Enhancing Dialogue Summarization with Topic-Aware Global- and Local-Level Centrality

Xinnian Liang[1]*, Shuangzhi Wu[2], Chenhao Cui[2], Jiaqi Bai[1], Chao Bian[2], Zhoujun Li[1]†
[1]State Key Lab of Software Development Environment, Beihang University, Beijing, China
[2]Lark Platform Engineering-AI, Beijing, China
{xnliang,cuich,bjq,lzj}@buaa.edu.cn,{wufurui,zhangchaoyue.0}@bytedance.com

Abstract

Dialogue summarization aims to condense a given dialogue into a simple and focused summary text. Typically, both the roles' viewpoints and conversational topics change in the dialogue stream. Thus how to effectively handle the shifting topics and select the most salient utterance becomes one of the major challenges of this task. In this paper, we propose a novel topic-aware Global-Local Centrality (GLC) model to help select the salient context from all sub-topics. The centralities are constructed at both the global and local levels. The global one aims to identify vital sub-topics in the dialogue and the local one aims to select the most important context in each sub-topic. Specifically, the GLC collects sub-topic based on the utterance representations. And each utterance is aligned with one sub-topic. Based on the sub-topics, the GLC calculates global- and local-level centralities. Finally, we combine the two to guide the model to capture both salient context and sub-topics when generating summaries. Experimental results show that our model outperforms strong baselines on three public dialogue summarization datasets: CSDS, MC, and SAMSUM. Further analysis demonstrates that our GLC can exactly identify vital contents from sub-topics. [1]

1 Introduction

Online conversations have become essential to communication in our daily work and life. Due to the information explosion, dialogue summarization has become a vivid field of research in recent years, which is meaningful for many applications, e.g. online customer service (Liu et al., 2019; Zhu et al., 2020) and meeting summary (Feng et al., 2021).

Dialogue summarization aims to condense crucial information in a long dialogue into a short text like traditional summarization tasks. Differently, the main challenges of dialogue summarization are the viewpoints of multiple speaker roles (Lin et al., 2021, 2022; Qi et al., 2021; Zhang et al., 2022) and shifting topics (Chen and Yang, 2020; Zou et al., 2021; Liu et al., 2021) during the conversation process. As shown in Figure 1, summaries not only depend on the overall context but also needs the identification and selection of salient context in crucial sub-topics. We can see that the blue text in summaries is about sub-topic #1 "The reason why the product is shipped yet" and the orange text is about sub-topic #2 "The user decided to refund", which are aligned to the two salient sub-topics from dialogue utterances in the first and second block. The sub-topic #3 is useless for summaries. This example shows the necessity to model the salient context and sub-topics in the dialogue.

In this paper, we propose a novel topic-aware Global-Local Centrality (GLC) model to select salient contexts from all sub-topics. The centrality is an effective technique to measure the importance of sentences in a given document from unsupervised extractive summarization (Zheng and Lapata, 2019; Liang et al., 2021, 2022). The GLC contains global- and local-level centrality, which are used to capture the salience of sub-topics and content in each sub-topic respectively. Based on these centralities, we can guide the model to focus on the salient context and sub-topics when generating summaries. Specifically, we employ utterance-level representations to cluster utterances and obtain sub-topic centers and assign each utterance to one sub-topic. Then, we compute the global centrality over sub-topic centers to measure the importance of each sub-topic and the local centrality over utterances of each sub-topic to measure the importance of sub-topic content. Finally, we combine the two to re-weight the dialogue context representations for the decoder to generate summaries.

To evaluate the effectiveness of our proposed GLC, we apply the GLC to three different types of

*Contribution during internship at ByteDance Inc.
†Corresponding Authors.
[1]https://github.com/xnliang98/bart-glc

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 27–38
May 2-6, 2023

	User: 我买的电话手表怎么还没发货？(Why hasn't the smart watch I bought been shipped yet?) **Agent:** 您好，还请您稍等，正在为您查询~，请问是这个商品吗？[商品快照]（Hello, please wait a moment. I am inquiring for you, is this product? [PRODUCT SNAPSHOT]) **User:** 是的。[数字]号下的单，一点反应都没有。(Yes. I placed an order at [Date], but no response at all.) **Agent:** [商品快照] 这个商品缺货的亲。([PRODUCT SNAPSHOT] This product is out of stock.) **User:** 那为什么我可以下单呢？也不告知我一下。(So why can I place an order? Don't tell me either) **Agent:** 这个是赠品哦亲。您电话对吗。妹子联系采销核实下具体时间。然后[数字]个小时内告知您。您看可以吗。(This is a giveaway. Is your phone number correct? I will contacts the purchase and sales to verify the specific time. Then [NUMBER] hours to let you know. Do you think this is ok?)	**Topic #1: The reason why the product is shipped yet.**
	User: 那直接申请退款吧，不买了。(Then I will apply for a refund directly, I will not buy it.) **Agent:** 哦哦，不要了吗？这边查看是在退款了，取消订单就无法恢复了呢。(Ok, don't you? The check here is a refund, and the cancellation of the order cannot be restored.) **User:** 嗯，退了吧。(Well, refund it.) **Agent:** 储蓄卡支付的吗？储蓄卡支付，原路返回的，周期是[数字]-[数字]个工作日内的。(Are you paying by debit card? Debit card payment, return the same way, the cycle is within [number]-[number] working days.)	**Topic #2: The user decided to refund.**
	User: 是的。(Yes.) **Agent:** 辛苦注意查收。请问还有其他还可以帮到您的吗? (Please pay attention to check the refund. Is there anything else I can help you with?) **User:** 好的没有了。(I have no more question.) **Agent:** 感谢您对我们的支持，祝您生活愉快，再见! (Thanks for your support. I wish you have a happy life, bye!)	**Topic #3: Useless information**
User Summary	用户询问购买的电话手表为什么还没发货，得知缺货后直接要求退款。(The user asks why the purchased smart watch has not been shipped, and directly asks for a refund after learning that it is out of stock.)	
Agent Summary	客服回复电话手表是赠品缺货，并准备向采销核实具体时间后告知用户。用户申请退款，客服提醒用户取消订单后无法恢复，随后告知储蓄卡支付会在一定时间内自动返还。(The customer service replies to the smart watch is a giveaway and is out of stock, and is ready to inform the user after verifying the specific time. When the user applies for a refund, the customer service reminds the user that the order cannot be restored after canceling the order, and then informs that the debit card payment will be automatically returned within a certain period of time.)	
Final Summary	用户询问购买的电话手表为什么还没发货。客服查询后回答赠品缺货，并准备向采销核实具体时间后告知用户。用户要求直接申请退款。客服提醒用户取消订单后无法恢复，随后告知储蓄卡支付会在一定时间内返还。(The user asks why the purchased phone watch has not been shipped. After customer service inquiries, they will answer that the gift is out of stock, and prepare to inform the user after verifying the specific time. The user requests to apply for a refund directly. The customer service reminded the user that the cancellation of the order cannot be restored, and then informed that the debit card payment will be returned within a certain period of time.)	

Figure 1: An example from the CSDS dataset. The dialogue contains 3 different sub-topics. The blue text represents sub-topic #1 and the red text represents sub-topic #2. The sub-topic #3 is useless information.

seq2seq structure: PGN, BERTAbs, and BART, and verify them on three public dialogue summarization datasets: CSDS, MC, and SAMSUM. CSDS and MC are two Chinese role-oriented summarization datasets that not only need to generate the overall summary of the dialogue but also need to generate role-oriented summaries for specific speakers in the dialogue as shown in Figure 1. SAMSUM is a widely used English dialogue summarization dataset. To generate role-oriented summaries, in this paper, we directly employ role prompts to guide the model to generate proper summaries. And the representations of role prompts can add role information to the centrality computation. Experimental results show that our GLC can improve the performance of all these seq2seq structures on three datasets. And the GLC-based BART model obtains new state-of-the-art results on the CSDS and MC.

Our contributions can be summarized as 1) We propose a novel topic-aware Global-Local Centrality (GLC) model to guide the model to identify the salient contexts and sub-topics in the dialogue. 2) Our GLC can bring improvement to different seq2seq models by easily plugging in and does not add any extra parameters to the seq2seq models. 3) The GLC-based BART model achieves new state-of-the-art results on CSDS and MC. Besides, extension studies prove our GLC can effectively capture vital sub-topics.

2 Methodology

Figure 2 shows the main structure of our proposed topic-aware global-local centrality (GLC) model. The seq2seq framework with GLC is on the left of Figure 2, which consists of the bi-directional encoder, global-local centrality model, and auto-regression decoder. The detail of GLC is on the right of Figure 2, which consists of global centrality and local centrality. In this section, we introduce them step by step.

2.1 Task Formulation

Firstly, we formulate the dialogue summarization task and role-oriented summarization task. Given a dialogue $\mathcal{D}$ with N utterances $\{u_1, \ldots, u_N\}$ with M roles $\{r_1, \ldots, r_M\}$. Each utterance u_i contains a speaker role r_j and sentence s_i. We simply concatenate them by ":" and get utterance $u_i = r_j : s_i$. For role-oriented summarization tasks, the data contains different summary y^{r_j} for different speaker roles r_j. In this paper, we employ y^{user} and y^{agent} to represent summaries of two different roles and y^{final} to represent the overall summary of the whole dialogue. It is deserved to mention that our method can also be easily applied to datasets with more than two speaker roles by introducing different role prompts. Normal dialogue summarization

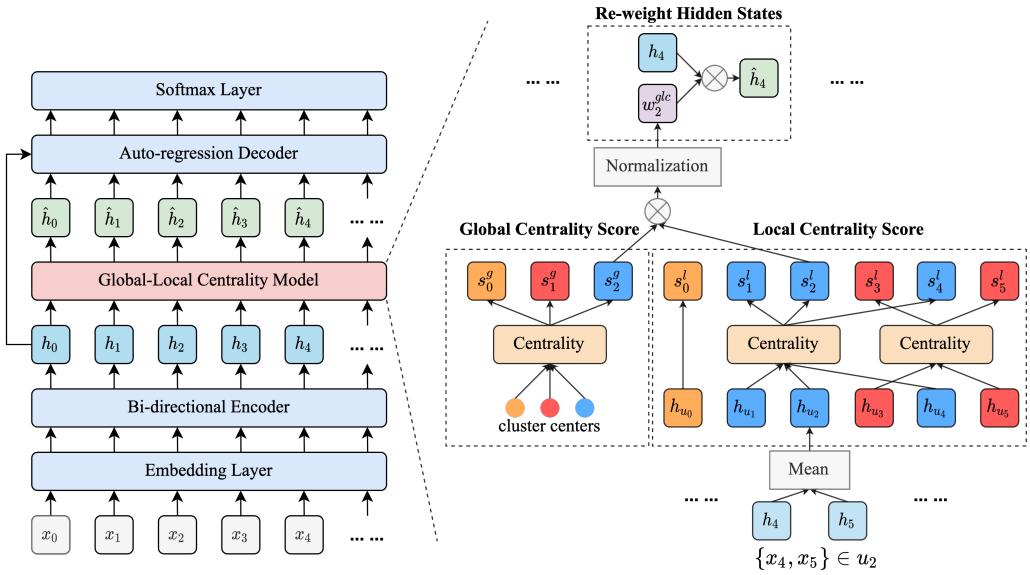

Figure 2: The main structure of our proposed method. The left is the framework of seq2seq with the GLC model. The right is the detailed process of our proposed GLC model.

task aims to generate overall summaries y^{final} and role-oriented summarization task aims to generate role-specific summaries $y^{[user|agent|final]}$ from the input dialogue $\mathcal{D} = \{u_1, \ldots, u_N\}$ according to the given role.

2.2 Role Prompts

For role-oriented summarization tasks, previous works train multiple independent models for different role summaries, which is proven to hurt the performance of model (Lin et al., 2022) and needs more computation resources. In this paper, we employ a simple but effective trick to ensure that we only need to train a single model to obtain different role-specific summaries and overall summaries. Specifically, we use the prompts to control the generation of different kinds of summaries, which attach "[User Summary]", "[Agent Summary]", and "[Final Summary]" to the start of each dialogue as input to guide the model to generate required summaries. After that, the input context is re-formalized as "[Role Prompt] Dialogue Contexts" and then tokenized as T tokens/words $\{x_t\}_{t=1}^{T}$ for the encoder of seq2seq model.

2.3 Bi-directional Encoder

The bi-directional encoder is used to get tokens the semantic vector representations $\{h_t\}_{t=1}^{T}$ by capture bi-directional context information from tokens $\{x_t\}_{t=1}^{T}$ as follows:

$$\{h_t\}_{t=1}^{T} = \texttt{Encoder}(\{x_t\}_{t=1}^{T}) \quad (1)$$

Then, we use the average of tokens vectors in each utterance as the semantic representations of dialogue utterances as follows:

$$h_{u_i} = \frac{1}{|u_i|} \sum_{t} x_t, x_t \in u_i \quad (2)$$

After that, we can get the token-level semantic representations $\{h_t\}_{t=1}^{T}$ and the utterance-level semantic representations $\{h_{u_i}\}_{i=0}^{N}$, where h_{u_0} is the vector representation of the attached role prompt, if role prompt is used.

2.4 Global-Local Centrality Model

Before feeding the representations into the decoder to generate the final summaries, we employ our proposed global-local centrality (GLC) model to re-weight the vector representations to identify salient facts in sub-topics over previous utterance-level semantic representations $\{h_{u_i}\}_{i=0}^{N}$.

Firstly, our GLC obtains several cluster center points $\{c_k\}k = 1^K$, which represent the center of sub-topics in the vector space. Then each utterance is assigned to the nearest center. As shown in Figure 2, utterances with the same color belong to the same sub-topic. We compute the global centrality score based on the cluster center representations to measure the importance of sub-topics and the local centrality score based on the utterance representations to measure the importance of each utterance belonging to the same sub-topic. Then, we employ their combination to get global-local

centrality weights, which are used to re-weight the token-level vector representations. Finally, the re-weighted token-level vector representations are fed into the decoder to generate the summary. Our GLC can be directly plug-in any seq2seq structures, which makes it flexible.

2.4.1 Obtain Cluster Centers

To obtain the clusters, we directly call the K-Means algorithm, which is effective and widely used for cluster tasks. And we all know setting the number of cluster centers for the K-Means algorithm is crucial and hard for the final results. However, we empirically find that we can set it as the number of utterances $(N+1)$ and then assign each utterance to the nearest cluster center point in the vector space. After that, we find that many cluster centers have no assigned utterances and can be dropped. Based on this, we assume $K < (N+1)$ cluster centers $\{c_i\}_{i=1}^{K}$ are kept and note the vector representations of them as $\{h_i^c\}_{i=1}^{K}$.

$$\{h_i^c\}_{i=1}^{K} = \texttt{KMeans}(\{h_{u_i}\}_{i=0}^{N}) \tag{3}$$

And after the assignment of utterances, we can get K clusters $\{\mathcal{C}_k\}_{i=k}^{K}$, which contain utterances with similar sub-topics. Each $\mathcal{C}_k$ contains several utterances and one cluster center point c_k. Through the previous method, we do not need to manually set the number of cluster centers for the K-Means algorithm.

2.4.2 Global Centrality

The global centrality score aims to measure the importance of each sub-topic by computing degree centrality based on the cluster center representations $\{h_k^c\}_{k=1}^{K}$. Each cluster center can be seen as one node on the graph, and the edge value between nodes k and j is $(h_k^c)^T \cdot h_j^c$. Then, the degree centrality of each cluster can be computed as follows:

$$\texttt{Cen}(c_k) = \sum_j (h_k^c)^T \cdot h_j^c \tag{4}$$

where $\texttt{Cen}(c_k)$ represents the importance of the cluster/sub-topic k in the dialogue. Then we normalize the score $\texttt{Cen}(c_k)$ by $\frac{\texttt{Cen}(c_k)}{||\{\texttt{Cen}(c_k)\}_{k=1}^{K}||_2}$.

2.4.3 Local Centrality

The local centrality score aims to measure the importance of utterances in each cluster by computing the degree centrality. Each utterance can be seen as one node on the graph, and the edge value between nodes i and j is $(h_{u_i})^T \cdot h_{u_j}$. Then, the centrality of each utterance in the same cluster $\mathcal{C}_k$ can be computed as follows:

$$\texttt{Cen}(u_i) = \sum_j (h_{u_i})^T \cdot h_{u_j}, u_i, u_j \in \mathcal{C}_k \tag{5}$$

where $\texttt{Cen}(u_i)$ represents the importance of utterances in the k-th cluster/sub-topic. Then we normalize the score $\texttt{Cen}(u_i)$ the same as the previous global centrality score.

2.4.4 Global-Local Centrality Weight

We can obtain the importance of each cluster (global centrality score) and the importance of utterances in each cluster (local centrality score) by the previous two steps. The most important utterance in the most important sub-topic should be assigned more attention when generating the summary. So we obtain global-local centrality weight for each utterance in the dialogue by simply multiplying two centrality scores as follows:

$$w_i^{glc} = \texttt{Cen}(u_i) \cdot \texttt{Cen}(c_k), u_i \in \mathcal{C}_k \tag{6}$$

Finally, we employ the global-local centrality weights to re-weight the token-level vector representations $\{h_t\}_{t=1}^{T}$ as follows:

$$\hat{h}_t = w_i^{glc} \cdot h_t, x_t \in u_i \tag{7}$$

Where each token uses the global-local centrality weight w_i^{glc} of its utterance u_i to re-weight the vector representation h_t. The token level representations $\{h_t\}_{t=1}^{T}$ are converted into $\{\hat{h}_t\}_{t=1}^{T}$.

2.5 Auto-regression Decoder

The auto-regression decoder generates the final summary based on the re-weighted context representations $\{\hat{h}_t\}_{t=1}^{T}$ as follows:

$$P(\hat{y}) = \texttt{Decoder}(\{\lambda \cdot \hat{h}_t + (1-\lambda) \cdot h_t\}_{t=1}^{T}) \tag{8}$$

where λ is a hyper-parameter to control the influence of GLC, the default value of λ is 0.5. In the training stage, the model learns the optimal parameters θ by minimizing the negative log-likelihood.

3 Experiments

3.1 Datasets and Metrics

We evaluate our method on three public datasets: CSDS (Lin et al., 2021)[2], MC (Song et al., 2020)[3],

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

[3]https://github.com/cuhksz-nlp/HET-MC

CSDS	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
PGN	55.58/53.55/50.20	39.19/37.06/35.12	53.46/51.05/47.59	30.03/29.64/28.25	77.96/78.68/76.13
PGN-both	57.20/56.08/51.62	40.37/39.10/36.50	55.14/53.85/49.12	32.58/33.54/29.78	78.69/79.52/76.74
PGN-GLC	57.94/57.14/52.85	40.97/39.55/37.14	55.68/54.25/49.86	32.95/33.87.30.15	78.93/79.86/76.98
BERT	53.87/52.72/49.57	37.59/36.39/33.82	52.40/50.44/46.83	29.90/30.17/26.99	78.52/79.23/76.39
BERT-both	57.24/54.36/51.92	40.12/40.70/36.37	54.87/55.17/49.52	32.13/32.04/29.23	79.85/80.70/77.23
BERT-GLC	57.59/55.14/52.34	41.28/41.84/36.48	55.74/55.86/50.16	32.75/32.64/29.81	79.89/80.71/77.28
BART	59.07/58.78/53.89	43.72/43.59/40.24	57.11/56.86/50.85	34.33/34.26/31.88	79.74/80.67/77.31
BART-both	59.21/58.93/54.01	43.88/43.69/**40.32**	57.32/57.28/51.10	34.75/34.49/32.30	79.72/80.64/77.30
BART-GLC	**60.07/61.42/54.59**	**44.67/45.83**/40.02	**58.10/59.25/52.43**	**35.89/36.43/32.58**	**80.10/81.83/77.61**

Table 1: Results on the CSDS dataset test set.

MC	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
PGN	85.32/94.82/82.56	81.25/94.32/77.91	84.34/94.77/81.47	71.50/87.66/68.10	92.90/97.60/91.74
PGN-both	85.98/95.10/83.37	81.93/94.59/78.78	84.94/95.06/82.20	72.77/87.82/69.63	93.23/97.71/92.15
PGN-GLC	86.57/95.31/83.97	82.04/94.88/79.16	85.37/96.48/82.84	73.02/88.11/70.04	93.47/97.95/92.36
BERT	84.07/95.10/81.53	79.90/94.48/76.78	83.04/95.06/80.30	68.19/87.20/64.09	92.68/97.86/91.71
BERT-both	84.69/95.18/82.02	80.76/94.62/77.54	83.68/95.14/80.84	69.33/87.40/65.40	93.02/97.90/91.91
BERT-GLC	85.64/95.49/82.87	81.44/94.97/78.05	84.16/96.10/81.57	69.84/87.94/66.01	93.15/97.92/92.36
BART	88.37/95.42/86.33	84.75/94.99/82.33	87.38/95.37/85.30	73.68/90.29/68.93	93.65/97.94/92.63
BART-both	88.52/95.63/87.06	85.22/95.42/82.89	87.75/95.91/85.78	73.87/90.70/69.31	93.69/97.88/92.69
BART-GLC	**89.55/96.84/88.47**	**86.47/96.14/84.62**	**88.56/96.23/86.77**	**74.19/91.32/70.18**	**94.17/98.25/92.96**

Table 2: Results on the MC dataset test set.

and SAMSUM (Gliwa et al., 2019)[4]. The statistical information of them is shown in the appendix. CSDS is the first role-oriented dialogue summarization dataset, which provides separate summaries for user and agent (customer service). MC is a Chinese medical inquiry dataset containing question summaries of patients and suggestion summaries of doctors. We note them as the user and agent summary. For the MC dataset, we follow the data process and data split from RODS (Lin et al., 2022). SAMSUM is a widely used English dialogue summarization dataset to evaluate the performance of models.

We employ lexical-level and semantic-level metrics to evaluate the performance of all models. Specifically, we use lexical level ROUGE-1/2/L (Lin, 2004)[5] and BLEU (Papineni et al., 2002)[6], which measure the similarity of references and generated summaries by computing the n-gram overlap of them. We use semantic level BERTScore (Zhang* et al., 2020)[7] and MoverScore (Zhao et al., 2019)[8], which employ pre-trained language models to map the text into low-dimensional vectors in semantic space and then measure the similarity by computing the similarity by cosine similarity or word mover distance. We can evaluate the performance of each model comprehensively through the previous metrics. And all reported results are the average results of three different model checkpoints. The results of Moverscore on three datasets can be found in the appendix.

3.2 Baselines

We applied our GLC on three widely used seq2seq models: PGN (See et al., 2017), BERTAbs (Liu and Lapata, 2019), and BART (Lewis et al., 2020; Shao et al., 2021). **PGN** model is an LSTM-based seq2seq model without pre-training. **BERTAbs** is a BERT-based model, which employs BERT as the encoder and adds several transformer blocks as the decoder to generate summaries. We note it as BERT. **BART** is a pre-trained transformer-based seq2seq model, which achieves the best results on many generation tasks. We add our proposed GLC into the previous three models and note them as **PGN-GLC**, **BERT-GLC**, and **BART-GLC**. We also compare our method with previous SOTA models: **PGN-both** and **BERT-both** from (Lin et al.,

[4]https://huggingface.co/datasets/samsum
[5]https://pypi.org/project/rouge-score/
[6]https://github.com/mjpost/sacreBLEU
[7]https://github.com/Tiiiger/bert_score
[8]https://github.com/AIPHES/emnlp19-moverscore

SAMSUM	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
PGN	40.08	15.28	36.63	37.49	80.67
PGN-GLC	41.11	16.24	37.31	38.10	81.54
BERT	50.34	24.71	46.63	46.98	88.72
BERT-GLC	51.18	25.26	47.07	47.66	89.64
BART	53.12	27.95	49.15	49.28	92.14
BART-GLC	**53.74**	**28.83**	**49.62**	**50.36**	**92.77**

Table 3: Results on the SAMSUM dataset test set.

2022), which proposed a role-interaction attention mechanism for the decoder. We reproduce it in the BART model as **BART-both**. For SAMSUM, we do not compare with BART-both due to this dataset does not contain role-oriented summaries.

3.3 Implementation Details

We use Chinese-BART-base[9] and BART-large[10] to initialize our transformer-based seq2seq model for Chinese and English datasets respectively. We train all BART models on 4xV100 GPUs and PGN/BERT-based models on 1xV100 GPU. For all models, the maximum input length is 512, the maximum generated summary length is 150, and the beam size is 3. For BART-based models, the learning rate is 1e-4 with 10% warmup steps, the total batch size is 64, and the training epochs are 5. For PGN/BERT-based models, we follow the settings from (Lin et al., 2022).

3.4 Results

The main results of the two role-oriented dialogue summarization datasets are shown in Table 1-2. Each block has three values, representing the final summary/user summary/agent summary from left to right. We can see that our proposed GLC can bring significant improvement to PGN, BERTAbs, and BART on the two datasets and BART-GLC achieves new state-of-the-art results. It is deserved to mention that our model does not need to modify any structure of the seq2seq structure and only needs to train one model for different summaries. We can see that the gain of metrics on the CSDS is better than on the MC, due to the summary of the MC dataset being highly similar to the input dialogue contexts. The results of the BERT-based model sometimes is worse than the PGN-based, we guess the reason is the prior knowledge learned in the pre-training stage of BERT is not suitable for the generation tasks. The improvement of lexical level metrics is more conspicuous than semantic level metrics due to the change of several words that may not affect the semantics of generated sentences. Overall, our proposed GLC is proved effective for the role-oriented dialogue summarization task with results on the two datasets.

The main results of the English dialogue summarization dataset are shown in Table 3. Because the SAMSUM does not provide role-specific summaries, we only report the performance of overall final summaries. From the results, we can see that our GLC can also bring significant improvements to three different seq2seq structures. We can see that the BERTScore is very high on SAUSUM, we guess that because the gold reference of this dataset is very short and this makes the semantic similarity between generated summaries and gold summaries close. The results of SAMSUM demonstrated the effectiveness and generalization of our proposed method.

4 Discussion

We conduct many external experiments on the CSDS dataset to further analyze the effectiveness of our proposed GLC. And more discussions are shown in the appendix.

4.1 Ablation Study

To understand the impact of each component of our proposed GLC model, we compare the full BART-GLC with the following variants: (1) **BART:** three fine-tuned BART models for different summaries (final/user/agent); (2) **BART+Prompt:** singe BART model with role prompts; (3) **BART+GC:** three BART models using global centrality scores

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

[10]https://huggingface.co/facebook/bart-large

	ROUGE-1
BART	59.07/58.78/53.89
+Prompt	59.42/58.96/54.03
+GC	59.64/59.55/54.24
+LC	59.37/59.47/54.11
+GC,LC	59.84/60.91/54.43
BART-GLC	**60.07/61.42/54.59**

Table 4: Ablation study on the CSDS dataset.

	Win	Loss	Tie
CSDS&MC	56.4	2.4	41.2
SAMSUM	51.8	3.2	45.3

Table 5: Human evaluation results.

to re-weight hidden states; (4) **BART+LC:** three BART models using local centrality scores to re-weight hidden states; (5) **BART+GC,LC:** three BART models using global-local centrality scores to re-weight hidden states. The results of these models are shown in Table 4. From the results, we can see that all three components can bring improvement to the BART model, and the global-local centrality brings the greatest improvement. Interesting, The improvement brought by the combination of global and local centrality is far greater than the improvements they bring separately. This proves that global and local centrality are mutually beneficial.

4.2 Human Evaluation

We use human evaluation (Fang et al., 2022) to verify that our model outperforms the baseline. Specifically, we randomly sample 100 examples from three datasets and ask five NLP researchers to give a comparison between our model and baseline models. The evaluation results are represented as win, loss, and tie, respectively indicating that the quality of generated summary by BART-GLC is better, weaker, or equal to the strong baselines. Annotators were asked to judge from two aspects: fluency (whether contains grammatical and factual errors) and coverage (whether contains salient sub-topic information in the dialogue). For two role-oriented dialogue summarization datasets CSDS and MC, our model is compared with BART-both. For SAMSUM, our model is compared with BART. From

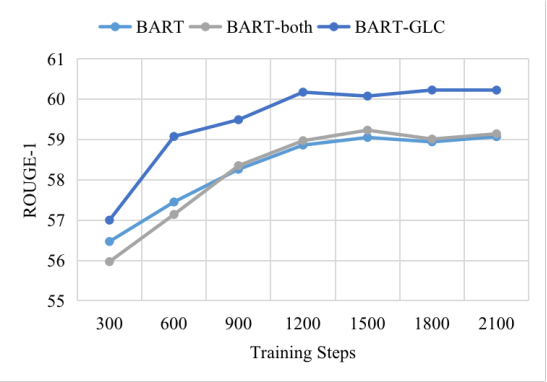

Figure 3: The ROUGE-1 score of different training step checkpoints.

the results in Table 5, we can see that our model is better than the baseline. Annotators tend to give ties on SAMSUM dataset. This may be caused by the length of summaries is short, which makes it hard to judge whether the summary is better or worse than the baseline model.

4.3 Convergence of Training

We also compare the training convergence speed with BART and BART-both to prove our proposed GLC can bring effective prior knowledge for the seq2seq model. As shown in Figure 3, we can see that BART-GLC achieves comparable performance at 900 steps during training and reaches the SOTA results at 1,200 steps. This phenomenon demonstrates that our GLC brings prior knowledge into the model and speeds up the model training.

4.4 Case study

We select one example from the test set to show the ability of our proposed GLC in Figure 4. On the upper-left of this figure are the GLC weights and the corresponding utterances. In the bottom-left of the figure is generated summary of our proposed BART-GLC. On the right of the figure is the input dialogue and each color refers to one sub-topic. From this case, we can see that the final summary focus on two sub-topics: "How to modify user's order" and "Questions about refunds". And from the color on the right of this figure, we can see that our GLC can catch them accurately. Interestingly, generic utterances are aggregated into one topic (e.g. hello). In the upper-left of this figure is the GLC weights and we can see that utterances, which are related to the final summary and belong to the vital sub-topics, are assigned high weights. This proves the global-local centrality exactly identified

Figure 4: One case from the CSDS test set. Each color refers to one sub-topic. In the upper-left of this figure are the GLC weights and the corresponding utterances. In the bottom-left of the figure is generated summary of our proposed BART-GLC. On the right of the figure is the input dialogue.

salient topics and utterances.

5 Related Work

Dialogue summarization has caught more and more attention in recent years and is widely used in various domains, e.g. meeting summarization (Carletta et al., 2006; Feng et al., 2021), daily dialogue summarization (Krishna et al., 2021; Chen et al., 2021; Zhong et al., 2021), etc. Different from traditional summarization tasks, dialogue summarization needs to identify the role of speakers and capture the change of sub-topics during the dialogue. Besides, the dialogue summarization task has less labeled data and longer inputs. All of these make dialogue summarization harder to solve (Chen and Yang, 2020; Zhang et al., 2021b; Feng et al., 2021; Lin et al., 2022).

Recent dialogue summarization models can be categorized into three types: 1) data augmentation methods (Feng et al., 2021; Chen and Yang, 2021; Khalifa et al., 2021), which attempt to construct more pseudo-data to train a better model; 2) topic-based models (Zou et al., 2021; Liu et al., 2021; Qi et al., 2021), which track the change of topic information in the dialogue to generate more focused summary; and 3) semantic structure-based models (Liu and Chen, 2021; Fu et al., 2021; Zhang et al., 2021a; Lei et al., 2021; Zhao et al., 2021; Zhang et al., 2022), which employs semantic structures to enhance the summarization model.

However, they ignored the sub-topics information in the dialogue utterances, which is crucial for dialogue summarization. Recently, Zhao et al. (2020) modified the attention mechanism to focus on the topic words, which can force the model to learn the topic information. Zou et al. (2021) employed Neural Topic Model to model the global level topic information. Liu et al. (2021) tried to model the change of sub-topics by introducing contrastive learning. Differently, in this paper, we bring the centrality, that has been widely used in unsupervised summarization (Zheng and Lapata, 2019; Liang et al., 2021, 2022), into the dialogue summarization task and proposed a novel topic-aware Global-Local Centrality model to capture salient dialogue utterances and sub-topics at the same time. Our proposed method is effective and more flexible.

6 Conclusion

In this paper, we bring the centrality into dialogue summarization tasks and proposed a novel topic-aware Global-Local Centrality (GLC) model for better capturing the sub-topic information in the dialogue utterances. Our GLC can be easily applied to any seq2seq structure and bring improvement to

their performance. Experiments and further analysis demonstrated that GLC can effectively identify vital sub-topics and salient content in the dialogue. In future work, we will try to extend our work to datasets with longer inputs.

Limitations

Our model also has some limitations: 1) The computation of sub-topic centers brings extra inference time into the basic seq2seq models. 2) We did not try to evaluate our model on longer dialogue summarization datasets. 3) We did not build a specific mechanism for different roles in role-oriented dialogue summarization task. We will try to solve these limitations in future work.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081), the 2022 Tencent Big Travel Rhino-Bird Special Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18).

References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jiaao Chen and Diyi Yang. 2020. 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. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yue Fang, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Bo Long, Yanyan Lan, and Yanquan Zhou. 2022. From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3859–3869, Seattle, United States. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

Xiyan Fu, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Changlong Sun, and Zhenglu Yang. 2021. RepSum: Unsupervised dialogue summarization based on replacement strategy. 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 6042–6051, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

Xinnian Liang, Jing Li, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2022. Improving unsupervised extractive summarization by jointly modeling facet and redundancy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1546–1557.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page 1957–1965, New York, NY, USA. Association for Computing Machinery.

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. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2021a. Unsupervised abstractive dialogue summarization for tete-a-tetes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14489–14497.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021b. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lulu Zhao, Weihao Zeng, Weiran Xu, and Jun Guo. 2021. Give the truth: Incorporate semantic slot into abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2435–2446, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

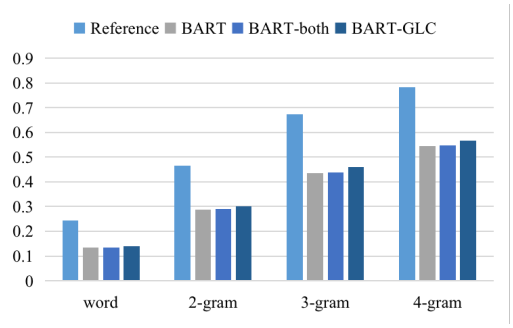

Figure 5: Percentage of novel words/n-grams in the reference and generated summaries of the CSDS test set.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14665–14673.

A Datasets

	CSDS	MC	SAMSUM
Train Size	9,101	29,324	14,732
Val. Size	800	3,258	818
Test Size	800	8,146	819
Input Length	321.92	292.21	94.52
User Sum. Length	37.28	22.37	-
Agent Sum. Length	48.08	95.32	-
Final Sum. Length	83.21	114.54	20.34

Table 6: Statistical information of three datasets.

The statistical information of three datasets is shown in Table 6.

B Moverscore Results

For Moverscore, we employ chinese-bert-wwm-ext[11] to get the contextual embeddings of Chinese text input. Because Lin et al. (2021) did not provide they use what Chinese representation model, we use chinese-bert-wwm-ext to re-evaluate all their results and report in Table 7.

B.1 How abstractive is our model?

An abstractive model can be innovative by using words that are not from the input document in the summary. We measure the abstractive by the ratio of novel words or n-gram phrases in the summary. A higher ratio means a more abstractive model. We show the results in Figure 5. We can see that

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

MoverScore	CSDS	MC	SAMSUM
PGN	59.00/58.68/58.23	80.90/93.84/79.69	59.87
PGN-both	59.48/59.32/58.64	81.67/94.04/80.52	-
PGN-GLC	59.67/59.51/58.85	81.97/94.45/80.84	60.04
BERT	58.23/58.10/57.79	81.28/93.90/80.48	61.17
BERT-both	59.52/59.55/58.46	82.26/94.20/81.02	-
BERT-GLC	59.74/59.62/58.90	82.64/94.49/81.44	61.59
BART	60.11/59.86/58.75	82.35/94.17/81.27	62.04
BART-both	60.12/59.86/58.73	82.32/94.02/81.40	-
BART-GLC	**60.32/61.03/59.02**	**82.94/95.35/82.10**	**62.27**

Table 7: MoverScore on three datasets.

our BART+GLC is more attractive than BART and BART-both. However, all of them have a big margin compared with references. It means more research is needed for generating more abstractive summaries.