Dialogue Summarization with Mixture of Experts based on Large Language Models

Yuanhe Tian♠♥, **Fei Xia**♥, **Yan Song**♠†
♠University of Science and Technology of China ♥University of Washington
♥{yhtian, fxia}@uw.edu ♠clksong@gmail.com

Abstract

Dialogue summarization is an important task that requires to generate highlights for a conversation from different aspects (e.g., content of various speakers). While several studies successfully employ large language models (LLMs) and achieve satisfying results, they are limited by using one model at a time or treat it as a black box, which makes it hard to discriminatively learn essential content in a dialogue from different aspects, therefore may lead to anticipation bias and potential loss of information in the produced summaries. In this paper, we propose an LLM-based approach with role-oriented routing and fusion generation to utilize mixture of experts (MoE) for dialogue summarization. Specifically, the role-oriented routing is an LLM-based module that selects appropriate experts to process different information; fusion generation is another LLM-based module to locate salient information and produce finalized dialogue summaries. The proposed approach offers an alternative solution to employing multiple LLMs for dialogue summarization by leveraging their capabilities of in-context processing and generation in an effective manner. We run experiments on widely used benchmark datasets for this task, where the results demonstrate the superiority of our approach in producing informative and accurate dialogue summarization.[1]

1 Introduction

Dialogue summarization is a crucial task that aims to extract essential information from a dialogue, which attracts much attention from existing studies in recent years (Gurevych and Strube, 2004; Gliwa et al., 2019). Different from documents that are monographs from one writer, dialogues involve contents from different roles and thus summarizing them needs to consider the interactions among them. Table 1 presents an example dialogue with its summary from two speakers (i.e., S1 and S2), where key information (in green and blue) of different roles are drawn from their interactions on discussing a concerned topic (in purple).

DIALOGUE
S1: *Hi, Mr. Smith. I'm Doctor Hawkins. Why are you here today?*
S2: *I found it would be a good idea to get a check-up.*
S1: *Yes, well, you haven't had one for 5 years. You should have one every year.*
S2: *I know. I figure as long as there is nothing wrong, why go see the doctor?*
S1: *Well, the best way to avoid serious illnesses is to find out about them early. So try to come at least once a year for your own good.*
S2: *Ok.*
S1: *Let me see here. Your eyes and ears look fine. Take a deep breath, please. Do you smoke, Mr. Smith?*
S2: *Yes.*
S1: *Smoking is the leading cause of lung cancer and heart disease, you know. You really should quit.*
S2: *I've tried hundreds of times, but I just can't seem to kick the habit.*
S1: *Well, we have classes and some medications that might help. I'll give you more information before you leave.*
S2: *Ok, thanks doctor.*
SUMMARY
Mr. Smith's getting a check-up, and Doctor Hawkins advises him to have one every year. Hawkins'll give some information about their classes and medications to help Mr. Smith quit smoking.

Table 1: An example dialogue between two speakers (i.e., S1 and S2) and its corresponding summary, where essential content shared in the dialogue and the summary are highlighted in the same colors.

Existing studies for dialogue summarization (Li et al., 2023b; Gao et al., 2023; Chen et al., 2023a; Hua et al., 2023; Ouyang et al., 2023) tend to utilize end-to-end approaches to produce dialogue summaries, where advanced text encoders, such as large language models (LLMs), are used to identify the key content in the dialogue. To further enhance the capability of models to identify essential information in the dialogue, additional information is applied with the interaction among dialogue participants, the structure of the dialogue, and the topics they are discussing, etc. (Kano et al., 2020; Song et al., 2020; Krishna et al., 2021; Zou et al., 2021; Liu et al., 2021; Zhang et al., 2022; Lin et al., 2023; Liang et al., 2023; Liu and Xu, 2023). Although these studies obtain promising results, they

†Corresponding author.

[1]Materials related to the paper is available at https://github.com/synlp/DiaSum-MoE.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7143–7155
August 11-16, 2024

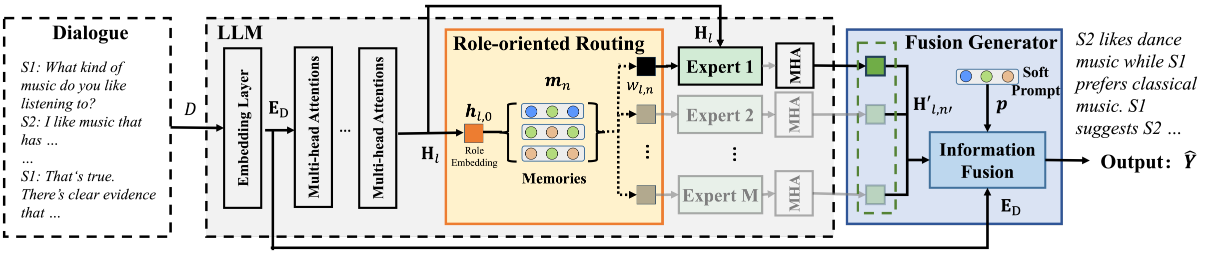

Figure 1: The overall architecture of our approach with an example input dialogue and output summary. The role-oriented routing and fusion generation are illustrated in the middle and right parts of the figure, respectively.

are mainly performed with a single-model or black box design, thus is potentially limited in generating biased output in the dialogue summary where diversified information are not comprehensively processed. Therefore, it is inevitable to consider whether there are more reasonable model designs to overcome such limitation.

Since mixture-of-experts (MoE) with LLMs demonstrates their effectiveness in many tasks (Chen et al., 2023b; Shen et al., 2023a,b; Li et al., 2023a), it is inspired to leverage such design for dialogue summarization to fully leverage the potential of LLMs to understand and generate summaries from different aspects. However, applying MoE to LLMs for dialogue summarization is not a trivial task, where careful design of the "mixture" is essential to prevent issues such as information dilution, so that essential contents are processed and preserved by appropriate experts, and the outputs from all experts retains a sufficient information-bearing without being distracted. Therefore, in this paper, we propose an MoE-based approach for dialogue summarization that alters the inner structure of LLMs. Specifically, we perform an utterance based processing with role-oriented routing, which is a part of LLM, to effectively identify the mapping of different experts for a particular utterance with the support of the entire dialogue, so as to avoid "*Blind Men and the Elephant*" phenomenon[2] when comprehending the content and extracting key information from the utterance. Then use another parts of LLM as experts and use the routing to selected some of them to take the utterance and generate outputs based on essential content in it. The fusion generation uses another LLM to combine the output of each expert and chooses the most valuable ones from them to form the final summary with all essential information from various aspects included in the dialogue. We run experiments on four widely used benchmark English and Chinese datasets. The results and further analyses demonstrate the superiority of our approach, which outperforms strong baselines and achieves state-of-the-art performance on all datasets, also show the validity of our design for each component.

2 The Approach

The overall architecture of our approach for dialogue summarization is illustrated in Figure 1, where the MoE framework consists of three main components, namely, role-oriented routing (RoR), expert processing, and fusion generation (FG). Specifically, RoR is based on the first K layers of an LLM that takes the input dialogue $\mathcal{D} = [(\mathcal{R}_1, \mathcal{U}_1), \cdots (\mathcal{R}_L, \mathcal{U}_L)]$ with L utterances associated with their speakers (i.e., $\mathcal{R}_l$ denotes the speaker of the l-th utterance $\mathcal{U}_l$), then selects appropriate experts to process each utterance $\mathcal{U}_l$ and its $\mathcal{R}_l$. The selected experts (each of which is the last $(K_{LLM} - K)$ layers of the LLM, where K_{LLM} is the number of layers in the LLM) generate key content of $\mathcal{U}_l$ with the guidance of $\mathcal{R}_l$, which is repeated L times to process all utterance-speaker pairs. The FG combines and emphasizes the content procured by the experts to form the final summary $\widehat{\mathcal{Y}}$. Details regarding RoR, MoE, and FG in our approach for dialogue summarization are elaborated in the subsequent subsections.

2.1 Role-oriented Routing

Existing studies demonstrate that task-related information is helpful for improving model performance (Zhang et al., 2019a,b; Kano et al., 2020; Chen et al., 2020; Chen and Yang, 2020a; Song, 2022; Lin et al., 2022; Tian et al., 2024). Consider the characteristics of dialogue, e.g., contents are from different roles (speakers), there are interactions among multiple roles, key information

[2]This anecdote originates from an ancient Indian parable, in which a group of blind men, each positioned around an elephant, attempt to understand the nature of the animal by touching a specific part, thus resulting an incomplete and weird figure of the elephant.

from them are imbalanced to contribute to the summarization process, it is crucial to have a routing design for LLM activation in MoE that incorporates such characteristics for better matching and preparation of input utterances for later expert processing. Given different experts in our approach are expected to produce appropriate information from various aspects, the speaker role is then encoded and contributes to dynamically determining which experts are suitable to generating particular contents. We propose RoR that is built upon an LLM with a memory module to perform the routing, where the LLM is used to encode the entire dialogue and the memory module is designed to select appropriate experts for each utterance.

Specifically, we feed the entire dialogue $\mathcal{D}$ into the embedding layer of LLM and obtain the embedding matrix $\mathbf{E}_D$. Next, the LLM takes $\mathbf{E}_D$ and compute the dialogue representation $\mathbf{H}_D$ following the standard LLM process procedure. $\mathbf{H}_D$ contains a list of matrixes $\mathbf{H}_1 \cdots \mathbf{H}_l \cdots \mathbf{H}_L$, where $\mathbf{H}_l$ denotes the representation for the l-th utterance and the speaker role in the dialogue. This process is formulated as

$$\mathbf{H}_1 \cdots \mathbf{H}_l \cdots \mathbf{H}_L = f_{LLM}(\mathbf{E}_D) \tag{1}$$

where each column in $\mathbf{H}_l$ corresponds to a particular token or the role representation in the input. Thus, $\mathbf{H}_l = [\mathbf{h}_l^r, \mathbf{H}_l']$ where $\mathbf{h}_l^r$ denotes the vector representation of speaker role $\mathcal{R}_l$ and $\mathbf{H}_l'$ the matrix representation of the utterance. Then, for each utterance $\mathcal{U}_l$ and its speaker $\mathcal{R}_l$, we feed the role representation $\mathbf{h}_l^r$ into the memory module to select appropriate experts. Specifically, for each expert $f_{e,n}$ $(1 \leq n \leq N)$, we associate it with a memory vector $\mathbf{m}_n$, which is used to store the aspects of summary that the corresponding expert is designed to address. For the current input utterance $\mathcal{U}_l$ with the role $\mathcal{R}_l$, the matching scores $w_{l,n}$ of the n-th expert $(f_{e,n})$ to $\mathcal{U}_l$ is calculated by

$$w_{l,n} = \mathbf{h}_l^r \cdot \mathbf{W}_a \cdot \mathbf{m}_n \tag{2}$$

where $\mathbf{W}_a$ is a trainable parameter matrix. Note that, $\mathbf{h}_l^r$ is a contextualized representation that contains both roles and their related context information. Therefore, the matching score $w_{l,n}$ is determined by both the role associated with the context (i.e., the current utterance).

2.2 The Experts

The expert system in our approach employs N experts, which are Transformer decoders (e.g., the last $(K_{LLM} - K)$ layers of LLM) and denoted as $f_{e,1} \cdots f_{e,n} \cdots f_{e,N}$. For each utterance $\mathcal{U}_i$, we collect the contribution scores $w_{l,n}$ from RoR, rank them in descending order, and select the corresponding experts with top N' scores (denoted as $f_{e,l,1} \cdots f_{e,l,n'} \cdots f_{e,l,N'}$). For each selected experts $f_{e,l,n'}$, it processes $\mathbf{H}_l$ $(1 \leq l \leq L)$ and generate a representation matrix $\mathbf{H}_{l,n'}'$ that carry important content about the dialogue, so that covers one or more essential aspects of the key information in each utterance, which is formulated as

$$\mathbf{H}_{l,n'}' = f_{e,l,n'}(\mathbf{H}_l) \tag{3}$$

Particularly, the entire dialogue information is also considered in producing $\mathbf{H}_{l,n'}'$ since $\mathbf{H}_l$ is directly obtained from the LLM f_{LLM}, which has the dialogue as the input. Finally, we perform the same process for all experts and all utterances, which leads to $\mathbf{H}_{1,1}' \cdots \mathbf{H}_{1,N'}' \cdots \mathbf{H}_{L,1}' \cdots \mathbf{H}_{L,N'}'$.

2.3 Fusion Generator

Once the information is processed by different experts, we use FG (denoted as f_{FG}) to collect the representations $\mathbf{H}_{l,1}' \cdots \mathbf{H}_{l,N'}'$ $(1 \leq l \leq L)$ produced from them, in association with the entire dialogue $\mathcal{D}$, to predict the final dialogue summary $\widehat{\mathcal{Y}}$. Specifically, FG is also an LLM-based generator that takes prompts (i.e., vectors) to perform a standard LLM generation process. We feed $\mathbf{E}_D$ and all $\mathbf{H}_{l,1}' \cdots \mathbf{H}_{l,N'}'$ $(1 \leq l \leq L)$ from experts into FG and generate the final summary $\widehat{\mathcal{Y}}$ by

$$\widehat{\mathcal{Y}} = f_{FG}(\mathbf{p}, \mathbf{E}_D, \mathbf{H}_{1,1}' \cdots \mathbf{H}_{l,N'}', \cdots \mathbf{H}_{L,1}' \cdots \mathbf{H}_{L,N'}') \tag{4}$$

where $\mathbf{p}$ is a soft prompt to instruct FG to generate the summary that designed specifically for all $\mathbf{H}_{l,n'}$ on the condition of $\mathbf{D}$, who provides global information to guide FG generation. During training, we compare the generated summary $\widehat{\mathcal{Y}}$ with the gold standard summary $\mathcal{Y}^*$ to compute the cross-entropy loss, and follow the standard procedure to update model parameters accordingly. The FG ensures effective combination of contents from different experts and the regularization of producing the final summary for each dialogue.

3 Experimental Settings

3.1 Datasets

In our experiments, we use four benchmark datasets, namely, DialogSum (Chen et al., 2021), SAMSum (Gliwa et al., 2019), CSDS (Lin et al.,

Datasets		# Dial.	Avg. Len.	Avg. Turns
DialogSum	Train	12,460	131.0	9.5
	Valid	500	129.3	9.4
	Test	1,500	134.5	9.7
SAMSum	Train	14,732	93.8	11.2
	Valid	818	91.6	10.8
	Test	819	95.5	11.3
CSDS	Train	9,101	401.1	26.0
	Valid	800	396.3	25.9
	Test	800	387.1	25.1
MC	Train	35,987	311.9	9.6
	Test	8,996	313.3	9.5

Table 2: The statistics of the datasets in the train, valid, and test sets. "# Dial.", "Avg. Len.", and "Avg. Turns" are the number of dialogues, the average number of characters/tokens, and the number of turns in a dialogue, respectively.

2021), and MC (Song et al., 2020) to evaluate our approach and different baselines for dialogue summarization, where the first two are English datasets and the rest are in Chinese. Specifically, DialogSum is a large-scale dialogue summarization dataset from daily life topics, where the dialogues are manually annotated with overall summaries and topics. SAMSum contains dialogues between two or more persons under different scenarios such as meetings, phone calls, online posts and replies, etc., where each dialogue is associated with a summary. CSDS dataset contains Chinese customer service dialogues and their summaries, where every dialogue has three different types of summaries for customer, agent, and the entire dialogue, where each summary of the entire dialogue is the concatenation of its customer and agent summaries.[3] MC is a Chinese medical conversation dataset that contains dialogues between patients and physicians with two types of summaries for each of the role, i.e., patients or physicians, respectively. To facilitate dialogue summarization for MC, we concatenate patient and doctor summaries for each dialogue to form the overall summary and use it in our experiments. For all four datasets, we follow their standard train, valid, and test splits[4]. The statistics of the datasets are illustrated in Table 2, where the number of dialogues, the average number of characters or tokens in each dialogue, and the number of turns in every dialogue are reported.

[3]We perform experiments on the summaries of the entire dialogue for CSDS, not the role-based ones.

[4]There is no official validation set for MC.

3.2 Implementation Details

Deep modeling of text representation plays an essential role in text understanding (Song et al., 2017, 2018; Devlin et al., 2019; Lewis et al., 2020; Diao et al., 2020; Tian et al., 2023; Touvron et al., 2023a) and thus determines the quality of the generated summary. In the experiments, we use the LLaMA-2 (Touvron et al., 2023b) and Ziya (Gan et al., 2023) that achieve state-of-the-art performance on natural language processing tasks as the LLMs in our approach for English and Chinese processing, respectively, following their default configurations. There are 40 layers of multi-head attention in the LLMs. By default, the number of multi-head attention layers in the expert is set to 5, the number of experts N is set to 4, and the number of selected experts N' is set to 2. It is worth noting that compared with the standard LLM, our approach has more Transformer layers. For example, following the default settings, the standard LLM with 13B parameters has 40 layers of Transformer; our approach needs to compute over 35 + 5 * 2 = 45 layers of Transformer, and the FG model needs more computation on 40 layers of Transformer.

We tune hyper-parameters on the validation set and use the setting that achieves the best performance to train our final models.[5] In evaluation, we use both automatic and human evaluations. Following existing studies, the automatic evaluation includes **ROUGE** (Lin, 2004) (i.e., the F-scores

[5]For MC that does not have the official validation set, we randomly select 10% dialogues from its training set and use them to tune hyper-parameters, which are used on the final model for the entire training set.

	R-1	R-2	R-L	BL	BS	MS
LLAMA-2	$44.88_{\pm 0.11}$	$21.87_{\pm 0.10}$	$44.64_{\pm 0.10}$	$16.25_{\pm 0.09}$	$62.79_{\pm 0.10}$	$50.12_{\pm 0.11}$
+ MOE	$45.35_{\pm 0.11}$	$22.34_{\pm 0.10}$	$45.11_{\pm 0.10}$	$16.72_{\pm 0.09}$	$63.26_{\pm 0.10}$	$50.59_{\pm 0.11}$
+ ROR	$47.31_{\pm 0.10}$	$23.44_{\pm 0.08}$	$46.99_{\pm 0.10}$	$17.43_{\pm 0.09}$	$64.58_{\pm 0.09}$	$52.30_{\pm 0.07}$
+ FG	$47.56_{\pm 0.13}$	$23.69_{\pm 0.10}$	$47.03_{\pm 0.10}$	$17.73_{\pm 0.11}$	$65.90_{\pm 0.11}$	$52.49_{\pm 0.12}$
+ ROR + FG	$\mathbf{49.82}_{\pm 0.10}$	$\mathbf{24.80}_{\pm 0.11}$	$\mathbf{47.37}_{\pm 0.09}$	$\mathbf{18.41}_{\pm 0.13}$	$\mathbf{68.48}_{\pm 0.08}$	$\mathbf{53.86}_{\pm 0.10}$

(A) DIALOGSUM

	R-1	R-2	R-L	BL	BS	MS
LLAMA-2	$52.48_{\pm 0.14}$	$28.90_{\pm 0.11}$	$50.10_{\pm 0.12}$	$23.55_{\pm 0.10}$	$72.95_{\pm 0.09}$	$58.47_{\pm 0.12}$
+ MOE	$52.95_{\pm 0.14}$	$29.37_{\pm 0.11}$	$50.57_{\pm 0.12}$	$24.02_{\pm 0.10}$	$73.42_{\pm 0.09}$	$58.94_{\pm 0.12}$
+ ROR	$53.98_{\pm 0.09}$	$29.80_{\pm 0.11}$	$51.61_{\pm 0.12}$	$25.83_{\pm 0.11}$	$74.79_{\pm 0.09}$	$61.32_{\pm 0.10}$
+ FG	$54.58_{\pm 0.13}$	$30.15_{\pm 0.10}$	$51.42_{\pm 0.10}$	$25.80_{\pm 0.11}$	$74.96_{\pm 0.11}$	$61.46_{\pm 0.12}$
+ ROR + FG	$\mathbf{55.93}_{\pm 0.11}$	$\mathbf{30.86}_{\pm 0.11}$	$\mathbf{52.02}_{\pm 0.12}$	$\mathbf{26.03}_{\pm 0.13}$	$\mathbf{75.66}_{\pm 0.10}$	$\mathbf{62.76}_{\pm 0.09}$

(B) SAMSUM

	R-1	R-2	R-L	BL	BS	MS
ZIYA	$58.42_{\pm 0.13}$	$46.37_{\pm 0.08}$	$56.46_{\pm 0.09}$	$29.52_{\pm 0.07}$	$80.81_{\pm 0.05}$	$59.40_{\pm 0.10}$
+ MOE	$58.89_{\pm 0.13}$	$46.84_{\pm 0.08}$	$56.93_{\pm 0.09}$	$29.99_{\pm 0.07}$	$81.28_{\pm 0.05}$	$59.96_{\pm 0.10}$
+ ROR	$59.11_{\pm 0.11}$	$47.35_{\pm 0.07}$	$56.88_{\pm 0.11}$	$30.54_{\pm 0.05}$	$82.46_{\pm 0.10}$	$60.78_{\pm 0.07}$
+ FG	$59.06_{\pm 0.12}$	$46.72_{\pm 0.08}$	$57.77_{\pm 0.10}$	$31.61_{\pm 0.06}$	$82.29_{\pm 0.06}$	$61.56_{\pm 0.06}$
+ ROR + FG	$\mathbf{61.86}_{\pm 0.12}$	$\mathbf{47.07}_{\pm 0.09}$	$\mathbf{60.04}_{\pm 0.11}$	$\mathbf{32.10}_{\pm 0.08}$	$\mathbf{83.26}_{\pm 0.14}$	$\mathbf{61.94}_{\pm 0.06}$

(C) CSDS

	R-1	R-2	R-L	BL	BS	MS
ZIYA	$91.35_{\pm 0.13}$	$87.45_{\pm 0.11}$	$87.79_{\pm 0.12}$	$76.31_{\pm 0.08}$	$91.28_{\pm 0.11}$	$83.80_{\pm 0.10}$
+ MOE	$91.82_{\pm 0.13}$	$87.92_{\pm 0.11}$	$87.86_{\pm 0.12}$	$76.78_{\pm 0.08}$	$91.75_{\pm 0.11}$	$83.27_{\pm 0.10}$
+ ROR	$92.83_{\pm 0.11}$	$88.71_{\pm 0.07}$	$90.72_{\pm 0.08}$	$79.35_{\pm 0.07}$	$93.67_{\pm 0.06}$	$87.29_{\pm 0.12}$
+ FG	$93.05_{\pm 0.11}$	$89.03_{\pm 0.07}$	$90.91_{\pm 0.09}$	$79.62_{\pm 0.04}$	$94.16_{\pm 0.05}$	$87.43_{\pm 0.07}$
+ ROR + FG	$\mathbf{93.45}_{\pm 0.11}$	$\mathbf{89.40}_{\pm 0.07}$	$\mathbf{91.71}_{\pm 0.10}$	$\mathbf{80.47}_{\pm 0.06}$	$\mathbf{95.67}_{\pm 0.13}$	$\mathbf{88.72}_{\pm 0.05}$

(D) MC

Table 3: Experiment results of different models on the test set of DialogSum, SAMSum, CSDS, and MC, respectively, where "+ MOE" denote the model with standard MoE, and "+ ROR" and "+ FG" means that RoR and FG are added on top of the "MOE" baseline. We also report the average and standard deviation over five runs with different random seeds. Metrics "R-1", "R-2", and "R-L" correspond to the F-scores of ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Similarly, "BL", "BS" and "MS" denote BLEU, BERT-Score, and Mover-Score, respectively.

	R-1	R-2	R-L
OUYANG ET AL. (2023)	47.94	21.67	45.10
CHEN ET AL. (2023A)	48.29	23.65	46.23
GAO ET AL. (2023)	48.02	21.68	45.88
OURS	**49.82**	**24.80**	**47.37**

Table 4: Experiment results from previous studies and ours (LLaMA-2 + RoR + FG) for the dialogue summaries on the test set of DialogSum.

	R-1	R-2	R-L
OUYANG ET AL. (2023)	53.56	28.66	50.04
CHEN ET AL. (2023A)	53.76	28.04	50.56
GAO ET AL. (2023)	54.97	30.01	**56.27**
OURS	**55.93**	**30.86**	52.02

Table 5: Experiment results from previous studies and ours (LLaMA-2 + RoR + FG) for the dialogue summaries on the test set of SAMSum.

of **ROUGE-1**, **ROUGE-2** and **ROUGE-L**) and **BLEU** (Papineni et al., 2002) that measures the n-gram overlap between the model output and reference summaries, as well as **BERT-Score** (Zhang et al., 2019c) and **Mover-Score** (Zhao et al., 2019) that computes the text similarity based on BERT embeddings (Devlin et al., 2019). Human evaluation metrics include **informativeness** that measures the coverage of the key points, **non-redundancy** that evaluates whether the generated summary contains redundant or repeated information, and **fluency** that examines whether the generated summary is fluent and grammatically correct. All human evaluation metrics have three levels: 0, 1, and 2, where 0 denotes the worst and 2 the best.

4 Results and Analysis

4.1 Overall Results

We report experiment results (i.e., the average and the standard deviation of five runs) of baselines and

	R-2	R-L	BL	BS	MS
*See et al. (2017)	39.19	47.94	32.31	78.40	28.58
*Chen and Bansal (2018)	41.39	47.07	33.04	79.57	29.78
*Liu and Lapata (2019)	37.03	45.30	24.59	78.45	27.00
*Zou et al. (2021)	33.19	42.43	20.24	76.84	24.29
Liang et al. (2022)	44.25	58.64	35.09	80.92	60.29
Liang et al. (2023)	45.83	59.25	**36.43**	81.83	61.03
Ours	**47.07**	**60.04**	34.30	**83.15**	**61.96**

Table 6: Experiment results from previous studies and ours (Ziya + RoR + FG) for the dialogue summaries on the test set of CSDS. The results marked by "*" come from Lin et al. (2021).

	Patient						Doctor					
	R-1	R-2	R-L	BL	BS	MS	R-1	R-2	R-L	BL	BS	MS
Song et al. (2020)	91.01	87.38	91.01	-	-	-	80.87	72.07	80.84	-	-	-
Lin et al. (2022)	95.19	94.63	95.14	87.40	97.90	90.72	82.11	77.49	80.92	65.40	91.91	68.95
Liang et al. (2022)	96.78	95.86	96.12	91.22	98.13	95.10	88.21	84.58	86.56	70.08	92.84	81.95
Liang et al. (2023)	**96.84**	**96.14**	**96.23**	91.32	98.25	95.35	88.47	84.62	86.77	70.18	92.96	82.10
Ours	96.60	95.82	95.40	**92.84**	**98.51**	**95.94**	**89.71**	**85.93**	**88.34**	**74.57**	**93.43**	**84.62**

Table 7: Experiment results from previous studies and ours (Ziya + RoR + FG) for dialogue summarization on the test set of MC. We follow the convention of existing studies to generate the summaries for different speakers (i.e., patients and physicians) separately and compare them with the gold standard.

our approach on the four benchmark datasets in Table 3. "LLaMA-2" and "Ziya" are baselines that directly applying LLMs to the task without using MoE, RoR, or FG, where "MoE" means the standard MoE approach on LLMs; "+RoR" and "+FG" stand for RoR and FG are added on top of "MoE", respectively; "+RoR+FG" is our full model. There are observations as follows. First, compared with the vanilla LLaMA-2 and Ziya, models with MoE achieve better results, which indicates that utilizing different experts allows models to learn important information for dialogue summarization from different aspects and thus leads to better summaries. Second, models with RoR or FG outperform the ones with only "MoE" setting, which demonstrates the effectiveness of using RoR or FG to select appropriate experts to generate essential content or combine different information so as to improve dialogue summarization, respectively. Third, our full model with both RoR and FG achieves the best performance on all datasets with different LLMs, which indicates that RoR and FG collaborate well and are able to be complementary to each other, thus further improve the quality of dialogue summarization.

We further compare the performance of our approach with existing studies on the four benchmark datasets and report the results in Table 4, Table 5, Table 6, and Table 7. It is observed that our approach outperforms existing studies on all datasets,

	Info.	NR	Flu.	Info.	NR	Flu.
	DialogSum			SAMSum		
LLaMA-2	1.40	1.35	1.33	1.27	1.18	1.19
+ MoE	1.42	1.40	1.52	1.38	1.27	1.33
+ RoR	1.46	1.46	1.50	1.42	1.41	1.32
+ FG	1.50	1.47	1.56	1.40	1.31	1.33
+ Zero-shot	1.48	1.41	1.42	1.42	1.30	1.36
+ One-shot	1.51	1.46	1.53	1.49	1.43	1.50
+ Zero-shot CoT	1.58	1.53	1.60	1.50	1.46	1.54
+ RoR + FG	**1.64**	**1.68**	**1.73**	**1.57**	**1.50**	**1.61**
	CSDS			MC		
Ziya	1.48	1.33	1.52	1.40	1.33	1.43
+ MoE	1.50	1.35	1.56	1.43	1.31	1.42
+ RoR	1.56	1.41	1.50	1.42	1.30	1.45
+ FG	1.58	1.42	1.53	1.47	1.34	1.50
+ Zero-shot	1.39	1.42	1.46	1.44	1.31	1.40
+ One-shot	1.42	1.47	1.52	1.47	1.43	1.51
+ Zero-shot CoT	1.51	1.50	1.57	1.54	1.42	1.50
+ RoR + FG	**1.60**	**1.55**	**1.69**	**1.64**	**1.58**	**1.70**

Table 8: Human evaluation scores (the higher the better) of different models on the test set of all datasets. "Info", "NR", and "Flu" refer to "informativeness", "non-redundancy", and "fluency", respectively.

where these studies mainly use a single model to process the text and generate summaries accordingly. This observation indicates that "*many hands make light work*", where using MoE with RoR and FG allows the model to effectively use different experts to capture various key information and smartly optimize the output of the experts to produce summaries, therefore is a more reasonable solution than that only uses a single model.

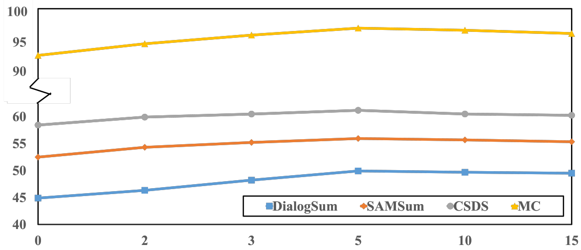

Figure 2: The performance (R-1 scores) of our approach with respect to the number of LLM layers used in the experts on four benchmark datasets.

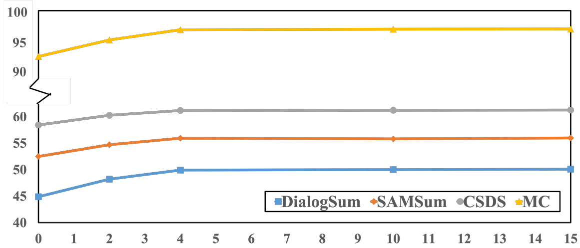

Figure 3: The R-1 scores of our approach with respect to the number of experts on four benchmark datasets.

	R-1	R-2	R-L	BL
ALL	47.53	23.70	47.13	17.60
RANDOM	47.30	23.63	46.98	17.52
NO ROLE	47.70	23.81	46.82	17.97
(A) DIALOGSUM				
ALL	54.44	30.07	51.02	25.65
RANDOM	54.39	29.96	50.95	25.53
NO ROLE	54.63	30.20	51.38	25.85
(B) SAMSUM				
ALL	59.03	46.58	57.16	30.87
RANDOM	58.94	46.40	56.96	30.69
NO ROLE	59.19	46.75	57.54	31.08
(C) CSDS				
ALL	92.85	88.64	90.68	79.47
RANDOM	92.73	88.58	90.59	79.32
NO ROLE	93.10	88.97	90.84	79.59
(D) MC				

Table 9: The performance of different baseline models on the test sets of the benchmark datasets. "All" indicates all experts are used; "Random" means the experts are randomly selected in routing; and "No Role" indicates the role information is not included in the router.

4.2 Human Evaluation

To further evaluate whether MoE is truly useful in dialogue summarization, we also perform human evaluation on different baselines and our approaches following the evaluation criteria specified in Section 3.2. The results are reported in Table 8. In addition to the baselines in Table 3, we also evaluate the performance of directly using LLMs under zero-shot, one-shot, and zero-shot chain-of-thought (CoT) (Wei et al., 2022) settings. It is observed that, the results show a similar trend to the results in Table 3 with our approach outperforming all baselines, which confirms the effectiveness of our approach for dialogue summarization.

4.3 Effect of Different Expert Settings

To investigate the effect of the experts, we run two groups of experiments, where the first one tries to set each expert by adjusting the numbers of LLM layers used in it, and the second one explores the relations of total expert number with model performance. For the first one, we try 2, 3, 5, 10, 15 layers in Transformer for the experts and present the curve of performance against such settings on different datasets in Figure 2. The results show that, when the number of layers is small, increasing the number of layers leads to improvements, which is intuitive since more layers enable the experts to capture more essential information from each particular aspect and thus connects to better performance. However, when the number reaches a threshold, further increasing the number brings fewer improvements. The reason is that, with more layers used for experts, fewer layers are left for RoR, which makes the router hard to understand and process each utterance and select appropriate experts. For the second one, we try 2, 4, 10, 15 experts in experiments and present the results in Figure 3. We find that the performance increases with more experts when their number is smaller than 10. When the number goes beyond 10, adding more experts does not lead to improvements. The explanation is that, when the number is small, the number of experts is not enough for them to learn essential information from different aspects in each dialogue. Therefore, increasing the number of experts allows MoE to gradually learn sufficiently and thus results in better performance. On the contrary, when the number of experts reaches a certain amount, adding new experts does not further help so that the performance is converged.

4.4 Effect of Role-oriented Routing

To illustrate the effect of RoR, we run three baselines, namely, 'ALL", "RANDOM", "NO ROLE" as comparison to our approach, where the first one selects all experts, the second randomly se-

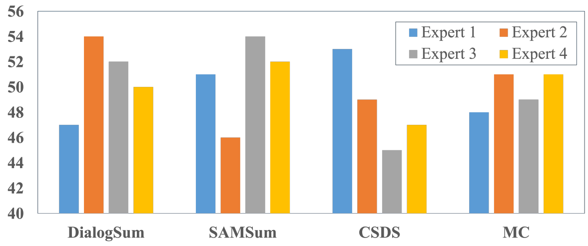

Figure 4: The distribution of top experts in our default setting (select 2 experts out of 4) for different datasets.

lect experts, and the third one does not use role information in router inputs. The results of the aforementioned models are reported in Table 9 with the following observations. First, comparing our approach (in Table 3) with the "ALL" and the "RANDOM" baselines, our approach achieves better performance, which complies with our intuition because that "ALL" and "RANDOM" actually do not select experts to process the input features and thus face problems of utilizing inappropriate experts to process the essential content of the dialogue, which introduces noise that leads to inferior results. Second, when the role information is not included in the router, the model's performance is also worse than our full model, which indicates that the role information is important to understand the key content of dialogue as we hypothesized in our motivation, so that it helps the router to better associate some contents to particular speakers and perform appropriate expert selection.

In addition, we explore the distribution of top experts (i.e., the number of times an expert is selected to process an utterance based on the score from Eq. (2), divided by the total number of times all experts are selected) under the default setting, i.e., selecting 2 experts from total 4 experts in processing each utterance. The results are presented in Figure 4. We observe that experts contribute differently on all tasks and have their own preference of being selected, which indicates our router is able to select appropriate experts in different scenarios.

4.5 Effect of Fusion Generation

In our main experiments, we use dialogue as the condition for generating the final summary in the FG. To explore the effect of using such condition, we run experiments without using it, where the results are ported in Table 10. It clearly shows that, compared with the models without using the entire dialogue, our approach is able to generate better summaries, which emphasizes the contribution of the entire dialogue, for the reason that it provides global or environmental information to guide FG identifying useful content produced by the experts.

DATASETS	R-1	R-2	R-L	BL
DIALOGSUM	49.04	24.25	46.90	18.13
SAMSUM	55.40	30.47	51.89	25.95
CSDS	60.87	46.93	59.09	31.87
MC	93.32	89.17	91.42	80.33

Table 10: The performance of our approach without using the entire dialogue as condition in FG.

4.6 Case Study

To further demonstrate the effectiveness of our approach, we use a case study to compare the original dialogue with the gold standard summary, and the final summaries generated by our approach in Figure 5. For better illustration, we also decode the hidden matrixes $\mathbf{H}'_{1,n} \cdots \mathbf{H}'_{L,n}$ produced by the experts into intermediate summaries and present them in Figure 5. The following are some observations. First, the intermediate summaries generated by different experts illustrate that these experts do learn to extract key information from different aspects. For example, expert 1 learns to focus on the information of speaker S2; expert 2 learns to focus on the interactions between speakers. This observation confirms the effectiveness of the router in selecting appropriate experts to process different utterances. Second, the final summary generated by our model includes the essential content covered by the gold standard and it also shows a better combination of intermediate summaries with the duplicate and unimportant content being filtered out, which further confirms the effectiveness of FG to optimize and refine the results produced by experts.

5 Related Work

Dialogue modeling has attracted attention from many existing studies (Li et al., 2018; Wang et al., 2018; Yu et al., 2019), especially dialogue summarization. A large body of dialogue summarization studies is devoted to leveraging advanced text encoders, such as BART (Lewis et al., 2020), to achieve a more nuanced modeling of dialogue content and thus optimize role-specific summarization (Chen and Yang, 2020a; Lin et al., 2022; Liang et al., 2022). To enhance the quality and relevance of generated summaries, many studies adopt particular elements in dialogues or extra features, such as important utterances (Song et al., 2020; Krishna et al., 2021), dialogue topics (Zou et al., 2021; Liu

S1: Don't tell me you're bored already?
S2: Well, yeah, I am.
S1: We just got here. You will have to find a way to entertain yourself then.
S2: What can I do? There are just a bunch of boring people giving boring speeches.
S1: Too bad. I have to sit through it, so you do, too. Even if it's a little boring, it's important for me to be here. I have to listen carefully, take notes and then interview some of the speakers afterward. I work for the Daily News after all. Didn't you bring anything to read?
S2: brought my iPod and some headphones. Would it be rude to listen to some music?
S1: Yes, it would. I might have a few old newspapers in my bag. You could read them. They also have sections with games in them, like crossword puzzles. That will keep you busy for a while.
S2: hate crosswords. Do the papers have comic sections in them?
S1: You'll probably find some. Now be quiet! Another man is getting up to speak. I really have to pay attention to this one. It's the head of the charity organization.

Dialogue

S2 tells S1 that S2 is bored and explains the reasons, so S1 suggests S2 read some old newspapers and be quiet.

Reference Summary

Expert 1:

S2 says he is bored
S2 feel the people and speech are boring
S2 has iPad and headphones
S2 asks if there are any comic sections in them

Expert 2:

S1 finds S2 might be bored
S1 want S2 to find a way to entertain himself
S1 asks if S2 brings anything to read
S1 suggest S2 to read some newspapers
S1 asks S2 to be quite since the next speaker is the head of the charity organization.

Intermediate Summary

S2 says he is bored because the people and speech are boring. S1 suggests S2 read some newspapers and be quiet, since the next speaker is the head of the charity organization

Final Summary

Figure 5: An example dialogue with the reference summary, intermediate summaries produced by experts, and the final summary from the full model. The reference summary is highlighted in yellow and blue, where the content in the intermediate and final summary that matches the reference is highlighted in the same color.

et al., 2021; Lin et al., 2023; Liang et al., 2023), and semantic relations among sentences (Kano et al., 2020; Zhang et al., 2022; Liu and Xu, 2023) to extend the capability of their summarization models. With the growing recognition of the importance of the structures and interactions in dialogues, summarization is thus performed by incorporating them as core components in several state-of-the-art studies (Chen and Yang, 2020b; Zhu et al., 2020; Feng et al., 2020; Joshi et al., 2020; Chowdhury et al., 2020; Lei et al., 2021; Qi et al., 2021; Chen and Yang, 2021; Zhang et al., 2021; Zhong et al., 2022; Jia et al., 2022; Zhu et al., 2023; Zou et al., 2023). Although summarization performance is promoted accordingly with such methodology improvements, these studies mainly use a single-model design to capture various types of essential information in the dialogue and generate summaries in an end-to-end manner. As a choice of using multiple models, MoE offers a solution to separately model different aspects of the input and process them accordingly, achieve remarkable success in handling complex tasks, such as language modeling, natural language inference, question answering, etc., (Fedus et al., 2022; Zoph et al., 2022; Chen et al., 2023c; Shen et al., 2023a,b; Li et al., 2023a), where they cover various applications, such as textual-only (Jacobs et al., 1991; Shazeer et al., 2017; Lepikhin et al., 2021; Du et al., 2022; Fedus et al., 2022) as well as cross-modal scenarios (Mustafa et al., 2022), and so far few are used for dialogue summarization. Therefore, compared with previous studies, this paper proposes a way of applying multiple models for dialogue summarization and introduce a novel design that improves MoE, where RoR is proposed to address the challenge of effectively selecting appropriate experts in the particular dialogue circumstance, and the FG highlights the salient content generated by different experts.

6 Conclusion

In this paper, we propose an MoE approach that alter LLMs for dialogue summarization, where a specific router and fusion generator are designed to facilitate the mixture process of experts. Specifically, the routing effectively organizes the matching of different experts to utterances, and the fusion generator further optimizes the information processed by experts and then utilizes appropriate contents from them to provide final summaries with essential information from different aspects. Experiment results and analysis on four English and Chinese benchmark datasets for dialogue summarization illustrate the effectiveness of our approach, which outperforms strong baselines and existing studies on all datasets, and show that MoE successfully distinguishes different contents in each dialogue with processing using appropriate experts.

Acknowledgements

This paper is supported by the National Key Research and Development Program of China under the grant (2021YFF0901600).

References

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279.

Jiaao Chen, Mohan Dodda, and Diyi Yang. 2023a. Human-in-the-loop Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9176–9190, Toronto, Canada.

Jiaao Chen and Diyi Yang. 2020a. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*.

Jiaao Chen and Diyi Yang. 2020b. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online.

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, Online.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. 2023b. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. In *The Eleventh International Conference on Learning Representations*.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023c. Lifelong Language Pretraining with Distribution-Specialized Experts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5383–5395.

Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. *arXiv preprint arXiv:1805.11080*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online.

Tanya Chowdhury, Sachin Kumar, and Tanmoy Chakraborty. 2020. Neural abstractive summarization with structural attention. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3716–3722.

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

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization. *Dialogue*, 1:U2.

Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Ping Yang, Qi Yang, Jiaxing Zhang, et al. 2023. Ziya2: Data-centric Learning is All LLMs Need. *arXiv preprint arXiv:2311.03301*.

Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao, and Rui Yan. 2023. Dialogue Summarization with Static-Dynamic Structure Fusion Graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13858–13873, Toronto, Canada.

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.

Iryna Gurevych and Michael Strube. 2004. Semantic Similarity Applied to Spoken Dialogue Summarization. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving Long Dialogue Summarization with Semantic Graph Representation. In *Findings of* the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Zhu. 2022. Post-Training Dialogue Summarization using Pseudo-Paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1660–1669, Seattle, United States.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online.

Ryuji Kano, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2020. Identifying Implicit Quotes for Unsupervised Extractive Summarization of Conversations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 291–302.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques. 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 4958–4972, Online.

Yuejie Lei, Fujia Zheng, Yuanmeng Yan, Keqing He, and Weiran Xu. 2021. A Finer-grain Universal Dialogue Semantic Structures based Model For Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1354–1364, Punta Cana, Dominican Republic.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, pages 1–19.

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

Jiamin Li, Qiang Su, Yitao Yang, Yimin Jiang, Cong Wang, and Hong Xu. 2023a. Adaptive Gating in Mixture-of-Experts based Language Models. *arXiv preprint arXiv:2310.07188*.

Jing Li, Yan Song, Haisong Zhang, and Shuming Shi. 2018. A manually annotated chinese corpus for non-task-oriented dialogue systems. *arXiv preprint arXiv:1805.05542*.

Yu Li, Baolin Peng, Pengcheng He, Michel Galley, Zhou Yu, and Jianfeng Gao. 2023b. DIONYSUS: A Pretrained Model for Low-Resource Dialogue Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1368–1386, Toronto, Canada.

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 (Volume 2: Short Papers)*, pages 43–50, Online.

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, Dubrovnik, Croatia.

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, Feifei Zhai, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2023. Topic-Oriented Dialogue Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1797–1810.

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 58th Annual Meeting of the Association for Computational Linguistics*.

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, Punta Cana, Dominican Republic.

Xingxian Liu and Yajing Xu. 2023. Learning to Rank Utterances for Query-Focused Meeting Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8496–8505, Toronto, Canada.

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.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts. *arXiv preprint arXiv:2206.02770*.

Siru Ouyang, Jiaao Chen, Jiawei Han, and Diyi Yang. 2023. Compositional Data Augmentation for Abstractive Conversation Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1471–1488, Toronto, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

MengNan Qi, Hao Liu, YuZhuo Fu, and Ting Liu. 2021. Improving Abstractive Dialogue Summarization with Hierarchical Pretraining and Topic Segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1121–1130, Punta Cana, Dominican Republic.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. pages 1–19.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. 2023a. Mixture-of-experts Meets Instruction Tuning: A Winning Combination for Large Language Models. *arXiv preprint arXiv:2305.14705*.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023b. Scaling Vision-Language Models with Sparse Mixture of Experts. *arXiv preprint arXiv:2303.07226*.

Yan Song. 2022. Composing Ci with Reinforced Non-autoregressive Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7229.

Yan Song, Chia-Jung Lee, and Fei Xia. 2017. Learning Word Representations with Regularization from Prior Knowledge. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 143–152.

Yan Song, Shuming Shi, and Jing Li. 2018. Joint Learning Embeddings for Chinese Words and Their Components via Ladder Structured Networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4375–4381.

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.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *arXiv preprint arXiv:2311.06025*.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Learning Multimodal Contrast with Cross-modal Memory and Reinforced Contrast Recognition. In *The 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. LLaMA 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Nan Wang, Yan Song, and Fei Xia. 2018. Coding structures and actions with the costa scheme in medical conversations. In *Proceedings of the BioNLP 2018 workshop*, pages 76–86.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. *arXiv preprint arXiv:1909.00421*.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019a. Incorporating Context and External Knowledge for Pronoun Coreference Resolution. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 872–881.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019b. Knowledge-aware Pronoun Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876, Florence, Italy.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022. Hierarchical Cross-modality Semantic Correlation Learning Model for Multimodal Summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11676–11684.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019c. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. *arXiv preprint arXiv:1909.02622*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 11765–11773.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6825–6845.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing Stable and Transferable Sparse Expert Models. *arXiv preprint arXiv:2202.08906*.

Yicheng Zou, Kaitao Song, Xu Tan, Zhongkai Fu, Qi Zhang, Dongsheng Li, and Tao Gui. 2023. Towards Understanding Omission in Dialogue Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14268–14286.

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.