Multi-view Contrastive Learning for Medical Question Summarization

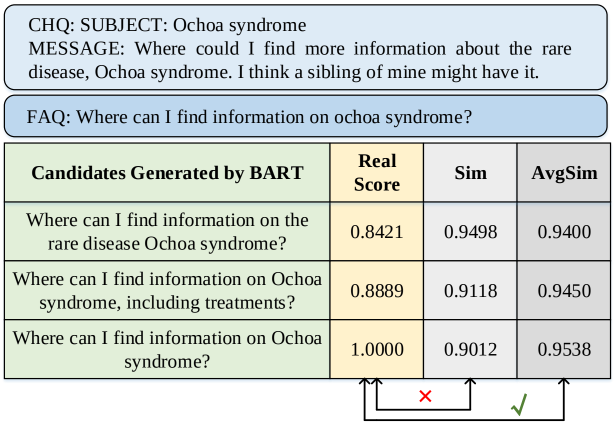
Sibo Wei*, Xueping Peng†, Hongjiao Guan*(✉), Lina Geng‡, Ping Jian§, Hao Wu§, Wenpeng Lu*(✉),
* Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
† Australian Artificial Intelligence Institute, University of Technology Sydney, Australia
‡ Department of Blood Purification, Qilu Hospital of Shandong University, Jinan, Shandong, China
§ School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
* Shandong Provincial Key Laboratory of Computer Networks,
Shandong Fundamental Research Center for Computer Science, Jinan, China
(✉) Corresponding author email: guanhongjiao2008@163.com, wenpeng.lu@qlu.edu.cn

***Abstract*—Most Seq2Seq neural model-based medical question summarization (MQS) systems have a severe mismatch between training and inference, i.e., exposure bias. However, this problem remains unexplored in the MQS task. To bridge this research gap and alleviate the problem of exposure bias, we propose a novel re-ranking training framework for MQS called Multi-view Contrastive Learning (MvCL). MvCL simultaneously considers the similarity scores between medical questions and candidate summaries as well as the average similarity scores between candidate summaries and other candidates within the same group, and utilizes contrastive learning to optimize the model's ranking ability. Additionally, we propose a new multi-level inference approach to adapt to this training strategy. The approach first filters out candidate summaries that are dissimilar to the original medical question, and then selects the summary with the highest average similarity to other candidate summaries from the remaining candidates as the final output. We conducted extensive experiments, and the results demonstrate that our proposed MvCL framework achieves state-of-the-art results on the majority of evaluation metrics across four datasets.** [1]

***Index Terms*—Medical Question Summarization, Contrastive Learning, Re-ranking Framework**

I. Introduction

Collaboration across disciplines is a key focus for researchers in the fields of deep learning and healthcare services [1]–[4]. Retrieval-based automatic medical question-answering systems (MQAS) are a significant application resulting from this interdisciplinary collaboration. MQAS not only provides convenient consultation services to consumers but also helps alleviate the shortage of healthcare professionals. Questions submitted by consumers often contain redundant or irrelevant information, as well as non-specialized or inaccurately expressed queries, making it challenging for MQAS to retrieve relevant answers. To provide accurate answers, MQAS must accurately understand the core intent of the questions [5], which is a crucial and challenging task. To address this problem, researchers have proposed solutions such as medical question entailment [6]–[8], query relaxation [9], and medical question summarization [1], [5], [7], [8]. Among these approaches, medical question summarization has demonstrated the most promising results and received more attention.



CHQ: SUBJECT: Ochoa syndrome
MESSAGE: Where could I find more information about the rare disease, Ochoa syndrome. I think a sibling of mine might have it.

FAQ: Where can I find information on ochoa syndrome?

Candidates Generated by BART	Real Score	Sim	AvgSim
Where can I find information on the rare disease Ochoa syndrome?	0.8421	0.9498	0.9400
Where can I find information on Ochoa syndrome, including treatments?	0.8889	0.9118	0.9450
Where can I find information on Ochoa syndrome?	1.0000	0.9012	0.9538

Fig. 1. A sample from the MeQSum dataset, where CHQ refers to a consumer health question and FAQ represents a reference summary written by an expert. The candidate summary is generated by a fine-tuned BART model. The term **Real Score** represents the actual score acquired from labels, while **Sim** denotes the similarity between a candidate summary and the CHQ. Additionally, **AvgSim** signifies the average similarity between a candidate summary and other candidate summaries.

The objective of medical question summarization (MQS) is to condense consumer health questions (CHQs), which may contain redundant or irrelevant information, as well as non-specialized or inaccurately expressed content, into concise frequently asked questions (FAQs). The resulting FAQs eliminate unnecessary and redundant information from CHQs, enabling them to capture the essential meaning of the original questions. This facilitates MQAS in retrieving accurate answers more effectively.

[1] Our code is available at: https://github.com/yrbobo/MvCL

The MQS task was introduced by Ben Abacha et al. in 2019 [10]. They manually annotated the MeQSum dataset and utilized pointer generation (PG) networks to generate summaries for consumer health questions. Most existing MQS methods rely on the Seq2Seq transformer model and employ various techniques to enhance model performance. Commonly used Seq2Seq transformer models for MQS include T5 [11], ProphetNet [12], PEGASUS [13], and BART [14], with BART achieving better performance. Previous studies have attempted to improve MQS from different perspectives. Some studies enhance the generative transformer models by incorporating structured knowledge or lexical resources [15], [16]. Others strengthen MQS using different strategies, including multi-task learning [7], [8], transfer learning [17], [18], reinforcement learning [19], and contrastive learning [1], [5]. Despite the certain success achieved by existing MQS works, most of them have overlooked the issue of exposure bias in Seq2Seq neural models.

Previous MQS systems based on Seq2Seq neural models have a severe mismatch between training and inference, i.e., exposure bias. Exposure bias has garnered considerable attention in open-domain abstractive summarization [20]–[23], where re-ranking systems are frequently employed to address this problem. However, this problem has not been investigated in MQS task. To bridge this research gap, we employ SimCLS [20], the popular framework in open-domain abstractive summarization, to the MQS task, and demonstrate its effectiveness in enhancing the performance of MQS systems. However, SimCLS solely considers on the relationship between the original medical question and its candidate summaries, thereby restricting its performance. We believe that when evaluating the quality of candidate summaries, the relationship between a candidate summary and other candidate summaries (belonging to the same CHQ) is also crucial because high-quality candidate summaries are likely to share more features with other candidate summaries. As shown in Fig.1, we use fine-tuned BART to generate three candidate summaries for the CHQ and then use the trained SimCLS model to score these three candidate summaries. We observe that the first two candidate summaries, which contain redundant information and have a higher lexical overlap with the original medical question, receive higher scores for 'Sim' than the third candidate summary. Consequently, the model chooses the first summary, which possesses lower abstractness and a lower real score. However, upon considering the relationship between candidate summaries and other candidate summaries, we discover that the average similarity score between a candidate summary and its counterparts aligns with the real score.

Based on the aforementioned issues, we propose a novel re-ranking training framework for MQS called Multi-view Contrastive Learning (MvCL). MvCL simultaneously considers the similarity scores between medical questions and candidate summaries (outer score view) as well as the average similarity scores between candidate summaries and other candidates within the same group (inner score view). This framework utilizes contrastive learning to optimize the model's ranking ability. Additionally, we propose a new multi-level inference approach to adapt to this training strategy. The approach filters candidate summaries dissimilar to the original medical question based on the outer score view and selects the summary with the highest shared feature degree from the inner score view as the final output. Extensive experiments demonstrate that our re-ranking training framework, MvCL, outperforms the SimCLS framework and significantly improves the performance of MQS systems. Furthermore, it can seamlessly integrate as a flexible component into existing Seq2Seq-based summarization frameworks.

Accordingly, this paper makes the following major contributions:

- We present a novel re-ranking training framework, named MvCL, for medical question summarization. Our MvCL framework simultaneously takes into account the similarity scores between medical questions and candidate summaries (outer score view) and the average similarity scores among candidate summaries within the same group (inner score view). It employs contrastive learning to optimize the model's ranking ability.
- In addition, we introduce a novel multi-level inference approach that, when combined with our proposed training framework, further improves the selection capability of the re-ranking model.
- Extensive experiments were conducted to evaluate the performance of our proposed MvCL framework, and the results demonstrate its state-of-the-art performance on most evaluation metrics across four datasets.

II. Methodology

A. The Proposed Framework

The model framework we propose is illustrated in Fig. 2. The framework consists of two stages: one for generating multiple candidate summaries and another for re-ranking the candidate summaries to select the final summary. In the first stage, we fine-tune a pre-trained Seq2Seq model to generate multiple candidate summaries for medical questions. In the second stage, we train a re-ranker that considers both the similarity score between the original medical question and the candidate summary from an outer view perspective and the average similarity score between the candidate summary and other candidate summaries from an inner view perspective. During the inference stage, we propose a multi-level rank inference method that first filters out some candidate summaries based on the outer view score and then selects the best summary from the remaining candidate summaries using the inner view score.

B. Candidate Summary Generation

First, we need to fine-tune a pre-trained Seq2Seq language model. There are various models to choose from, such as T5 [11], ProphetNet [12], PEGASUS [13], and BART [14]. Among these models, BART has demonstrated superior performance in the MQS task. Therefore, we will use BART as the base generative model. BART is trained using the

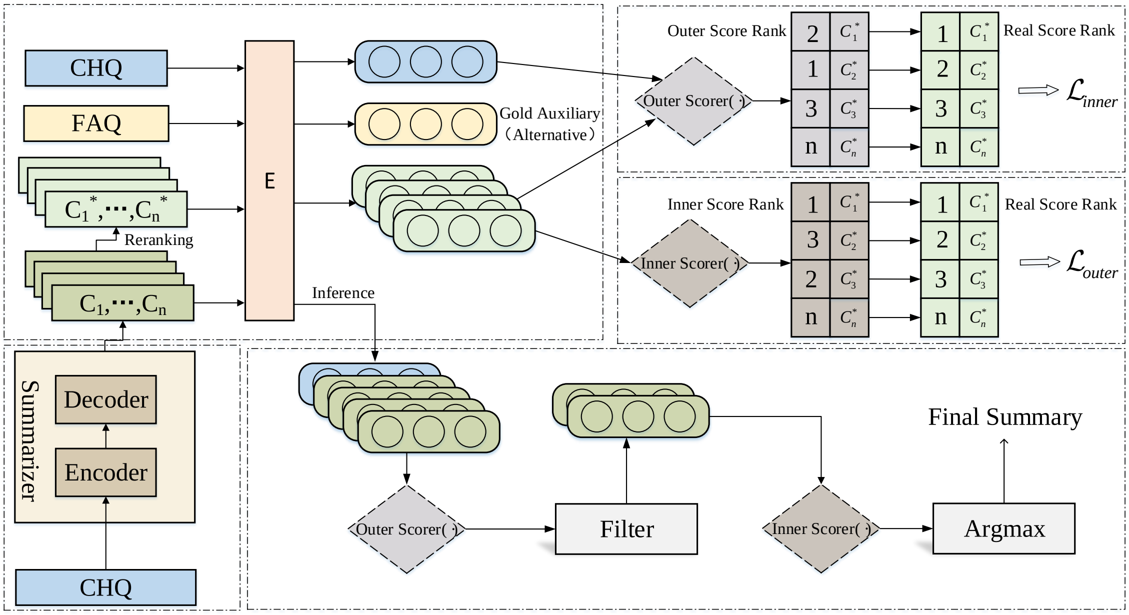


Fig. 2. The framework of our proposed model, which contains two stages, one for generating multiple candidate summaries with Seq2Seq models, another for re-ranking them to select the best one as final summary. In the first stage, we fine-tune a pre-trained Seq2Seq model to generate multiple candidate summaries for medical questions. In the second stage, we train a re-ranker that considers both the similarity score between the original medical question and the candidate summary from an outer score view and the average similarity score between the candidate summary and other candidate summaries from an inner score view. Additionally, reference summaries can be used as golden auxiliary during the training stage to improve the training effectiveness of the model. During the inference stage, we propose a multi-level rank inference method that first filters out some candidate summaries based on the outer view score and then selects the best summary from the remaining candidate summaries using the inner view score.

maximum likelihood estimation (MLE) algorithm. For the i-th training sample $\{Q, S\}$, where Q, S denote the consumer health question and gold reference summary, respectively, MLE is equivalent to minimizing the sum of the negative likelihood of the l tokens $\{s_1, \cdots, s_j, ..., s_l\}$ in the reference summary S, i.e., to optimize the cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{j=1}^{l}\sum_{s^*} p_{true}(s^*|Q, S_{<j}) \log p_{f_\theta}(s^*|Q, S_{<j}; \theta) \quad (1)$$

where s^* represents the token currently generated by the model. $S_{<j}$ refers to the partial reference sequence $\{s_0, \cdots, s_{j-1}\}$ and s_0 is a pre-defined start token. p_{true} denotes the one-hot distribution in the standard MLE framework. θ refers to the parameters of f and p_{f_θ} is the probability distribution entailed by these parameters.

After the fine-tuning of the model, for a specific sample $\{Q, S\}$, we employ the beam search algorithm to generate a set of candidate summaries, denoted as $\mathbb{C} = \{C_1, \cdots, C_n\}$. In order to train the re-ranker in the next stage, we score these candidate summaries based on the reference summary S using the ROUGE evaluation metric, which is widely used in the MQS task. The resulting set of candidate summaries after scoring is represented as follows:

$$\mathbb{C}^* = \{C_1^*, \cdots, C_i^*, \cdots, C_n^*\}, \quad (2)$$

where $C_1^*, \cdots, C_n^*$ are sorted in descending order based on their ROUGE score. Please note that these scored candidate summaries are only used as supervised signals during the training of the re-ranker model and are not used during the inference, as reference summaries cannot be used.

C. Train a Re-ranker with Multi-view Contrastive Learning

Next, we will apply Multi-view Contrastive Learning (MvCL) to train a re-ranker $E(\cdot)$ that can measure the quality of candidate summaries. The goal of $E(\cdot)$ is to assign different scores to the candidate summaries $C_1^*, \cdots, C_i^*, \cdots, C_n^*$ in set $\mathbb{C}^*$. Unlike SimCLS [20], which relies solely on the similarity score between the source document and the candidate summary, MvCL considers scores from two views simultaneously. One view is the similarity between the medical question and the candidate summary (Outer Score View), while the other view is the average similarity between the candidate summary and other candidate summaries belonging to the same medical question (Inner Score View). In this work, we instantiate $E(\cdot)$ as a large pre-trained self-attention model, RoBERTa [24]. It is used to encode the medical question and its candidate summaries separately and calculate similarity scores from different views.

1) Contrastive Loss on Outer Score View: For a specific sample $\{Q, S\}$ and its ordered set of candidate summaries $\mathbb{C}^*$, a high-quality candidate summary should have a higher semantic similarity with the medical question Q. Therefore, the outer view score considers the similarity score between candidate summary C_i^* and medical question Q. The outer scorer, denoted as $E_O(\cdot)$, is defined as the cosine similarity between medical question Q and its candidate summary C_i^*. Our goal is to make the scores assigned to candidate summaries $C_1^*, \cdots, C_i^*, \cdots, C_n^*$ by $E_O(\cdot)$ rank as closely as possible to the ranks of the real scores. To achieve this, we introduce the following contrastive loss:

$$\mathcal{L}_{outer} = \sum_i \sum_{j>i} \max(0, E_O(Q, C_j^*) - E_O(Q, C_i^*) + \lambda_{ij}) \tag{3}$$

where $\lambda_{ij} = (j-i) * \lambda$ is the corresponding margin following [25], and λ is a hyper-parameter.

2) Contrastive Loss on Inner Score View: For a specific sample $\{Q, S\}$ and its ordered set of candidate summaries $\mathbb{C}^*$, we believe that the quality of candidate summary C_i^* is not only related to the medical question Q but also to the other summaries in $\mathbb{C}^*$. This idea is inspired by the work of Seonwoo et al. [26], who proposed that the semantic textual similarity between two input sentences is not only determined by the sentences themselves but also by the nearest-neighbor sentences that are similar to the input sentences. The sentences in the candidate summary set have a natural neighbor relationship, so we incorporate this idea into our model. Specifically, for a candidate summary C_i^* in $\mathbb{C}^*$, if it has higher quality, it should share more common features with the other candidate summaries in $\mathbb{C}^*$. Therefore, the outer view score considers the average similarity score between C_i^* and the summaries in $\mathbb{C}_{\neq i}^*$, and the inner scorer $E_I(\cdot)$ is defined as follows:

$$E_I(\mathbb{C}_{\neq i}^*, C_i^*) = \frac{\sum\limits_{k \neq i}^{n} sim(C_k^*, C_i^*)}{n-1} \tag{4}$$

where $sim(\cdot)$ is cosine similarity.

Similar to the outer view, the goal of the inner view is to align the ranking of candidate summary $C_1^*, \cdots, C_i^*, \cdots, C_n^*$ assigned scores by $E_I(\cdot)$ as closely as possible with the ranking of real scores. To achieve this, we introduce the following contrastive loss:

$$\mathcal{L}_{inner} = \sum_i \sum_{j>i} \max(0, E_I(\mathbb{C}_{\neq j}^*, C_j^*) - E_I(\mathbb{C}_{\neq i}^*, C_i^*) + \mu_{ij}) \tag{5}$$

where $\mu_{ij} = (j-i) * \mu$ is the corresponding margin , and μ is a hyper-parameter.

3) Overall Objective Function: We jointly train the reranker $E(\cdot)$ using a combination of outer score view contrastive learning and inner score view contrastive learning. The overall objective function is formulated as follows:

$$\mathcal{L}_{rank} = \partial \cdot \mathcal{L}_{outer} + (1 - \partial) \cdot \mathcal{L}_{inner} \tag{6}$$

where ∂ is a weight parameter.

D. Inference with Multi-level Rank

To adapt to our training strategy, we propose a multi-level inference approach that complements multi-view contrastive learning. For a specific medical question Q in the test dataset, we first employ fine-tuned BART to generate n candidate summaries $\mathbb{C} = \{C_1, \cdots, C_n\}$. In the re-ranking stage, we initially score $C_1, \cdots, C_n$ using the outer scorer $E_O(\cdot)$ and rank them in descending order to obtain $\mathbb{C}' = \{C'_1, \cdots, C'_n\}$. Subsequently, we filter out $n-m$ candidates with lower scores, resulting in the following candidate set:

$$\mathbb{C}'' = \{C'_1, \cdots, C'_m\} \tag{7}$$

The first stage above filters out summaries that are not very similar to the original medical question itself, retaining summaries that have relatively high similarity to the medical question and are difficult to distinguish. In the second stage, the inner scorer $E_I(\cdot)$ is used to internally score the remaining summaries in set $\mathbb{C}''$, and the candidate summary with the highest score is selected as the final output:

$$C_{final} = \arg\max_i (E_I(\mathbb{C}''_{\neq i}, C'_i)) \tag{8}$$

III. Experiments

A. Datasets and Experimental Settings

We conduct the experiments on four medical question summarization datasets, i.e., MeQSum [10], CHQ-Summ [27], iCliniq [28], and HealthCareMagic [28]. Among them, the iCliniq and HealthCareMagic datasets are the modified fair version by Wei et al [1]. Table I displays the statistics of the four datasets.

TABLE I
Datasets Statistics

Datasets	Examples			Avg. Words	
	Train	Valid	Test	CHQ	FAQ
MeQSum	400	100	500	70	12
CHQ-Summ	800	300	407	176	13
iCliniq	16,556	2,069	2,071	114	13
HealthCareMagic	180,697	22,587	22,588	93	11

We fine-tune BART-large [14] to generate candidates summaries. The learning rate is set to 1e-5, and the batch size is set to 16.

For re-ranking model, we utilize RoBERTa-base [24] as the encoder. The maximum learning rate is set to 2e-3, and the batch size is set to 16. The number of candidate summaries is set to 16. λ in Eq.(3) is set to 0.01 for MeQSum and CHQ-Summ, and 0.001 for iCliniq and HealthCareMagic. μ in Eq.(4) is set to 0.01 for CHQ-Summ, and 0.001 for MeQSum, iCliniq and HealthCareMagic. The loss weight ∂ is set to 0.4 for MeQSum, and 0.7 for CHQ-Summ, iCliniq and HealthCareMagic.

We adopt ROUGE [29] as the evaluation metric, which is used to measure the similarity between automatically generated summaries and reference summaries. R1, R2, and RL metrics represent the F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L, respectively. All experiments are conducted with one NVIDIA A100 40GB GPU.

TABLE II
EXPERIMENTAL RESULTS ON FOUR MQS DATASETS.

Model	MeQSum			CHQ-Summ			iCliniq			HealthCareMagic		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
T5	34.47	17.18	30.54	35.17	18.87	32.33	39.25	20.84	34.92	38.09	18.41	34.99
PEGASUS	43.18	26.15	40.87	35.89	18.86	33.27	36.09	18.30	31.45	35.17	15.75	30.15
LLaMA-Vicuna-13B	43.75	22.64	39.76	36.02	15.95	31.40	34.84	14.10	28.87	33.36	12.17	28.45
PG + Data Augmentation*	44.16	27.64	42.78	-	-	-	-	-	-	-	-	-
ProphetNet	44.40	26.93	41.57	40.46	22.80	38.13	39.13	20.15	34.01	30.34	12.00	26.00
ProphetNet + QTR + QFR*	45.52	27.54	48.19	-	-	-	-	-	-	-	-	-
Data-Augmented Joint Learning*	48.50	29.70	44.90	-	-	-	-	-	-	-	-	-
RQE + MTL + Data Augmentation*	49.20	29.50	44.80	-	-	-	-	-	-	-	-	-
BART	51.79	34.94	49.16	42.40	24.00	39.79	40.66	21.87	35.83	43.64	23.60	40.55
QFCL	51.48	34.16	49.08	42.18	23.48	39.81	40.93	22.07	36.27	43.36	23.39	40.44
ECL	52.85	**36.06**	50.48	43.16	24.26	40.46	41.31	22.27	36.68	43.52	**23.75**	40.56
SimCLS	52.41	33.88	49.90	42.46	23.06	39.58	41.85	22.18	36.97	44.58	23.21	41.24
MvCL (Ours)	**53.56**	34.91	**50.76**	**43.76**	**24.70**	**40.81**	**42.65**	**23.02**	**37.88**	**45.22**	23.58	**41.60**

The results marked with * are taken from the original paper, while all other results are obtained by running the corresponding code ourselves. The best performance is **boldfaced**.

TABLE III
ABLATION STUDY OF MVCL

Model	MeQSum			CHQ-Summ			iCliniq			HealthCareMagic		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
MvCL	**53.56**	**34.91**	**50.76**	**43.76**	**24.70**	**40.81**	**42.65**	**23.02**	**37.88**	**45.22**	**23.58**	**41.60**
w/o $\mathcal{L}_{inner}$	53.14	34.05	49.79	43.32	23.90	40.33	42.31	22.82	37.37	45.05	23.34	41.44
w/o *Multi-level Inference*	52.41	33.88	49.90	42.46	23.06	39.58	41.85	22.18	36.97	44.58	23.21	41.24

w/o $\mathcal{L}_{inner}$ represents removing $\mathcal{L}_{inner}$ and retaining $\mathcal{L}_{outer}$ and multi-level inference. w/o *Multi-level Inference* represents further removing multi-level inference, and the model degrades to the baseline model. The best performance is **boldfaced**.

B. Experimental Results

We compared our proposed MvCL framework with popular and state-of-the-art baseline models to demonstrate its effectiveness on the MQS task. The overall experimental results are shown in Table II. According to the table, we have the following observations.

First, the previous models are almost one-stage models. Although these one-stage models have achieved good results, they overlook the exposure bias issue faced by Seq2Seq neural models. Specifically, they employ beam search to generate a single summary for CHQ, but there may exist summaries with higher evaluation metric scores in the vast search space. This means that summaries generated in this way may not necessarily be of the highest quality in the search space. Therefore, it is necessary to expand the search space of such models and generate more candidate summaries, and then use a re-ranking model to score and re-rank these candidate summaries.

Second, SimCLS is a popular model for open-domain abstractive summarization, and we applied it to the MQS task, finding that it can achieve results comparable to previous state-of-the-art works. This demonstrates the effectiveness of similarity-based re-ranking models. However, we point out that SimCLS still has limitations as it solely considers the relationship between the original medical question and its candidate summaries. To address this issue, we propose MvCL based on SimCLS. MvCL simultaneously considers the similarity scores between medical questions and candidate summaries, as well as the average similarity scores between candidate summaries and other candidates within the same group. Combined with our proposed multi-level inference approach, MvCL outperforms SimCLS comprehensively, mainly because MvCL measures the quality of candidate summaries from multiple views during training and inference.

Finally, MvCL significantly improves the performance of the BART model and achieves new state-of-the-art results on most metrics across the four datasets. This demonstrates the superiority and effectiveness of our proposed MvCL framework. Furthermore, our framework can be flexibly applied to any Seq2Seq neural model, and it can have a greater impact as the performance of the Seq2Seq model becomes stronger.

C. Ablation Study

We conducted ablation experiments to demonstrate the effectiveness of different components in MvCL. The experimental results are shown in Table III. Firstly, we removed $\mathcal{L}_{inner}$ while keeping the multi-level inference. From the experimental results, it can be observed that after removing $\mathcal{L}_{inner}$, all evaluation metrics decreased, which confirms the effectiveness of $\mathcal{L}_{inner}$. Furthermore, we removed the multi-level inference from the model as well, and almost all evaluation metrics decreased again, demonstrating the effectiveness of the proposed multi-level inference method.

Overall, these experimental results indicate that the two key components in MvCL, including $\mathcal{L}_{inner}$ and multi-level inference, are necessary for the outstanding performance of MvCL.

IV. CONCLUSION

In this paper, in order to alleviate the exposure bias problem in medical question summarization (MQS), we first explore the performance of the popular framework SimCLS for open-domain abstractive summarization on MQS and identify one of the key factors limiting its performance. Addressing the issues in SimCLS, we propose a novel re-ranking training framework for MQS called Multi-view Contrastive Learning (MvCL). MvCL simultaneously considers the similarity scores between medical questions and candidate summaries as well as the average similarity scores between candidate summaries and other candidates within the same group, and utilizes contrastive learning to optimize the model's ranking ability. Additionally, we propose a new multi-level inference approach to adapt to our training strategy. Experimental results demonstrate the effectiveness of the proposed MvCL framework. In future work, we will further explore the application of re-ranking framework in MQS to alleviate the problem of exposure bias within it.

Acknowledgment

This work was supported by the National Key R&D Program of China (Grant No.2022YFC3302101) and the National Natural Science Foundation of China (Grant No.62376130), Shandong Provincial Natural Science Foundation (Grant No.ZR2022MF243), Program of New Twenty Policies for Universities of Jinan (Grant No.202333008), Program of Innovation Improvement of Shandong (Grant No.2023TSGC0182).

References

[1] W. Lu, S. Wei, X. Peng, Y.-F. Wang, U. Naseem, and S. Wang, "Medical question summarization with entity-driven contrastive learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2024.

[2] X. Zhang, X. Peng, H. Guan, L. Zhao, X. Qiao, and W. Lu, "Fusion of dynamic hypergraph and clinical event for sequential diagnosis prediction," in *Proceedings of the 29th IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, 2024.

[3] Y. Cheng, Y. Gong, Y. Liu, B. Song, and Q. Zou, "Molecular design in drug discovery: a comprehensive review of deep generative models." *Briefings in Bioinformatics*, vol. 22, no. 6, pp. bbab344:1–11, 2021.

[4] Y. Zhang, W. Lu, W. Ou, G. Zhang, X. Zhang, J. Cheng, and W. Zhang, "Chinese medical question answer selection via hybrid models based on cnn and gru," *Multimedia Tools and Applications*, vol. 79, pp. 14 751–14 776, 2020.

[5] M. Zhang, S. Dou, Z. Wang, and Y. Wu, "Focus-Driven contrastive learning for medical question summarization," in *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 2022, pp. 6176–6186.

[6] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–23, 2019.

[7] K. Mrini, F. Dernoncourt, W. Chang, E. Farcas, and N. Nakashole, "Joint summarization-entailment optimization for consumer health question understanding," in *Proceedings of the 2nd Workshop on Natural Language Processing for Medical Conversations (NLPMC)*, 2021, pp. 58–65.

[8] K. Mrini, F. Dernoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole, "A gradually soft multi-task and data-augmented approach to medical question understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 1505–1515.

[9] C. Lei, V. Efthymiou, R. Geis, and F. Özcan, "Expanding query answers on medical knowledge bases," in *Proceedings of the 2020 International Conference on Extending Database Technology*, 2020, pp. 567–578.

[10] A. B. Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2228–2234.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[12] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "ProphetNet: Predicting future n-gram for sequence-to-sequencepre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2401–2410.

[13] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 11 328–11 339.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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 (ACL)*, 2020, pp. 7871–7880.

[15] M. Sänger, L. Weber, and U. Leser, "WBI at MEDIQA 2021: Summarizing consumer health questions with generative transformers," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 86–95.

[16] Y. He, M. Chen, and S. Huang, "damo_nlp at MEDIQA 2021: Knowledge-based preprocessing and coverage-oriented reranking for medical question summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 112–118.

[17] K. Mrini, F. Dernoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole, "UCSD-Adobe at MEDIQA 2021: Transfer learning and answer sentence selection for medical summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 257–262.

[18] S. Yadav, M. Sarrouti, and D. Gupta, "NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 291–301.

[19] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Reinforcement learning for abstractive question summarization with question-aware semantic rewards," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 249–255.

[20] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 1065–1072.

[21] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing order to abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 2890–2903.

[22] J. Sul and Y. S. Choi, "Balancing lexical and semantic quality in abstractive summarization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[23] J. Xie, Q. Su, S. Zhang, and X. Zhang, "Alleviating exposure bias via multi-level contrastive learning and deviation simulation in abstractive summarization," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 9732–9747.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X.-J. Huang, "Extractive summarization as text matching," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 6197–6208.

[26] Y. Seonwoo, G. Wang, C. Seo, S. Choudhary, J. Li, X. Li, P. Xu, S. Park, and A. Oh, "Ranking-enhanced unsupervised sentence representation learning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[27] S. Yadav, D. Gupta, and D. Demner-Fushman, "CHQ-Summ: A dataset for consumer healthcare question summarization," *arXiv preprint arXiv:2206.06581*, 2022.

[28] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang *et al.*, "MedDialog: Large-scale medical dialogue datasets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9241–9250.

[29] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.