ConvGQR: Generative Query Reformulation for Conversational Search

Fengran Mo[1], **Kelong Mao**[2], **Yutao Zhu**[1], **Yihong Wu**[1], **Kaiyu Huang**[3*], **Jian-Yun Nie**[1*]
[1]University of Montreal, Quebec, Canada
[2]Gaoling School of Artificial Intelligence, Renmin University of China
[3]Institute for AI Industry Research, Tsinghua University, Beijing, China
{fengran.mo,yihong.wu}@umontreal.ca, yutaozhu94@gmail.com
nie@iro.umontreal.ca, mkl@ruc.edu.cn, huangkaiyu@air.tsinghua.edu.cn

Abstract

In conversational search, the user's real search intent for the current conversation turn is dependent on the previous conversation history. It is challenging to determine a good search query from the whole conversation context. To avoid the expensive re-training of the query encoder, most existing methods try to learn a rewriting model to de-contextualize the current query by mimicking the manual query rewriting. However, manually rewritten queries are not always the best search queries. Thus, training a rewriting model on them would lead to sub-optimal queries. Another useful information to enhance the search query is the potential answer to the question. In this paper, we propose **ConvGQR**, a new framework to reformulate conversational queries based on generative pre-trained language models (PLMs), one for query rewriting and another for generating potential answers. By combining both, ConvGQR can produce better search queries. In addition, to relate query reformulation to the retrieval task, we propose a knowledge infusion mechanism to optimize both query reformulation and retrieval. Extensive experiments on four conversational search datasets demonstrate the effectiveness of ConvGQR.

1 Introduction

Conversational search (Gao et al., 2022) is a rapidly developing branch of information retrieval, which aims to satisfy complex information needs through multi-turn conversations. The main challenge is to determine users' real search intents based on the interaction context and formulate good search queries accordingly. Existing methods can be roughly categorized into two groups. The first group directly uses the whole context as a query and trains a model to determine the relevance between the long context and passages (Qu et al., 2020; Hashemi et al., 2020; Yu et al., 2021; Lin et al., 2021b; Mao et al., 2022a,b; Kim and Kim, 2022; Mo et al., 2023). This approach requires additional training of retriever to take the long context as input, which is not always feasible (Wu et al., 2021). What is available in practice is a general retriever (*e.g.*, adhoc search retriever) that uses a stand-alone query. The second group of approaches aims at producing a de-contextualized query using query reformulation techniques (Elgohary et al., 2019). Such a query can be submitted to any *off-the-shelf* retrievers. We focus on this second approach.

Two types of query reformulation techniques have been widely studied in the literature, *i.e.*, *query rewriting* and *query expansion*. The former trains a generative model to rewrite the current query to mimic the human-rewritten one (Yu et al., 2020; Vakulenko et al., 2021a), while the latter focuses on expanding the current query by relevant terms selected from the context (Kumar and Callan, 2020; Voskarides et al., 2020). Although both approaches achieve promising results, they are all studied separately. Two important limitations are observed: (1) Query rewriting and query expansion can produce different effects. Query rewriting tends to deal with ambiguous queries and add missing tokens, while query expansion aims to add supplementary information to the query. Both effects are important for query reformulation. It is thus beneficial to use both of them. (2) Previous query rewriting models have been optimized to produce human-rewritten queries, independently from the passage ranking task. Even though human-rewritten queries usually perform better than the original queries, existing studies have shown that they may not be the best search queries alone (Lin et al., 2021b; Wu et al., 2021). Therefore, it is useful to incorporate additional criteria directly related to ranking performance when reformulating a query. As shown in Fig. 1 (left), although the human-rewritten query recovers the crucial missing information (i.e. "goat") from the context, it is

[*] Corresponding authors.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4998–5012
July 9-14, 2023 ©2023 Association for Computational Linguistics

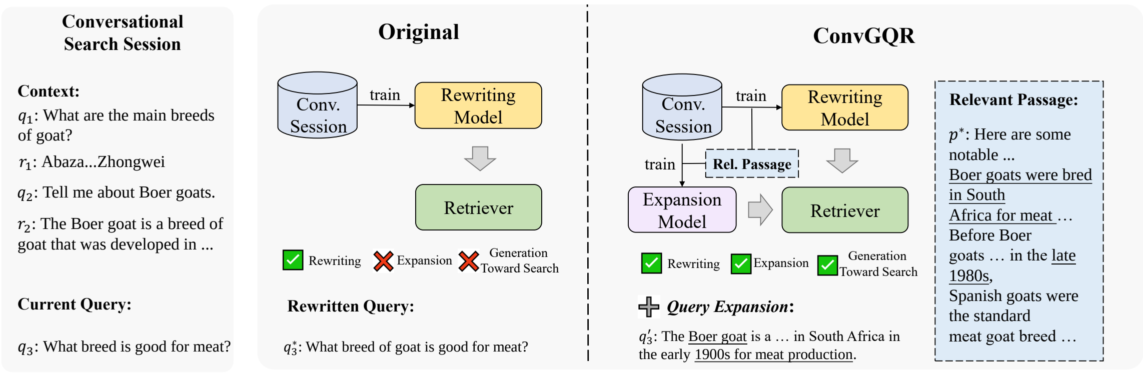
Conversational Search Session
Context:
q_1: What are the main breeds of goat?
r_1: Abaza...Zhongwei
q_2: Tell me about Boer goats.
r_2: The Boer goat is a breed of goat that was developed in ...
Current Query:
q_3: What breed is good for meat?
Original
Conv. Session
train
Rewriting Model
Retriever
✓ Rewriting
✗ Expansion
✗ Generation Toward Search
Rewritten Query:
q_3^*: What breed of goat is good for meat?
ConvGQR
Conv. Session
train
Rewriting Model
train
Rel. Passage
Expansion Model
Retriever
✓ Rewriting
✓ Expansion
✓ Generation Toward Search
Query Expansion:
q_3': The Boer goat is a … in South Africa in the early 1900s for meat production.
Relevant Passage:
p^*: Here are some notable ... Boer goats were bred in South Africa for meat … Before Boer goats … in the late 1980s, Spanish goats were the standard meat goat breed …

Figure 1: An example of conversational search session and the high-level comparison between the original method and our ConvGQR. The dashed box illustrates the potential connection (underline) between the relevant passage and expansion terms.

still possible to further improve the search query.

To tackle these problems, we propose ConvGQR, a new **G**enerative **Q**uery **R**eformulation framework for **Conv**ersational search. It combines query rewriting with query expansion. The right side of Fig. 1 illustrates the differences between ConvGQR and the existing query rewriting method. In addition to query rewriting based on human-rewritten queries, ConvGQR also learns to generate the potential answer of the query (*e.g.*, the answer in the downstream question-answering task) and uses it for query expansion. This strategy is motivated by the fact that a passage containing the generated potential answer is more likely a relevant passage, because either the generated answer is the right answer, or it may co-occur with the right answer in the same passage. The final query reformulation model is trained by combining both query rewriting and query expansion criteria in the loss function. Moreover, the learning of both query rewriting and expansion are guided by the relevant passage information through our knowledge infusion mechanism to encourage query generation toward better search performance. We carry out extensive experiments on four conversational search datasets using both dense and sparse retrievers, and the results show that our method outperforms most existing query reformulation methods. Our further analysis confirms the complementary contributions of query rewriting and query expansion.

Our contributions are summarized as follows: (1) We propose ConvGQR to integrate query rewriting and query expansion for conversational search. In particular, query expansion is performed by adding the generated potential answer by a generative PLM. This is a way to exploit PLM's capability of capturing rich world knowledge. (2) We further design a knowledge infusion mechanism to optimize query reformulation with the guidance of passage retrieval. (3) We demonstrate the effectiveness of ConvGQR with two off-the-shelf retrievers (sparse and dense) on four datasets. Our analysis confirms the complementary effects of both components in conversational search.

2 Related Work

Conversational Query Reformulation The intuitive idea is that a well-formulated search query from the conversation context can be submitted to an *off-the-shelf* retriever for search without modifying it. Query rewriting and query expansion are two typical query reformulation methods. Query rewriting aims to train a rewriting model to mimic human-rewritten queries. This approach is shown to be able to solve the ambiguous problem and recover some missing elements (e.g. anaphora) from the context (Yu et al., 2020; Lin et al., 2020; Vakulenko et al., 2021a; Mao et al., 2023a). However, Wu et al. (2021) and Lin et al. (2021b) argue that the human-rewritten queries might not necessarily be the optimal queries. Wu et al. (2021) enhances the rewriting model by leveraging reinforcement learning. However, it turns out that reinforcement learning requires a long time for training. To be more efficient, Lin et al. (2021b) proposes a query expansion method by selecting the terms via the normalization score of their embeddings but still needs to re-train a retriever. Some earlier query expansion methods (Kumar and Callan, 2020; Voskarides et al., 2020) also focus on selecting useful terms from conversational context. The previous studies show that query rewriting and

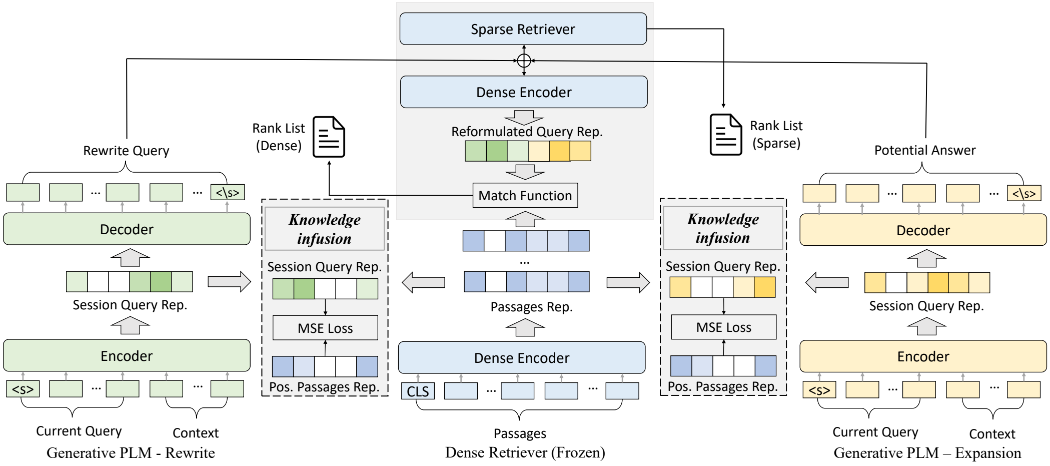

Figure 2: Overview of ConvGQR. Two generative PLMs are used to generate a rewritten query and expansion terms for both training and inference. The knowledge infusion mechanism (dashed boxes) is only applied during training.

query expansion can both enhance the search query and produce better retrieval results. However, these approaches have been used separately. Our ConvGQR model thus integrates both query rewriting and query expansion to reformulate a better conversational query. Moreover, a new knowledge infusion mechanism is used to connect query reformulation with retrieval.

Query Expansion via Potential Answers Earlier studies on question answering (Ravichandran and Hovy, 2002; Derczynski et al., 2008) demonstrate that an effective way to expand a query is to extract answer patterns or select terms that could be possible answers as expansion terms. Recently, some generation-augmented retrieval methods (Mao et al., 2021; Chen et al., 2022) focus on exploiting the knowledge captured in PLMs (Roberts et al., 2020; Brown et al., 2020) to generate the potential answer as expansion terms. We draw inspiration from these studies and apply the idea to conversational search.

3 Methodology

3.1 Task Formulation

We formulate the conversational search task in this paper as retrieving the relevant passage p from a large passage collection C for the current user query q_i given the conversational historical context $H^k = \{q_k, r_k\}_{k=1}^{i-1}$, where the q_k and r_k denote the query and the system answer of the k^{th} previous turn, respectively. In this paper, we aim to design a query reformulation model to transform the current query q_i together with the conversational historical context H^k into a de-contextualized rewritten query for conversational search.

3.2 Our Approach: ConvGQR

A first desired behavior of query reformulation is to produce a similar rewritten query as a human expert. This will solve some ambiguities arisen in the current query (*e.g.*, omission and coreference). So, query rewriting will be an integral part of our approach. Query rewriting can be cast as a text generation problem: given the query in the current turn and its historical context, we aim to generate a rewritten query. Inspired by the large capability of PLM, we rely on a PLM for query rewriting to mimic the human query rewriting process.

However, as the human-rewritten query might not be optimal (Yu et al., 2020; Anantha et al., 2021) and the standard query rewriting models are agnostic to the retriever (Lin et al., 2021b; Wu et al., 2021), a query rewriting model alone cannot produce the best search query. Therefore, we also incorporate a component to expand the query by adding additional terms that are likely involved in relevant passages. Several query expansion methods can be used. In this paper, we choose to use the following one which has proven effective in question answering (Mao et al., 2021; Chen et al., 2022): we use the current query and its context to generate a potential answer to the question (query). The generated answer is used as expansion terms. This approach leverages the large amount of world knowledge implicitly captured in a large PLM[1]. The generated potential answer can be useful for

[1] As shown by the recent success of ChatGPT, PLMs can generate correct answers to a large variety of questions.