History-Aware Conversational Dense Retrieval

Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, Jian-Yun Nie
1University of Montreal, Quebec, Canada; 2University of Massachusetts Amherst, USA
3Renmin University of China; 4Beihang University, China
5University of Copenhagen, Denmark; 6Beijing Jiaotong University, China
fengran.mo@umontreal.ca, kyhuang@bjtu.edu.cn, nie@iro.umontreal.ca

Abstract

Conversational search facilitates complex information retrieval by enabling multi-turn interactions between users and the system. Supporting such interactions requires a comprehensive understanding of the conversational inputs to formulate a good search query based on historical information. In particular, the search query should include the relevant information from the previous conversation turns. However, current approaches for conversational dense retrieval primarily rely on fine-tuning a pre-trained ad-hoc retriever using the whole conversational search session, which can be lengthy and noisy. Moreover, existing approaches are limited by the amount of manual supervision signals in the existing datasets. To address the aforementioned issues, we propose a History-Aware Conversational Dense Retrieval (HAConvDR) system, which incorporates two ideas: context-denoised query reformulation and automatic mining of supervision signals based on the actual impact of historical turns. Experiments on two public conversational search datasets demonstrate the improved history modeling capability of HAConvDR, in particular for long conversations with topic shifts.

1 Introduction

Conversational search is expected to be the next generation of search engines (Gao et al., 2022). It aims to satisfy complex user information needs via multi-turn interactions between a user and the system. In single-turn ad-hoc search, users typically employ stand-alone queries to convey their information requirements (Bajaj et al., 2016) in a brief and clearly-expressed manner. In conversational search, however, queries are usually context-dependent, which highlights the necessity of understanding the search intent within the conversational context.

To uncover the user’s information need, conversational query rewriting (CQR) (Yu et al., 2020; Wu et al., 2022; Mo et al., 2023a) employs human-rewritten queries to train a rewriting model that generates de-contextualized queries. However, obtaining large-scale manual annotations for this purpose is challenging in practice. Besides, CQR models cannot be directly optimized for the downstream retrieval task (Wu et al., 2022; Mo et al., 2023a).

In comparison, a more desirable approach is to perform end-to-end conversational dense retrieval (CDR) by training a query encoder that incorporates conversation history (Qu et al., 2020; Yu et al., 2021). Since human annotations are usually not available to indicate which previous conversation turns are relevant to the current query, a common practice is to utilize all historical turns to reformulate the current query as the input to the model.

However, the conversation history can be lengthy and often includes a substantial amount of noise, i.e., historical turns that are irrelevant to the current query. Despite the observation (Adlakha et al., 2022) that conversational sessions often center around a specific topic (e.g., sports), it is worth noting that different turns may focus on different aspects (e.g., match results, or player statistics). Some of them are relevant to the current turn, while others may not. This is especially the case when conversations are long. This problem can give rise to the issue of shortcut history dependency (Kim and Kim, 2022; Fang et al., 2022), i.e. the reformulated query depends excessively on the historical turns while neglecting the current query. We illustrate this issue by an example in Figure 1. Given the current query $q_4$, instead of retrieving the passage $p_4^*$ (addressing the current information need) in top-ranked positions, the retriever ranks $p_3^*$ (addressing historical information needs) higher than $p_4^*$.

To tackle the aforementioned challenge, we put forward HAConvDR, a new History-Aware
Conversational Dense Retrieval method, aiming to leverage the useful information from the history as much as possible to reformulate the current query. Our approach consists of two prongs of enhancements as detailed in the following sections.

The first prong is to incorporate an explicit denoising mechanism into the model training process so that the model is less affected by the noisy history while being history-aware. To achieve a similar purpose, recent studies (Mao et al., 2022a, 2023c; Mo et al., 2023b) typically assess whether a historical turn is relevant to the current turn based on the historical query. However, these approaches are inherently lacking because historical queries alone are often not sufficient to fully cover the historical context. To address this shortcoming, we additionally leverage the passages associated with historical queries to better evaluate the intent of a historical turn. Specifically, we use a pseudo-labeling approach to assess the relevance and usefulness of the historical turns—whether they contribute to improving the retrieval effectiveness of the current query. We then retain the relevant historical turns for context-denoised query reformulation.

The second prong is to mine additional supervision signals to further alleviate the pitfalls of shortcut history dependency. Despite having context-denoised queries, a single ground-truth passage (given by the dataset) is often indirect and insufficient to guide the training of conversational retrieval due to the remaining noise in the formulated query. Thus, mining additional supervisions, either positive (Mao et al., 2022b) or negative (Kim and Kim, 2022), can enhance the original supervision signal and reduce the negative impact by the distractors in the conversation history. Different from the aforementioned work that acquires additional supervisions by human annotation or retrieval, we mine pseudo positive and hard negative supervisions from the conversation history based on the same relevance judgment of historical turns used for query reformulation. Intuitively, among the top-ranked historical ground-truth passages in Figure 1, some of them can be highly relevant to the current query, which resembles the pseudo relevant documents in Pseudo Relevance Feedback (PRF) (Xu and Croft, 1996), while others are less relevant and can serve as hard negatives for training. These additional supervisions enable the model to be aware of the usefulness or harmfulness of historical ground-truth passages and leverage them in a history-aware contrastive learning process.

We carry out extensive experiments on two conversational search datasets to test the effectiveness of HACoNDR. The results show that our method outperforms most existing strong baselines, demonstrating how relevance judgments of historical turns can benefit conversational retrieval.

Our contributions are summarized as follows: (1) We propose HACoNDR to train a history-aware conversational dense retriever by using the ground-truth passage from historical turns as additional supervision signals. (2) We conduct pseudo relevance judgment on selecting historical turns to denoise the context for query reformulation, whose results are the foundation of mining additional supervision signals. (3) We demonstrate the effectiveness of HACoNDR by outperforming different types of strong baselines on two public datasets. A series of analyses are conducted to understand how historical ground-truth passages work well to solve the conversation with lots of topic shifts.

2 Related Work

Conversational Query Reformulation. This approach aims to reformulate an explicit query via training a CQR model. Typical methods include query rewriting (Yu et al., 2020; Lin et al., 2020; Vakulenko et al., 2021; Qian and Dou, 2022; Mao et al., 2023a,b) and query expansion (Kumar and Callan, 2020; Voskarides et al., 2020), which aim to mimic human query rewriting or selection of useful terms from historical context for expansion. However, the manual annotations needed for training are difficult to obtain in practice and the human-rewritten queries might not necessarily be
the optimal search queries (Wu et al., 2022; Mo et al., 2023a). Some recent studies (Ye et al., 2023; Mao et al., 2023a; Jang et al., 2023) leverage large language models (LLMs) to generate reformulated queries via prompting but the generated queries are not optimized for search.

**Conversational Dense Retrieval.** Another research direction is to perform conversational dense retrieval, which leverages conversational search data to fine-tune a well-trained ad-hoc retriever. Existing studies (Yu et al., 2021; Lin et al., 2021; Mao et al., 2022b; Mo et al., 2024; Chen et al., 2024) usually focus on few-shot scenarios or rely on external resources, but without context denoising. On context denoising, some recent work (Mao et al., 2022a, 2023c; Mo et al., 2023b; Mao et al., 2024) designs sophisticated mechanisms to enhance the denoising ability explicitly and implicitly for the models. However, they do not take into account historical feedback. To perform context-denoising more effectively, our method explicitly selects the useful historical turns, as well as their ground-truth passage via pseudo relevant judgment before model training.

**Supervision Signals in Dense Retrieval.** Robinson et al. (2021) demonstrates that sufficient supervision signals, either positive or negative (especially hard negatives), are important for contrastive learning. For dense retrieval, hard negatives are usually mined by BM25 (Karpukhin et al., 2020) or a vanilla backbone model (Xiong et al., 2020). In the conversational scenario, Kim and Kim (2022) uses the CQR model to construct hard negatives and Mao et al. (2022a) relies on human annotators to generate augmented positives, but the amount of generated data is limited. Differently, our method leverages additional supervision signals from the historical ground-truth passages to enhance the model’s history-awareness (e.g., enjoying the efficiency and avoiding the harmfulness).

## 3 Methodology

### 3.1 Task Definition

We are given a conversation session that contains the current query \( q_n \), and \( n - 1 \) historical turns preceding \( q_n \). The \( i \)-th historical turn is denoted as \( (q_i, p^*_i) \), where \( q_i \) is a historical query and \( p^*_i \) is the historical ground-truth passage corresponding to \( q_i \). Our task is to retrieve the passage \( p^*_n \) from a passage collection \( \mathcal{D} \) to satisfy the information need in \( q_n \). Our utilization of historical ground-truth passages \( \mathcal{P}_h = \{p^*_i\}_{i=1}^{n-1} \) is consistent with the settings adopted in previous work on conversational search (Choi et al., 2018; Qu et al., 2019), i.e. we assume that the relevant passages for the previous turns are known. In some real-world applications, if \( \mathcal{P}_h \) is not available, it can be replaced with a set of top-ranked passages for those turns. We discuss and analyze such adaptation in Sec. 4.5 for generalizability.

### 3.2 Method Overview

As illustrated in Figure 2, HAConvDR consists of three stages. The first stage is to generate the pseudo relevance judgments (PRJs) for historical turns by evaluating whether a given turn \( (q_i, p^*_i) \) is relevant to the current query \( q_n \). This is achieved by a pseudo-labeling approach presented in Sec. 3.3. In the second stage, we leverage the generated PRJs of historical turns for two purposes. The first purpose is to use the relevant historical turns to perform a context-denoised query reformulation (Sec. 3.4), while the second purpose is to create additional positive and negative training pairs by leveraging historical passages according to their PRJs. Given the reformulated queries and the augmented training pairs from conversation history, we train a dense retriever based on dual-encoder by history-aware contrastive learning in the third stage (Sec. 3.5). We highlight that, in our approach, the conversation history is considered as a source of not only context information, but also supervision signals. We describe each stage as follows.

### 3.3 Relevance Judgement for Historical Turns

A common practice to obtain a conversational dense retriever is to adapt models for ad-hoc retrieval to a conversational setting by concatenating the entire conversation history to the current query. In theory, the attention mechanism within the backbone transformer should allow the adapted retriever to implicitly conduct history modeling. In practice, however, the attention can be easily distracted by the irrelevant information in the conversation history. Therefore, we argue that it is essential to judge whether a historical turn is relevant to the current turn as part of the history modeling process.

In the literature of information retrieval, *relevance* is used to denote how well a document meets the information need of a query. Here, we take the
liberty of using the same term to describe whether a historical turn is relevant to the current query.

Learning to judge the relevance of historical turns is non-trivial because conversation datasets rarely contain such labels. Mo et al. (2023b) addresses this issue by adopting a simple and effective approach based on real impact on retrieval to derive pseudo labels – a historical query \( q_i \) is judged relevant if concatenating it to the current query \( q_n \) leads to an improved retrieval performance for \( q_n \) (similar to selecting query expansion terms as in Cao et al. (2008)). This pseudo-labeling approach is referred to as pseudo relevance judgment for historical turns. Despite the direct association with the retrieval task, this approach is limited by the fact that it only considers the queries in the historical turns, while ignoring the relevant or retrieved passages for them. To leverage the full conversational IR context, we also include the corresponding passages for each historical turn in our approach. We use a similar idea to Mo et al. (2023b) to label if a relevant passage \( p_i^r \) to a previous turn \( i \) is also relevant to the current turn, by assessing the impact of it on retrieval when it is concatenated, together with the historical query, to the current query.

The algorithm is illustrated in Algorithm 1. It divides the historical ground-truth passages \( \mathcal{P}_h = \{p_i^r\}_{i=1}^{n-1} \) into two disjoint groups:

\[
\mathcal{P}_h^+ = \{p_j^r\}_{j=1}^r, \quad \mathcal{P}_h^- = \{p_k^r\}_{k=1}^{n-r}
\]

where \( \mathcal{P}_h^+ \) denotes the relevant passage group and \( \mathcal{P}_h^- \) denotes the irrelevant passage group. For the use case where historical ground-truth passages are not available, we demonstrate that top-retrieved passages can serve as a substitute in Sec. 4.5.

**Algorithm 1 Generating pseudo relevance judgments for historical turns**

**Require:** current query \( q_n \), historical turn \( (q_i, p_i^r) \), retriever \( \phi \), retrieval evaluation metric \( M \)

1: RankList-raw \( \leftarrow \phi(q_n) \)
2: RankList-reform. \( \leftarrow \phi(q_n \circ q_i \circ p_i^r) \)
3: Score-raw \( \leftarrow M(\text{RankList-raw}) \)
4: Score-reform. \( \leftarrow M(\text{RankList-reform}) \)
5: if Score-reform. \( > \) Score-raw then
6: PRJ\((q_n, (q_i, p_i^r))\) \( \leftarrow \) relevant
7: else
8: PRJ\((q_n, (q_i, p_i^r))\) \( \leftarrow \) irrelevant
9: end if
10: Output PRJ\((q_n, (q_i, p_i^r))\)

### 3.4 Context-Denoised Query Reformulation

Based on the PRJs of historical turns derived in Sec. 3.3, we reformulate the current query \( q_n \) to obtain the context-denoised query \( q_n^r \):

\[
q_n^r = q_n \circ \cdots p_i^r \circ q_i \cdots
\]

where \( q_i \) and \( p_i^r \) are from relevant historical turns, and \( \circ \) denotes concatenation.

Since the reformulated query contains historical passages \( \mathcal{P}_h^+ \), a potential concern arises regarding the length of the reformulated query – it might exceed the input length limitations of some pre-trained language models. However, the analysis of the generated PRJ statistics, as presented later in Sec. 4.3, reveals that only a small portion of historical turns are deemed relevant and used for query reformulation. This indicates the practical feasibility of our approach. Nonetheless, in our future work, we will consider developing a more sophisticated mechanism to make a more strict selection of relevant passages in \( \mathcal{P}_h^+ \).
3.5 History-Aware Contrastive Learning

Contrastive learning is a prevalent approach to train dense retrievers (Karpukhin et al., 2020). This approach first projects queries and passages into an embedding space with dual encoders $F_Q$ and $F_P$. It then evaluates the relevance of any given pair of query and passage $(q, p)$ by taking the dot product similarity $S(q, p) = F_Q(q)^T F_P(p)$. Finally, supervision signals are derived from the positive and negative passages so that the distance between a query and a relevant passage (positive pair) should be closer than that between the same query and an irrelevant passage (negative pair). These supervision signals are back-propagated to train the encoders.

In a research setting, for the current query $q_n$, the relevant passage (positive passage) is the ground-truth passage $p_{n}^*$ given by the dataset. For the irrelevant passages (negative passages), one option is to simply take the passages other than $p_{n}^*$ found in the same training batch. These negative passages are referred to as in-batch negatives, here denoted as $\mathcal{P}_b^-$. In addition to in-batch negatives, another commonly adopted approach is to leverage retrieved hard negatives $\mathcal{P}_r^-$ (Lin et al., 2021; Kim and Kim, 2022; Karpukhin et al., 2020). One way to obtain such negatives is to use the top-ranked passages retrieved with $q_n$ by an off-the-shelf retriever (e.g., BM25) after removing $p_{n}^*$ (if present). Supervision signals generated from these retrieved negatives are believed to be more meaningful than those from in-batch negatives. The power of retrieved negatives suggests that the effectiveness of supervision signals could be heavily impacted by the quality and quantity of the positive and negative pairs.

Given the insight that augmenting positive and negative pairs can boost retrieval performance, we propose to mine additional pairs to further enhance the contrastive learning process. For this very purpose, we found the PRJs of historical turns derived in Sec. 3.3 come in handy.

Intuitively, $\mathcal{P}_h^+$ contains historical passages from the historical turns that are deemed relevant to $q_n$. Although $\mathcal{P}_h^+$ may not directly address the information need of $q_n$, $\mathcal{P}_h^+$ helps enhance or complement $q_n$. We believe this relationship can serve as a proxy to claim a certain level of relevance between $\mathcal{P}_h^+$ and $q_n$. Therefore, we use $\mathcal{P}_h^+$ as pseudo positives. Similarly, passages in $\mathcal{P}_h^-$ are less relevant to $q_n$ as demonstrated by the irrelevant PRJs. So we use $\mathcal{P}_h^-$ as additional negatives. More importantly, $\mathcal{P}_h^-$ resembles retrieved negatives $\mathcal{P}_r^-$ in the sense that both are hard negatives that can generate more meaningful supervisions. We refer to $\mathcal{P}_h^-$ as historical hard negatives.

By leveraging these pseudo positives and historical hard negatives mined from the conversation history, we upgrade traditional contrastive learning to history-aware contrastive learning. Formally, we denote the final positive and negative passages used for training as follows:

\[
\mathcal{P}_n^+ = \{p_n^*\} \cup \mathcal{P}_h^+, \quad |\mathcal{P}_n^+| = N \\
\mathcal{P}_n^- = \mathcal{P}_b^- \cup \mathcal{P}_r^- \cup \mathcal{P}_h^-, \quad |\mathcal{P}_n^-| = M
\]

The final training objective is illustrated in Eq. 4, where $p_i^+ \in \mathcal{P}_n^+$ and $p_j^- \in \mathcal{P}_n^-$.

\[
\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} e^{S(q_n, p_i^+)} + \sum_{j=1}^{M} e^{S(q_n, p_j^-)}
\]

4 Experiments

Datasets We evaluate our methods on two widely-used conversation datasets. The first is the TopiOCQA (Adlakha et al., 2022) dataset that contains complex topic-switch phenomena within each conversational session. These sessions have the potential to conceal a wealth of supervision signals in historical turns. The other dataset we use is QReCC (Anantha et al., 2021), where most queries in a conversational session are on the same topic. The selection of the datasets assures we verify the model performance on conversations with different intrinsic characteristics and enables more informative analyses. The statistics and more details of the datasets are provided in Appendix A.1.

Evaluation metrics For an adequate comparison with previous studies, we use four standard evaluation metrics: MRR, NDCG@3, Recall@10, and Recall@100 to evaluate the retrieval results.

Baselines We compare our method with two lines of conversational search approaches. The first line (CQR) performs conversational query reformulation based on generative rewriter models and off-the-shelf retrievers, including PLM-based GPT2+WS (Yu et al., 2020), QuReTeC (Voskarides et al., 2020), CQE-Sparse (Lin et al., 2021), T5QR (Lin et al., 2020), CONQRR (Wu et al., 2022), and ConvGQR (Mo et al., 2023a), and LLM-based
Table 1: Performance of different dense retrieval methods on two datasets. † denotes significant improvements with t-test at $p < 0.05$ over the main competitors, all CDR methods. Bold indicate the best results.

IterCQR (Jang et al., 2023), and LLM-Aided IQR (Ye et al., 2023). The second line (CDR) conducts conversational dense retrieval based on ad-hoc search dense retrievers to learn the latent representation of the reformulated query, including Conv-ANCE (Mao et al., 2023c) using the original contrastive ranking loss, InstructorR (Jin et al., 2023) utilizing LLMs to predict the relevance score between the session and passages then conduct the training of the retriever, ConvDR (Yu et al., 2021) relying also on human-rewritten queries as supervision signals and SDRConv (Kim and Kim, 2022) that includes mining additional hard negatives. The LLM-based methods employ ChatGPT or LLaMa as backbone models.

Implementation details The backbone model for conversational dense retriever training is ANCE (Xiong et al., 2020) and the dense retrieval is performed using Faiss (Johnson et al., 2019). During training, we only update the parameters of the query encoder while keeping the passage encoder frozen. The number of mined positives and negatives from historical turns can vary across different query turns. Instead of trying to utilize all of them, we randomly select one historical pseudo positive and one historical hard negative (along with the top retrieved hard negative) for each training instance to strike a balance between effectiveness and efficiency. More details are provided in Appendix A.2 and our code.1

4.1 Main Results

The main evaluation results on TopiOCQA and QReCC datasets are reported in Table 1.

We find that our method achieves a significantly better performance on both datasets compared with other methods on most metrics. In particular, it improves MRR by 10.7% and NDCG@3 by 8.0% on TopiOCQA over the second-best results ConvDR. The superior effectiveness can be attributed to the following two aspects. (1) The context-denoised query reformulation and history-aware contrastive learning with mined supervision signals enhance the ranking ability of our HAConvDR. (2) Conversational dense retrieval tends to be more effective compared with conversational query rewriting pipelines, including those leveraging the powerful generation capacity of LLMs. Besides, the improvements achieved over Conv-ANCE serve as additional validation of the effectiveness of exploiting supplementary supervision signals derived from ground-truth information of past interactions and confirm our underlying assumption.

Moreover, we find that performance improvements are more pronounced on TopiOCQA. This can be attributed to the characteristics of the datasets: the session context is longer in TopiOCQA, and contains more noise. This comparison indicates that our method has a greater potential for longer sessions with topic shifts. In contrast, the turns in QReCC are usually on the same topic, and the ground-truth passages of historical turns can also properly address the information needs of

---

1https://github.com/fengranMark/HAConvDR
the current query. In such a situation, most previous turns can be relevant, making it less critical to select the relevant turns. Notice that TopiOCQA provides a better simulation of real-world scenarios, where a conversation (or search) session is expected to be on related but different topics. Our results demonstrate that our approach is better at addressing this practical situation. More analysis on this is provided in Sec. 4.3 and Sec. 4.4.

4.2 Ablation Study

Compared to the contrastive learning technique in conversational dense retrieval, our proposed method introduces two extra components, i.e., context-denoised query reformulation and history-aware contrastive signals comprising historical pseudo positives and historical hard negatives. To assess the effectiveness of these individual components, we conduct an ablation study and present the analysis in Table 2.

We observe that, on both datasets, removing pseudo positives can cause a more pronounced performance degradation compared with removing hard negatives. This suggests that, although both hard negatives and pseudo positives are useful, the latter serves as a more effective supervision. This insight complements the currently prevalent studies on negative mining. On the other hand, we observe the decrease is more prominent on TopiOCQA, which is true for both removing hard negatives and pseudo positives. This can be attributed to the prevalence of topic-switch phenomena within the sessions in TopiOCQA, where historical supervision can and should be leveraged to boost performance as illustrated in our approach.

4.3 Investigation of PRJs of Historical Turns

The PRJs of historical turns are the foundation of context-denoised query reformulation and history-aware contrastive learning. The ablation study in Sec. 4.2 has shown the effectiveness of the approach. In this section, we take a deeper look to reveal the reasons behind the performance gain.

4.4 Impact of Historical Supervision Signals

We analyze how HACovDR alleviates the issue of models favoring retrieving historical turn passages over current ones by examining the effect of historical supervision signals.

Quantitative analysis The quantitative anal-
Table 3: Performance on TopiOCQA and QReCC for the adapted use case of historical ground-truth passage substitution. The $k$ denotes the top-k passages in pseudo relevance feedback.

![Figure 4](image1.png)

Figure 4: The percentage of the queries whose retrieved list has the ground-truth passage of the historical turns ranked higher than its own.

![Figure 5](image2.png)

Figure 5: T-SNE visualization of query, ground-truth passage, and pseudo positives and history hard negatives embeddings via two ANCE models with and without HAConvDR training.

The computation of PRJs for historical turns relies on having access to historical ground-truth passages $\{p^*_i\}_{i=1}^{n-1}$. In many real-world applications, identifying ground-truth passages can be accomplished by analyzing user clicks, engagement, and feedback. However, we acknowledge that there are applications where historical ground-truth passages are difficult to obtain. In such cases, we can use top-retrieved passages as a substitute. This simple substitution allows us to perform the proposed approach described in Sec. 3 with only minor modifications. Specifically, in Alg. 1 and Eq. 2, $p^*_i$ is approximated by the concatenation of the top-$k$ retrieved passages for $q_i$, where $k$ is a hyperparameter. This retrieval is completed with the same backbone model of the conversational dense retriever. Meanwhile, $P^-_{\text{b}} \cup P^-_{\text{r}}$. The rest of the approach is kept as is.

We conduct an ablation study to verify the effectiveness of our approach under this adaptation, with results presented in Table 3. We see the PRJ information still contributes to the retrieval performance of the reformulated query, further indicating its ef-
fectiveness. Besides, we find model performance degrades as $k$ increases, suggesting that longer contexts are more likely to contain noise, which cannot be entirely compensated by our approach. This suggests the potential for more advanced context-denoising approaches. Finally, we find that using the full model with history-aware contrastive learning under the adapted setting continues to yield better results on top-ranking positions and outperforms most existing systems in Table 1.

5 Conclusion

In this paper, we present a new history-aware contrastive learning strategy for conversational dense retriever training, HACConvDR, which is based on context-denoised query reformulation and additional supervision signals mining from historical turns. Extensive experimental results on public datasets demonstrate the effectiveness of our model. Furthermore, we conduct comprehensive analyses to gain insights into the impact of each component of HACConvDR on enhancing search performance and provide valuable insights on how they can work well for conversations with topic shifts.

Limitations

Our work demonstrates the feasibility of using historical ground-truth passages for query reformulation and contrastive supervision signals. Within our proposed HACConvDR, the context used for query reformulation includes selected historical passages, which are usually longer than hundreds of tokens. Thus, an explicit selection mechanism on raw text or an implicit fusion method on the latent representation could be designed to reduce the risk of information loss and the effect of noise. Besides, an LLM-aided mechanism could be designed for query reformulation, e.g., selecting part of each historical passage that is helpful and with less noise as better supervision signals. In addition, the historical supervised signals for model training might not be as important as the original annotation. Thus, a regulatory mechanism can be added to adjust the weight for pseudo positives within the history-aware conversational dense retrieval.

Acknowledgements

This work is supported by a discovery grant from the Natural Science and Engineering Research Council of Canada and a Talent Fund of Beijing Jiaotong University (2024JBRC005).

References


Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023a. Large language models know your contextual search intent: A prompting framework for conversational search.


Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023b. Learning to relate to previous turns in conversational search. In *29th ACM SIGKDD Conference On Knowledge Discover and Data Mining (SIGKDD)*.


A More Detailed Experimental Setup

A.1 Datasets

The statistics of each dataset are presented in Table 4 where we eliminate the samples without gold passages in QReCC. The details of each dataset are in the following:

**TopiOCQA** addresses the novel issue of topic switching, a common occurrence in realistic scenarios. In typical conversations, there are usually over 10 turns and a minimum of 3 topics. Furthermore, turns related to the same topic tend to have similar gold passages, thus we could leverage them as additional supervision signals.

**QReCC** primarily addresses the task of query rewriting by attempting to reformulate the query to approach the human-rewritten query. In comparison to TopiOCQA, QReCC involves conversations with a smaller number of turns, and most of these conversations revolve around the same topic. As a result, turns within the same conversation often yield identical gold passage results, making it possible to extract only a limited number of additional supervision signals.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Split</th>
<th>#Conv.</th>
<th>#Turns(Qry.)</th>
<th>#Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopiOCQA</td>
<td>Train</td>
<td>3,509</td>
<td>45,450</td>
<td>25M</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>205</td>
<td>2,514</td>
<td></td>
</tr>
<tr>
<td>QReCC</td>
<td>Train</td>
<td>10,823</td>
<td>29,596</td>
<td>54M</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>2,775</td>
<td>8,124</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Statistics of conversational search datasets.

A.2 Implementation Details

We implement all models by PyTorch (Paszke et al., 2019) and Huggingface’s Transformers (Wolf et al., 2019). The experiments are conducted on one Nvidia A100 40G GPU. For conversational dense retriever training, we use Adam optimizer with 3e-5 learning rate and set the batch size as 32. The maximum length of the reformulated query and the passage as model input is 512 and 384 for TopiOCQA and both 256 for QReCC, respectively. For the compared baseline systems, we implement the main competitor SDRConv with the same number of hard negatives and batch size as ours and use the ANCE+InstructoR QRPG version in InstructoR for fair comparison. All the dense retrievers are initiated with ANCE. For evaluation, we adopt the pytrec_eval tool (Van Gysel and...
B Qualitative Example

Table 5 presents a qualitative example corresponding to the T-SNE visualization in Figure 5, which gives a comprehensive understanding of how historical ground-truth passage can benefit current query retrieval as supervision signals.
Conversation (id:4-13)

q1: who sang all i want for christmas in 1995? (irrelevant)
p1: All I Want for Christmas Is You is a Christmas song by American singer-songwriter ... (536, -, -)
q2: who is she? (relevant)
p2: Mariah Carey (born March 27, 1969 or 1970) is an American singer-songwriter ... (5, 20, 17)
q3: what was her early days like? (irrelevant)
p3: Mariah Carey was born in Huntington, New York, on March 27, 1969 or 1970 ... (614, -, -)
q4: what are some famous songs she performed during 2010? (relevant)
p4: It missed out on the top one-hundred in the United Kingdom by one position ... (-, -, -)
q5: who composed the former mentioned one? (irrelevant)
p5: Cox plated the keyboard and percussion. The background vocals were sung by ... (-, -, -)
q6: how did it perform in the charts? (relevant)
p6: In the United States, Oh Santa! became a record-breaking entry on ... (-, -, -)
q7: how was it received critically? (relevant)
p7: Mike Diver of the BBC wrote that Oh Santa! is a “boisterous” song ... (-, -, -)
q8: what was her other song about? (irrelevant)
p8: Auld Lang Syne (The New Year’s Anthem) is a re-write of Auld Lang Syne ... (-, -, -)
q9: how was it received critically? (relevant)
p9: Auld Lang Syne (The New Year’s Anthem) garnered a negative response from critics ... (937, -, 322)
q10: what are some philanthropic activities this singer is associated with? (relevant)
p10: Carey is a philanthropist who has been involved with several ... (197, 502, 31)
q11: what does the latter mentioned foundation do? (relevant)
p11: The Make-A-Wish Foundation is a 501(c)(3) nonprofit organization founded in ... (-, -, -)
q12: what is her style of music? (relevant)
p12: Love is the subject of the majority of Carey’s lyrics, although she has written ... (6, 68, 29)
q13: what are some awards she has received?

Gold Passage (107, 68, 2)

Throughout her career, Carey has earned numerous awards and honors, including the World Music Awards’, Best Selling Female Artist of the Millennium, the Grammy Award for Best New Artist in 1991, and Billboard Special Achievement Award for the Artist of the Decade during the 1990s. In a career spanning over 20 years, Carey has sold over 200 million records worldwide, making her one of the best-selling music artists of all time. Carey is ranked as the best-selling female artist of the Nielsen SoundScan era, with over 52 million copies sold. Carey was ranked first in MTV and Blender magazine’s 2003 countdown of the 22 Greatest Voices in Music, and was placed second in Cove magazine’s list of The 100 Outstanding Pop Vocalists. Aside from her voice, she has become known for her songwriting.

Table 5: A qualitative example of how historical ground-truth passage can benefit current query retrieval as supervision signals within HACovDR. The brackets following each historical query indicate whether it is relevant or irrelevant to the current turn. The brackets with three numbers after each historical gold passage indicate its rank position by ANCE, Conv-ANCE, and our HACovDR within top-1000, where “-” means it is ranked outside the top-1000.