss $\tau$ is important for the model performance. As the weight of the contrastive loss increases, the model's performance first improves and then degrades. Assigning small weights to the contrastive objective could not help the model fully capture interaction between different roles. However, assigning large weights to the contrastive objective will force the model to prefer the auxiliary task and ignore the primary generation task.

**Summary Completeness Analysis** To capture the key information of a role-oriented summary, the model often needs to integrate the content from other roles, especially for the agent summary. To verify whether our method can capture interaction information, we evaluate our methods on these incomplete cases as Lin et al. (2022). Following the setting in Lin et al. (2021), we divide the test agent summaries from CSDS into incomplete samples and complete samples according to whether other role's information are needed to be integrated[8] and compare the summary quality of different types of samples.

As shown in Table 7, BART-CIAM outperforms BART-GLC on incomplete samples, which proves that our method can help capture the interaction-related information. After removing the contrastive objective (IACL), the results drop substantially, demonstrating the interaction-aware contrastive

[8]It is judged by leveraging the key utterance indices labeled in the CSDS.

| CSDS | ROUGE-2 Type A/B | ROUGE-L Type A/B |
|---|---|---|
| **BART-GLC** | 43.57/45.06 | 57.20/60.91 |
| **BART-CIAM** | **45.05/47.68** | **58.44/61.93** |
| **w/o IACL** | 43.77/45.67 | 57.60/60.19 |

Table 7: The performance on different types of samples. Type A represents incomplete samples that need to integrate other role's information, and Type B represents complete samples that do not. All the metrics here are recall scores.

learning is effective. Besides, the results of complete samples are also improved as interaction information can help capture the role's main content.

**Case Study** We use the same example shown in Figure 1 to demonstrate the superiority of our method. Figure 3 shows the role-oriented summaries generated by different methods for the example. As for the user summary, BART-GLC ignores the user's issue about "canceling all orders" and BART generates a redundant sentence. The summary generated by our method is similar to the reference summary and is able to concisely contain all key information. As for the agent summary, the summary generated by our method still contains more key information and less redundancies than other methods. Moreover, only our method could generate the key information "address" which needs to be integrated from the utterances of the users. Therefore, our method can help the model capture interaction information and generate more informative role-oriented summaries.

## 6. Conclusion

In this paper, we focus on the role-oriented dialogue summarization task. To fully grasp interaction information, we propose an interaction-aware contrastive objective, which could help the model capture interaction patterns between different roles and obtain interaction-aware dialogue representations. And we propose a role-oriented decoder, which utilizes different start tokens to control the generation. Thus we can generate different role-oriented summaries by training a single model and fully utilize the relatedness between different summaries. Experiments on two public datasets prove our methods can surpass previous SOTA methods and powerful LLMs. In the future, we will try to apply our method to multi-role dialogue scenarios and other dialogue-related tasks.

| **User Summary** | |
|---|---|
| **BART** | 用户**询问购买的商品是否可以更换地址**。**用户表示刚买的那些订单已经取消了**。用户询问**能否取消订单**。(The user asks **whether the address of the purchased product can be changed**. **The user says that the order purchased has been cancelled**. The user asks **whether the order could be canceled**.) |
| **BART-GLC** | 用户询问**购买的商品是否可以更换收货地址**。(The user asks **whether the address of the purchased product can be changed**.) |
| **BART-CIAM** | 用户询问**购买的商品是否可以更换地址**。用户询问能否**全部取消订单**。(The user asks **whether the address of the purchased product can be changed**. The user asks whether **all orders could be canceled**.) |
| **Reference** | 用户询问**购买的商品能否更换地址**。用户希望**客服帮助取消全部订单**。(The user asked **whether the address of the purchased product could be changed**. The user **wants the agent to help cancel all orders**.) |
| **Agent Summary** | |
| **BART** | 客服回答是**改不了**的。**客服回答支付订单是[数字]的，只取消了一部分**。(The agent answers that it **cannot be changed**. **The agent replies that the payment order was [number] and only a part of it is cancelled**.) |
| **BART-GLC** | 客服回答**下单之后是改不了**的。**客服回答只能取消一部分**。客服回答是**三方商家的**，需要**商家审核**一下。(The agent answers that it **cannot be changed after placing the order**. **The agent replies that only a part of it is cancelled**. The agent replies **that the order is from a third-party merchant** and **needs to be reviewed by the merchant**.) |
| **BART-CIAM** | 客服回答**下单之后无法更换地址**。客服回答**订单是三方商家的**，需要**商家审核**一下。(The agent replies that **the address cannot be changed after placing the order**. The agent replies **that the order is from a third-party merchant** and **needs to be reviewed by the merchant**.) |
| **Reference** | 客服表示**下单后不可以更换地址**。客服表示**需要商家审核**。(The agent states that **the address cannot be changed after placing the order**. The agent says it **needs to be reviewed by the merchant**.) |

Figure 3: The generated summaries for the example shown in Figure 1. Blue texts represent key information and green texts represent redundant content.

## 7. Ethical Considerations

In this paper, we use two public datasets, CSDS and MC. CSDS is constructed based on a public customer service dialogue dataset JDDC. All private information is anonymized. MC is constructed based on a public Chinese medical inquiry website. All private information of the patient has been anonymized by the website. We acquire the dataset by using the official script and preprocess the dataset by following the setting in the original paper strictly.

## 8. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62272092, No. 62172086, No. 62262045) and the Fundamental Research Funds for the Central Universities of China (No. N2116008).

## 9. Bibliographical References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The

ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Zhichao Geng, Ming Zhong, Zhangyue Yin, Xipeng Qiu, and Xuan-Jing Huang. 2022. Improving abstractive dialogue summarization with speaker-aware supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6540–6546.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung-won Hwang, and Jinyoung Yeo. 2022. Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6285–6300.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

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.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. 2023. Structure-aware language model pretraining improves dense retrieval on structured data. *arXiv preprint arXiv:2305.19912*.

Xinnian Liang, Chao Bian, Shuangzhi Wu, and Zhoujun Li. 2022. Towards modeling role-aware centrality for dialogue summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 43–50.

Xinnian Liang, Shuangzhi Wu, Chenhao Cui, Jiaqi Bai, Chao Bian, and Zhoujun Li. 2023. Enhancing dialogue summarization with topic-aware global-and local-level centrality. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 27–38.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Csds: A fine-grained chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558.

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-guided contrastive learning of unsupervised dialogue embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7272–7282.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Jingbo Wang, Yumeng Song, Yu Gu, Xiaohua Li, and Fangfang Li. 2023. Clnie: A contrastive learning based node importance evaluation method for knowledge graphs with few labels. In *International Conference on Database Systems for Advanced Applications*, pages 689–705. Springer.

Weichao Wang, Shi Feng, Kaisong Song, Daling Wang, and Shifeng Li. 2022. Informative and diverse emotional conversation generation with variational recurrent pointer-generator. *Frontiers of Computer Science*, 16(5):165326.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11556–11565.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2021. Unsupervised abstractive dialogue summarization for tete-a-tetes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14489–14497.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arxiv.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14665–14673.

# Appendix

## A. Dataset Statistics

The statistical information of the two datasets is shown in Table 8.

| | CSDS | MC |
|---|---|---|
| Train Size | 9101 | 29324 |
| Val Size | 800 | 3258 |
| Test Size | 800 | 8146 |
| Dial Length | 321.92 | 292.21 |
| UserSum. Length | 37.28 | 22.37 |
| AgentSum. Length | 48.08 | 95.32 |
| FinalSum. Length | 83.21 | 117.69 |

Table 8: Statistics of the two datasets.

## B. Implementation Details

For BART-based methods, the learning rate is 3e-05, the warm up steps is 800 and the training epochs is 5. At the inference process, the beam size is 3 and the maximum generated summary length is 150. For BERTAbs-based methods, the warm up steps is 1000 and the training steps is 8000. The learning rate of the encoder and the decoder are set to 0.002 and 0.02, respectively. At the inference process, the beam size is 5 and the maximum generated summary length is 200.

## C. Prompt used for LLMs

The prompt used for ChatGPT and GPT4 is shown in Figure 4. Four examples are provide in prompt. The examples are randomly sampled for each test sample. The user summary, the agent summary and the overall summary are generated simultaneously through the prompt. We apply a regex expression to extract different summaries from the generated response.

给定一段中文客服对话，请分别生成用户视角的摘要，客服视角的摘要和整段对话的总体摘要。(Given a Chinese customer service dialogue, please generate a summary for the user, a summary for the agent, and an overall summary for the whole dialogue.)
下面是四个示例：(The Following are four examples:)
示例一：(Example 1:)
对话：（Dialogue:）
{ Dialogue }
用户摘要：(User Summary:)
{ User Summary }
客服摘要：（Agent Summary:）
{ Agent Summary }
总体摘要：(Overall Summary:)
{ Final Summary }
.....
请为下面的测试对话生成用户摘要，客服摘要和总体摘要。(Please generate the user summary, the agent summary, the overall summary for the following test dialogue:)
对话：(Dialogue:)
{ Dialogue }
用户摘要：(User Summary:)
客服摘要：（Agent Summary:）
总体摘要：(Overall Summary:)

Figure 4: Prompt used for LLMs.