Topic-Oriented Dialogue Summarization

Haitao Lin, Junnan Zhu, Lu Xiang, Feifei Zhai, Yu Zhou, Jiajun Zhang, *Senior Member, IEEE*, and Chengqing Zong, *Fellow, IEEE*

***Abstract*—A multi-turn dialogue often contains multiple discussion topics. In several scenarios (e.g., customer service dispute, public opinion monitoring), people are only interested in the gist of a specific topic in the dialogue. Therefore, we propose a novel summarization task, i.e., Topic-Oriented Dialogue Summarization (TODS). Given a dialogue with a topic label, TODS aims to produce a summary covering the main content of the given topic in the dialogue. To model the relationship between dialogues and topics, three key abilities are needed for TODS: (1) Learning the semantic information of different topics. (2) Locating the topic-related content in the dialogue. (3) Distinguishing summaries for different topics in the same dialogue. Thus, we propose three topic-related auxiliary tasks to make the summarization model learn the three abilities above. First, the topic identification task aims at generating all the topics in the dialogue. Second, the topic attention restriction task tries to constrain the attention distribution on topic-related utterances. Third, the topic summary distinguishing task focuses on increasing the difference of summaries for different topics in the same dialogue. Experimental results on two public TODS datasets show that all auxiliary tasks are critical for TODS and help generate high-quality summaries. We also point out the expansions and challenges in TODS for future research.**

***Index Terms*—Dialogue summarization, abstractive summarization, controllable text generation, natural language processing.**

I. Introduction

DIALOGUE summarization is an information compression task to compress a long dialogue into a short text while maintaining the key information of the dialogue [1]. Since dialogues often have long turns and many redundant utterances, using dialogue summarization could help filter out the unimportant content in the dialogue and present the most critical information to readers, alleviating the time costs and reading difficulties.

Since a dialogue consists of multiple speakers' utterances, the discussion topics in a dialogue may shift as different speakers exchange their viewpoints. According to the existing study [2], in a single dialogue, the discussion topic shifts every 12 turns on average (one turn means one utterance from a speaker). It demonstrates that topic shifting is a common phenomenon in the dialogue. With multiple topics scattered in the dialogue, people may only focus on the content related to a specific topic. Thus, in this work, we aim to generate the summary for a given topic in the dialogue, which we define as **Topic-Oriented Dialogue Summarization (TODS)**.

We present an example in Fig. 1 to illustrate the TODS task. Given a whole dialogue and a specific topic, the summarization model extracts the main content related to the topic in the dialogue and summarizes it in a short text. To complete this task, we could simply apply the controllable summarization method [3], [4], [5] to it by using the topic label as guidance. However, the existing controllable summarization methods do not learn to model the relationship between dialogue and topic, thus performing poorly on TODS. To handle this task, the summarization model needs to have three abilities:

1) *Learning the semantic information of topics:* To summarize the information for a specific topic, the first step is to learn what the topic represents and figure out what kind of information is relevant to the topic. With the increase in the topic numbers and semantic range, it is increasingly difficult for the model to distinguish the semantics of different topics.
2) *Extracting the information of a given topic in the dialogue:* After learning the topic labels, the model needs to match the topic with the dialogue by extracting related utterances. It is challenging due to two reasons. First, the discourse structure of dialogue is usually complex, which can lead to crossed dialogue segments for different topics, such as the topic *"attribute consultation"* and *"discount consultation"* in the given example. Second, it is difficult to judge the topic for an utterance due to some interfering words. In the example, the 15th utterance *"How to participate in the discount activity"* belongs to the topic *"payment-related"*, while it also has the word *"discount"*, related to the topic *"discount consultation"*. It may mislead the model to classify the utterance into inappropriate topics.
3) *Distinguishing summaries for different topics in the same dialogue:* The last step is generating summaries according to the dialogue. For a dialogue with multiple topics, how

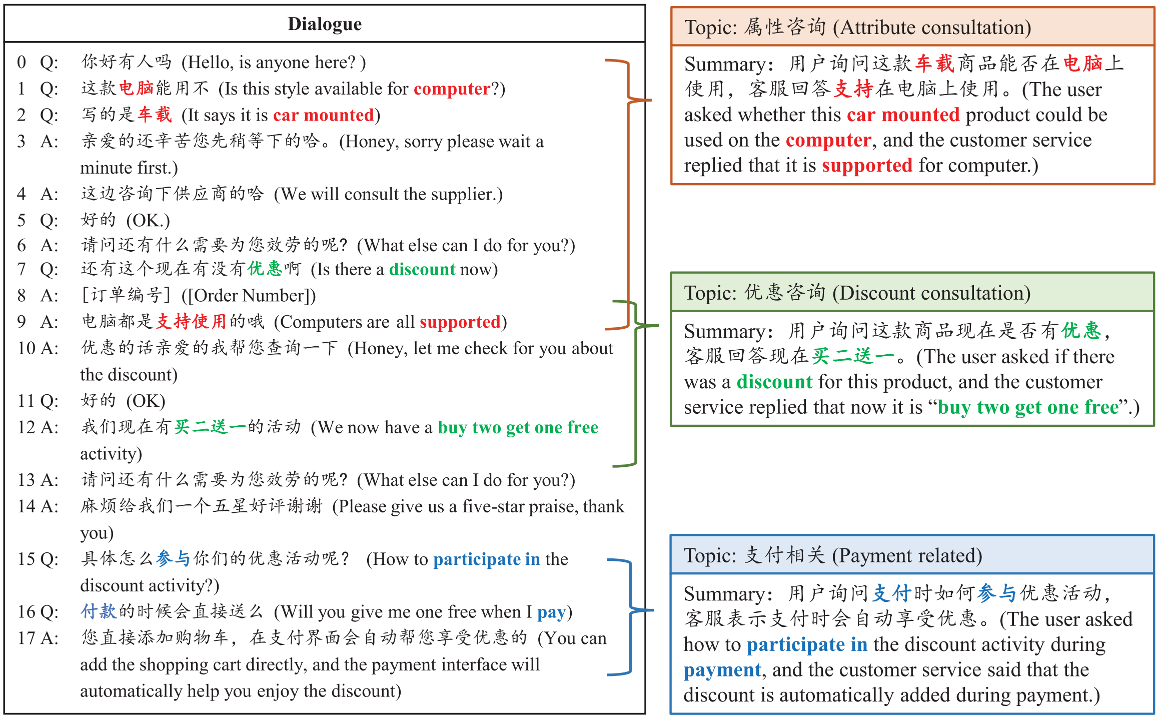
Manuscript received 21 July 2022; revised 19 January 2023 and 13 April 2023; accepted 19 April 2023. Date of publication 4 May 2023; date of current version 11 May 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0108600 and in part by the Natural Science Foundation of China under Grant 62106263. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nancy F. Chen. *(Corresponding author: Yu Zhou.)*

Haitao Lin, Junnan Zhu, Lu Xiang, Jiajun Zhang, and Chengqing Zong are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: haitao.lin@nlpr.ia.ac.cn; junnan.zhu@nlpr.ia.ac.cn; lu.xiang@nlpr.ia.ac.cn; jjzhang@nlpr.ia.ac.cn; cqzong@nlpr.ia.ac.cn).

Feifei Zhai is with the Fanyu AI Research, Zhongke Fanyu Technology Co., Ltd, Beijing 100190, China (e-mail: zhaifeifei@zkfy.com).

Yu Zhou is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Fanyu AI Research, Zhongke Fanyu Technology Co., Ltd, Beijing 100190, China (e-mail: yzhou@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASLP.2023.3271118



Dialogue

0 Q: 你好有人吗 (Hello, is anyone here?)
1 Q: 这款**电脑**能用不 (Is this style available for **computer**?)
2 Q: 写的是**车载** (It says it is **car mounted**)
3 A: 亲爱的还辛苦您先稍等下的哈。(Honey, sorry please wait a minute first.)
4 A: 这边咨询下供应商的哈 (We will consult the supplier.)
5 Q: 好的 (OK.)
6 A: 请问还有什么需要为您效劳的呢? (What else can I do for you?)
7 Q: 还有这个现在有没有**优惠**啊 (Is there a **discount** now)
8 A: [订单编号] ([Order Number])
9 A: 电脑都是**支持使用**的哦 (Computers are all **supported**)
10 A: 优惠的话亲爱的我帮您查询一下 (Honey, let me check for you about the discount)
11 Q: 好的 (OK)
12 A: 我们现在有**买二送一**的活动 (We now have a **buy two get one free** activity)
13 A: 请问还有什么需要为您效劳的呢? (What else can I do for you?)
14 A: 麻烦给我们一个五星好评谢谢 (Please give us a five-star praise, thank you)
15 Q: 具体怎么**参与**你们的优惠活动呢? (How to **participate in** the discount activity?)
16 Q: **付款**的时候会直接送么 (Will you give me one free when I **pay**)
17 A: 您直接添加购物车，在支付界面会自动帮您享受优惠的 (You can add the shopping cart directly, and the payment interface will automatically help you enjoy the discount)

Topic: 属性咨询 (Attribute consultation)

Summary: 用户询问这款**车载**商品能否在**电脑**上使用，客服回答**支持**在电脑上使用。(The user asked whether this **car mounted** product could be used on the **computer**, and the customer service replied that it is **supported** for computer.)

Topic: 优惠咨询 (Discount consultation)

Summary: 用户询问这款商品现在是否有**优惠**，客服回答现在**买二送一**。(The user asked if there was a **discount** for this product, and the customer service replied that now it is "**buy two get one free**".)

Topic: 支付相关 (Payment related)

Summary: 用户询问**支付**时如何**参与**优惠活动，客服表示支付时会自动享受优惠。(The user asked how to **participate in** the discount activity during **payment**, and the customer service said that the discount is automatically added during payment.)

Fig. 1. An example of topic-oriented dialogue summarization. The original data is in Chinese, and we translate it into English for better illustration.

to prevent the summaries for different topics from being similar and highlight the topic-related content in each summary is a tough issue.

Based on these three abilities, we propose three topic-related auxiliary tasks relatively.

1) *What are the topics?* Topic Identification Task is proposed to predict all the topic labels in the dialogue, thus learning the semantic information of each topic.
2) *What is the topic-related dialogue content?* We wish the model to focus more on the topic-related dialogue content in the decoding phase. Thus, **Topic Attention Restriction Task** is added by punishing the model when attending to other unrelated dialogue utterances.
3) *What is the content of topic-oriented summaries?* The summaries for different topics in the same dialogue should be different by highlighting the topic labels. We induce this requirement in the model as **Topic Summary Distinguishing Task**. It learns this ability by enlarging the probability discrepancy of generating the same summary given different topics.

Elaborate experiments are conducted on two dialogue summarization datasets, CSDS [6] and DIALOGSUM [7] with a few modifications for TODS. For comparison, we adopt BART [8], a strong model for dialogue summarization, as the backbone structure and report several automatic and human evaluation metrics results. Results have shown that all three auxiliary tasks can bring significant improvements for topic-oriented dialogue summarization, and learning three tasks together achieves the highest performance in most cases. The results are consistent on all the metrics. Besides, fine-grained results show that these tasks work better on dialogues with multiple topics and crossed topic segments. Further analysis shows that the auxiliary tasks could improve the summarization model on the three abilities and help generate the overall dialogue summary as well.[1] Finally, we discuss the scalability of TODS in a variety of ways and the remaining challenges for existing methods. These findings could open up new avenues for research and motivate future studies.

The main contributions of our work include:

- We are the first to study how to generate a topic-oriented summary for the dialogue scenario.
- According to the abilities needed in TODS, we design three topic-related auxiliary tasks to enhance the summarization model, including predicting existing topics in the dialogue, constraining the decoder attention distribution, and distinguishing summaries for different topics in the same dialogue.
- Experimental results verify that all the auxiliary tasks help improve the summary quality and work better for the dialogue with multiple topics and crossed topic segments. We also point out the expansion of TODS and the limitations of existing methods for future research.

II. Related Work

A. Dialogue Summarization

Dialogue summarization aims at summarizing the main content of a dialogue and plays an important role in many scenarios, including meetings [9], [10], [11], daily chatting [12], [13], [14], medical conversations [15], [16], [17] and customer services [6],

[1] Our codes and processed datasets used in our experiments are available at https://github.com/xiaolinAndy/TODS.

[18], [19], [20]. Since dialogue often contains multiple topics, some works try to enhance the summarization model by using topic information explicitly or implicitly. Zhao et al. [21] extract topic words with an LDA topic model and construct a graph modeling the relationship between topic words and utterances, which could help encode the dialogue content. Liu et al. [18] also extract topic information first and take each topic as guidance to generate summary sentences. They define the topic information as key points representing different dialogue stages. Both of the above researches explicitly extract topic information for the enhancement. Meanwhile, some works implicitly add topic information to the summarization model. Chen et al. [22] adopt the idea of topic segmentation and segment the dialogue into multiple views in an unsupervised way. Multiple views are considered during decoding to generate better summaries. Zou et al. [23] import the neural topic model and use the learned topic vectors for different roles to generate summaries with the correct topic information. Liu et al. [24] apply the contrastive learning method to consider topic coherence and summary-dialogue mappings for summarization.

Although the above methods incorporate topic information into the summarization model, they all focus on summarizing the whole dialogue but not on a specific topic. We focus on a different scenario where people are only interested in the content of a single topic and explore methods to generate summaries that could condense the key information about the topic in the dialogue.

B. Aspect-Based Summarization

A similar task with TODS is aspect-based summarization [25], [26], [27], [28], which summarizes the key information of a specific aspect in a document. Some works [29], [30] first extract existing aspects in the document and summarize each aspect afterward. Other works only focus on the summarization step given an aspect. Frermann and Klementiev [25] design three fusion methods to add the aspect into the summarization model, thus learning only to extract information related to the aspect. Tan et al. [26] adopt the controllable summarization method to summarize for any given aspect by concatenating the document and the aspect as the input of summarization.

Compared with the above works, our works differ in two aspects: (1) We experiment on dialogue texts instead of news, where the discourse structure is more complicated for topic-oriented dialogue summarization. We also experiment on an authentic dataset, different from the synthesized ones used in [25], [26]. (2) Our work focuses on modeling the relationship between the dialogue and the topic by designing three topic-related auxiliary tasks. The experiment results also show that these auxiliary tasks help achieve better performance than existing aspect-based summarization methods.

C. Query-Focused Summarization

Query-focused summarization is another similar summarization task that tries to provide a summary answer to a question given a document. It could be served as a blending task of question answering and summarization, and multiple datasets are proposed to tackle this issue [31], [32], [33]. When the input document is long, an extract-then-summarize pipeline method is adopted to first extract query-related sentences and summarize later [31]. With the improvement of model structure for long document summarization [34], [35], [36], some end-to-end approaches are proposed to solve the problem of long input and used in query-focused summarization [37] by concatenating the query and the document as model inputs.

TODS differs from query-focused summarization in the query format since the query in TODS is a topic label rather than a question. Furthermore, we concentrate on dialogue texts in which different topics are spread in a complicated structure, making this task more difficult. In addition, our proposed method outperforms pipeline and concatenation-based end-to-end methods commonly employed in query-focused summarization.

III. Methods

A. Task Definition

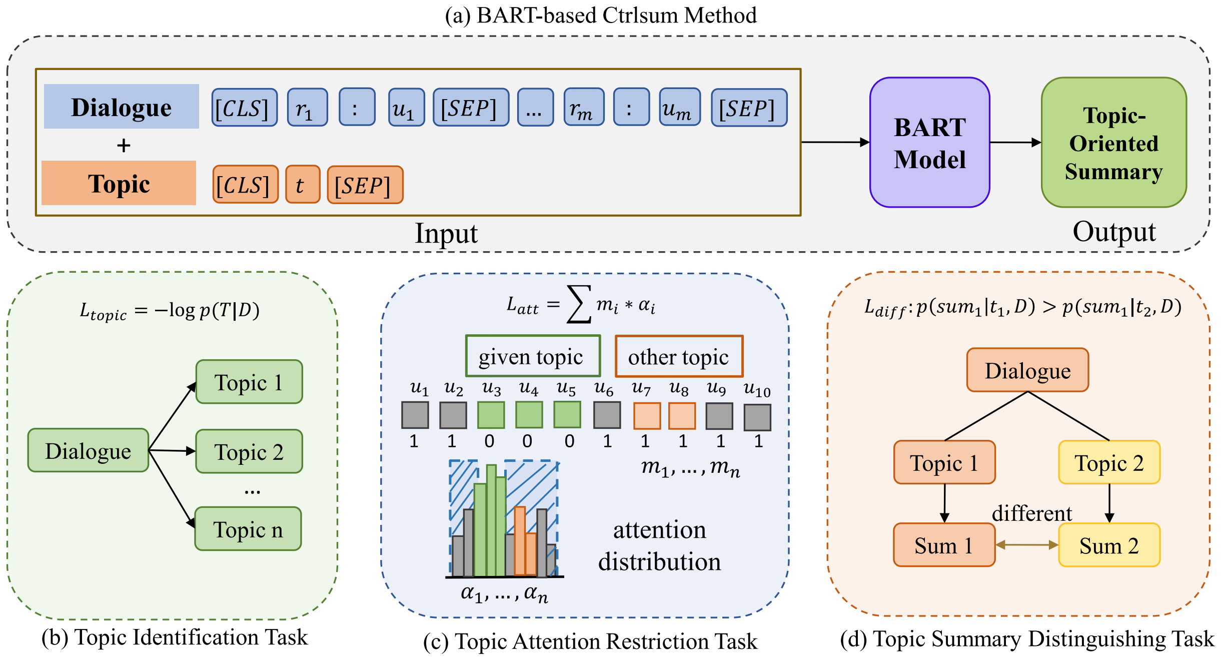
Given a dialogue D containing m utterances $\{u_1, \ldots, u_m\}$ and k topics $T = \{t_1, \ldots t_k\}$, there exist k sub-summaries $S = \{s_1, \ldots s_k\}$ and each sub-summary S_i refers to the main content of the dialogue related with topic t_i. For each utterance u_i, its speaker role is noted as r_i. The input of TODS is dialogue D and one of the topic $t \in T$; the output is the related sub-summary $y = \{y_1, \ldots, y_n\} \in S$ containing n tokens.

B. Our Approach

Since we aim to make the model learn the relationship between dialogues and topics, we propose three topic-related auxiliary tasks to solve the three questions mentioned above: (1) What are the topics? (2) What is the topic-related dialogue content? (3) What is the content of topic-oriented summaries? The architecture of our proposed method is given in Fig. 2, and we will introduce each task as follows.

1) Backbone Structure: We choose BART as our backbone structure for our follow-up experiments. BART [8] is a denoising sequence-to-sequence pretrained language model, and its pretraining goal is to reconstruct the original text from noisy input. The added noises include token masking, token deletion, sentence permutation, etc. BART has shown its effectiveness on many natural language generation tasks [38], [39]. In the dialogue summarization community, BART has achieved state-of-the-art results on multiple datasets [40]. Besides, some controllable summarization methods [3], [26] also adopt BART as their backbone structures. Considering these studies, we run our experiments on BART. It is also worth mentioning that our method is also available on other sequence-to-sequence models.

Fig. 2(a) presents the BART-based controllable summarization model. Taking the Chinese input as an example, we first concatenate each utterance and the speaker's role in sequence with an "[SEP]" token to compose the dialogue input. Next, we append the given topic to the end of the dialogue with a "[CLS]" token. The final input I is structured as "[CLS] r_1 : u_1 [SEP] r_2 : u_2... r_m : u_m [SEP] [CLS] t [SEP]" and sent to the BART encoder. The input for English dialogues is similar



(a) BART-based Ctrlsum Method

(b) Topic Identification Task

(c) Topic Attention Restriction Task

(d) Topic Summary Distinguishing Task

Fig. 2. The structure of our proposed method. Part (a) refers to the baseline method. Part (b), (c) and (d) represent three auxiliary tasks, respectively.

except for different special tokens due to different tokenizers used in pretrained BART models.[2] The model is trained to output the topic-oriented summary by maximizing the likelihood of generating the ground truth y. The loss function is calculated as:

$$\mathcal{L}_{\text{sum}} = -\sum_{i=1}^{|y|} \log p(y_i | y_{<i}, I) \tag{1}$$

2) What are the Topics: First, the summarization model should understand the semantic information of each topic label. This ability could make the model easier to learn the related information of the given topic for summarization. Therefore, we add a **Topic Identification Task** to the training process by predicting all the topics discussed in the dialogue. As shown in part (b) of Fig. 2, for a given dialogue D, we let the summarization model generate all the existing topics $\{t_1, \ldots, t_k\}$ in D. The input of this task is the whole dialogue, noted as "$[CLS] r_1 : u_1 [SEP] r_2 : u_2 \ldots r_m : u_m [SEP]$". The output of this task is the concatenation of k topics, noted as "$[CLS] t_1, t_2, \ldots, t_k [SEP]$" and we use T to represent it. The order in T is consistent with the the order of appearance of the first related utterance in dialogue D. The **Topic Identification Loss** is defined as:

$$\mathcal{L}_{\text{topic}} = -\sum_{i=1}^{|T|} \log p(T_i | T_{<i}, D) \tag{2}$$

Note that we predict all the topics in a generative way (outputting each word of the topic) instead of serving it as a multi-label classification task based on three considerations: (1) The classification task is unable to train when the number of topic labels is unfixed. (2) Some of the topics are similar and share some words. Generating each word in the topic could help the model learn the semantic information of each topic better. (3) In the summarization task, we take the natural language description of the topic as input. By generating topics in natural language description as well, the model could connect these two tasks and model the relationship between the dialogue and the topic better.

This task is trained together with the summarization task through multi-task training. If the input does not contain a topic label, the model is trained to decode all the topic labels in the dialogue. Otherwise, the model will generate the summary related to the given topic.

3) What is the Topic-Related Dialogue Content: After identifying topics in the dialogue, the model should also have the ability to locate the dialogue utterances related to the given topic for generating the summary. Based on this consideration, we propose a **Topic Attention Restriction Task** by forcing the model to only focus on the topic-related utterances in the decoding process. Specifically, we determine the topic-related utterances by extracting several most related dialogue utterances according to the ground-truth summary. If the dataset has the related key utterance labels, they could be directly served as extracted topic-related utterances for training. Otherwise, we could obtain topic-related utterances by calculating and ranking the ROUGE scores between each utterance and the summary. More details are given in Section IV-C. According to the extracted utterance indexes, we could assign a mask m_i for each utterance u_i, representing whether u_i is related to the given topic. Here 0 denotes that the utterance is related to the topic, and 1 denotes that the utterance is unrelated. To make the model only attend to the topic-related utterances, as presented in part (c) of

[2] For English version, the input is "<s > $r_1 : u_1$ | $r_2 : u_2 \ldots r_m : u_m$ </s> < s> t </s>".

Fig. 2, we design an **Attention Restriction Loss** as:

$$\mathcal{L}_{\text{att}} = \sum_{i=1}^{m} m_i * \alpha_i \tag{3}$$

Here α_i is the sum of the last layer encoder-decoder cross attention weights of all the words in u_i.[3] If the decoder attends to utterances unrelated to the topic, this information is irrelevant and could harm the summary. Thus we need to penalize this case by inducing this loss since the higher weight it attends to wrong utterances, the higher loss it achieves. One alternative method is first training an utterance extraction model to extract the topic-related utterances and then using these utterances for summarization. This pipeline structure may lead to error propagation since the utterance extraction model could provide the wrong utterances or omit key utterances. We thus use this loss function to learn how to extract the utterances implicitly. We will also compare the pipeline method in the experiment section.

4) What is the Content of Topic-Oriented Summaries: The last ability is to generate a good summary related to the topic. When the dialogue has multiple topics, the summaries for each topic should only focus on the topic-related content. Therefore, we are unwilling to see different topic-oriented summaries become similar and have little discrimination. To achieve this goal, we adopt the method of contrastive learning and design a **Topic Summary Distinguishing Task**. As shown in part (d) of Fig. 2, the task tries to increase the discrepancy between summaries for different topics. The **Topic Difference Loss** is given as:

$$\mathcal{L}_{\text{diff}} = \max\{\log p(s_i|t_j, D) - \log p(s_i|t_i, D) + \eta, 0\}, i \neq j \tag{4}$$

In the function, t_i and t_j are two different topics in D, and s_i is the reference summary of t_i. η is a hyper-parameter standing for the margin of two probabilities. The probability of generating s_i is formulized as:

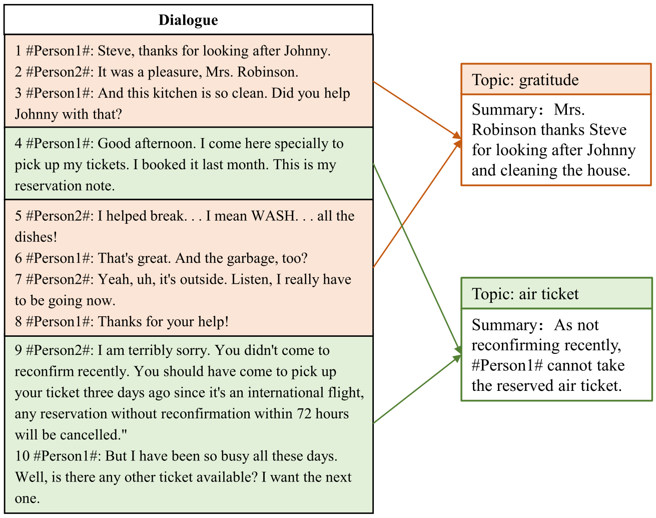
$$\log p(s_i|t_i, D) = \frac{1}{|s_i|} \sum_{k=1}^{|s_i|} \log p(s_{i,k}|s_{i,<k}, t_i, D) \tag{5}$$

This loss function increases the probability of generating the summary with the correct topic and decreases the probability with the wrong topic. Note that in this task, we only train on the dialogues with multiple topics.

5) Training and Inference: We simultaneously train the three tasks with the original summarization task. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{sum}} + \alpha\mathcal{L}_{\text{topic}} + \beta\mathcal{L}_{\text{att}} + \gamma\mathcal{L}_{\text{diff}} \tag{6}$$

α, β, γ are three hyperparameters that control the weights of different auxiliary tasks. During inference, all the auxiliary tasks are abandoned, and we only send the dialogue and topic to the model. The model will generate a text as the summary prediction of the given topic.



1 #Person1#: Steve, thanks for looking after Johnny.
2 #Person2#: It was a pleasure, Mrs. Robinson.
3 #Person1#: And this kitchen is so clean. Did you help Johnny with that?
4 #Person1#: Good afternoon. I come here specially to pick up my tickets. I booked it last month. This is my reservation note.
5 #Person2#: I helped break. . . I mean WASH. . . all the dishes!
6 #Person1#: That's great. And the garbage, too?
7 #Person2#: Yeah, uh, it's outside. Listen, I really have to be going now.
8 #Person1#: Thanks for your help!
9 #Person2#: I am terribly sorry. You didn't come to reconfirm recently. You should have come to pick up your ticket three days ago since it's an international flight, any reservation without reconfirmation within 72 hours will be cancelled."
10 #Person1#: But I have been so busy all these days. Well, is there any other ticket available? I want the next one.

Summary: Mrs. Robinson thanks Steve for looking after Johnny and cleaning the house.

Summary: As not reconfirming recently, #Person1# cannot take the reserved air ticket.

Fig. 3. An example of DIALOGSUM-topic dataset construction process.

IV. Experiments

A. Dataset

Among the existing dialogue summarization datasets, few have the summary for a given topic. The most appropriate dataset is CSDS [6]. It is a Chinese fine-grained customer service dialogue summarization dataset where each summary is split into several QA pairs with a topic label. Most of the dialogues contain more than one topic, thus requiring the summarization model to have the ability of distinguishing the content among different topics. We preprocess the samples in CSDS to obtain the appropriate data for our task. First, we filter out the summaries with the null topic labels. Second, we merge the summaries with the same topic labels and serve them as the final summary for the topic. At last, we take a dialogue in CSDS, one of its topics, and the corresponding summary as a sample for the TODS task. The new dataset is named as CSDS-topic. CSDS also provides all the topics for the dialogue and the key dialogue utterances for each topic, which are useful in our auxiliary tasks.

Apart from CSDS, we also want to examine the effectiveness of our proposed auxiliary tasks on other datasets and in other languages. Here we choose DIALOGSUM [7], an English real-life scenario dialogue summarization dataset, in our experiment since it contains a topic label for each dialogue. However, each dialogue in DIALOGSUM contains only one topic, making this task similar to the traditional dialogue summarization. To simulate the multi-topic scenario, we refer to the process in [25] by randomly selecting two to four dialogues from the dataset and interleaving these selected dialogues to construct a multiple-topic dialogue. In detail, as presented in Fig. 3, dialogues are randomly split into several segments and merged following the original order. After constructing new dialogues, we also take a dialogue, one of its topics, and the corresponding summary as a sample for TODS, naming the new dataset DIALOGSUM-topic. The construction process is executed for train, validation and test sets. As there are three summaries for each dialogue in the test set, we choose summary 1 as the reference in our synthesized

[3] We average the attention scores of all the heads.

TABLE I

STATISTICS OF CSDS-TOPIC AND DIALOGSUM-TOPIC. ALL THE LENGTHS IN CSDS-TOPIC ARE COUNTED ON CHINESE CHARACTERS, AND THE LENGTHS IN DIALOGSUM-TOPIC ARE COUNTED ON ENGLISH WORDS

	CSDS-topic	DIALOGSUM-topic
Train Size	14,430	12,460
Val Size	1,277	500
Test Size	1,232	500
Turns Num.	27.99	30.60
Dial. Length	391.07	595.61
Topic Length	5.50	2.17
Sum. Length	47.85	31.65
Topic Labels Num.	209	7,434
Topic Num. per Sample	2.25	3.23
Examples of Topic Label	attribute consultation discount consultation payment related	bus route wash the dishes shopping

dataset. During the construction, we could obtain the utterance indexes for each topic and all the topics of the dialogue for training our auxiliary tasks.

As DIALOGSUM-topic is a synthesized dataset, we propose the following measures to ensure its effectiveness on evaluating the TODS task.

1) *Coherent speaker information:* We maintain the consistency of speaker names in the new dialogue. In DialogSum, speakers are denoted as “#Person_1#” and “#Person_2#”. Therefore, after merging, “#Person_1#” in the new dialogue could represent all instances of “#Person_1#” in the original dialogue. The same applies to “#Person_2#”. This ensures coherence in the new dialogue and prevents the model from easily discerning different topics based on speaker information.
2) *Consistent domain and expression style:* The dialogues in DialogSum dataset are derived from real-life scenarios, ensuring little domain gap between them. Furthermore, the dataset exhibits consistent expression style across different dialogues, which makes the synthesized dialogue reasonable enough to comprehend and summarize.
3) *Evaluating dataset quality:* We conduct a human evaluation experiment to assess the readability of synthesized dialogues. For each dialogue, we score it based on the level of difficulty in comprehending all the discussed topics and their respective content. We assign the score on a three-point scale, with 0 indicating difficulty in comprehension, 1 for moderate, and 2 for ease. The average result on 50 randomly selected dialogues from the DialogSum-topic test set is 1.32, indicating that most synthesized dialogues, despite containing multiple topics, are readily comprehensible. This evaluation confirms the dataset’s suitability for empirical verification.

Besides, we also discuss about the limitations of DIALOGSUM-topic in Section VI-C.

Detailed statistics of CSDS-topic and DIALOGSUM-topic datasets are given in Table I. Labels in DIALOGSUM-topic are more detailed compared with those in CSDS-topic. Thus, the number of topic labels in DIALOGSUM-topic is also much larger. We also present the proportion of samples with different numbers of topics, and the result is presented in Fig. 4.

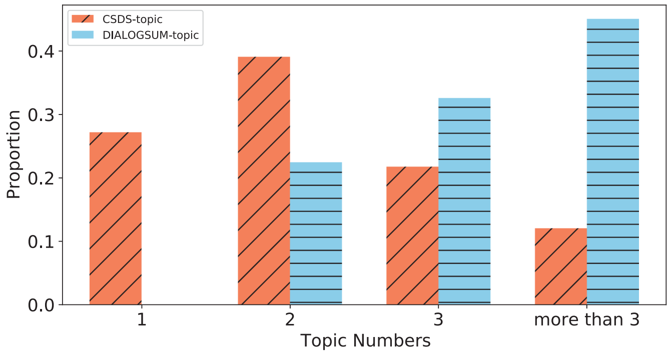


Fig. 4. The proportion of samples with different numbers of topics in CSDS-topic and DIALOGSUM-topic.

B. Baselines

We employ the BART model [8] as the backbone structure in our experiment. For CSDS-topic, we adopt the pretrained Chinese BART model provided by [41], including two versions (bart-base[4], bart-large[5]). For DIALOGSUM-topic, we use the original widely-used bart-base[6] and bart-large[7] checkpoints. To validate the effectiveness of our method, we experiment on both versions. For each version, the baselines to compare include:

Non-topic: Only sending the dialogue to the BART model. This could serve as a lower bound baseline.

Fusion: Fremann and Klementiev [25] provide three fusion methods to add the aspect label into the source document, and they perform similarly in their experiment. Thus, we choose one fusion method (encoder fusion) and transfer it to the BART model. Note that we do not implement this baseline on DIALOGSUM-topic since the topic label number is too large, and more than half of the samples in the test set have new topic labels that are not present in the training set. The model is unable to learn the topic semantic information through aspect fusion in our task settings.

Ctrlsum: A simple but effective controllable summarization method. We concatenate the topic to the dialogue with a special token and send it to the BART model, similar to the methods for aspect-based summarization [3], [26].

Pipeline: We propose an extract-then-summarize pipeline method to achieve the TODS task. First, we train an utterance extraction model to extract the topic-related utterances from the dialogue. For CSDS-topic, the topic-related utterances are all the utterances in a dialogue segment. We select the minimum index of the critical utterances as the start of the segment and the maximum index as the end. For DIALOGSUM-topic, the topic-related utterances are all the utterances that belong to the original dialogue according to the dataset building-up process. Afterward, we train a summarization model by taking the topic-related dialogue utterances as input and generating the topic-oriented summary. At the inference time, we first employ

[4] [Online]. Available: https://huggingface.co/fnlp/bart-base-chinese
[5] [Online]. Available: https://huggingface.co/fnlp/bart-large-chinese
[6] [Online]. Available: https://huggingface.co/facebook/bart-base
[7] [Online]. Available: https://huggingface.co/facebook/bart-large

the utterance extraction model to predict the topic-related utterances and then generate the topic-oriented dialogue summary according to the summarization model.

Ctrlsum+ L_{topic}*:* Adding the topic identification loss to Ctrlsum.

Ctrlsum+ L_{topic} + L_{att}*:* Adding the topic identification and attention restriction losses to Ctrlsum.

Ctrlsum+ L_{topic} + L_{att} + L_{diff}*:* Adding the topic identification, attention restriction and topic difference losses to Ctrlsum.

C. Experiment Settings

In the Fusion method, we achieve the topic representation from the encoder output of the "[CLS]" token. Then we train an attention model to re-weight the encoder output of the dialogue. In the Pipeline method, the utterance extraction model is a BART encoder. We add a linear layer on the encoder output of each "[SEP]" token representing an utterance.[8] We calculate the probability of each utterance being topic-related and use these utterances for summarization.

For all the summarization models, we keep the hyperparameters consistent. We train the model for 10 epochs with a batch size of 24 and save the checkpoints every 400 steps. The optimizer is Adam, and the learning rate is 2e-5. The best checkpoint is chosen based on the lowest validation set loss. For CSDS-topic, we use the BERT tokenizer with a maximum input length of 512 characters, following the settings of the pretrained Chinese BART model. Since some dialogues have long turns, we truncate each turn to 30 characters and then concatenate. This process is effective since it could preserve more turns in the dialogue, and turns with many tokens in CSDS are often meaningless. The maximum output length is 200 characters, and we employ beam search with a beam size of 5. We adjust the non-repeat n-gram size as 6 since there exist some repeated 5-grams in the summary of CSDS. For DIALOGSUM-topic, we use the BART tokenizer with a maximum input length of 1024 tokens. The maximum output length is 200 tokens. The non-repeat n-gram size is set as 3 to prevent generating repeated tokens.

For the topic attention restriction task, the topic-related utterances are obtained differently according to different dataset annotation ways. For CSDS-topic, it contains the critical utterance indexes for each topic. However, there may also exist topic-related utterances that are not critical but relevant in the dialogue. Therefore, we regard a dialogue segment as the topic-related content by selecting the minimum and maximum indexes of the critical utterances as the start and end, respectively. All the utterances in this segment are topic-related.[9] Through this process, if two topic-related segments are crossed, attending to the utterances in the crossed segment is allowed and will not be punished for summarizing both topics. For DIALOGSUM-topic, we obtain the related topic label for each utterance according to the dataset constructing process and thus extract the topic-related utterances straightforwardly.

For our proposed three auxiliary tasks, the hyper-parameters are set as $\alpha = 0.2, \beta = 1, \gamma = 1, \eta = 0.5$, except for $\eta = 0.1$ when experimenting bart-large on DIALOGSUM-topic. The parameter settings are chosen based on the performance of the validation set. The weight of topic identification loss is rather small since we found that increasing α may harm the summarization results. It could lead the model to focus more on topic identification and weaken the summarization ability of the decoder.

D. Evaluation Metrics

1) Automatic Evaluation: Six automatic evaluation metrics are adopted in our experiment, and we divide them into two types:

N-gram overlapping metrics: **ROUGE** [42] and **BLEU** [43]. Here we apply the files2rouge[10] toolkit to calculate the F1 scores of ROUGE-1, ROUGE-2 and ROUGE-L. BLEU score is calculated with the NLTK package.[11]

Distributed representation matching metrics: **BERTScore**[12] [44] and **MoverScore**[13] [45]. We use the official codes to calculate both scores.

2) Human Evaluation: We also let human evaluators rate the summaries generated from different methods through three aspects to evaluate different methods from the human perspective. The rating aspects include:

Informativeness: Does the generated summary contain all the key information in the ground-truth summary?

Non-redundancy: Does the generated summary not contain any unimportant content of the dialogue or any repeated content?

Topic Relevance: Is the generated summary relevant to the given topic?

Three graduate student volunteers are recruited and trained to rate summaries in the three aspects with a five-point rating scale, where 1 represents the worst and 5 represents the best. We randomly select 100 samples from the CSDS-topic test set and let them evaluate according to rating criteria. We calculate the average scores of each aspect among different samples to achieve the final results. Besides, another "Overall" metric is calculated by averaging three scores of different aspects and could reflect the summary quality in general.

V. Results and Analysis

A. Automatic Evaluation Results

We present the automatic evaluation results in Tables II and III. All the methods are divided into two parts based on different backbone structures (BART-base or BART-large), and the results are similar for the two structures and two datasets. First, Ctrlsum outperforms Non-Topic to a large extent. This is especially obvious in DIALOGSUM-topic since all the samples have multiple topics. The results show that the topic label is critical to this task. In CSDS-topic, we are surprised that Fusion performs similarly

[8] For DIALOGSUM-topic, "<s>" represents the topic information; "|" and "'</s>" represent the utterance information.

[9] We also try to only use the critical utterances as topic-related, while it achieves worse results.

[10] [Online]. Available: https://github.com/pltrdy/files2rouge

[11] [Online]. Available: https://www.nltk.org

[12] [Online]. Available: https://pypi.org/project/bert-score/0.2.1/

[13] [Online]. Available: https://github.com/AIPHES/emnlp19-moverscore

TABLE II
THE AUTOMATIC EVALUATION OF DIFFERENT METHODS ON CSDS-TOPIC. BS AND MS DENOTE BERTSCORE AND MOVERSCORE, RESPECTIVELY. ALL THE SCORES ARE PRESENTED IN PERCENTAGES AND HIGHER SCORES REPRESENT BETTER PERFORMANCE

Methods		RG-1	RG-2	RG-L	BLEU	BS	MS
BART-base	Non-Topic	49.55	33.83	47.81	22.30	76.16	21.27
	Fusion	49.32	33.62	47.55	22.10	76.03	21.18
	Ctrlsum	54.99	39.06	53.12	26.73	78.79	26.75
	Pipeline	55.39	39.48	53.19	27.63	79.06	28.48
	Ctrlsum+L_{topic}	55.65	39.58	53.65	27.33	79.02	27.67
	Ctrlsum+L_{topic}+L_{att}	56.07	39.97	54.11	27.80	79.20	28.11
	Ctrlsum+L_{topic}+L_{att}+L_{diff}	**57.05***	**41.07***	**55.07***	**28.44***	**79.81***	**28.80**
BART-large	Non-Topic	49.48	33.91	47.76	22.63	76.07	21.53
	Fusion	50.12	34.14	48.26	22.78	76.35	21.98
	Ctrlsum	57.00	40.89	54.83	28.79	79.64	28.90
	Pipeline	56.68	40.74	54.46	28.69	79.59	29.57
	Ctrlsum+L_{topic}	57.15	41.23	55.13	29.17	79.77	29.17
	Ctrlsum+L_{topic}+L_{att}	57.40	41.47	55.37	29.08	79.92	29.54
	Ctrlsum+L_{topic}+L_{att}+L_{diff}	**57.94***	**42.01***	**55.79***	**29.51***	**80.15***	**30.06**

TABLE III
THE AUTOMATIC EVALUATION OF DIFFERENT METHODS ON DIALOGSUM-TOPIC. THE ABBREVIATIONS ARE KEPT THE SAME AS THE ONES IN TABLE II

Methods		RG-1	RG-2	RG-L	BLEU	BS	MS
BART-base	Non-Topic	26.45	7.36	21.63	7.05	88.48	16.77
	Ctrlsum	40.58	16.48	33.51	11.63	90.89	28.04
	Pipeline	41.00	16.39	33.13	11.55	90.93	28.07
	Ctrlsum+L_{topic}	40.58	16.56	33.27	11.67	90.91	27.93
	Ctrlsum+L_{topic}+L_{att}	41.29	16.75	33.85	11.98	90.97	28.42
	Ctrlsum+L_{topic}+L_{att}+L_{diff}	**41.62***	**17.54***	**34.48***	**12.04**	**91.11***	**28.49**
BART-large	Non-Topic	26.80	8.14	21.93	6.08	88.67	17.50
	Ctrlsum	44.07	19.84	36.19	14.05	91.34	30.54
	Pipeline	43.74	18.74	35.36	13.26	91.33	30.09
	Ctrlsum+L_{topic}	44.83	20.24	37.36	14.72	91.61	31.14
	Ctrlsum+L_{topic}+L_{att}	**45.28***	**20.47**	**37.62***	14.79	**91.68***	**31.95***
	Ctrlsum+L_{topic}+L_{att}+L_{diff}	44.97	20.45	37.49	**14.88**	91.55	31.32

to Non-Topic, demonstrating that this fusion method cannot learn the topic information and thus struggles in generating topic-oriented summaries. Based on its poor performance, we do not experiment with it on DIALOGSUM-topic as this dataset has larger topic numbers.

Pipeline achieves comparable results with Ctrlsum (better on BART-base while worse on BART-large). This result shows that using a pipeline method to extract topic-related content is not very effective, and the error from content extraction[14] may influence the quality of summaries. Another disadvantage of the pipeline method is that it needs to train an extra utterance selection model, which increases the cost of training and inference.

After adding our proposed three auxiliary tasks one by one, the performance keeps increasing in most circumstances. In CSDS-topic, adding all three tasks achieves the best performance with statistically significant improvements over Ctrlsum and Pipeline on most metrics. For the BART-base structure, adding three auxiliary tasks could bring 2.06 points of ROUGE-1 and 2.05 points of MoverScore improvements. For the BART-large structure, our proposed method could also achieve 1.12 points of ROUGE-2 and 1.06 points of MoverScore improvements. Compared with the two backbone structures, BART-large methods perform better than BART-base ones, mainly due to the larger network structure. The results are similar in DIALOGSUM-topic. For the BART-base structure, adding three auxiliary tasks could bring 1.06 points of ROUGE-2 and 0.97 points of ROUGE-L improvements. While there is an exception for BART-large structure, adding topic identification loss and attention restriction loss yields the best result. It improves ROUGE-L by 1.43 points and MoverScore by 1.41 points. We infer that this model is good enough to distinguish different topics due to the larger difference between topics in DIALOGSUM-topic and the stronger comprehension ability of the model structure. As a result, adding the topic summary distinguishing loss does not bring extra improvement.

B. Human Evaluation Results

We also conduct human evaluation analysis on CSDS-topic to further validate the effectiveness of our proposed methods. Here we directly compare the results of Ctrlsum and adding three auxiliary tasks. We conduct the inter-annotator agreement study on three volunteers' scores to ensure consistency among different volunteers. The kappa value for each aspect is 0.56, 0.44, and 0.44, and the kappa value for all the scores is 0.52 on average, showing that the three volunteers have reasonable consistency with each other.

We present the results in Table IV. First, we could easily discover that adding three auxiliary tasks brings improvements

[14] The coverage of topic-related segments is only 60% on utterance level for CSDS-topic.

TABLE IV
THE HUMAN EVALUATION OF DIFFERENT METHODS ON CSDS-TOPIC

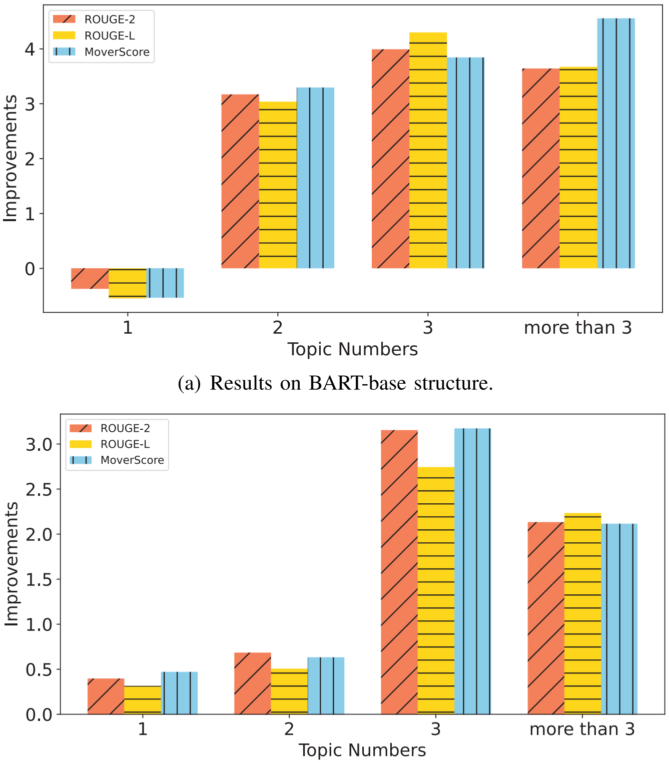
Methods		Informativeness	Non-Redundancy	Topic Relevance	Overall
BART-base	Ctrlsum	3.14	3.35	4.03	3.51
	Ctrlsum+L_{topic}+L_{att}+L_{diff}	**3.33***	**3.63***	**4.29***	**3.75***
BART-large	Ctrlsum	3.37	3.41	4.13	3.64
	Ctrlsum+L_{topic}+L_{att}+L_{diff}	**3.54**	**3.58**	**4.30***	**3.81***

on all the metrics with two different backbone structures, which directly proves the effectiveness of our proposed auxiliary tasks. Most of the improvements are statistically significant. Comparing two backbone structures, BART-large still achieves better results than BART-base after adding the auxiliary tasks, consistent with the automatic evaluation results. When focusing on different aspects, we conclude that they perform comparably on non-redundancy and topic relevance. Thus, the larger model has the only advantage on the informativeness metric after adding the auxiliary tasks.

C. Impact of Topic Numbers

Since our proposed auxiliary tasks have shown great improvements on many metrics, we are curious about in which cases our proposed methods work. Therefore, we divide the test set into several parts according to the topic numbers of the dialogue in the sample. The topic numbers include 1, 2, 3, and more than 3. We choose the CSDS-topic dataset to evaluate since it is a natural multi-topic dataset and has samples with only one topic as well. We calculate the results of adding three auxiliary tasks on all the metrics for different sample parts and present the improvements of our proposed method over the Ctrlsum baseline in Fig. 5.

The horizontal axis stands for different topic numbers, and the vertical axis represents the improvements of adding three auxiliary tasks over the Ctrlsum baseline. Here we present the results of three metrics, and they all exhibit similar tendencies. When the dialogue has only one topic, the provided topic label contains little information, and the summary could be easily generated without the topic. Thus, there is no apparent change after adding the auxiliary tasks on these samples. As the topic number grows, the summarization model needs to distinguish the difference between each topic and generate the content most related to the topic. In these samples, adding auxiliary tasks could significantly enhance all the metrics greatly. The most significant improvement occurs on the samples with three topics. For dialogues with more than three topics, the improvements are smaller, mainly due to their greater difficulty and small sample sizes. Overall, the outcome demonstrates the priority of our proposed auxiliary tasks on the dialogue with multiple topics.



(a) Results on BART-base structure.

(b) Results on BART-large structure.

Fig. 5. Improvements of our proposed auxiliary tasks on samples with different topic numbers in CSDS-topic.

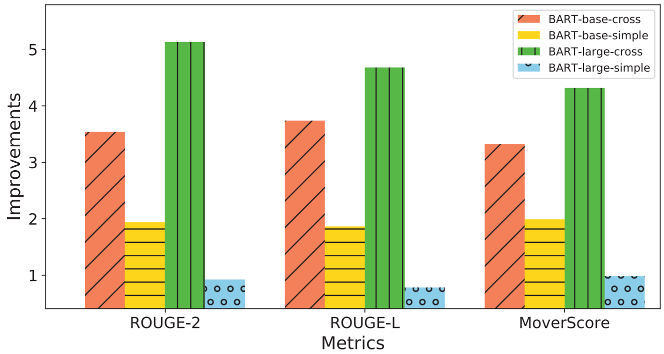


Fig. 6. The improvements of our proposed auxiliary tasks on "cross" samples and "simple" samples in CSDS-topic.

D. Impact of Crossed Topic Segments

The discourse structure in a dialogue is sometimes complicated, and different topics may have crossed dialogue segments. This phenomenon makes TODS difficult since it is hard to distinguish different topics in the dialogue. In the CSDS-topic dataset, some samples have crossed topic segments while others do not. We thus try to examine in which cases our proposed method works better. We divide the test set into "cross" and "simple". "Cross" means that the topic-related dialogue utterances of the sample have crossed ranges compared with other topics in the same dialogue; "simple" means that the utterances for different topics do not have crossed ranges.

We present the experiment results in Fig. 6. The horizontal axis represents different metrics, and the vertical axis stands for

TABLE V
THE TOPIC IDENTIFICATION RESULTS

Methods	Micro			Macro		
	P	R	F1	P	R	F1
BART-base	**50.88**	46.20	48.43	52.10	48.78	48.63
+aux tasks	50.74	**49.24**	**49.98**	**52.99**	**52.70**	**50.94**
BART-large	54.92	53.56	54.23	56.94	56.75	55.07
+aux tasks	**55.85**	**56.51**	**56.18**	**57.77**	**59.85**	**57.03**

the improvements of adding three auxiliary tasks over the Ctrlsum baseline. A bar of the same color represents a method with a single sample type. First, all the improvements have positive values, indicating that our method is effective on both types of samples. Second, comparing cross samples (in orange and green) and single samples (in yellow and blue), the improvements on cross samples are significantly larger than the ones on simple samples. It demonstrates that our method is more effective at helping extract the correct content when different topics have crossed dialogue segments.

E. Extra Abilities

Our proposed three auxiliary tasks derive from solving three topic-related key questions: (1) What are the topics? (2) What is the topic-related dialogue content? (3) What is the content of topic-oriented summaries? In this section, we want to examine whether our trained model has learned these three abilities. All the analysis below is based on the CSDS-topic dataset.

1) Predicting Topics in the Dialogue: To compare the ability to predict topic labels, we employ the backbone structure to train a topic identification model as the baseline. We send the dialogue as input and let the model generate all the topics in the dialogue. Then we use our trained summarization model with the auxiliary tasks to generate all the topics as well and compare the result with the ground truth topics. Here we calculate the precision, recall, and F1 scores using exact match evaluation. Since one dialogue may have multiple topics, we calculate both micro and macro scores. The micro scores count the correctly predicted topics in all the samples and calculate the precision, recall, and F1 scores. The macro scores first calculate these three scores for each dialogue sample and then obtain the average scores among different samples.

The results are given in Table V. After adding the auxiliary tasks, the topic prediction results are improved on both micro and macro scores. It shows that our trained summarization model has a stronger ability to learn the semantic information of topics than learning the topic identification task alone. Common errors include omitting topic labels in long dialogues and confusing similar topic labels, such as *"delivery tracking"* and *"delivery contacting"*.

Meanwhile, we also conduct cross-domain topic label prediction experiment to verify whether existing models have the domain adaptation ability. In detail, we translate the dialogues and topic labels in DIALOGSUM-topic into Chinese and use the model trained on CSDS-topic to predict. The result is rather poor. The model predicts topic labels in CSDS-topic for 30 percent of samples in DIALOGSUM-topic. For the other 70 percent, it outputs a summary-like sentence instead of topic labels. It shows that two datasets have huge differences, and cross-domain topic label prediction is difficult to learn under existing methods.

TABLE VI
THE RESULTS OF KEY UTTERANCES ATTENTION RATIO

Methods	Attention Ratio
BART-base	0.400
+auxiliary tasks	**0.946**
BART-large	0.323
+auxiliary tasks	**0.982**

TABLE VII
THE SIMILARITY OF SUMMARIES IN THE SAME DIALOGUE WITH DIFFERENT TOPICS

Models	ROUGE-1	ROUGE-2	ROUGE-L
BART-base	62.84	53.04	62.01
+auxiliary tasks	**49.13**	**35.31**	**47.91**
BART-large	55.19	43.26	53.73
+auxiliary tasks	**46.93**	**33.30**	**45.50**
Reference	38.21	21.72	36.18

2) Locating the Topic-Oriented Dialogue Utterances: The topic-related content is crucial for generating topic-oriented summaries. We want to find out whether our trained summarization models focus on this content in the decoding step. Therefore, we sum up the encoder-decoder attention distribution of all the decoding steps, normalize it into a new probability distribution, and count the attention ratio on the topic-related key utterances for different models.[15]

We present the results in Table VI. The method with training auxiliary tasks achieves a higher attention ratio on key utterances than the Ctrlsum baseline. It proves that our trained summarization model has the ability to attend to more important dialogue content that helps improve the summary quality.

3) Distinguishing Summaries for Different Topics: Since topic-oriented summaries need to focus on the topic-related content, the summaries of different topics in the same dialogue should be different enough to present their topic relevance. Unfortunately, many summaries are highly overlapped when topics are similar. Here we try to validate whether our trained summarization model has the ability to generate diverse summaries for different topics. We calculate the ROUGE scores of summaries of different topics in the same dialogue, and the higher score represents less diversity and poorer topic-oriented summary quality.

We present the results in Table VII. There is a huge drop in ROUGE scores after adding the auxiliary tasks on both BART-base and BART-large. However, it still has an obvious gap with the scores of the ground truth. In summary, it proves that our trained summarization model can distinguish different input topics and generate more diverse summaries for different topics.

F. Summaries for all Topics

In some cases, we are interested in all the important content in the dialogue. Although our method focuses on topic-oriented dialogue summarization, we try to enlarge the usage of our

[15] The key utterances are labeled in CSDS-topic dataset.

TABLE VIII
RESULTS ON GENERATING SUMMARIES FOR ALL THE TOPICS

Models	RG-1	RG-2	RG-L	BLEU	BS	MS
BART-base	57.39	40.86	55.12	28.76	79.43	30.44
+aux tasks	**58.41**	**41.87**	**56.28**	**29.99**	**79.80**	**30.84**
BART-large	58.73	42.07	56.47	29.92	**80.08**	**31.68**
+aux tasks	**58.99**	**42.54**	**57.01**	**30.70**	80.03	31.40

method by adding another topic label named "all topic". We employ the overall summary in CSDS as the target summary and add these samples to the training set. The training process is kept the same, and we compare the results of generating the overall summaries.

The results are presented in Table VIII. We compare the result of our method with directly training a summarization model given the dialogue alone on the same backbone structures. The method with training auxiliary tasks performs better than the baseline on most metrics. It proves that our method also has the ability to generate the overall summary, and the topic-oriented training process is beneficial in extracting the information of all the topics as well.

G. Case Study

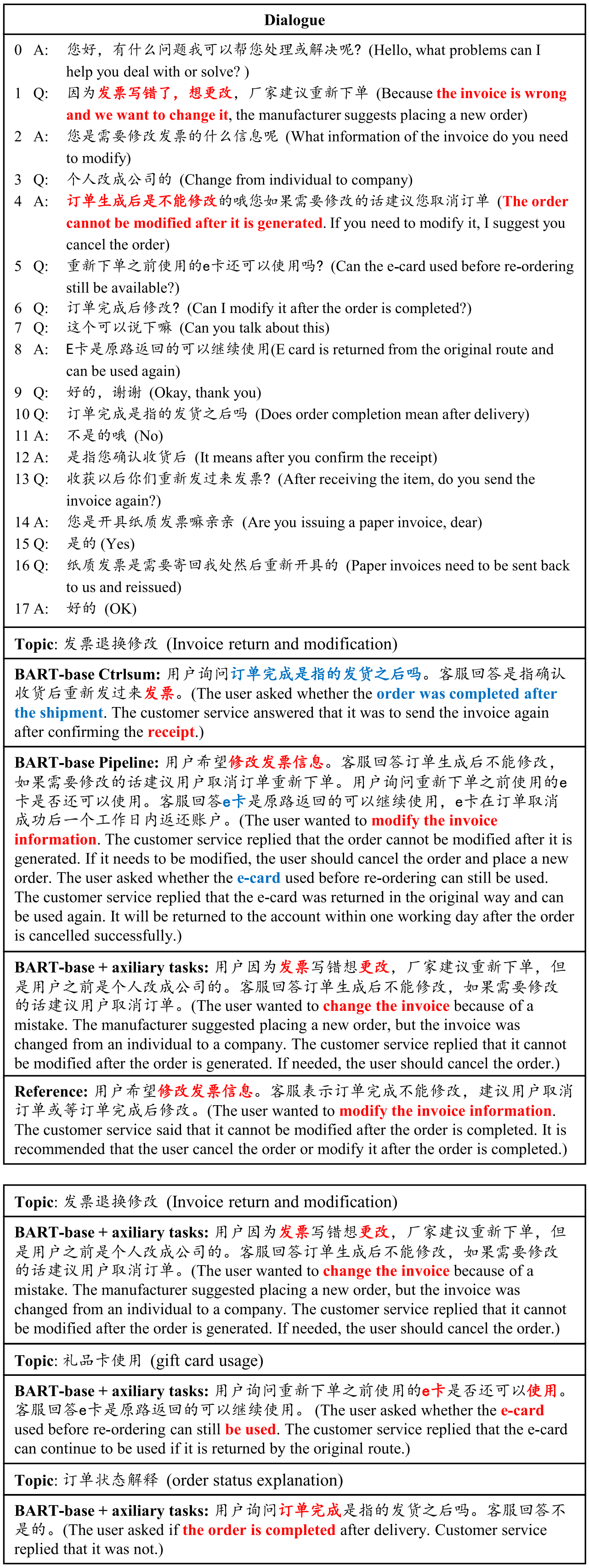
We present an example from CSDS-topic in Fig. 7 to illustrate the effectiveness of our methods directly. In this example, the dialogue has three topics, including *"invoice return and modification"*, *"gift card usage"*, and *"order status explanation"*. For topic *"invoice return and modification"*, the Ctrlsum baseline wrongly extracts another issue about *"order completion"*, mainly influenced by the keyword *"receipt"* existing in this issue. Although the Pipeline method extracts the right content about invoice modification, it also extracts other issues about the *"e-card"*. It may be due to the error of extracting topic-related utterances. Only our proposed method accurately summarizes the key information in the reference and does not generate other redundant content. Meanwhile, we also list the summaries provided by our method on all three topics. Given different topics, our method could distinguish the semantic differences and only summarize the content related with the topic.

VI. Discussion

A. The Scalability of TODS

Our proposed TODS task focuses on generating the summary for a given topic. However, considering the complex relationship between dialogues and topics, TODS could be expanded into more complicated cases. We list three reasonable expansion directions for TODS.

Generating both topics and related summaries: Each dialogue in TODS is accompanied by a particular preference topic. However, readers may have no topic preferences in some situations, such as interviews and meetings. To achieve the core information of the dialogue with multiple topics, a structural summary is more welcome by presenting all the topics mentioned in the dialogue and the related summaries for each topic.



Dialogue
0 A: 您好，有什么问题我可以帮您处理或解决呢？(Hello, what problems can I help you deal with or solve?)
1 Q: 因为**发票写错了，想更改**，厂家建议重新下单 (Because **the invoice is wrong and we want to change it**, the manufacturer suggests placing a new order)
2 A: 您是需要修改发票的什么信息呢 (What information of the invoice do you need to modify)
3 Q: 个人改成公司的 (Change from individual to company)
4 A: **订单生成后是不能修改**的哦您如果需要修改的话建议您取消订单 (**The order cannot be modified after it is generated**. If you need to modify it, I suggest you cancel the order)
5 Q: 重新下单之前使用的e卡还可以使用吗？(Can the e-card used before re-ordering still be available?)
6 Q: 订单完成后修改？(Can I modify it after the order is completed?)
7 Q: 这个可以说下嘛 (Can you talk about this)
8 A: E卡是原路返回的可以继续使用(E card is returned from the original route and can be used again)
9 Q: 好的，谢谢 (Okay, thank you)
10 Q: 订单完成是指的发货之后吗 (Does order completion mean after delivery)
11 A: 不是的哦 (No)
12 A: 是指您确认收货后 (It means after you confirm the receipt)
13 Q: 收获以后你们重新发过来发票？(After receiving the item, do you send the invoice again?)
14 A: 您是开具纸质发票嘛亲亲 (Are you issuing a paper invoice, dear)
15 Q: 是的 (Yes)
16 Q: 纸质发票是需要寄回我处然后重新开具的 (Paper invoices need to be sent back to us and reissued)
17 A: 好的 (OK)
Topic: 发票退换修改 (Invoice return and modification)
BART-base Ctrlsum: 用户询问**订单完成是指的发货之后吗**。客服回答是指确认收货后重新发过来**发票**。(The user asked whether the **order was completed after the shipment**. The customer service answered that it was to send the invoice again after confirming the **receipt**.)
BART-base Pipeline: 用户希望**修改发票信息**。客服回答订单生成后不能修改，如果需要修改的话建议用户取消订单重新下单。用户询问重新下单之前使用的e卡是否还可以使用。客服回答**e卡**是原路返回的可以继续使用，e卡在订单取消成功后一个工作日内返还账户。(The user wanted to **modify the invoice information**. The customer service replied that the order cannot be modified after it is generated. If it needs to be modified, the user should cancel the order and place a new order. The user asked whether the **e-card** used before re-ordering can still be used. The customer service replied that the e-card was returned in the original way and can be used again. It will be returned to the account within one working day after the order is cancelled successfully.)
BART-base + axiliary tasks: 用户因为**发票**写错想**更改**，厂家建议重新下单，但是用户之前是个人改成公司的。客服回答订单生成后不能修改，如果需要修改的话建议用户取消订单。(The user wanted to **change the invoice** because of a mistake. The manufacturer suggested placing a new order, but the invoice was changed from an individual to a company. The customer service replied that it cannot be modified after the order is generated. If needed, the user should cancel the order.)
Reference: 用户希望**修改发票信息**。客服表示订单完成不能修改，建议用户取消订单或等订单完成后修改。(The user wanted to **modify the invoice information**. The customer service said that it cannot be modified after the order is completed. It is recommended that the user cancel the order or modify it after the order is completed.)

Topic: 发票退换修改 (Invoice return and modification)
BART-base + axiliary tasks: 用户因为**发票**写错想**更改**，厂家建议重新下单，但是用户之前是个人改成公司的。客服回答订单生成后不能修改，如果需要修改的话建议用户取消订单。(The user wanted to **change the invoice** because of a mistake. The manufacturer suggested placing a new order, but the invoice was changed from an individual to a company. The customer service replied that it cannot be modified after the order is generated. If needed, the user should cancel the order.)
Topic: 礼品卡使用 (gift card usage)
BART-base + axiliary tasks: 用户询问重新下单之前使用的**e卡**是否还可以**使用**。客服回答e卡是原路返回的可以继续使用。 (The user asked whether the **e-card** used before re-ordering can still **be used**. The customer service replied that the e-card can continue to be used if it is returned by the original route.)
Topic: 订单状态解释 (order status explanation)
BART-base + axiliary tasks: 用户询问**订单完成**是指的发货之后吗。客服回答不是的。(The user asked if **the order is completed** after delivery. Customer service replied that it was not.)

Fig. 7. A case study of summaries generated by different methods.

Dealing with nonexistent topics: In real scenarios, sometimes the dialogue does not discuss the provided topic at all. Our proposed task does not include this scenario since all the topics in CSDS-topic and DIALOGSUM-topic are related to their corresponding dialogues. If the dialogue is unrelated to the given topic, the summarization model should be able to figure out and generate a null summary or special token to indicate that no relevant content could be found.

Different topic granularities: The topics in CSDS-topic are derived from dialogue utterance intents, while the topics in DIALOGSUM-topic are manually labeled short descriptions (around three tokens). The granularity of topics in each dataset is kept the same. In reality, people may be concerned about various granularities of topics, sometimes as big as sports or news, sometimes as small as the World Cup finals or a particular TV show. More complicated datasets with specialized designed topics and related summaries are required to let the model deal with topics at multiple granularities.

B. Remaining Challenges in TODS

We analyze the results of several baselines and our proposed methods and summarize three challenges that still remain in existing methods.

Dialogue with more topics: We calculate the summary quality of samples with different topics separately in DIALOGSUM-topic. The results demonstrate that existing methods perform worse on dialogues with more topics.[16] Dialogues with more topics are usually longer and have more intricate discourse structures, making TODS more challenging as well.

Wide-ranging topics: Unlike the specific topics provided in DIALOGSUM-topic, topics in CSDS-topic are more abstract and general, such as “attribute consultation” or “order status explanation”. These topics could cover a wide range of issues and have little lexical coverage with the related summaries. Existing methods are prone to omitting or wrongly extracting the critical issues related to these topics.

Semantically similar topics: It is also hard for summarization methods to recognize the subtle differences between similar topics when topics are too specific. Taking topics in DIALOGSUM-topic as an example, “singing” and “English songs” are two topics discussed in a dialogue. The former emphasizes the action while the latter emphasizes the content of songs. It requires summarization methods to understand the semantic information of dialogues and topics more accurately.

C. Limitations of DIALOGSUM-Topic

Since DIALOGSUM-topic is a synthesized dataset, its dialogues may differ from real multi-topic dialogues, potentially introducing biases into experimental results. Thus, we analyze the potential implications of using DIALOGSUM-topic in our experiments.

Dialogue context incoherence: During the merging process, we do not add any cohesive words, such as “in addition” or “by the way”, to facilitate topic transitions. This approach makes synthesized dialogues different from real ones and may potentially hurt the dialogue coherence. Moreover, summarization models may struggle with recognizing topic shifts in the new dialogue. This bias may lead to inconsistent experiment results on DIALOGSUM-topic comparing with other real multi-topic dialogue datasets.

Dialogue segmentation problem: When segmenting original dialogues, we randomly split them without considering the dialogue flow. Although it is common that different topics have crossed segments in a dialogue, particularly in an online conversation, this segmentation approach may disrupt the dialogue flow and significantly increase the comprehension difficulty of synthesized dialogues. Our choice of segmentation is aimed at constructing dialogues with more complex topic structures that could widen the gap between different summarization methods on TODS. Due to this approach, the performance gap observed on DIALOGSUM-topic may differ from the one on other datasets.

VII. Conclusion and Future Work

In this paper, we focus on the topic-oriented dialogue summarization task. To learn the relationship between dialogues and topics better, we design three topic-related auxiliary tasks to enhance the abilities of the model. The topic identification task helps learn the semantic information of each topic label; the topic attention restriction task helps extract accurate content from the dialogue; the topic summary distinguishing task helps generate diverse summaries for different topics and focus on the topic-related information. Detailed experiment analysis proves that all proposed auxiliary tasks could significantly improve summary qualities and bring additional abilities to the model (e.g., topic prediction, generating summaries for all topics). In the future, we will try to expand the application scenarios for TODS and solve the challenges provided in the discussion section.

References

[1] C. Zong, R. Xia, and J. Zhang, *Text Data Mining*. vol. 711, Berlin, Germany: Springer, 2021.

[2] M. Soni et al., “An empirical study of topic transition in dialogue,” in *Proc. 3rd Workshop Comput. Approaches Discourse*, 2022, pp. 92–99.

[3] J. He, W. Kryściński, B. McCann, N. Rajani, and C. Xiong, “CTRLsum: Towards generic controllable text summarization,” 2020, *arXiv:2012.04281*.

[4] A. Fan, D. Grangier, and M. Auli, “Controllable abstractive summarization,” in *Proc. 2nd Workshop Neural Mach. Transl. Gener.*, 2018, pp. 45–54.

[5] H. Li, J. Zhu, J. Zhang, C. Zong, and X. He, “Keywords-guided abstractive sentence summarization,” in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 8196–8203.

[6] H. Lin et al., “CSDS: A fine-grained chinese dataset for customer service dialogue summarization,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4436–4451.

[7] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, “DialogSum: A real-life scenario dialogue summarization dataset,” in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 5062–5074.

[8] M. Lewis et al., “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[9] I. McCowan et al., “The AMI meeting corpus,” in *Proc. Measuring Behav., 5th Int. Conf. Methods Techn. Behav. Res.*, 2005, pp. 137–140.

[16] The ROUGE-2 scores of dialogues with different topic numbers on bart-large Ctrlsum are: 2 topics: 21.49, 3 topics: 19.53, 4 topics: 19.19.

[10] A. Janin et al., “The ICSI meeting corpus,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, vol. 1, pp. 364–367.

[11] C. Zhu, R. Xu, M. Zeng, and X. Huang, “A hierarchical network for abstractive meeting summarization with cross-domain pretraining,” in *Proc. Findings Assoc. Comput. Linguistics*, 2020, pp. 194–203.

[12] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, “SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization,” in *Proc. 2nd Workshop New Front. Summarization*, 2019, pp. 70–79.

[13] X. Feng, X. Feng, L. Qin, B. Qin, and T. Liu, “Language model as an annotator: Exploring DialoGPT for dialogue summarization,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, vol. 1, pp. 1479–1491.

[14] Z. Liu and N. Chen, “Controllable neural dialogue summarization with personal named entity planning,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 92–106. [Online]. Available: https://aclanthology.org/2021.emnlp-main.8

[15] Z. Liu, A. Ng, S. Lee, A. T. Aw, and N. F. Chen, “Topic-aware pointer-generator networks for summarizing spoken conversations,” in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 814–821.

[16] Y. Song, Y. Tian, N. Wang, and F. Xia, “Summarizing medical conversations via identifying important utterances,” in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 717–729.

[17] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, “Generating SOAP notes from doctor-patient conversations using modular summarization techniques,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, vol. 1, pp. 4958–4972.

[18] C. Liu, P. Wang, J. Xu, Z. Li, and J. Ye, “Automatic dialogue summary generation for customer service,” in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 1957–1965.

[19] G. Feigenblat, C. Gunasekara, B. Sznajder, S. Joshi, D. Konopnicki, and R. Aharonov, “TWEETSUMM - A dialog summarization dataset for customer service,” in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 245–260.

[20] H. Lin, J. Zhu, L. Xiang, Y. Zhou, J. Zhang, and C. Zong, “Other roles matter! enhancing role-oriented dialogue summarization via role interactions,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, vol. 1, pp. 2545–2558. [Online]. Available: https://aclanthology.org/2022.acl-long.182

[21] L. Zhao, W. Xu, and J. Guo, “Improving abstractive dialogue summarization with graph structures and topic words,” in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 437–449.

[22] J. Chen and D. Yang, “Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 4106–4118.

[23] Y. Zou et al., “Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 14665–14673.

[24] J. Liu et al., “Topic-aware contrastive learning for abstractive dialogue summarization,” in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 1229–1243.

[25] L. Frermann and A. Klementiev, “Inducing document structure for aspect-based summarization,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6263–6273.

[26] B. Tan, L. Qin, E. Xing, and Z. Hu, “Summarizing text on any aspects: A knowledge-informed weakly-supervised approach,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 6301–6309.

[27] R. K. Amplayo, S. Angelidis, and M. Lapata, “Aspect-controllable opinion summarization,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6578–6593.

[28] O. Ahuja, J. Xu, A. Gupta, K. Horecka, and G. Durrett, “ASPECTNEWS: Aspect-oriented summarization of news documents,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, 2022, vol. 1, pp. 6494–6506.

[29] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2004, pp. 168–177.

[30] A.-M. Popescu and O. Etzioni, “Extracting product features and opinions from reviews,” in *Proc. Conf. Hum. Lang. Technol. Empirical Methods Natural Lang. Process.*, 2005, pp. 339–346.

[31] M. Zhong et al., “QMSum: A new benchmark for query-based multi-domain meeting summarization,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 5905–5921.

[32] P. J. Liu et al., “Generating wikipedia by summarizing long sequences,” in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hyg0vbWC-

[33] S. Kulkarni, S. Chammas, W. Zhu, F. Sha, and E. Ie, “AQuaMuSe: Automatically generating datasets for query-based multi-document summarization,” 2020, *arXiv:2010.12694*.

[34] L. Huang, S. Cao, N. Parulian, H. Ji, and L. Wang, “Efficient attentions for long document summarization,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 1419–1436.

[35] Y. Zhang et al., “Summ^{N}: A multi-stage summarization framework for long input dialogues and documents,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, vol. 1, pp. 1592–1604.

[36] Z. Mao et al., “DYLE: Dynamic latent extraction for abstractive long-input summarization,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, vol. 1, pp. 1687–1698.

[37] J. Vig, A. R. Fabbri, W. Kryściński, C.-S. Wu, and W. Liu, “Exploring neural models for query-focused summarization,” in *Proc. Findings Assoc. Comput. Linguistics*, 2022, pp. 1455–1468.

[38] Y. Liu, Y. Wan, L. He, H. Peng, and P. S. Yu, “KG-BART: Knowledge graph-augmented BART for generative commonsense reasoning,” in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 6418–6425.

[39] H. Lin, L. Xiang, Y. Zhou, J. Zhang, and C. Zong, “Augmenting slot values and contexts for spoken language understanding with pretrained models,” in *Proc. Interspeech*, 2021, pp. 4703–4707.

[40] X. Feng, X. Feng, and B. Qin, “A survey on dialogue summarization: Recent advances and new frontiers,” in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 5453–5460.

[41] Y. Shao et al., “CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation,” 2021, *arXiv:2109.05729*.

[42] C.-Y. Lin and E. Hovy, “Manual and automatic evaluation of summaries,” in *Proc. Workshop Autom. Summarization Phildadelphia*, 2002, pp. 45–51.

[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTscore: Evaluating text generation with BERT,” in *Proc. 8th Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[45] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, “MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance,” in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 563–578.

Haitao Lin received the B.S. degree in 2018 from the University of Chinese Academy of Sciences, Beijing, China, where he is currently working toward the Ph.D. degree with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation. His research interests include summarization, dialogue understanding, and natural language processing. He was ACL-IJCNLP 2021 as Student Research Workshop Co-Chair.

Junnan Zhu received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently an Assistant Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include summarization, text generation, and multimedia.

Lu Xiang received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2011 and the M.S. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2014, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include dialogue systems, text generation, and natural language processing.

Feifei Zhai received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He completed postdoc research with the City University of New York, New York, NY, USA, in 2015 and was at IBM Research Center as a Research Member. After that, he lead the Sogou Machine Translation Team and built Sogou Translation System. He is currently the Director of Fanyu AI Research, Beijing Fanyu Technology Company LTD.

Yu Zhou received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in January 2008. She is currently a Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA. Her research interests include machine translation, text mining, summarization, and natural language processing.

Jiajun Zhang (Senior Member, IEEE) received the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently a Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include machine translation, natural language processing, multilingual and multimodal analysis.

Chengqing Zong (Fellow, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in March 1998. He is currently a Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. He has authored a book titled *Statistical Natural Language Processing* and coauthored a book titled *Text Data Mining*. His research interests include machine translation, dialog systems, and linguistic cognitive computing. Dr. Zong is a Fellow of *Association for Computational Linguistics*, Chinese Association for Artificial Intelligence, and China Computer Federation. He has served many top-tier international conferences, such as, ACL-IJCNLP 2021 as the Conference Chair, ACL-IJCNLP 2015 and COLING 2020 as PC Co-Chair. He is also a Member of the Editorial Board of theIEEE INTELLIGENT SYSTEMS.