	CAsT-19		CAsT-20	
	MRR	NDCG@3	MRR	NDCG@3
Transformer++	69.6	**44.1**	29.6	18.5
Query Rewriter	66.5	40.9	37.5	25.5
CQE-Sparse	67.1	39.9	42.3	27.1
QuReTeC	68.9	43.0	43.0	28.7
T5QR	70.1	41.7	42.3	29.9
ConvGQR	**70.8**‡	43.4	**46.5**‡	**33.1**‡
Human-Rewritten	74.0	46.1	59.1	42.2

Table 3: Zero-shot dense retrieval performance of different query reformulation methods. ‡ denotes significant improvements with t-test at $p < 0.05$ over all compared methods. **Bold** indicates the best result (except Human-Rewritten).

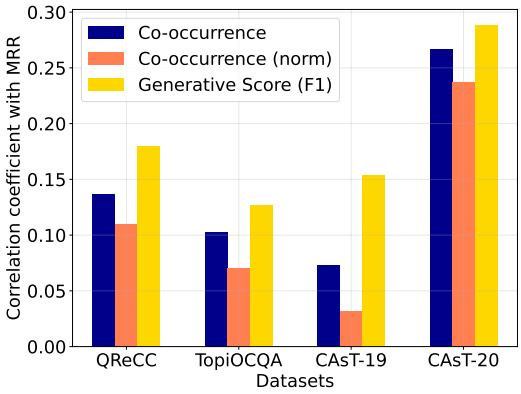

Figure 3: Pearson correlation coefficient (PCC) between three generative evaluation metrics with MRR scores.

4.3 Zero-Shot Analysis

The zero-shot evaluation is conducted on CAsT datasets to test the transferability of ConvGQR. By comparing with the other strongest QR methods in Table 3, we have the following main findings.

The ConvGQR outperforms all the other methods on the more difficult dataset CAsT-20 and matches the best results on CAsT-19, which demonstrates its strong transferability to new datasets. The human-rewritten queries in CAsT datasets achieve the highest retrieval scores, because they have been formulated carefully by experts for search. This observation is different from the results of QReCC in Table 1, for which query rewriting has been done by crowd-sourcing. However, this observation should not lead to the conclusion that human-rewritten queries should be used as the gold standard for the training of query rewriting, because it is difficult to obtain a large number of high-quality human-rewritten queries as in the CAsT datasets. As one can see in Table 7, these datasets only contain a very limited number of queries. Therefore, the generated expansion terms based on the knowledge captured in PLM is still a valuable means to obtain superior performance for new queries.

In addition, combining Table 1 and Table 3, we notice that the effectiveness of ConvGQR for dense retrieval varies with datasets. A potential reason is the different degrees of co-occurrence of generated expansion terms within their relevant passages. This will be further analyzed in Section 4.4.

4.4 Impact of Generated Answer for Retrieval

The aforementioned hypothesis of the ConvGQR for query expansion is that the PLM-generated potential answers might contain useful expansion terms that co-occur with the right answer in the relevant passages. To see how expansion terms are related to retrieval performance, we use three metrics to analyze their correlation with the retrieval score.

Correlation Analysis Specifically, for each rewritten query with expansion terms, we first calculate the token overlaps between the generated answers and the relevant passages, which can measure their co-occurrence. However, the potential problem is that the generated answers or relevant passages are of variable lengths. Therefore, we further normalize it by the length of its corresponding relevant passage. Besides, we compute the F1 scores between the generated answers and the gold answers to explore if the generation quality has an impact on retrieval effectiveness. Finally, we calculate the Pearson Correlation Coefficient (PCC) for all these three generative evaluation metrics with the respective MRR scores of every reformulated query.

The results are shown in Fig. 3. The relative PCC value can reflect the helpfulness of generated answers for different datasets to some extent. For example, the PCC of QReCC and CAsT-20 are higher than TopiOCQA and CAsT-19, suggesting that the potential answers are more useful in the first datasets. This is consistent with our previous experimental observations that QReCC and CAsT-20 have larger improvements by ConvGQR compared to TopiOCQA and CAsT-19. Thus, the co-occurrence between generated answer and the relevant passage is crucial for the retrieval effectiveness for ConvGQR.

The PCC of generative score F1 is the highest

	Query Form	QReCC				TopiOCQA			
		MRR	NDCG@3	R@10	R@100	MRR	NDCG@3	R@10	R@100
Dense	Rewritten Query	36.4	33.5	56.6	76.0	23.4	22.5	39.8	56.2
	Generated Answer	33.4	30.6	51.9	70.4	3.7	3.2	6.9	14.4
	Concatenation	**41.5**	**38.7**	**63.7**	**81.4**	**25.0**	**23.7**	**42.3**	**57.9**
Sparse	Rewritten Query	33.8	30.6	54.3	86.7	11.3	9.8	22.1	**44.7**
	Generated Answer	33.7	31.3	49.2	69.6	2.0	1.7	3.9	9.6
	Concatenation	**43.4**	**40.6**	**63.8**	**88.1**	**11.6**	**10.2**	**22.5**	42.8

Table 4: Performance of both dense and sparse retrieval on different reformulated query forms.

among the three metrics, which indicates its strong correlation with retrieval effectiveness. However, utilizing generated answers alone as search queries could produce false positive results as we will demonstrate in the subsequent analysis. As a result, it may not reflect the genuine correlation strength in comparison to the co-occurrence metric.

Effects of Different Generated Forms We show the performance of using three different forms of generated queries, i.e. the rewritten query, the generated answer, and the concatenation of them, as the reformulated query for retrieval in Table 4. We find that using the concatenation of both significantly outperforms the two other forms alone, indicating that these two forms can complement each other to achieve better retrieval performance, which confirms again our initial hypothesis. Besides, we find that using the rewritten query alone performs better than using the generated answer, especially on TopiOCQA. The potential reason is the different forms of answers in the datasets: QReCC is more related to factoid questions than TopiOCQA. The correct answer with non-factoid question type is more difficult for a PLM to directly generate. So, the generated answers may be of less utility.

4.5 Impact of Knowledge Infusion Loss

We conduct an analysis of the impact of two knowledge infusion loss functions trying to approach the query representation to that of the relevant passage: contrastive learning (CL) loss and mean square error (MSE) loss. They correspond to Eq. 6 and Eq. 4. The difference between them is that the MSE loss only considers positive passages $\mathbf{h}_{p_+}$ while the CL loss also considers negative passages $\mathbf{h}_{p_-}$ for model training as follows:

$$\mathcal{L}_{\text{CL}} = -\log \frac{e^{(\mathbf{h}_S \cdot \mathbf{h}_{p_+})}}{e^{(\mathbf{h}_S \cdot \mathbf{h}_{p_+})} + \sum_{\mathbf{P}^-} e^{(\mathbf{h}_S \cdot \mathbf{h}_{p_-})}}. \quad (6)$$

We compare the conversational search results of the reformulated queries training by these two loss functions on QReCC and report the results in Table 5. We can find that the reformulated queries trained by CL loss are slightly worse than those with MSE loss. In most previous literature (Xiong et al., 2020; Karpukhin et al., 2020), the CL loss usually performs better for dense retrieval training, thus we expected similar results. The reason for the opposite result might be as follows: since ConvGQR is mainly a generation task rather than a retrieval task, a positive passage can provide a clear signal to instruct the right direction for the target generation, while the additional negative passages used in CL loss only suggest the wrong directions to avoid. Intuitively, the generation objective has only one correct optimization direction but many wrong directions in the high dimensional latent space. This may make it difficult for the knowledge infusion mechanism to determine the correct direction to follow, resulting in sub-optimal queries. Note that despite the above observation, our method ConvGQR trained with CL loss still outperforms most of the existing baselines.

	Type	MRR	NDCG@3	R@10	R@100
CL	Dense	41.7	38.9	62.8	80.9
MSE	Dense	42.0	39.1	63.5	81.8
CL	Sparse	43.9	40.9	64.0	87.5
MSE	Sparse	44.1	41.0	64.4	88.0

Table 5: Retrieval performance of two knowledge infusion loss functions on QReCC.

4.6 Case Study

We finally show a case in Table 6 to help understand more intuitively the impact of expansion terms on ConvGQR. The model is expected to rewrite the query and generate the potential answer toward the human-rewritten query and the gold answer. Although the model produces the same rewritten query as the human, which solves the anaphora

Context: (QReCC Session 2)
q_1: What are the main breeds of goat?
r_1: Abaza...Zhongwei
q_2: Tell me about boer goats.
r_2: The Boer goat is a breed of goat that was developed ... Their name is derived from the Afrikaans (Dutch) ...
Current Query: q_3: What breed is good for meat?
Human-Rewritten: q_3^*: What breed of goat is good for meat?
ConvGQR Reformulated Query:
$\hat{q_3}$: What breed of goat is good for meat? The Boer goat is a breed of goat that was developed in South Africa in the early 1900s for meat production.
Relevant Passage:
p^*: Here are some notable breeds ... Boer goats were bred in South Africa for meat ... *Before Boer goats became available in the United States in the late 1980s, Spanish goats were the standard meat goat breed ...*
Dense Score: 0.06 (Human-Rewritten) **1.00 (Ours)**
Sparse Score: 0.03 (Human-Rewritten) **0.13 (Ours)**

Table 6: A successful example illustrating the reformulated query by ConvGQR. Rewritten and expanded query are in blue and orange, respectively. The expansion terms and gold answer are underlined and *italicized* in the relevant passage.

problem of "goat" with the context, the query expansion generated by ConvGQR with the knowledge of "Boer goat" can still improve the performance for both dense and sparse retrieval. In this case, even though the generated answer is not a correct answer to the question, there is a strongly similar description (underlined) that co-occurs with the right answer in the relevant passage. This example shows a typical case where the generated answer can be highly useful expansion terms. More cases are provided in Appendix B.

5 Conclusion

In this paper, we present a new conversational query reformulation framework, ConvGQR, which integrates query rewriting and query expansion toward generating more effective search queries through a new knowledge infusion mechanism. Extensive experimental results on four public datasets demonstrate the superior effectiveness of our model for conversational search. We also carried out detailed analyses to understand the effects of each component of ConvGQR on the performance improvements.

Limitations

Our work demonstrates the feasibility of combining query rewriting and query expansion to reformulate a conversational query for passage retrieval. Within our proposed ConvGQR, the rewriting and expansion are based on two PLMs trained with different data, which introduce additional training load and model parameters for storage. Thus, designing an integrated model that can simultaneously generate the query rewrite and the expanded terms would be a promising improvement to our method. Another limitation is that the potential answer acting as expansion terms could be generated from more resources (*e.g.*, pseudo-relevant feedback and knowledge graph) rather than only relying on the generative PLMs. Besides, more alternative methods for knowledge infusion can be tested to connect query reformulation with the search task.

Acknowledgements

This work has been partly supported by the China Scholarship Council, the Institute for AI Industry Research, Tsinghua University and a discovery grant from the Natural Science and Engineering Research Council of Canada.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Haonan Chen, Zhicheng Dou, Yutao Zhu, Zhao Cao, Xiaohua Cheng, and Ji-Rong Wen. 2022. Enhancing user behavior sequence modeling by generative tasks for session search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 180–190.