ConQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning

Anonymous ACL submission

Abstract

Compared to standard retrieval tasks, passage retrieval for conversational question answering (CQA) poses new challenges in understanding the current user question, as each question needs to be interpreted within the dialogue context. Moreover, it can be expensive to re-train well-established retrievers such as search engines that are originally developed for non-conversational queries. To facilitate their use, we develop a query rewriting model ConQRR that rewrites a conversational question in the context into a standalone question. It is trained with a novel reward function to directly optimize towards retrieval using reinforcement learning and can be adapted to any fixed retriever. We show that ConQRR achieves state-of-the-art results on a recent open-domain CQA dataset containing conversations from three different sources, and is effective for two different fixed retrievers. Our extensive analysis also shows the robustness of ConQRR to out-of-domain dialogues as well as to limited query rewriting supervision.

1 Introduction

Conversational question answering (CQA) systems (Reddy et al., 2019; Choi et al., 2018) allow information-seeking users to ask a sequence of questions interactively. In an open-domain setting (Anantha et al., 2021), we often want the answer to be grounded in trustworthy, external evidence. How do we find this evidence? Compared to standard retrieval tasks (Voorhees and Tice, 2000; Nguyen et al., 2016), passage retrieval for CQA poses new challenges in understanding the current user question, as each question needs to be interpreted within the dialogue context.

The task of question-in-context rewriting or query rewriting (QR) in a conversation (Elgohary et al., 2019; Dalton et al., 2020) is to convert a context-dependent question into a self-contained question. It enables the use of a standard retriever

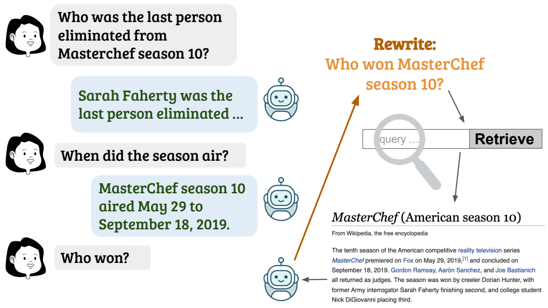

Figure 1: A CQA agent rewrites the current user question into a more effective one (in orange) for the given retriever to find the passage that answers the question.

like BM25 (Robertson and Zaragoza, 2009) or a search engine (Komeili et al., 2021) without fine-tuning it on conversation-specific labeled data, which can be expensive in practice. Therefore, in this paper, we focus on the task of *query rewriting* for *conversational passage retrieval* in a CQA dialogue, with a *fixed* (i.e., not-to-be-fine-tuned) retriever. We seek to build a QR model that rewrites a user query into the retriever's input, in such a way that optimizes for passage retrieval performance. For example, in Figure 1, the agent rewrites the current user query "Who won?" into "Who won MasterChef season 10?", in order to have the retriever retrieve the best answer passage for the question.

Recent work that leverages QR for conversational passage retrieval (Anantha et al., 2021; Dalton et al., 2020) collects human-rewritten queries to train a supervised QR model. However, humans usually rewrite conversational queries to be unambiguous to a human outside the dialogue context, but not necessarily to optimize the retrieval performance. We conduct comprehensive experiments in Section 4.5 to confirm these human rewrites indeed sometimes omit information from the dialogue context that is useful to the retrieval system. This limitation of human query rewrites impacts supervised training. In addition, prior supervised QR models are agnostic to downstream retrievers as they are separately trained before their predicted rewrites being used for retrieval at inference.

To overcome the shortcomings of prior work, We design a reinforcement learning (RL)-based model CONQRR (**Con**versational **Q**uery **R**ewriting for **R**etrieval) that directly optimizes the rewritten query towards retrieval performance, using only weak retrieval supervision. As performing retrieval to calculate the reward for every training step can be time-consuming, we adopt a novel reward function that computes an approximate but effective retrieval performance metric on in-batch passages. Our reward function does not assume any specific retriever model design, and is generic enough for CONQRR to adapt to any fixed retriever.

We show CONQRR outperforms supervised QR models on a recent and the first large-scale open-domain CQA dataset QReCC (Anantha et al., 2021) by over 12% and 14% for BM25 and a neural dual encoder retriever model (Ni et al., 2021) trained on the standard MSMARCO retrieval dataset (Nguyen et al., 2016) respectively, averaging over three retrieval metrics. We observe the performance boost on all three QReCC subsets from different conversation sources, including one that only appears in the test set (i.e., out-of-domain). CONQRR also demonstrates robustness to limited QR labels, topic shifts and longer dialogue contexts, compared to the supervised model.

To conclude, our contributions are as follows. 1) We conduct a novel quantitative study to analyze both the limitations and utility of human rewrites, as well as the importance of QR for conversational passage retrieval in a CQA dialogue, which are largely under-explored in prior work. 2) We introduce a RL-based model CONQRR for the task of QR for conversational retrieval, that can optimize towards and adapt to any fixed retriever using a novel reward function. 3) We demonstrate that CONQRR achieves state-of-the-art results on the public dataset QReCC with conversations from three sources, and is effective for two retrievers including BM25 and a dual encoder model. 4) Our analysis shows CONQRR is robust to out-of-domain dialogues, topic shifts, longer dialogue contexts and limited QR labels.

2 Related Work

Conversational Question Answering (CQA) Most existing CQA datasets (Choi et al., 2018; Reddy et al., 2019) are designed for the task of reading a document to answer questions in a conversation, which does not require the retrieval step. In contrast, QReCC (Anantha et al., 2021) is a recent open-domain CQA dataset where a conversational agent retrieves the most relevant passage(s) before generating an answer to the question.

Conversational Retrieval A few recent works (Dalton et al., 2020; Qu et al., 2020) collect retrieval datasets for conversational search tasks (Belkin et al., 1995; Solomon, 1997) which usually do not have answer utterances in a conversation. Dalton et al. (2020) annotate 80 conversations for the TREC CAsT-19 task and Qu et al. (2020) derive their dataset based on QuAC by removing all answer turns and propose to fine-tune a dual encoder retriever (Guu et al., 2020; Karpukhin et al., 2020).

For such conversational search tasks, Yu et al. (2020) propose a supervised QR model trained with a large number of weak QR supervisions from additional non-conversational data resources. Kumar and Callan (2020) develop a retrieval framework that focuses on the passage re-ranker instead of the first-step retrieval model. Yu et al. (2021) propose a framework to adapt dual encoder retrievers to conversational queries by training a separate query encoder. In contrast, QReCC (Anantha et al., 2021) is a large-scale open-domain CQA dataset, with each conversation containing both user and agent utterances, and also fits the focus of our work. The authors use a supervised QR model based on GPT2 (Radford et al., 2019) followed by a BM25 retriever for the retrieval task, while we show the limitations of human rewrites used as QR supervision and design a RL-based QR model. Conversational retrieval is also leveraged as an intermediate component in some social chat agents to address factual hallucination and user engagement (Shuster et al., 2021; Komeili et al., 2021).

Query Rewriting (QR) Conversational QR is initially proposed to help a model understand the dialogue context (Elgohary et al., 2019), and gets recently adopted for downstream tasks like conversational retrieval and question answering (Anantha et al., 2021; Dalton et al., 2020; Yu et al., 2020). There are also studies in IR research on query reformulation or suggestion that consider non-conversational queries only (Chen et al., 2018; Ahmad et al., 2019; Das et al., 2019).

RL for Text Generation Prior work applies RL approaches to address text generation tasks like machine translation (Ranzato et al., 2016; Wu et al., 2016), text summarization (Paulus et al., 2018; Celikyilmaz et al., 2018) and image captioning (Rennie et al., 2017; Fisch et al., 2020) by training a model directly optimized towards generation quality metrics like BLEU, ROUGE or CIDEr. Buck et al. (2018) use RL to train a QA model that reformulates a non-conversational query into multiple different inputs to a fixed QA system and aggregate returned results to be the final answer. Nogueira and Cho (2017) apply RL based on gold passage labels to reformulate non-conversational user queries in order to effectively improve the downstream retrieval task. Adolphs et al. (2021) apply RL with a restricted action space using multiple rounds of query reformulation and retrieval to respond to a non-conversational query. In contrast, we focus on more challenging conversational queries, and only use weak supervision for the downstream task passage retrieval and an approximate retrieval metric for computational efficiency.

3 Approach

Problem Definition In this work, we focus on the task of *query rewriting (QR)* for *conversational passage retrieval* in a CQA dialogue, with a *fixed* retriever. The inputs to this task include a dialogue context x consisting of a sequence of previous utterances $(u_1, u_2, \ldots, u_{n-1})$ and the current user question u_n, a passage corpus P and a fixed retriever R. R returns a ranked list of top-k passages when given a query string and a passage corpus. The task aims to rewrite x into a query q such that R can take q as the input query to retrieve passages relevant to x from P. Specifically, a passage p is relevant to x if p provides enough information to answer u_n in the context of $(u_1, u_2, \ldots, u_{n-1})$.

In this section, we first introduce a T5-based QR model (T5QR) that applies a generic Seq2Seq training objective with QR labels (Section 3.1). Then we introduce our RL-based framework CONQRR (**Con**versational **Q**uery **R**ewriting for **R**etrieval) that trains a QR model to optimize towards retrieval and is adaptable to any given retriever, with weak retrieval supervision (Section 3.2).

3.1 T5QR

T5 is an encoder-decoder model that is pretrained on large textual corpora (Raffel et al., 2020). We fine-tune T5 to rewrite a conversational query with the input as the concatenation of utterances in the dialogue context x and the output as the human rewrite $\hat{q}$. Note that we concatenate the utterances in a reversed order such that u_n becomes the first one in the input string and any truncation impacts more distant context. Utterances are separated with a seperator token "[SEP]" in the concatenated string. The model is then trained with a standard cross entropy (CE) loss to maximize the likelihood of generating $\hat{q}$, which is a self-contained version of the query u_n that can be interpreted without knowing previous turns $(u_1, u_2, \ldots, u_{n-1})$ in x.

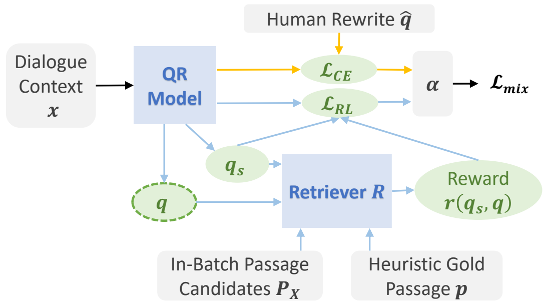

Figure 2: Our CONQRR framework. Yellow and blue arrows mark the flow of CE and RL loss calculation respectively. During inference, only q (with the dashed border) is generated as the final rewrite.

3.2 CONQRR

QR models trained with a standard CE loss are agnostic to the retriever. In addition, human rewrites are not necessarily the most effective ones for passage retrieval (See Section 4.5 for an exploration).

This motivates us to design our RL-based framework CONQRR (Figure 2) that trains a QR model directly optimized for the retrieval performance and can be adapted to any given fixed retriever.

To be comparable with supervised QR models that do not use gold passages in training, we first describe how we obtain weak retrieval supervision for RL reward calculation in CONQRR. Then we introduce the RL training details of CONQRR.

Weak Retrieval Supervision In a CQA dialogue, each question naturally comes with an answer in its following conversational utterance. For each x, we mark its weak passage label p as the one having a string span with the highest token overlap F1-score with the following answer string u_{n+1}:

$$p = \arg\max_{p' \in P} \left[\arg\max_{s \in p'} sim(s, u_{n+1}) \right] \quad (1)$$

where s is a string span and $sim()$ calculates the token overlap score between two strings.[1] Tokens

[1]If multiple passages have the highest score, we randomly choose one.

are lower-cased from the NLTK tokenizer.[2] However, as searching within all candidates in P is very time-consuming, we instead first use BM25 to retrieve the top 100 passages from P with the BM25 input being the human rewrite, and then locate the best passage p from these 100 candidates.

RL Training CONQRR also has T5 as the base model architecture. Following prior work on RL for text generation (Paulus et al., 2018; Fisch et al., 2020), we first initialize it with a supervised model (T5QR) as a warm-up.[3]

For each training example with the dialogue context x, we use the concatenated utterances in x as the model input. For each input, we generate m sampled rewritten queries $(q_{s_1}, \ldots, q_{s_m})$ as well as a baseline generated rewrite q. To generate each sampled rewrite q_s, at time step t of the decoding process, a token q_s^t is drawn from the decoder probability distribution $Pr(w|x, q_s^{1:t-1})$ The baseline rewrite q is the output of greedy decoding, which is also applied for query rewriting during inference. We then apply a self-critical sequence training algorithm (Rennie et al., 2017) to calculate the reward for each q_s relative to q as $r(q_s, q) = score(q_s) - score(q)$. Ideally, the $score()$ function should be some retrieval evaluation metric like mean reciprocal rank (MRR) or Recall@K. However, as it is very costly to run actual retrieval for each training step, we instead use an approximate scoring function described below.

To compute $score(q)$ for a rewrite q, we first use q to do retrieval from the in-batch passage candidates P_X defined as follows, instead of from the full passage corpus P. We pre-compute one positive and one negative passage (p and p_n) for each training example x where p_n is a randomly selected passage that is different from p, 50% of the time from the top 100 BM25-retrieved candidates (with the BM25 input being the human rewrite) and remaining 50% of the time from P. We define the set of all such positive and negative passages of input examples in a batch X as the in-batch passage candidates P_X. Formally, we define $P_X = \{p^i, p_n^i | x_i \in X\}$ as the set of in-batch passage candidates for the batch X. Then for a generated rewritten query q of $x \in X$, we calculate $score(q)$ as a binary indicator of whether the retriever R ranks the assigned positive passage p highest from P_X. We denote $R(q, P_X, k)$ as the k-th most relevant passage retrieved by R from the candidate pool P_X, and define:

$$score(q) = \mathbb{1}\left[R(q, P_X, 1) = p\right] \qquad (2)$$

Then the RL training loss for x becomes:

$$\mathcal{L}_{RL} = -\frac{1}{m}\sum_{i=1}^{m} r(q_{s_i}, q) \log Pr(q_{s_i}|x)$$

$$Pr(q_{s_i}|x) = \prod_{t=1}^{|q_{s_i}|} Pr(q_{s_i}^t | x, q_{s_i}^{1:t-1})$$

Following prior work (Paulus et al., 2018; Celikyilmaz et al., 2018), we experiment with both a pure RL loss ($\mathcal{L}_{RL}$) and a mixed RL and CE training loss:

$$\mathcal{L}_{mix} = \alpha\mathcal{L}_{RL} + (1-\alpha)\mathcal{L}_{CE} \qquad (3)$$

where $\alpha \in [0, 1]$ is a tunable parameter.

3.3 Retriever Models

We evaluate the effectiveness of CONQRR in experiments with two retrieval systems.

BM25 We follow Anantha et al. (2021) using Pyserini (Yang et al., 2017) with the default parameters $k_1 = 0.82$ and $b = 0.68$. These values were chosen based on retrieval performance on MS MARCO (Nguyen et al., 2016), which contains non-conversational queries only. During the RL training of CONQRR, due to the complexity of applying Pyserini to calculate rewards on-the-fly, we instead use a Pyserini approximate called BM25-light. The only differences between them are that BM25-light (1) uses T5's subword tokenization instead of whole word tokenization and (2) does not use special operations (e.g., stemming) as applied in Pyserini. After training, we still run inference and report retrieval performance on BM25.

Dual Encoder (DE) We use a shared T5-base query and passage encoder. For each query and passage pair, their relevance is decided by the dot product similarity between their encodings. The architecture is the same as the recent DE model (Ni et al., 2021). We use a model fine-tuned on MS MARCO, and keep it fixed for our experiments.

3.4 Inference

At inference time, both T5QR and CONQRR work in the same way. The trained QR model is used to greedily generate the rewritten query given a dialogue context. Then, the predicted rewrite is given to the provided retriever to perform retrieval.

[2] https://www.nltk.org

[3] In Section 4.5, we show that although initializing with T5QR works better than T5, both setups generally work well.

4 Experiment

4.1 Dataset and Evaluation Metrics

Dataset QReCC (Anantha et al., 2021) is a dataset of 14k open-domain English conversations in the format of alternating user questions and agent-provided answers with 80k question and answer pairs in total. The conversations are collected from different sources: QuAC (Choi et al., 2018), Natural Questions (Kwiatkowski et al., 2019) and TREC CAsT-19 (Dalton et al., 2020) with additional annotations by crowd workers. See more details and statistics in Appendix A.1. Therefore, QReCC can be divided into three subsets for evaluation. We name them as *QuAC-Conv*, *NQ-Conv* and *TREC-Conv* respectively to differentiate them from the original datasets from which they are derived. TREC-Conv only appears in the test set. Each user question comes with a human-rewritten query. For each agent turn, gold passage labels are provided if any. The entire text corpus for retrieval contains 54M passages, segmented in the released data.[4]

Evaluation Metrics Following (Anantha et al., 2021), we use mean reciprocal rank (MRR), Recall@10 and Recall@100 to evaluate the retrieval performance by reusing the provided evaluation scripts.[5] Some agent turns in QReCC do not have valid gold passage labels,[6] and the original evaluation script assigns a score of 0 to all such examples. Their updated evaluation script calculates the scores by removing those examples from the evaluation set (roughly 50%), which results in 6396, 1442 and 371 test instances for QuAC-Conv, NQ-Conv and TREC-Conv, respectively. We use the *updated* evaluation script for most of our experiments, except that we also use the *original* version for calculating scores in Table 1 to compare with their reported QReCC baseline results . We note that these two evaluation scripts only differ by a scaling factor so they should lead to the same conclusions regarding model comparisons.

4.2 Implementation Details

Our models are implemented using JAX.[7] T5QR models are all initialized with T5-base (Raffel et al., 2020). For training, we set 64, 1k and 10k as the batch size, warm-up steps and total training steps respectively. We use e^{-3} and e^{-4} as the learning rate for supervised and RL training respectively. We use Adafactor (Shazeer and Stern, 2018) as our optimizer with the default parameters. Linear decay is applied after 10% of the total number of training steps, reducing the learning rate to 0 by the end of training. For supervised training, models are selected based on the best dev set Rouge-1 F1 score with the human rewrites, following Anantha et al. (2021). CONQRR is initialized with T5QR. For RL-based training of CONQRR, models are selected based on the average in-batch gold passage prediction accuracy as in Eq. (2) on dev set with greedily decoded rewrites. We experiment with CONQRR trained with either a mixed ($\mathcal{L}_{mix}$) or pure RL ($\mathcal{L}_{RL}$) loss. For the mixed loss, we observe that CONQRR works well when the RL loss weight α is large.[8] We tune its values in 0.9, 0.95, 0.97, 0.99, and use 0.99 as the final value. For the experiment with the pure RL loss and the retriever BM25, our results are obtained with the initialized model being fine-tuned with only 10% QR labels, as we find initializing with a model using 100% QR labels is unstable for BM25. Previous work (Wu et al., 2021) also had a similar observation that initializing with a less trained model leads to more stable RL training. More implementation and hyper-parameter details including input and output length limits are reported in Appendix A.2.

QR Model	Original Eval MRR	Original Eval R10	Original Eval R100	Updated Eval MRR	Updated Eval R10	Updated Eval R100
Transformer++	0.155	24.8	40.6	0.311	49.8	81.4
T5QR	0.164	26.2	42.3	0.328	52.5	84.7
CONQRR (mix)	0.186	29.2	**45.0**	0.373	58.5	**90.2**
CONQRR (RL)	**0.191**	**30.0**	44.4	**0.383**	**60.1**	88.9
Human	0.199	32.8	49.4	0.398	62.6	98.5

Table 1: Passage retrieval performance of QR models, comparable to scores in Anantha et al. (2021) by using the same BM25 retriever for QReCC test set. CONQRR achieves *state-of-the-art* results. Recall@10 and Recall@100 are abbreviated as R10 and R100.

4.3 Compared Systems

For QR models, we compare our supervised model **T5QR** and **CONQRR (mix/RL)** with a mixed ($\mathcal{L}_{mix}$) or pure RL ($\mathcal{L}_{RL}$) loss. We also compare to the original baseline **Transformer++**, which is

[4] Original QReCC data: https://zenodo.org/record/5115890#.YZ8kab3MI-Q.

[5] Both original and updated evaluation scripts: https://github.com/scai-conf/SCAI-QReCC-21.

[6] Missing gold labels for certain examples in the dataset has no effect on the training of CONQRR as we induce weak labels without using the provided labels.

[7] https://github.com/google/jax

[8] We also conduct experiments with $\alpha = 0.0$ for both retrievers and get similar results as T5QR.

QR Model	IR System	QReCC (Overall)			QuAC-Conv			NQ-Conv			TREC-Conv (OOD)*		
		MRR	R10	R100	MRR	R10	R100	MRR	R10	R100	MRR	R10	R100
T5QR	BM25	0.328	52.5	84.7	0.33	52.7	85.0	0.345	54.2	83.9	**0.230**	44.5	82.3
ConQRR (mix)	BM25	0.373	58.5	**90.2**	0.379	59.2	**90.9**	**0.385**	**58.8**	**88.9**	0.229	**44.7**	**82.7**
ConQRR (RL)	BM25	**0.383**	**60.1**	88.9	**0.395**	**61.6**	90.2	0.378	58.0	86.7	0.198	43.5	75.9
Human Rewrite	BM25	0.398	62.6	98.5	0.403	62.9	98.4	0.408	63.8	99.0	0.273	53.8	98.9
T5QR	DE	0.361	56.2	75.9	0.349	55.7	76.1	0.417	58.7	74.2	0.343	55.9	79.2
ConQRR (mix)	DE	0.395	61.9	81.8	0.387	62.0	82.4	0.439	62.2	79.0	**0.361**	**58.9**	**81.0**
ConQRR (RL)	DE	**0.418**	**65.1**	**84.7**	**0.416**	**65.9**	**85.8**	**0.453**	**64.1**	**80.9**	0.327	55.2	79.6
Human Rewrite	DE	0.422	64.8	84.0	0.409	64.5	84.1	0.483	65.8	83.2	0.411	66.0	86.5

Table 2: Passage retrieval performance on QReCC test set and 3 subsets. ConQRR (mix) beats the supervised T5QR model on all retriever system and test set combinations. * OOD (out-of-domain): only appear in the test set.

based on GPT2-medium that achieves the best retrieval performance in Anantha et al. (2021). Transformer++ has two language modeling heads that produce separate vocabulary distributions, which are then combined via a weighted sum for rewritten query generation. Similar to T5QR, it is a QR model trained in a standard supervised learning manner. For analysis purposes, we also report performance for directly using the concatenated dialogue context as the retriever input without any query rewriting in Section 4.5. We experiment with two retrievers, **BM25** and **DE** (Section 3.3).

4.4 Quantitative Results

The original baseline Transformer++ has numbers reported on the overall QReCC test set with BM25 as the retriever. As mentioned in Section 4.1, to have a direct comparison with Anantha et al. (2021), we first compare all QR models' downstream retrieval performance in Table 1, including both the original and updated versions of the evaluation script. T5QR and ConQRR outperform the baseline Transformer++ by 5% and 18% respectively, averaged on three metrics,[9] although Transformer++ is based on a larger base model - GPT2-medium. Therefore, ConQRR (RL) becomes the *state-of-the-art* QR model for conversational passage retrieval on QReCC.

Table 2 shows more comprehensive retrieval results comparing ConQRR and the supervised model T5QR, with the updated evaluation script. For the overall QReCC test set, ConQRR outperforms T5QR for all three metrics and both retrievers. For MRR and Recall@10, gains are roughly 15% with the RL loss and 9-14% with the mixed loss for both retrievers. Gains in Recall@100 vary more (4-12%). Breaking down the results by subset shows that the mixed loss is more robust. ConQRR (RL) is less effective for the TREC-Conv subset, which only appears in the test set. This suggests that RL loss alone does not generalize well to out-of-domain examples. Across all subsets, the best MRR and Recall@10 results are consistently from DE, whereas BM25 has better Recall@100 scores. See our explanation in Appendix A.3.

[9]We obtained prediction results from the authors and reran their evaluation script. The numbers we got are slightly lower than what they reported, but do not affect the conclusions.

4.5 Analysis

Effects of Topic Shift & Human Rewrites We hypothesize that a context involving a topic shift will present the greatest challenges for conversational passage retrieval. To explore this factor, we split the QReCC data into topic-concentrated and topic-shifted subsets as follows. A test example is considered *topic-shifted* if it has at least one previous turn besides the current user question and all previous turns have gold passages from a different document than the gold passage of the current question. All other examples (with at least one previous turn) are *topic-concentrated*. There are about 4.7k and 1.1k examples in the topic-concentrated and topic-shifted subsets respectively. We compare the retrieval performance of different retriever inputs: dialogue context (which uses the concatenated dialogue history without QR), the predicted rewrite from T5QR and ConQRR with two loss alternatives, and the human rewrite. Table 3 shows that the dialogue context outperforms even the human rewrite on the topic-concentrated set by 22% and 17%, averaging over three metrics, for BM25 and DE respectively, which shows the *limitation of human rewrites*. We also see that ConQRR (RL) surpass the human rewrite on the topic-concentrated set on MRR for BM25 and all three metrics for DE.

Input	IR	Topic-Concentrated MRR	R10	R100	Topic-Shifted MRR	R10	R100
Dial Context	BM25	**0.620**	**81.4**	**94.9**	0.154	39.1	68.6
T5QR	BM25	0.352	54.4	84.0	**0.252**	45.1	79.1
CONQRR (mix)	BM25	0.419	63.1	91.2	**0.252**	**45.9**	**82.1**
CONQRR (RL)	BM25	0.444	66.2	90.3	0.233	44.5	78.4
Human Rewrite	BM25	0.440	66.7	98.8	0.318	56.7	98.4
Dial Context	DE	**0.551**	**78.1**	**93.2**	0.179	35.7	61.4
T5QR	DE	0.353	55.7	75.4	0.329	50.8	69.2
CONQRR (mix)	DE	0.404	63.8	83.4	**0.334**	**53.2**	72.6
CONQRR (RL)	DE	0.445	69.3	87.8	0.303	50.4	**73.3**
Human Rewrite	DE	0.424	65.5	84.5	0.397	61.0	79.8

Table 3: Performance of using different retriever inputs for *Topic-Concentrated* or *Topic-Shifted* examples.

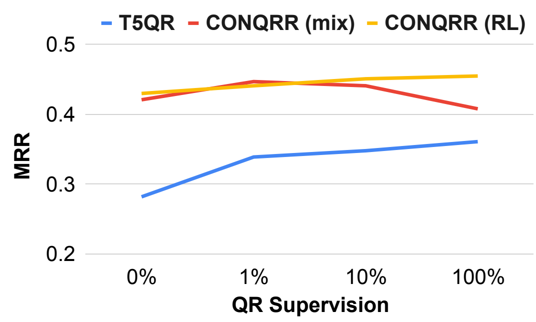

Figure 4: MRR on QReCC versus the percentage of QR supervision used for training, with DE as the retriever.

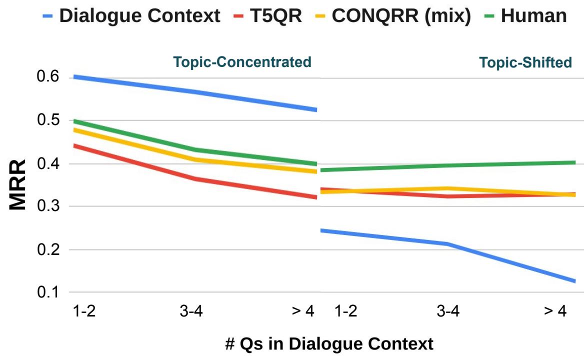

Figure 3: MRR versus the number of questions in the dialogue context, with DE as the retriever.

However, for the topic-shifted set, the human rewrite outperforms the dialogue context by 52% and 61%, averaging over three metrics, on BM25 and DE respectively. The predicted rewrite by CONQRR (mix) outperforms the dialogue context by 30% and 44% on BM25 and DE respectively. Therefore, compared with dialogue context, QR has great value in the aspect of *robustness to topic shifts*. When comparing with human rewrites, we also see improvement room for QR models.

These observations are *largely unexplored* in previous work, and they are actually the motivations for us to work on the task of QR for conversational passage retrieval, and to build CONQRR that optimizes directly towards retrieval and goes beyond the human rewrite limitations. In addition, although fine-tuning the retriever is not our focus, we discuss very different empirical observations in Appendix A.3 and show that QR may not be necessary if the retriever can be fine-tuned.

Effect of Dialogue Context Length Figure 3 shows the MRR score on topic-concentrated and topic-shifted subsets with DE as the retriever for various dialogue context lengths. Dialogue context lengths are grouped into 1-2, 3-4 and ≥ 4 previous utterances (including the current question). For topic-concentrated conversations, all compared models have similar robustness to the dialogue context length and CONQRR (mix) is slightly more robust than T5QR. For topic-shifted conversations, both QR models and human rewrites show little drop or even an increase in performance as the context length gets longer. In contrast, the robustness of the dialogue context worsens with longer contexts, which confirms the importance of QR discussed above. We have similar observations for other metrics as well as for the BM25 retriever.

Data Efficiency We investigate how sensitive CONQRR and T5QR are to the availability of QR labels. We experiment with training T5QR with 0%, 1%, 10% or 100% of QR labels in the QReCC train set. For the case of 0% examples, we simply use the original T5 checkpoint without fine-tuning. When training CONQRR, we mask out the CE loss in Eq. (3) for unused QR labels in training its initialized T5QR model, and we use dialogue context to induce gold and hard negative passages for each training example, instead of using human rewrites. Figure 4 plots the curve of MRR on the overall QReCC test data using DE as the retriever versus the percentage of QR labels used for training. We see that CONQRR can achieve good performance with even 0% or 1% of QR supervision. The slight difference in performance for the 100% QR label case with respect to Table 2 is due to the different mechanism (using human rewrite vs. the dialogue context) for choosing the positive and hard negative passages for RL training. Performance of the RL and mixed loss are similar when there is little supervision, roughly tracking the trends of the T5QR model that it is initialized with. The finding that performance degrades for the mixed

Dialogue Context	*Q*: What were **John Stossel**'s most popular publications? *A*: Give Me a Break: How I Exposed Hucksters, Cheats, and Scam Artists and Became ... ... *Q*: What was the response?	*Q*: What were some notable live performances at the Buena Vista Social Club? *A*: **Ibrahim Ferrer and Rubén González** ... ... *Q*: What other live performances are important?
Gold Passage	**Stossel** has written three books. Give Me a Break: ... It was a New York Times bestseller for 11 weeks ...	The first performances ... **Ibrahim Ferrer and Rubén González** performed together ... a 1999 Miami performance ...
CONQRR (mix)	What was the response to **John Stossel**'s book, Give Me a Break? (Rank=2)	What other live performances at the Buena Vista Social Club are important besides **Ibrahim Ferrer and Rubén González**? (Rank=2)
T5QR	What was the response to the book Give Me a Break? (Rank >100)	What other live performances are important at the Buena Vista Social Club? (Rank=18)
Human	What was the response to Give Me a Break: How I Exposed Hucksters, Cheats, and Scam Artists and Became the Scourge of the Liberal Media? (Rank >100)	What other live performances of the Buena Vista Social Club are important? (Rank=17)

Table 4: Examples of predicted rewrites and the gold passage ranks by using them as the DE retriever input.

QR Model	**QuAC-Conv** L	**QuAC-Conv** OL	**NQ-Conv** L	**NQ-Conv** OL	**TREC-Conv** L	**TREC-Conv** OL
T5QR	10.9	3.9	8.9	3.6	8.2	3.1
Ours (mix) w/ BM25	12.1	4.5	9.5	4.0	8.5	3.3
Ours (RL) w/ BM25	11.2	4.5	10.1	4.5	9.4	3.7
Ours (mix) w/ DE	12.1	4.5	9.6	4.0	8.7	3.4
Ours (RL) w/ DE	28.2	14.4	21.7	12.1	18.3	8.1
Human	12.1	4.5	9.3	4.0	8.4	3.5

Table 5: Average number of tokens (L) and overlapping tokens (OL) with the gold passage(s) in output rewrites.

loss with 100% supervision may be due to a mismatch in the CE and RL losses as minimizing the CE loss does not directly optimize the retrieval performance. T5QR is more sensitive to QR supervision but also does not require many QR labels for training, as its curve becomes flattened after 1% supervision. We see similar trends with Recall@100 (see Appendix A.3).

Quantitative Attributes of Rewrites Table 5 shows the average number of tokens per rewritten query, and overlapping tokens (excluding stopwords) between the rewrite and the gold passage(s). CONQRR generally generates longer rewrites with more overlapping tokens with gold passage(s), compared with T5QR. When having DE as the retriever, CONQRR (RL) generates more than double the length of T5QR, CONQRR (mix) and even human rewrites. We show in Appendix A.3 that T5QR still underperforms CONQRR (mix) even when we make it generate rewrites of similar lengths by applying a brevity penalty (Wu et al., 2016).

Rewrite Examples Table 4 shows two examples of generated rewritten queries of T5 and CONQRR (mix) trained with DE in the loop, as well as the human rewrites. In the left example, the rewrite of CONQRR is able to generate an entity "John Stossel" that is mentioned in the gold passage but not included by rewrites from T5QR or Human. Thus, even if the human rewrite is longer by containing the book's full name, CONQRR enables more efficient retrieval with a partial book name along with its author name. In the right example, CONQRR generates a longer rewritten query that contains much richer contextual information. See more examples in Appendix A.3

5 Conclusion and Discussion

To summarize, we introduce CONQRR to address query rewriting for conversational passage retrieval with a fixed retriever. Motivated by our analysis showing both the limitations and utility of human rewrites, which are under-explored by prior work, we adopt a RL approach with a novel reward function to train CONQRR directly towards retrieval. We show that CONQRR can be trained adaptively to any fixed retriever. The model achieves state-of-the-art retrieval performance on QReCC with conversations from 3 different sources.

A direction for future work includes leveraging QR to facilitate other tasks. For example, it can also be used for question answering and response generation in a full CQA system. Sentence rewriting can be used to understand context-dependent sentences in a document (Choi et al., 2021). As current CQA datasets have a restricted dialogue format of alternating questions and answers, future investigation is needed to explore conversations with discourse relations like asking for clarifications. We put more discussion in Appendix A.4.

Ethical Considerations

Our work is primarily intended to leverage query rewriting (QR) models to facilitate the task of conversational passage retrieval in an open-domain CQA system. Retrieving the most relevant passage(s) to the current user query in a conversation would help to generate a more appropriate agent response. Predicted rewrites from our QR model are mainly intended to be used as *intermediate* results (e.g., the inputs to the downstream retrieval system). They may also be useful for interpretability purposes when a final response does not make sense to the user in a full CQA system, but that introduces a potential risk of offensive text generation. In addition, to prevent the retriever from retrieving passages from unreliable resources, filtering of such passages in the corpus should be performed before any practical use.

References

Leonard Adolphs, Benjamin Boerschinger, Christian Buck, Michelle Chen Huebscher, Massimiliano Ciaramita, Lasse Espeholt, Thomas Hofmann, and Yannic Kilcher. 2021. Boosting search engines with interactive agents. *arXiv preprint arXiv:2109.00527*.

Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 385–394, New York, NY, USA. Association for Computing Machinery.

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, Online. Association for Computational Linguistics.

Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *International Conference on Learning Representations*.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2018. Attention-based hierarchical neural query suggestion. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1093–1096, New York, NY, USA. Association for Computing Machinery.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1985–1988, New York, NY, USA. Association for Computing Machinery.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan Clark, and Regina Barzilay. 2020. CapWAP: Image captioning with a purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8755–8768, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

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

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

Vaibhav Kumar and Jamie Callan. 2020. Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3971–3980, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, and Chris Alberti et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, Copenhagen, Denmark. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 539–548, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*. Accessed 22 March 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Solomon. 1997. Conversation in information-seeking contexts: A test of an analytical framework. *Library & Information Science Research*, 19(3):217–248.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, and Wolfgang Macherey et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, and Bill Dolan. 2021. Automatic document sketching: Generating drafts from analogous texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2102–2113, Online. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1933–1936, New York, NY, USA. Association for Computing Machinery.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. *Few-Shot Conversational Dense Retrieval*, page 829–838. Association for Computing Machinery, New York, NY, USA.

A Appendix

A.1 Additional Data Details

QReCC reuses questions in QuAC and TREC conversations and re-annotates answers. For each NQ-based conversation, they only use one randomly chosen question from NQ to be the starting question and then annotate the remaining conversation. In total, there are 63k, 16k and 748 question and answer pairs in the three subsets QuAC-Conv, NQ-Conv, TREC-Conv respectively, where TREC-Conv only appears in the test set. The original data is only divided into train and test sets. We randomly choose 5% examples from the train set to be our validation set.

In some conversations from QuAC-Conv, the first user query is ambiguous as it depends on some topical information from the original QuAC dataset. Therefore, in order to fix this issue, we follow Anantha et al. (2021) to replace all first user queries in QReCC conversations with the their corresponding human rewrites.

QReCC is a publicly available dataset that was released under the Apache License 2.0 and we use the same task set-up proposed by the original qrecc authors.

A.2 Additional Implementation Details

The maximum length of the dialogue context fed into the QR model is 384 (longer than 97.9% dialogue contexts in QReCC) and the maximum output rewrite length is 64 (longer than 99.9% human rewrites). To generate each sampled rewrite q_s (see Section 3.2), we apply top-k sampling where $k = 20$. For each training example, we sample 5 rewrites in total (i.e., $m = 5$ for the RL training explained in Section 3.2). Each training process is run on 8 TPU nodes. It takes about 2 and 9 hours for the supervised and RL-based training respectively. For each experiment, we observe similar performance or training curves for 2-3 runs and report numbers on a random run. Both T5QR and CONQRR are based on T5-base and have about 220M parameters. In contrast, the baseline Transformer++ is based on GPT2-medium and has about 345M parameters.

For retriever models, BM25 Pyserini simply encodes the whole query input and each passage without truncating. We set maximum query and passage length as 128 and 2000 for BM25-light, but only less than 0.1% cases require truncation with these thresholds. For the dual encoder, the maximum query or passage length is 384. The average passage length is 378, but we observe performance drop by further increasing the maximum length for the dual encoder.

A.3 Additional Analysis

Lower Recall@100 with DE Previous work (Karpukhin et al., 2020) shows that DE retrievers generally lead to better recall scores than BM25. However, in Table 2, we observe that across all subsets, the best MRR and Recall@10 results are consistently from DE, whereas BM25 has better Recall@100 scores. One reason to explain the observation difference is that we use a *fixed* retriever for our retrieval task while most previous work that compare BM25 and DE focuses on fine-tuning the DE model. Without being fine-tuned, a DE model may be more vulnerable to domain shift than BM25. On the other hand, prior work (Luan et al., 2021) proves that a DE model's performance would drop as the passage length increases. In the QReCC dataset, the average passage length is 378, which is relatively long according to Luan et al. (2021).

Analysis of Longer Rewrites We hypothesize that simply generating a longer rewritten query is not the only factor that contributes to better retrieval performance. We investigate this by applying a brevity penalty (Wu et al., 2016) during decoding for T5QR such that its average query length matches that of CONQRR (mix). Figure 5 shows that CONQRR (mix) still outperforms T5QR with the brevity penalty for all three evaluation metrics on QReCC.

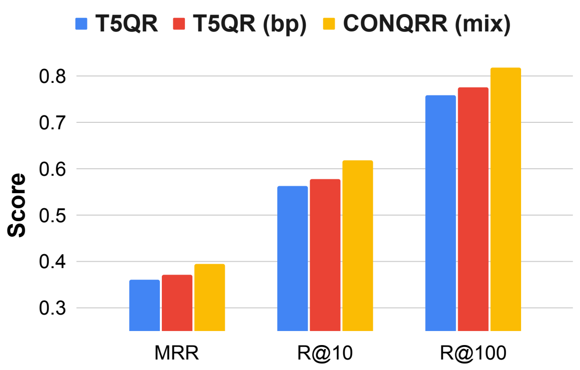

Figure 5: Evaluation scores on QReCC for T5QR w/ or w/o brevity penalty and CONQRR (mix), with DE as the retriever. Recall scores (R@k) are divided by 100.

Fine-tuned Retriever Although our work focuses on the fixed retriever setting, we also conduct an experiment of fine-tuning the DE retriever

Input	Topic-Concentrated MRR	R10	R100	Topic-Shifted MRR	R10	R100
Dial Context	**0.643**	**87.7**	**96.9**	**0.312**	**56.2**	**81.9**
CONQRR (mix)	0.588	84.0	96.9	0.259	48.3	77.2
Human Rewrite	0.510	79.9	95.2	0.380	61.3	86.0

Table 6: Results of using the dialogue context, predicted rewrite or human rewrite as the retriever input with the *finetuned* DE as the retriever.

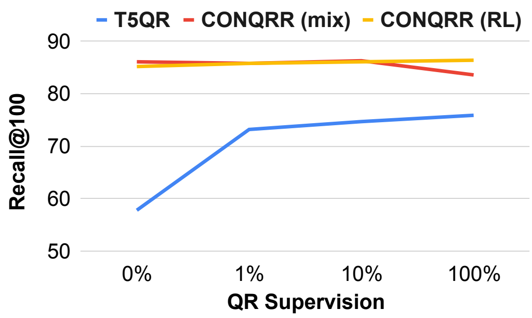

Figure 6: Recall@100 on QReCC versus the percentage of QR supervision used for training, with DE as the retriever.

with the concatenated dialogue context, the predicted rewrite from CONQRR (mix) or the human rewrite as the query input, with results in Table 6. The numbers are comparable to those in Table 3. Fine-tuning the DE retriever improves results for all scenarios, but the dialogue context benefits substantially, to the extent that it outperforms ConQRR in topic-shifted cases. However, there is still improvement room as we see benefits of human query-rewrites for topic shifts.

Additional Data Efficiency Figure Figure 6 shows the curve of Recall@100 on the overall QReCC test data using DE as the retriever versus the percentage of QR labels used for training.

Additional Rewrite Examples In addition to Table 4, we put more examples in Table 7. Different from Table 4, we put predicted rewrites from CONQRR (mix) that is trained towards BM25 instead of the DE retriever. We also put gold passage ranks in the table, by using the predicted rewrites as the BM25 retriever input.

A.4 Discussion

We first summarize the scenarios when leveraging QR for conversational passage retrieval may bring most benefits. As shown in Section 4.5 (Table 3), compared to directly use dialogue context without

Dialogue Context	*Q*: What is Get 'Em Girls? *A*: Jessica Mauboy's second studio album, **Get 'Em Girls** (**2010**). ... *Q*: Did she receive any awards or honors during these years?	*Q*: What is one actress who was a Bond girl? *A*: Ursula Andress in **Dr. No** is widely regarded as the first Bond girl. ... ... *Q*: Who was another Bond girl?
Gold Passage	... Mauboy performed "**Get 'Em Girls**" at the **2010** ... and won the award for ... **Get 'Em Girls** was re-released as a deluxe edition ...	... Ursula Andress (as Honey Ryder) in **Dr. No** (1962) is widely regarded as the first Bond girl, although she was preceded by both Eunice Gayson as Sylvia Trench and ...
CONQRR (mix)	Did Jessica Mauboy receive any awards or honors during the years she released **Get 'Em Girls**? (Rank=7)	Who was another Bond girl besides Ursula Andress in **Dr. No**? (Rank=7)
T5QR	Did Jessica Mauboy receive any awards or honors during these years? (Rank >100)	Who was another Bond girl? (Rank=68)
Human	Did Jessica Mauboy receive any awards or honors during the **2010**s? (Rank=24)	Who was another Bond girl, besides Ursula Andress? (Rank=12)

Table 7: Examples of predicted rewrites and the gold passage ranks by using them as the BM25 retriever input.

QR, a QR model has great values in robustness to topic shifts with a fixed retriever.

On the other hand, if most conversations of interest are topic-concentrated, we show that using the dialogue context itself can already work well. From Table 6, we also see that if the downstream retriever is allowed to be fine-tuned, our best QR model CONQRR (mix) underperforms the dialogue context in both topic-concentrated and topic-shifted scenarios.

Another downside of QR is that it requires additional labels. Although we show that CONQRR (RL) initialized with T5 does not require QR labels and can still work well on the overall QReCC test set, CONQRR (RL) does show worse robustness to out-of-domain and topic-shifted examples when compared with CONQRR (mix). Therefore, training a more robust CONQRR model may still require additional annotation efforts to collect human rewrites.

CONQRR has only been tested on the standard CQA dialogue format of alternating questions and answers. To facilitate more practical use cases with more diverse dialogue acts or discourse relations (e.g., the agent asks a clarification question to the user), further investigation is needed.