CONTRASTIVE LEARNING WITH HIGH-QUALITY AND LOW-QUALITY AUGMENTED DATA FOR QUERY-FOCUSED SUMMARIZATION

Shaoyao Huang[1], *Ziqiang Cao*[1*], *Luozheng Qin*[1], *Jun Gao*[1], *and Jun Zhang*[2]

[1] Soochow University, China [2] Changping Laboratory, China

ABSTRACT

Unlike general text summarization, Query-focused summarization (QFS) is severely limited by insufficient datasets, forcing previous research to transform datasets from other tasks into QFS format for data augmentation. However, this approach has resulted in two problems: the task and train-test gaps. To alleviate these gaps, we propose QFS-CL, a novel in-place data augmentation framework equipped with contrastive learning. Firstly, we design diverse prompts for ChatGPT to paraphrase the original QFS data into high-quality/low-quality document-summary pairs, filling the task gap. Then, instead of directly incorporating the augmented data into the training set, we train the QFS baseline model in a contrastive learning scheme. Specifically, our approach encourages the model to imitate high-quality pairs and distinguish itself from low-quality pairs, enabling the model to learn how to acquire reliable information and avoid extracting invalid information. Our method achieves state-of-the-art performance on Debatepedia and DUC datasets in ROUGE scores, GPT-4, and human evaluations.

Index Terms— Query-focused summarization, contrastive learning, data augmentation, ChatGPT

1. INTRODUCTION

Query-focused summarization (QFS) aims to generate summaries that address specific queries by extracting essential information from the source documents [1]. A typical application of QFS is to provide the users with a summary based on their query in search engines [2]. Although general text summarization has drawn significant attention in recent years, research on QFS is limited by the small amount of available data [3].

To alleviate the scarcity of datasets, data augmentation has been widely applied in the QFS task. Researchers introduced datasets from different tasks to augment the QFS data, including Document-based Question Answering [4], machine translation [5], Question Answering [6], Wikipedia [7], and CNN/Daily [3].

Although data augmentation from different tasks has achieved some effects, the task gap and train-test gap limit the upper bound of such methods [8, 9]. Firstly, the task gap refers to discrepancies caused by inconsistent tasks. For example, Wikipedia and CNN/Daily Mail are general text summarization datasets, where the summaries in the dataset are focused on the entire document, while the QFS's summaries are focused on the query. Secondly, the train-test gap arises from inconsistent input between training and testing due to data augmentation. During training, we have both the original and augmented data, where the quality of the augmented data varies. If we directly merge them into a single dataset for training, low-quality data can significantly impair the model's performance, and some exceptionally clean high-quality data may inadvertently undermine the model's robustness.

To alleviate these problems, we propose QFS-CL, a novel in-place data augmentation framework equipped with contrastive learning. Specifically, we design diverse prompts for ChatGPT [10] to paraphrase the original QFS data into high-quality/low-quality document-summary pairs, filling the task gap. In the high-quality pairs, the documents refer to the summary information in order to standardize them to standard English and correct errors. To address the problem of inconsistent input between training and testing, we design a contrastive learning scheme with query-document relevance. This approach encourages the model to emulate high-quality document-summary pairs and distinguish itself from low-quality ones, addressing input inconsistencies between training and testing effectively. We conduct experiments on the Debatepedia [11] and DUC [12] datasets and achieve new state-of-the-art performance.

Our contributions are as follows:

- We propose to utilize ChatGPT to generate high/low-quality document-summary pairs, performing in-place data augmentation to fill the task gap.
- We design a contrastive learning method with query-document relevance to mitigate the train-test gap and enhance model adaption for the QFS task.
- Experiments conducted on the Debatepedia and DUC datasets verify the effectiveness of our method, achieving state-of-the-art performance in ROUGE scores, GPT-4, and human evaluations.

*Corresponding Author.

Table 1: Prompt templates are used to generate two types of documents. The [SUMMARY] placeholder refers to the gold summary in the QFS datasets, while the [DOCUMENT] placeholder refers to the source document in the QFS datasets.

Document type	prompts template
High-Quality Document	Gold Summary: [SUMMARY]. Referring to the Gold Summary, normalize the following document to standard English and correct errors, keeping the original word order: [DOCUMENT].
Low-Quality Document	Gold Summary: [SUMMARY]. Referring to the Gold Summary, please rephrase the following document while masking certain keywords: [DOCUMENT].

2. DATA AUGMENTATION WITH CHATGPT

To mitigate the task gap, we utilize ChatGPT to perform data augmentation on the original QFS documents. Notably, a query typically consists of one or two concise sentences, which indicates that any modification may significantly alter its meaning. To avoid this issue, we only focus on augmenting the data of the source documents and do not modify the queries in any way.

To alleviate the train-test gap, we employ contrastive learning. Specifically, we design prompts to generate high/low-quality document-summary pairs, making the contrastive learning samples more distinguishable. The prompt templates are shown in Table 1. To generate a more credible high-quality document-summary pair, we design a prompt that allows the document to refer to summary knowledge. Conversely, we create a prompt that reduces the document's keywords, resulting in a low-quality document-summary pair.

3. SUMMARIZAITON MODEL

To fill the train-test gap, we design a contrastive learning strategy to utilize the augmented data, instead of directly adding them to the training set. In this section, we present our contrastive learning method between the augmented data and the original data. Firstly, we introduce the baseline model. Next, we describe our contrastive learning method, which incorporates query-document relevance to enhance model adaption for the QFS task. The overall framework of contrastive learning is shown in Figure 1.

3.1. Baseline Model

Our data augmentation framework is generic and can be applied to any general summarization model. In this paper, we utilize a popular seq2seq model, BART-LARGE [13], as our base model. BART-LARGE is highly effective for general text summarization.

3.2. Contrastive Learning with Query-Document Relevance

To standardize the input format for BART, we concatenate a query and source document as input, denoted as X_s. This enables query and document information to interact with each other in the self-attention layer of the BART encoder. The hidden state H_s is obtained from BART's encoder in the following way:

$$H_s = Encoder(X_s) \tag{1}$$

Before conducting contrastive learning, we obtain logits corresponding to three types of documents. During training, we use the BART's decoder to obtain logits O_s:

$$O_s = Decoder(H_s,\ Y) \tag{2}$$

where Y refers to the gold summary. Similarly, we use high-quality and low-quality documents to obtain logits O_h and O_l.

We incorporate query-document relevance into contrastive learning to enhance model adaption for the QFS task. Dense Passage Retrieval (DPR) [14] is a widely adopted model for question-answering retrieval. It can effectively retrieve the most relevant document from a large corpus in relation to a search query. We use the DPR model [1] to calculate the relevance score between the query and the document. This is accomplished through the following equation:

$$DPR_{Sim}\,(D, Q) = \frac{\vec{D} \cdot \vec{Q}}{||\vec{D}|| \cdot\ ||\vec{Q}||} \tag{3}$$

where $\vec{D}$ is the embedding vector of the input document, and $\vec{Q}$ represents the embedding vector of the input query.

We calculate the relevance scores R_s between the source document and query, R_h between the high-quality document and query, and R_l between the low-quality document and query. After scaling, the values of R_s, R_h, and R_l fall within the range of (0, 1).

We drew inspiration from the work of [15] for our contrastive learning approach. They first select clean and noisy subsets from the original training set and train expert and anti-expert models. Then, they steered the base model towards the expert model and away from the anti-expert model. In contrast, we no longer extract from the original dataset but perform data augmentation and only use the base model during testing. In this case, the contrastive learning formula we designed to calculate the comprehensive logits $O_{s\text{-}h}$/$O_{s\text{-}l}$ between O_s and O_h/O_l as follows:

$$O_{s\text{-}h} = (1 - \frac{2 \cdot \theta \cdot R_h}{R_s + R_h}) \cdot O_s + \frac{2 \cdot \theta \cdot R_h}{R_s + R_h} \cdot O_h \tag{4}$$

$$O_{s\text{-}l} = (1 + \frac{2 \cdot \theta \cdot R_l}{R_s + R_l}) \cdot O_s - \frac{2 \cdot \theta \cdot R_l}{R_s + R_l} \cdot O_l \tag{5}$$

where θ represents the temperature coefficient, the initial value of θ is 0.3. Since no data augmentation is performed

[1] https://huggingface.co/facebook/dpr-ctx_encoder-multiset-base

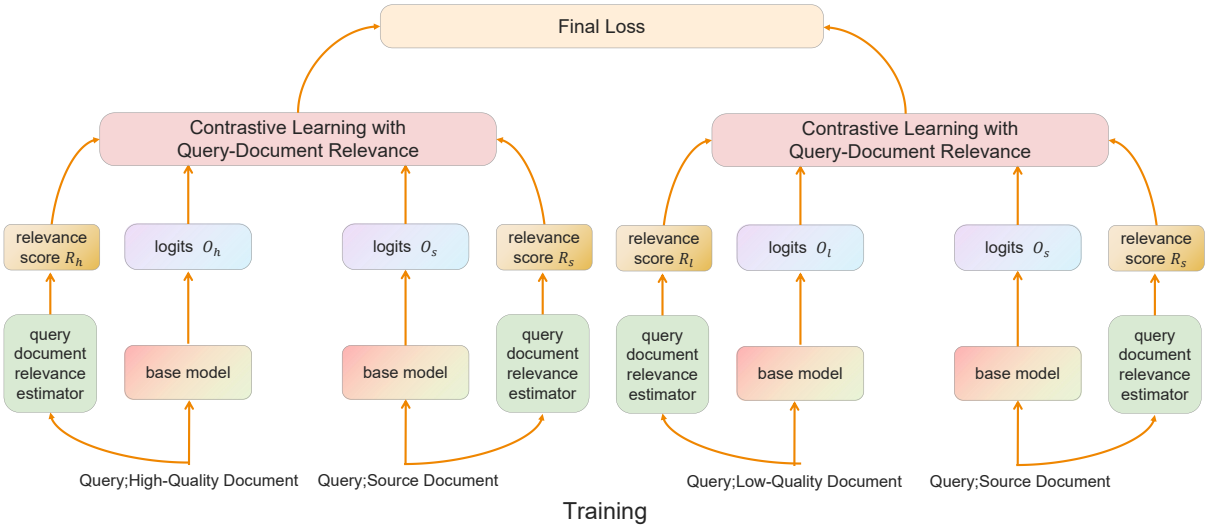

Fig. 1: The overall framework of contrastive learning. During training, we perform data augmentation and input three types of documents for contrastive learning. During testing, we only input the source document and use the base model.

during the testing process, gradually decreasing θ over time is necessary. This ensures that training data distribution in the final batch is consistent with the testing process, ultimately reducing the gap caused by the inconsistency between training and testing.

The query-document relevance indirectly reflects the characteristics of the QFS data, and strengthening the weight of high-relevance samples can enhance model adaption for the QFS task. Based on this, we calculate the loss between $O_{s\text{-}h}$/$O_{s\text{-}l}$ and the gold summary Y using the Cross-Entropy Loss function and multiply with query-document relevance score, denoted as $\mathcal{L}_{s\text{-}h}$/$\mathcal{L}_{s\text{-}l}$:

$$\mathcal{L}_{s\text{-}h} = (\frac{R_s + R_h}{2} + \xi) \cdot Cross - Entropy(Y, O_{s\text{-}h}) \tag{6}$$

$$\mathcal{L}_{s\text{-}l} = (\frac{R_s + R_l}{2} + \xi) \cdot Cross - Entropy(Y, O_{s\text{-}l}) \tag{7}$$

where ξ is a hyperparameter set to 0.5 to prevent ineffective learning when R_s, R_h and R_l approach 0.

Finally, since $O_{s\text{-}h}$ and $O_{s\text{-}l}$ already incorporate the logits of the source document, we no longer calculate the loss for the source document separately, and the final loss function is as follows:

$$\mathcal{L}_{Final} = \mathcal{L}_{s\text{-}h} + \mathcal{L}_{s\text{-}l} \tag{8}$$

In contrast, we do not perform data augmentation during the testing process. There are two main reasons for this: Firstly, the gold summaries are not visible [16], preventing us from applying data augmentation to all test data in advance. Secondly, data augmentation is slow, making real-time augmentation difficult to accomplish.

Table 2: ROUGE-F1 scores for Debatepedia and DUC 2007 dataset.

Datasets	Models	ROUGE-1	ROUGE-2	ROUGE-L
Debatepedia	ChatGPT one-shot	22.7	6.7	18.8
	ChatGPT three-shot	23.2	6.9	19.1
	SD2	41.3	18.8	40.4
	CSA Transformer	46.4	37.4	45.9
	QR-BERTSUM-TL	58.0	45.2	57.1
	QFS-BART	59.0	44.6	57.4
	PreQFAS	59.3	45.6	58.2
	BART-LARGE	64.6	52.1	63.2
	QSG BART	64.9	52.3	63.3
	QFS-CL	**67.1**	**54.9**	**65.6**
DUC 2007	ChatGPT one-shot	11.2	6.6	10.9
	ChatGPT three-shot	10.9	6.5	10.7
	QMDSIR	32.8	4.2	30.5
	Prefix-Merging	35.0	7.5	24.7
	HEROsumm	38.3	7.6	**35.7**
	BART-LARGE	38.7	14.3	21.6
	QFS-CL	**40.2**	**16.1**	23.5

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

We mainly evaluate the proposed approach for the QFS task in the single-document Debatepedia [11] and the multi-document DUC [1] datasets.

We use ROUGE [17] as fundamental evaluation metrics. Additionally, we conduct GPT-4 and human evaluation to assess the quality of the summaries based on four aspects: fluency, faithfulness, coverage, and overall performance.

4.2. Results

ROUGE Metrics: The experimental results for Debatepedia and DUC can be found in Table 2, with the highest scores marked in bold. Compared with the original baseline model, the proposed QFS-CL model has shown a signif-

Table 3: The evaluation of summary quality using GPT-4 is based on four metrics. We feed the query, document, and shuffled summaries as inputs to GPT-4 to determine which summary performs the best. Subsequently, we rank the number of summaries assessed as the best for each specific metric.

Summary Sources	Rank			
	fluency	**faithfulness**	**coverage**	**overall**
QFS-BART	3	3	3	3
QFS-CL	2	1	2	2
Gold	1	2	1	1

Table 4: Human evaluation of Debatepedia dataset. We shuffled the summaries and asked five text summarization researchers to score the summaries (maximum score of 5 points).

summary sources	fluency	faithfulness	coverage	overall
QFS-BART	4.23	3.50	3.67	3.59
BART-LARGE	4.35	3.61	3.46	3.64
QFS-CL	**4.48**	**4.03**	**3.85**	**3.92**

icant improvement in performance. Moreover, our model outperforms the comparison models on most metrics for both datasets, indicating the effectiveness of the QFS-CL model. It is worth mentioning that ChatGPT tends to generate excessively long summaries, which can result in poorer performance on ROUGE scores.

GPT-4 Evaluation: The results are presented in Table 3. The summaries generated by QFS-CL outperform QFS-BART in terms of all metrics. Our model even achieves higher faithfulness evaluation scores than gold summaries, which may be because the authors of the gold summaries were not meticulous enough. In contrast, our approach effectively avoids such issues by learning from high-quality document-summary pairs.

Human Evaluation: The results show in Table 4, our model outperforms the comparison methods across all metrics. We successfully enhanced model adaption for the QFS task and our summaries are comprehensible to humans.

4.3. Ablation Study

To conduct a more detailed analysis, we designed a comprehensive ablation study to explore the roles of different components in QFS-CL. The results are presented in Table 5. Using low-quality or high-quality documents directly as the training set does not produce satisfactory results. However, we can achieve satisfactory results by exclusively combining low-quality or high-quality documents with the original documents for contrastive learning. Moreover, the performance can be significantly enhanced when all three types of documents are used simultaneously.

4.4. Case Study

We present a case study comparing the QFS-BART summary, the QFS-CL summary, and the gold summary, shown in Table 6. Compared with the QFS-BART and gold summaries,

Table 5: Ablation Study for Debatepedia Dataset. **Input Document Types** indicates the types of input documents for contrastive learning.

Input Document Types	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Source Documents	64.6	52.1	63.2
Low-Quality Documents	59.7	49.3	58.4
High-Quality Documents	61.1	51.4	60.3
Low-Quality and Source Documents	66.2	53.8	64.6
High-Quality and Source Documents	66.5	54.3	65.1
All Documents	**67.1**	**54.9**	**65.6**

Table 6: An example taken from the Debatepedia test set. highlight indicates words that influence faithfulness. QFS-CL demonstrates superior faithfulness compared to the gold summary and QFS-BART in the GPT-4 evaluation.

Document: by granting amnesties congress has set a dangerous precedent that threatens homeland security. our normal immigration process involves screening to block potential criminals and terrorists from entering the united states. yet millions of illegal aliens have avoided this screening and an amnesty would allow them to permanently bypass such screening. **Query**: would it benefit national security ?
Gold Summary: colorado alliance for immigration reform cair.
QFS-CL Summary: amnesties sets a dangerous precedent that undermines homeland security.
QFS-BART Summary:illegal alien amnesty undermines homeland security controls.

QFS-CL is more faithful to the source document by finding and copying spans that can answer the query. QFS-BART also copies keywords from the source document, but the distant relationship between them reduces the relevance and leads to a drop in accuracy. On the other hand, the gold summary writers may have overlooked important details, resulting in the summary being unrelated to the source document and the query.

5. CONCLUSION

To address the gaps caused by previous QFS data augmentation methods, we propose QFS-CL, a novel in-place data augmentation framework equipped with contrastive learning. Firstly, we design diverse prompts for ChatGPT to paraphrase the original QFS documents into different qualities, filling the task gap. Then, we employ contrastive learning to alleviate the gap caused by the inconsistency between training and testing. Finally, we use query-document relevance to adjust the learning rate according to the relevance between queries and different types of documents, enhancing model adaption for the QFS task. The experimental results show that our model achieves state-of-the-art performance on Debatepedia and DUC datasets in ROUGE scores, GPT-4, and human evaluations.

Acknowledgements: The work described in this paper was supported by the National Natural Science Foundation of China (NSFC 62106165) and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

6. REFERENCES

[1] Hoa Trang Dang, "Overview of duc 2005 (draft)," in *Proceedings of Document Understanding Conferences*, 2005.

[2] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung, "Caire-covid: A question answering and multi-document summarization system for covid-19 research," *arXiv preprint arXiv:2005.03975*, 2020.

[3] Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang, "Domain adaptation with pre-trained transformers for query-focused abstractive text summarization," *Computational Linguistics*, vol. 48, no. 2, pp. 279–320, June 2022.

[4] Weikang Li, Xingxing Zhang, Yunfang Wu, Furu Wei, and Ming Zhou, "Document-based question answering improves query-focused multi-document summarization," in *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, Berlin, Heidelberg, 2019, p. 41–52, Springer-Verlag.

[5] Elozino Egonmwan, Vittorio Castelli, and Md Arafat Sultan, "Cross-task knowledge transfer for query-based text summarization," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Hong Kong, China, Nov. 2019, pp. 72–77, Association for Computational Linguistics.

[6] Dan Su, Tiezheng Yu, and Pascale Fung, "Improve query focused abstractive summarization by incorporating answer relevance," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 3124–3131, Association for Computational Linguistics.

[7] Haichao Zhu, Li Dong, Furu Wei, Bing Qin, and Ting Liu, "Transforming wikipedia into augmented data for query-focused summarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2357–2367, 2022.

[8] Boyu Wang, Jorge Mendez, Mingbo Cai, and Eric Eaton, "Transfer learning via minimizing the performance gap between domains," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[9] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman, "Learning bounds for domain adaptation," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. 2007, vol. 20, Curran Associates, Inc.

[10] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[11] Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran, "Diversity driven attention model for query-based abstractive summarization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017, pp. 1063–1072, Association for Computational Linguistics.

[12] Hoa Trang Dang, "Duc 2005: Evaluation of question-focused summarization systems," in *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, 2006, pp. 48–55.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[14] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 6769–6781, Association for Computational Linguistics.

[15] Prafulla Kumar Choubey, Alexander R Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Fatema Rajani, "Cape: Contrastive parameter ensembling for reducing hallucination in abstractive summarization," *arXiv e-prints*, pp. arXiv–2110, 2021.

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning, "A large annotated corpus for learning natural language inference," *CoRR*, vol. abs/1508.05326, 2015.

[17] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.