Interpreting Conversational Dense Retrieval by Rewriting-Enhanced Inversion of Session Embedding

Anonymous ACL submission

Abstract

Conversational dense retrieval has shown to be effective in conversational search. However, a major limitation of conversational dense retrieval is their lack of interpretability, hindering intuitive understanding of model behaviors for targeted improvements. This paper presents CONVINV , a simple yet effective approach to shed light on interpretable conversational dense retrieval models. CONVINV transforms opaque conversational session embeddings into explicitly interpretable text while faithfully maintaining their original retrieval performance as much as possible. Such transformation is achieved by training a recently proposed Vec2Text model (Morris et al., 2023) based on the ad-hoc query encoder, leveraging the fact that the session and query embeddings share the same space in existing conversational dense retrieval. To further enhance interpretability, we propose to incorporate external interpretable query rewrites into the transformation process. Extensive evaluations on three conversational search benchmarks demonstrate that CONVINV can yield more interpretable text and faithfully preserve original retrieval performance than baselines. Our work connects opaque session embeddings with transparent query rewriting, paving the way toward trustworthy conversational search. Our code is available at this anonymous repository.

1 Introduction

With the rapid development of language modeling, conversational search has emerged as a novel search paradigm and is garnering more and more attention. Different from the traditional ad-hoc search paradigm characterized by keyword-based queries and "ten-blue" links (Yu et al., 2020), conversational search empowers users to interact with the search engine through multi-turn natural language conversations to seek information, which brings a more intuitive and efficient search experience (Mao et al., 2022b; Gao et al., 2022).

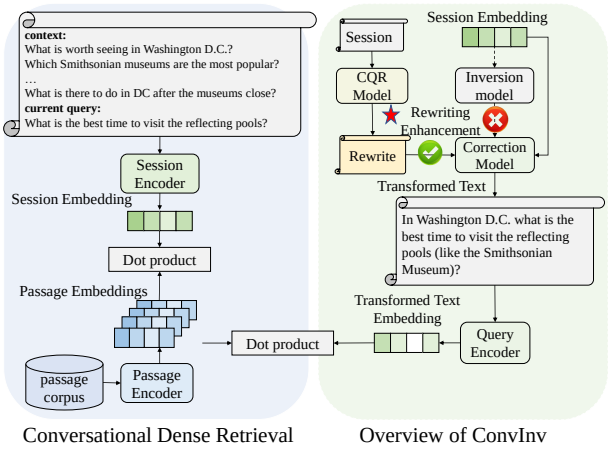


Figure 1: The blue section on the left signifies the conversational dense retrieval, and the green section on the right provides an overview of CONVINV.

In conversational search, the system input is a multi-turn natural language conversation, which may have many linguistic problems such as omissions, co-references, and ambiguities (Radlinski and Craswell, 2017), posing great challenges for accurately grasping the user's real information needs. Recently, conversational dense retrieval (CDR) (Yu et al., 2021; Lin et al., 2021; Kim and Kim, 2022; Mao et al., 2022a; Qian and Dou, 2022; Mo et al., 2023b), which directly encodes the whole conversational search session and the passages into a unified embedding space to perform matching, has shown to be a promising method to solve this complex search task. Compared to another type of method: conversational query rewriting (CQR) (Lin et al., 2020; Vakulenko et al., 2021a; Wu et al., 2022; Mo et al., 2023a), which is a two-step method that first reformulates the search session into a context-independent query rewrite and then feeds it into existing ad-hoc search models for search, the end-to-end CDR models can be directly optimized towards better search effectiveness (Yu et al., 2021) and is more efficient as it avoids the extra latency

caused by the rewriting step.

However, a notable drawback of conversational dense retrieval is that it inherently lacks interpretability (Mao et al., 2023d). By encoding conversations into dense vector embeddings rather than readable text, it becomes opaque how these CDR models comprehend search intent. The absence of interpretability becomes a severe obstacle for developers to comprehend the reasons behind the search results, hindering effective and targeted enhancements to the bad cases of the models (Mao et al., 2023d,a). Moreover, the absence of interpretability poses challenges in identifying and addressing potential biases or errors within the models, which could lead to unfair or misleading search results without the possibility of timely correction.

In this paper, we present CONVINV : a simple and effective approach aiming to shed light on the opacity problem of conversational dense retrieval. CONVINV demystifies the opaque conversational session embeddings by transforming them into explicitly interpretable text while faithfully maintaining their retrieval performance as much as possible. This transformation allows us to intuitively decipher the characteristics of behaviors of different conversational dense retrieval models.

Figure 1 provides an overview of CONVINV. Specifically, our approach is based on the recently proposed Vec2Text (Morris et al., 2023), which is a powerful method that can invert any text embedding into its original text given the corresponding text encoder. However, inverting the session embedding into the original session is meaningless as it brings no interpretability. We adapt Vec2Text to suit our interpretable inversion of conversational session embedding by taking specific advantage of how the conversational session encoders are trained: the session encoder starts from an ad-hoc query encoder and the passage encoder is frozen during the training. This makes the session and query embeddings finally share the same embedding space for retrieval. Therefore, we propose to train a Vec2Text model based on the ad-hoc query encoder to transform the session embedding so that the transformed text is different from the original session, but also maintains a similar retrieval performance when encoding it with the ad-hoc query encoder. To further enhance the interpretability of the transformed text, we directly incorporate well-interpretable external query rewrites into the Vec2Text transformation process, effectively guiding it to yield more interpretable text.

We conduct extensive evaluations on three conversational search benchmarks. Compared to baselines, the proposed CONVINV can transform conversational session embeddings into more interpretable text as well as faithfully restore the original retrieval performance of the session embeddings.

In summary, the contributions of our work are:

(1) We introduce a simple and effective approach CONVINV to shed light on the interpretability of conversational dense retrieval models by transforming opaque conversational session embeddings into interpretable text as well as faithfully maintain their original retrieval performance.

(2) We propose to incorporate the query rewrites into the transformation process to effectively enhance the interpretability of the transformed text.

(3) Our work connects opaque session embeddings with transparent query rewriting, paving the way toward trustworthy conversational search.

2 Related Work

2.1 Conversational Search

Currently, conversational search primarily relies on two main methods: conversational query rewriting (CQR) and Conversational dense retrieval (CDR). CQR (Yu et al., 2020; Wu et al., 2022; Kumar and Callan, 2020; Voskarides et al., 2020; Lin et al., 2020; Mao et al., 2023b; Liu et al., 2021; Vakulenko et al., 2021a,b; Mao et al., 2023c; Mo et al., 2023a) transforms the whole session into a context-independent query. The generated query rewrites can directly perform ad-hoc retrieval. In contrast, CDR (Yu et al., 2021; Mo et al., 2024; Krasakis et al., 2022; Mao et al., 2022a; Mo et al., 2023b; Mao et al., 2023d, 2022b; Dai et al., 2022) aims to train a session encoder that is capable of encoding the conversational context into a high-dimensional space for conducting dense retrieval. However, the session embedding encoded by the conversational query encoder lacks interpretability, hindering developers from comprehending the retrieval results obtained by the search engine.

2.2 Interpretable information retrieval

The interpretability issues have raised more and more attention in the field of information retrieval. (Ram et al., 2023) proposed to interpret the session embeddings produced by dual encoders by projecting them into the model's vocabulary space. (Mao et al., 2023d) proposed to augment the SPLADE model by incorporating multi-level denoising ap-

proaches, which can produce denoised and interpretable lexical session representations.

To explore the intricate interplay between embedded representations and their textual counterparts, a substantial body of research has focused on the task of inverting embeddings to coherent text. Representing the embedding of sentences as the initial token, (Li et al., 2023) trained a powerful decoder model to decode the entire sequence. (Morris et al., 2023) endeavored to produce text whose embedding closely approximates the given embedding. They achieved this by using the difference between hypothesis embeddings and actual embeddings.

3 Methodology

In this work, we present CONVINV, a new approach designed to demystify conversational session embeddings. Our approach focuses on transforming these opaque conversational session embeddings into explicitly interpretable text while maintaining their retrieval performance as much as possible. CONVINV aims to bridge the gap between the mysterious nature of dense embeddings and the necessity for clear, understandable insights in conversational search intent analysis.

3.1 Preliminaries

3.1.1 Conversational dense retrieval

Formally, conversational search involves a series of turns $\{(q_i, a_i)\}_{i=1}^{n}$, where the users express their information needs at i-th turn through q_i, and the system returns a relevant response a_i. This paper focuses on the conversational retrieval task, where the goal of conversational search models is to retrieve relevant passages p for the current query q_i, considering its historical context $H_i = \{(q_j, a_j)\}_{j=1}^{i-1}$. The idea of conversational dense retrieval is to jointly map the current query q_i along with the historical context H_i and passages into a unified embedding space, and use the similarity between the session embedding and the passage embedding as the retrieval score:

$$\mathbf{s_i} = E_{\text{s}}(q_i, H_i), \quad \mathbf{p} = E_{\text{p}}(p), \tag{1}$$

$$r = \cos(\mathbf{s_i}, \mathbf{p}), \tag{2}$$

where E_{s} and E_{p} are the session and passage encoders, respectively. cos is the cosine similarity used to compute the retrieval score r.

3.1.2 Task formulation

The encoded conversational session embedding $\mathbf{s_i}$, while effective, is inherently mysterious and lacks interpretability. Our goal is to transform the session embedding $\mathbf{s_i}$ into an explicit, interpretable text $\hat{q}_i$ while faithfully maintaining the original retrieval effectiveness of the session embedding in $\hat{q}_i$.

3.2 Our Approach

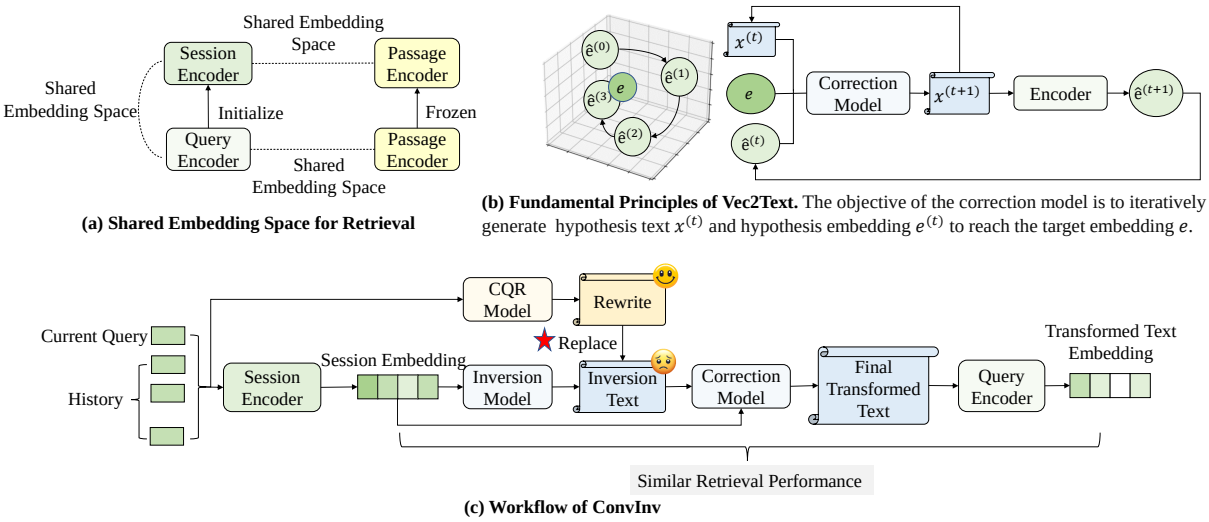
To achieve this transformation from session embeddings to interpretable text, we propose a simple yet effective approach, called CONVINV , which is built upon the Vec2Text model (Morris et al., 2023) with tailored adjustments for the interpretation of conversational dense retrieval. Specifically, our approach has two important steps: (1) Training a Vec2Text model based on the ad-hoc query encoder. (2) Enhancing interpretation with rewriting. Figure 2 shows an illustration of our approach.

3.2.1 Training Vec2Text based on Ad-hoc Query Encoder

Vec2Text (Morris et al., 2023) is a recently proposed method for transforming embeddings into text. Given any text encoder E and a large collection of texts $T = \{t_i\}$ where t_i is a text, a Vec2Text model ϕ is trained based on a large number of (embedding, text) pairs (i.e., $\langle E(t_i), t_i \rangle$) to learn to invert any text embedding $E(t_i)$ into a text $t_i^{'}$, where $E(t_i^{'})$ is very similar to $E(t_i)$. As reported in their original paper, $\cos(E(t_i^{'}), E(t_i))$ can reach up to 0.99. Motivated by the remarkable effectiveness of Vec2Text, we adapt it to suit our interpretable inversion of conversational session embedding by leveraging a specific training characteristic of conversational session encoders: *Shared Embedding Space for Retrieval*.

Shared embedding space for retrieval. For the training of conversational dense retrievers, it is common to initialize the conversational session encoder and the passage encoder from a pre-trained ad-hoc retriever, and only fine-tune the session encoder while freezing the passage encoder for facilitating the training (Yu et al., 2021; Lin et al., 2021; Mao et al., 2022a; Mo et al., 2023b). Therefore, we may assume that the session encoder and the ad-hoc query encoder share the same embedding space for retrieval as they share the same passage encoder. This characteristic is ideal for us to achieve more interpretable session embedding inversion as well as maintain its original retrieval effectiveness.

Interpretable query generation. For a session encoder E_{s} fine-tuned from an ad-hoc query encoder E_{q}, we train a Vec2Text model ϕ_{q} based on E_{q} but not based on E_{s}. Then, for a session em-



(a) Shared Embedding Space for Retrieval

(b) Fundamental Principles of Vec2Text. The objective of the correction model is to iteratively generate hypothesis text $x^{(t)}$ and hypothesis embedding $e^{(t)}$ to reach the target embedding e.

(c) Workflow of ConvInv

Figure 2: Architecture of our proposed ConvInv.

bedding $\mathbf{s_i} = E_s(q_i, H_i)$, we obtain its transformed text $\hat{q}_i = \phi_q(\mathbf{s}_i)$ through ϕ_q. Specifically, Vec2Text includes two models: the inversion model and the correction model, and the generation process of Vec2Text includes two steps: (1) The initial inversion step, where an inversion model first inverts the embedding into an initial inverted text t^{inv}. (2) The correction step, where a correction model then progressively refines this initial inverted text t^{inv} to be more accurate. Figure 2 shows an illustration of the whole generation process of Vec2Text. The detailed introduction of our Vec2Text model training is provided in Appendix A.1.

Since E_s and E_q share the same retrieval embedding space, the transformed query embedding $E_q(\hat{q}_i)$ is supposed to be highly similar to the original session embedding $\mathbf{s_i}$ and thus keep similar retrieval performance.

3.2.2 Interpretability Enhancement with Conversational Query Rewriting

While the transformed text $\hat{q}_i$ can attain retrieval performance comparable to that of the original session embedding $\mathbf{s_i}$ when encoded by the ad-hoc query encoder E_q, there is no assurance that $\hat{q}_i$ will form a coherent and interpretable sentence for human understanding.

We propose a simple method to leverage external query rewrites to enhance the interpretability. Specifically, we first employ a conversational query rewriting model R (for example, the T5QR (Lin et al., 2020) model) to transform the conversational search session $\{q_i, H_i\}$ into a standalone query rewrite $q_i^* = R(q_i, H_i)$. Then, in the generation process of Vec2Text, we discard the initial inversion process and directly use the query rewrite q_i^* as the initial inverted text t^{inv}.

The rewriting model R, trained on a vast dataset of human-crafted rewrites, ensures that the resultant query rewrite is coherent and understandable compared to the original inverted text produced by VecText's inversion model. The new inverted text q_i^*, serving as an improved starting point for the session embedding transformation, can help lead the whole generation process towards a more interpretable direction, and thus enhance the interpretability of the final transformed text $\hat{q}_i$.

4 Experimental Settings

This section presents our basic experimental settings. See Appendix A.2 for full details.

4.1 Datasets

We use four public conversational search datasets: QReCC (Anantha et al., 2021), TREC CAsT-19 (Dalton et al., 2020), TREC CAsT-20 (Dalton et al., 2021a), and TREC CAsT-21 (Dalton et al., 2021b). The QReCC dataset consists of 13.6K conversations, with an average of 6 turns per conversation. While the three CAsT datasets (19, 20, 21) only comprise 50, 25, and 26 conversations, respectively, but with more detailed relevance labeling. All four datasets provide human rewrites for each turn. Following existing works (Mao et al., 2023d; Mo et al., 2023a), we train CDR models on the QReCC dataset and conduct evaluations on the three CAsT datasets.

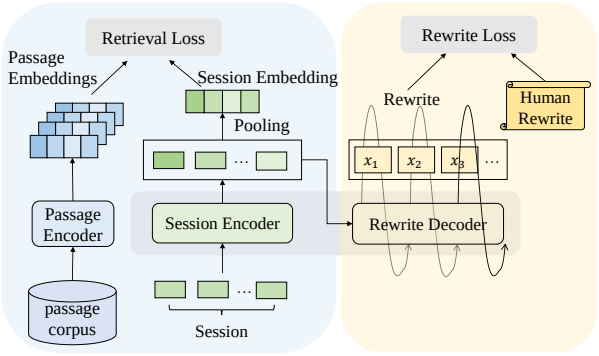


Figure 3: The workflow of UniCRR (**Uni**fying **C**onversational Dense **R**etrieval and Query **R**ewriting).

4.2 Conversational Dense Retrieval Models

Currently, there are mainly two paradigms to train conversational session encoders. The first is proposed by Yu et al. (2021) which employs an ad hoc query encoder as the teacher and learns the student session encoder by mimicking the teacher embeddings originating from human queries. The second is to use the classical ranking loss function (Karpukhin et al., 2020; Lin et al., 2021) to maximize the distance between the session and its positive passages and minimize the distance between the session and negative passages.

Our evaluation is based on both types of CDR models. We name the first type KD-*Retriever* and the second type Conv-*Retriever*, where *Retriever* can be replaced with any base ad-hoc retriever. Specifically, we mainly experiment with a popular ad-hoc retriever, i.e., GTR (Ni et al., 2022), and we investigate the universality of our method to different ad-hoc retrievers in Section 5.3.

4.3 Baselines

Our main goal is to demonstrate the interpretability and preserved retrieval performance of the transformed text generated by our ConvInv , compared to the original session embeddings of **KD-GTR** and **Conv-GTR**. To the best of our knowledge, there is no existing method that is completely suitable for our task, i.e., interpreting conversational session embeddings (see the full task definition in Section 3.1.2). Therefore, we propose a straightforward but strong baseline called **UniCRR**. Figure 3 illustrates UniCRR. Specifically, we unify the session encoder and the query rewriter in an encoder-decoder architecture and adopt multi-task learning to simultaneously train both. As such, the rewrite generated from the decoder part can interpret the session embedding generated from the encoder part to some extent.

In addition to the original KD-GTR, Conv-GTR, and our proposed UniCRR, we also use the following conversational search baselines mainly for the comparisons of retrieval performance: (1) **T5QR** (Lin et al., 2020): A conversational query rewriter based on T5 (Raffel et al., 2020), trained using human-generated rewrites. (2) **ConvGQR** (Mo et al., 2023a): A framework for query reformulation that integrates query rewriting with generative query expansion. (3) **LeCoRE** (Mao et al., 2023d): A conversational lexical retrieval model extending from the SPLADE model with two well-matched multi-level denoising methods.

4.4 Evaluation Metrics

Retrieval and inversion evaluation. Following existing works (Mo et al., 2023a; Mao et al., 2023d) and the official settings of the CAsT datasets (Dalton et al., 2021a), we choose MRR, NDCG@3, and Recall@100 to evaluate the retrieval performance. We use two metrics to quantify the fidelity of the embedding inversion: (1) The absolute difference in the retrieval performances between using the session embeddings and the transformed text. (2) Following Vec2Text (Morris et al., 2023), we also calculate the cosine similarity between the session embeddings and the transformed text embeddings.

Interpretability evaluation. We conduct a human evaluation for the interpretability of the transformed text from three aspects: (1) *Clarity*: evaluating the clarity of text expression and identifying the presence of ambiguity or vague expressions; (2) *Coherence*: examining the logical structure of the text; (3) *Completeness*: determining the extent to which the text comprehensively covers all historical information. Five information retrieval researchers are employed to assign scores ranging from 1 to 5. A larger score indicates better performance.

4.5 Implementations

For ConvInv , we train Vec2Text models on the large-scale MSMARCO (Nguyen et al., 2016) query and passage collections based on different ad-hoc query encoders. The inversion model is trained for 50 epochs with a batch size of 128 and the correction model is trained for 100 epochs with a batch size of 200 with 1e-3 learning rate. The maximum sequence length is set to 48. By default, we use the rewrites generated by T5QR to perform rewriting enhancement.

Method	CAsT-19			CAsT-20			CAsT-21		
	MRR	NDCG@3	Recall@100	MRR	NDCG@3	Recall@100	MRR	NDCG@3	Recall@100
T5QR	65.8	41.9	38.2	46.6	32.1	41.4	47.9	34.1	45.2
ConvGQR	66.7	39.3	33.7	39.7	25.9	33.8	40.6	25.3	37.3
LeCoRE	70.3	42.2	**49.4**	45.0	29.0	46.7	54.8	32.3	38.7
Conv-GTR	53.8	31.0	34.6	27.9	18.4	31.8	42.2	28.4	46.4
UniCRR	54.4 (+0.6)	31.9 (+0.9)	31.0 (-3.6)	36.0 (+8.1)	23.7 (+5.3)	33.3 (+1.5)	35.0 (-7.2)	23.2 (-5.2)	31.3 (-15.1)
ConvInv	56.4 (+2.6)	33.1 (+2.1)	37.0 (+2.4)	27.2 (-0.7)	18.5 (+0.1)	30.4 (-1.4)	41.9 (-0.3)	28.2 (-0.2)	41.7 (-4.7)
KD-GTR	**74.9**	**46.9**	41.9	**49.5**	**35.9**	**46.9**	**54.7**	36.4	**55.4**
UniCRR	65.1 (-9.8)	40.6 (-6.3)	37.0 (-4.9)	44.4 (-5.1)	32.3 (-3.6)	39.5 (-7.4)	41.0 (-13.7)	27.3 (-9.1)	39.5 (-15.9)
ConvInv	74.2 (-0.7)	44.9 (-2.0)	43.0 (+1.1)	47.6 (-1.9)	34.4 (-1.5)	44.0 (-2.9)	**54.7** (+0.0)	**37.4** (+1.0)	55.1 (-0.3)

Table 1: Retrieval performance comparisons. Our main competitor is UniCRR. The numbers in parentheses indicate the absolute difference between the original CDR model (i.e., Conv-GTR or KD-GTR) and the transformed text. In the comparison between ConvInv and UniCRR, a green background indicates that its performance gap with the original session embedding is smaller compared to its counterpart, while a red background indicates a larger gap. The best performance is bold.

Method	CAsT-19			CAsT-20			CAsT-21		
	MRR	NDCG@3	Recall@100	MRR	NDCG@3	Recall@100	MRR	NDCG@3	Recall@100
Conv-GTR	53.8	31.0	34.6	27.9	18.4	31.8	42.2	28.4	46.4
TX-Inversion	58.0(+4.2)	33.5 (+2.5)	37.1(+2.5)	28.2(+0.3)	18.8(+0.4)	29.5(-2.3)	40.7(-1.5)	26.6(-1.8)	43.1(-3.3)
TX-Human	55.6(+1.8)	33.0(+2.0)	36.0(+1.4)	27.3(-0.6)	18.5(+0.1)	30.9(-0.9)	42.6(+0.4)	26.5(-1.9)	41.1(-5.3)
ConvInv	56.4 (+2.6)	33.1 (+2.1)	37.0 (+2.4)	27.2 (-0.7)	18.5 (+0.1)	30.4 (-1.4)	41.9 (-0.3)	28.2 (-0.2)	41.7 (-4.7)
KD-GTR	**74.9**	**46.9**	41.9	**49.5**	**35.9**	**46.9**	**54.7**	36.4	55.4
TX-Inversion	71.6(-3.3)	44.2(-2.7)	42.3(+0.4)	48.1(-1.4)	33.6(-2.3)	44.8(-2.1)	53.8(-0.9)	36.1(-0.3)	**55.6**(+0.2)
TX-Human	73.1(-1.8)	44.1(-2.8)	42.3 (+0.4)	48.6(-0.9)	35.0(-0.9)	45.9(-1.0)	53.1(-1.6)	35.8(-0.6)	54.3(-1.1)
ConvInv	74.2 (-0.7)	44.9 (-2.0)	**43.0** (+1.1)	47.6 (-1.9)	34.4 (-1.5)	44.0 (-2.9)	**54.7** (+0.0)	**37.4** (+1.0)	55.1 (-0.3)

Table 2: Ablation results of the effect of rewriting-enhancement. The numbers in parentheses indicate the difference between the original (i.e., Conv-GTR or KD-GTR) and the transformed text. ConvInv uses T5QR for rewriting enhancement by default. In the comparison between TX-Inversion, TX-Human, and ConvInv , a green background indicates that its performance gap with the original session embedding is the smallest. The best performance is bold.

We train the conversational dense retrieval models on the QReCC dataset. The session encoder is initialized from an ad-hoc query encoder and the passage encoder is frozen during training. The input of the session encoder is the concatenation of all historical turns and the current query following existing works (Mao et al., 2023d; Mo et al., 2023a). For KD-Retriever, we follow Yu et al. (2021) using the Mean Squared Error (MSE) loss function to perform knowledge distillation. For Conv-Retriever, we use the contrastive ranking loss function with 48 batch size. The maximum input lengths of the session encoder and the passage encoder are set to 512 and 384, respectively. We generally train 2 epochs with 5e-5 learning rate for CDR models.

5 Experimental Results

5.1 Retrieval and Inversion Evaluation

Note that our work does not aim to achieve absolutely higher retrieval performance, but rather to faithfully restore the retrieval performance of the original session embeddings, so the main competitor of our ConvInv is only UniCRR. The retrieval performance comparisons on three CAsT datasets are shown in Table 1 and the similarity is shown in Table 3. We find:

Method	CAsT-19	CAsT-20	CAsT-21
UniCRR	94.10	92.80	87.90
ConvInv	**95.80**	**95.20**	**94.50**

Table 3: The similarity between the embeddings of texts generated by UniCRR and ConvInv, and the original session embeddings. The best performance is bold.

(1) Compared to UniCRR, **ConvInv achieves superior embedding restoration**. For example, for KD-GTR, the average absolute differences for ConvInv are 0.87 (MRR), 1.5 (NDCG@3), and 1.43 (Recall@100), and the average absolute differences for UniCRR are 9.53 (MRR), 6.3 (NDCG@3), and 9.4 (Recall@100). This indi-

cates that the transformed texts generated by CONVINV are closer to the original session embeddings. This aligns with the restoration similarity, which is shown in Table 3. The superior reconstruction performance of ConvInv compared to UniCRR may stem from the fact that UniCRR fails to establish a direct correlation between session embeddings during both the training and inference phases.

(2) We surprisingly notice that **the transformed text generated by CONVINV can sometimes even yield slightly better retrieval performance**. For example, on the CAsT-21 dataset, we observe 2.7% NDCG@3 relative gains over the original session embedding, respectively. This discovery could potentially pave the way for enhancing retrieval efficacy and interpretability through the collaborative optimization of CQR and CDR.

Ablation study for rewriting enhancement. We propose using external query rewrites generated by T5QR to improve the interpretability of transformed text, which matches the original session embedding's retrieval performance but may lack coherence and understandability. Building on this proposition, we compare three types of transformed text to investigate the effect of rewriting enhancement: (1) using T5QR rewrites for the rewriting enhancement, which is the default CONVINV. (2) *TX-Human*: using human rewrites for the rewriting enhancement. (3) *TX-Inversion*: not performing rewriting enhancement (i.e., just using the text generated by the inversion model for the correction step). The ablation results of the retrieval performance of transformed text are shown in Table 2. We observe that the utilization of rewriting enhancement brings the retrieval performance closer to the original. **Using rewriting enhancement generally leads to stronger overall retrieval performance compared to not**.

5.2 Interpretability Evaluation

We manually evaluate the interpretability of three types of transformed text generated by CONVINV. Evaluation results are shown in Figure 4 and we have the following observations:

(1) Using query rewrites as the initial inverted text improves the interpretability of the transformed text of KD-GTR and Conv-GTR across the CAsT-19, CAsT-20, and CAsT-21 datasets. This improvement can be attributed to the introduction of the rewrite as the initial inverted text, which essentially offers the corrector model a more informative and clear starting point. These notable enhancements underscore the necessity of our rewriting-enhancement approach in improving text interpretability.

Context: (CAsT-19 Session 54) q_1: What is worth seeing in Washington D.C.? ... q_4: Is the spy museum free? q_5: What is there to do in DC after the museums close?
Current Query(68.1): What is the best time to visit the reflecting pools?
CONVINV (68.1): In Washington D.C. what is the best time to visit the reflecting pools (like the Smithsonian Museum)? **TX-Human**(47.9): In Washington D.C., what is the best time to visit the reflecting pools by the Smithsonian and other DC museums? **TX-Inversion**(20.2): In Washington D.C., what is the best time to visit the reflecting pools (e.g. Smithsonian National Museum)?
Human Rewrite(36.1): What is the best time to visit the reflecting pools in Washington D.C.?

Table 4: A case illustrating the distinction in utilizing rewriting enhancement for transformed text. The numbers in parentheses indicate the retrieval performance NDCG@3 of the transformed text. Notably, the number in parentheses under **Current Query** represents the retrieval results of the original session embedding, not that of the current query statement.

(2) For both KD-GTR and Conv-GTR, the human evaluation scores of transformed text on CAsT-19 are higher, whether using rewriting-enhancement or not, compared to CAsT-20 and CAsT-21. This observation may be attributed to the absence of response information in the CAsT-19 dataset, which exclusively contains query content. Consequently, the session embedding on CAsT-19 is relatively simple, lacking the complexity introduced by response data.

(3) The lower human evaluation scores of transformed text for Conv-GTR compared to KD-GTR on three datasets may be due to the implications of contrastive learning. This method often introduces additional noise. Therefore, Conv-GTR's session embedding might be more prone to interference, potentially leading to its less effective performance in generating transformed text.

We provide a concrete example of the transformed texts in Table 4. More case studies are in Appendix A.4. We find that the transformed text CONVINV not only exhibits high interpretability, fully capturing the user's query intent about "in Washington D.C.", but also maintains the closest proximity of retrieval performance to the original session embedding. We notice that it includes an

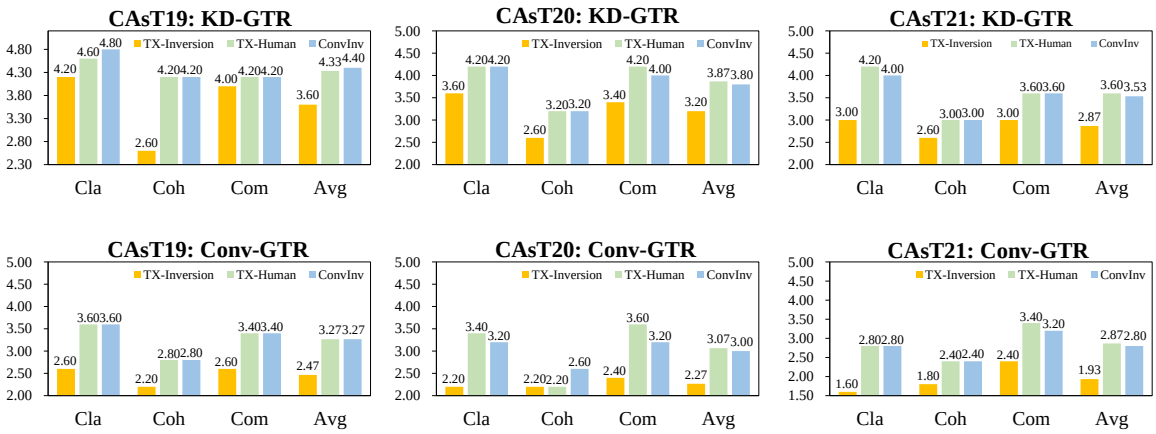


Figure 4: Results of human evaluations for interpretability. Cla, Coh, and Com represent Clarity, Coherence, and Completeness, respectively. The Avg indicates the average of these scores.

Retriever	Method	CAsT-19					CAsT-21				
		Retrieval Performance			Interpretablity		Retrieval Performance			Interpretablity	
		MRR	NDCG@3	Recall@100	similarity	hum eval	MRR	NDCG@3	Recall@100	similarity	hum eval
GTR	KD-GTR	**74.9**	**46.9**	41.9	-	-	54.7	36.4	55.4	-	-
	ConvInv	74.2 (-0.7)	44.9 (-2.0)	**43.0** (+1.1)	0.985	4.40	54.7(0.0)	37.4(+1.0)	55.1(-0.3)	0.945	3.53
ANCE	KD-ANCE	72.0	44.4	34.2	-	-	52.8	36.9	50.8	-	-
	ConvInv	72.0(0.0)	44.5(+0.1)	34.3(+0.1)	0.999	4.90	55.8(+3.0)	37.4(+0.5)	53.1(+2.3)	0.998	4.07
BGE	KD-BGE	69.5	44.0	41.2	-	-	57.9	**41.2**	**56.0**	-	-
	ConvInv	69.9(+0.4)	45.4(+1.4)	41.5(+0.3)	0.972	4.33	**59.8**(+1.9)	41.1(-0.1)	54.4(-1.6)	0.954	4.25

Table 5: Retrieval performance and interpretability of generated transformed text based on different ad-hoc retrievers on CAsT-19 and CAsT-21 datasets. The "hum eval" represents the human evaluation score. The numbers in parentheses indicate the difference between the original and the transformed text. The best performance is bold.

additional clue "(like the Smithsonian Museum)" in the query, which may just be additional knowledge reflected in the mysterious session embedding that can help retrieve passages about famous attractions in Washington D.C.

5.3 Experiments with Different Retrievers

We investigate the universality of our ConvInv by changing the base ad-hoc retriever of the CDR models. Specifically, we experiment with another two popular ad-hoc retrievers: ANCE (Xiong et al., 2021) and BGE (Xiao et al., 2023). Results are shown in Table 5. We find that:

(1) Regardless of the selected ad-hoc retriever, both retrieval similarity and text similarity metrics are observed to be high. To illustrate, on the CAsT-19 dataset, the average absolute differences for KD-ANCE on CAsT-19 dataset are 0.0 (MRR), 0.1 (NDCG@3), and 0.1 (Recall@100), and the cosine similarity is up to 99.9%.

(2) Across both CAsT-19 and CAsT-21 datasets, there is a sustained consistency between similarity scores and human evaluations, indicating that textual similarity is a reliable indicator of quality as perceived by human judges. However, this does not encapsulate all the factors considered in human evaluations, especially as similarity scores remain robust while human evaluations show a decline from CAsT-19 to CAsT-21. Although there is a noted decrease in human evaluation scores across all methods when moving from CAsT-19 to CAsT-21, the similarity scores remain high or even show marginal improvement.

6 Conclusion

In this paper, we present a novel approach ConvInv to shed light on the interpretability of conversational dense retrieval. By experimenting with two typical conversational dense retrieval models on three conversational search benchmarks, we demonstrate the effectiveness of our approach in providing interpretable text as well as faithfully restoring the original retrieval performance of session embeddings. Our work not only enhances interpretability in conversational dense retrieval but also lays a groundwork for future research toward trustworthy conversational search.

Limitations

Our work provides a simple but effective solution to enhance the interpretability of conversational dense retrieval models, bridging the gap between opaque session embeddings and transparent query rewriting. However, the necessity to train distinct Vec2Text models based on various retrievers demands a significant time investment. Additionally, for session embeddings trained using contrastive learning, the transformed text fails to achieve sufficiently high similarity to the original session embedding, suggesting an incomplete decoding of the session embedding. Besides, some of the transformed texts may not exhibit retrieval performance as effective as the original session embeddings. Some more sophisticated conversational dense retrievers have not been investigated.

References

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 520–534. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC cast 2019: The conversational assistance track overview. *CoRR*, abs/2003.13624.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021a. Cast 2020: The conversational assistance track overview. In *In Proceedings of TREC*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021b. TREC cast 2021: The conversational assistance track overview. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Sungdong Kim and Gangwoo Kim. 2022. Saving dense retriever from shortcut dependency in conversational search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10278–10287. Association for Computational Linguistics.

Antonios Minas Krasakis, Andrew Yates, and Evangelos Kanoulas. 2022. Zero-shot query contextualization for conversational search. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1880–1884. ACM.

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, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3971–3980. Association for Computational Linguistics.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14022–14040. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Frassetto Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *CoRR*, abs/2004.01909.

Hang Liu, Meng Chen, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Conversational query rewriting with self-supervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7628–7632. IEEE.

Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023a. Search-oriented conversational query

editing. In *ACL (Findings)*, volume ACL 2023 of *Findings of ACL*. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023b. Search-oriented conversational query editing. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4160–4172. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023c. Large language models know your contextual search intent: A prompting framework for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1211–1225. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022a. Curriculum contrastive context denoising for few-shot conversational dense retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 176–186. ACM.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022b. Convtrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2935–2946. Association for Computational Linguistics.

Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023d. Learning denoised and interpretable session representation for conversational search. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3193–3202. ACM.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023a. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4998–5012. Association for Computational Linguistics.

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 *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1722–1732. ACM.

Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024. History-aware conversational dense retrieval. *arXiv preprint arXiv:2401.16659*.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12448–12460. Association for Computational Linguistics.

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. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9844–9855. Association for Computational Linguistics.

Hongjin Qian and Zhicheng Dou. 2022. Explicit query rewriting for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4725–4737. Association for Computational Linguistics.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, pages 117–126. ACM.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What are you token about? dense retrieval as distributions over the vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2481–2498. Association for Computational Linguistics.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021a. Question rewriting for conversational question answering. In *WSDM '21,*

The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021, pages 355–363. ACM.

Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021b. A comparison of question rewriting methods for conversational passage retrieval. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 418–424. Springer.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 921–930. ACM.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10000–10014. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul N. 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 2020, Virtual Event, China, July 25-30, 2020*, pages 1933–1936. ACM.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 829–838. ACM.

A Appendix

A.1 Vec2Text

Due to the necessity of transforming session embeddings into explicit and interpretable text, we integrate the Vec2Text model into our architecture. The utilization of Vec2Text (Morris et al., 2023) is driven by its capability to effectively invert the full text represented in dense text embeddings, aligning with our goal to provide interpretability of session embeddings in conversational dense retrieval.

The Vec2Text model aims for the complete inversion of input text from its embedding; it leverages the difference between a hypothesis embedding and a ground-truth embedding to make discrete adjustments to the text hypothesis. Specifically, the Vec2Text model begins by proposing an initial hypothesis and subsequently refines this hypothesis through iterative corrections. The goal is to progressively bring the hypothesis's embedding $\hat{e}^t$ closer to the target embedding e.

The Vec2Text comprises two models: the inversion model and the corrector model. Firstly, the inversion model endeavors to invert encoder ϕ by learning a distribution of texts given embeddings $p\left(x \mid e, \theta\right)$. The training objective for the inversion model is to find θ using maximum likelihood estimation:

$$\theta = arg\max_{\hat{\theta}} E_{x\sim D}\left[p\left(x \mid \phi\left(x\right); \theta\right)\right]$$

On the basis of the simple learned inversion hypothesis x^0, the corrector model iteratively refines this hypothesis via marginalizing over intermediate hypotheses:

$$p\left(x^{(t+1)} \mid e\right) = \sum_{x^{(t)}} p\left(x^{(t)} \mid e\right) p\left(x^{(t+1)} \mid e, x^{(t)}, \hat{e}^{(t)}\right)$$

where $\hat{e}^{(t)} = \phi\left(x^{(t)}\right)$.

A.2 More Detailed Experimental Settings

A.2.1 Details of Datasets

The statistical data for each dataset are presented in Table 6 and a more detailed description is provided as follows:

Statistics	QReCC		CAsT-19	CAsT-20	CAsT-21
	Train	Test	Test	Test	Test
# Conv.	10823	2775	50	25	26
# Questions	63501	16451	479	216	239
# Documents	54M		38M	38M	40M

Table 6: Data statistics of conversational search datasets.

QReCC is a large dataset designed for the study of conversational search. Every query is accompanied by an answer and a human-generated rewrite. QReCC includes a total of 13,598 dialogues featuring 79,952 queries. Of these, 9.3K conversations

Context: (CAsT-19 Session 79) q_1: What is taught in sociology? q_2: What is the main contribution of Auguste Comte? q_3: What is the role of positivism in it? q_4: What is Herbert Spencer known for? q_5: How is his work related to Comte?
Current Query(35.2): What is the functionalist theory?
ConvInv (46.9): what is comte's functionalist theory in philosophy?
TX-Human(46.9): what is comte's functionalist theory in philosophy?
TX-Inversion(20.7): What is the functionalist theory?
Human Rewrite(38.3): What is the functionalist theory in sociology?

Table 7: An additional case illustrating the distinction in utilizing rewriting enhancement for transformed text. The numbers in parentheses indicate the retrieval performance NDCG@3 of the transformed text. Notably, the number in parentheses under **Current Query** represents the retrieval results of the original session embedding, not that of the current query statement.

originate from QuAC questions; 80 from TREC CAsT; and 4.4K from NQ. Additionally, 9% of the questions within QReCC lack corresponding answers.

CAsT-19, **CAsT-20**, and **CAsT-21** are three widely used conversational search datasets released by TREC Conversational Assistance Track (CAsT). For CAsT-19, relevance assessments are available for 173 queries within 20 test conversations. For CAsT-20, the majority of queries are accompanied by relevance judgments. For CAsT-21, there are relevance judgments for 157 queries within 18 test conversations. CAsT-19 and CAsT-20 share the same corpus, whereas CAsT-21 employs a different one.

A.2.2 Implementation Details

During the training process, we conduct the training experiments of the Vec2Text model on four Nvidia A100 40G GPUs. We use bf16 precision and AdamW optimizer with 0.001 as the initial learning rate. The strategy to adjust the learning rate is constant with warm-up. We choose T5 (Raffel et al., 2020) as the backbone model. The number of times to repeat embedding along the T5 input sequence length is set to 16.

During the inference process, the sequence beam width and the invert num steps are set to 10 and 30, respectively. The maximum input length and the maximum response length are set to 512 and 100, respectively. The dense retrieval is performed using Faiss (Johnson et al., 2021).

A.3 Examples of Human Evaluation

Examples of the three metrics for human evaluation are shown in Table 8.

A.4 Supplement of Case Study

In this section, We provide an additional case in Table 7 for analysis. The transformed text not only includes the keyword of the original query "functionalist theory", but also enriches it with additional information "comte" and "philosophy", thus yielding a retrieval performance that surpasses that of the human rewrite.

A.5 Experiments with Different Retrievers

Investigations of Based on Different Ad-hoc Retrievers on CAsT-19, CAsT-20, and CAsT-21 datasets are shown in Table 9, Table 10 and Table 11, separately.

Clarity	***Context:*** q_1: What is throat cancer? q_2: Is it treatable? q_3: Tell me about lung cancer. q_4: What are its symptoms? q_5: Can it spread to the throat? q_6: What causes throat cancer?
	Query: What is the first sign of it? ***Human Rewrite:*** What is the first sign of throat cancer? ***Positive Example:***What is throat cancer and what is the first sign of it? ***Negative Example:*** what is the first sign of throat or lung cancer?
Coherence	***Context:*** q_1: What are the different types of sharks? q_2: Are sharks endangered? If so, which species? q_3: Tell me more about tiger sharks.
	Query: What is the largest ever to have lived on Earth? ***Human Rewrite:*** What is the largest shark ever to have lived on Earth? ***Positive Example:***What's the largest sharks to have ever lived on earth? ***Negative Example:*** What is the largest ever to have lived on earth, shark sharks?
Completeness	***Context:*** q_1: What are the origins of popular music? q_2: What are its characteristics? q_3: What technological developments enabled it?
	Query: When and why did people start taking pop seriously? ***Human Rewrite:*** When and why did people start taking pop music seriously? ***Positive Example:***When did people start taking pop music seriously. and why? ***Negative Example:*** What causes pop music and when did it begin to be taken seriously?

Table 8: Examples of the criteria of three metrics of human evaluation.

Method	Retriever	Retrieval Performance			Interpretability	
		MRR	NDCG@3	Recall@100	similarity	human evaluation
KD	KD-GTR	**74.9**	**46.9**	41.9	-	-
	ConvInv	74.2(-0.7)	44.9(-2.0)	**43.0**(+1.1)	0.958	4.40
	KD-ANCE	72.0	44.4	34.2	-	-
	ConvInv	72.0(0.0)	44.5(+0.1)	34.3(+0.1)	0.999	4.90
	KD-BGE	69.5	44.0	41.2	-	-
	ConvInv	69.9(+0.4)	45.4(+1.4)	41.5(+0.3)	0.972	4.33
Conv	Conv-GTR	53.8	31.1	34.6	-	-
	ConvInv	56.4(+2.6)	33.1(+2.0)	37.0(+2.4)	0.778	3.27
	Conv-ANCE	62.8	34.5	29.6	-	-
	ConvInv	47.6(-15.2)	27.2(-7.3)	22.0(-7.6)	0.974	4.13
	Conv-BGE	59.6	35.1	36.4	-	-
	ConvInv	55.2(-4.4)	32.0(-3.1)	37.1(+0.7)	0.736	3.47

Table 9: Retrieval performance and interpretability of generated transformed text based on different ad-hoc retrievers on CAsT-19 Dataset. The best performance is bold.

Method	Retriever	Retrieval Performance			Interpretability	
		MRR	NDCG@3	R@100	similarity	human evaluation
KD	KD-GTR	49.5	**35.9**	**46.9**	-	-
	ConvInv	47.6(-1.9)	34.4(-1.5)	44.0(-2.9)	0.952	3.80
	KD-ANCE	**51.0**	35.8	38.6	-	-
	ConvInv	49.2(-1.8)	34.1(-1.7)	39.9(+1.3)	0.999	4.60
	KD-BGE	44.7	31.9	46.8	-	-
	ConvInv	43.3(-1.4)	30.5(-1.4)	45.3(-1.5)	0.966	4.25
Conv	Conv-GTR	27.9	18.4	31.8	-	-
	ConvInv	27.2(-0.7)	18.5(+0.1)	30.4(-1.4)	0.719	3.00
	Conv-ANCE	38.4	25.8	31.5	-	-
	ConvInv	27.8(-10.6)	18.6(-7.2)	22.8(-8.7)	0.972	2.93
	Conv-BGE	30.7	20.9	35.4	-	-
	ConvInv	31.5(+0.8)	21.4(+0.5)	34.0(-1.4)	0.733	3.13

Table 10: Retrieval performance and interpretability of generated transformed text based on different ad-hoc retrievers on CAsT-20 Dataset. The best performance is bold.

Method	Retriever	Retrieval Performance			Interpretability	
		MRR	NDCG@3	R@100	similarity	human evaluation
KD	KD-GTR	54.7	36.4	55.4	-	-
	ConvInv	54.7(0.0)	37.4(+1.0)	55.1(-0.3)	0.945	3.53
	KD-ANCE	52.8	36.9	50.8	-	-
	ConvInv	55.8(+3.0)	37.4(+0.5)	53.1(+2.3)	0.998	4.07
	KD-BGE	57.9	**41.2**	**56.0**	-	-
	ConvInv	**59.8**(+1.9)	41.1(-0.1)	54.4(-1.6)	0.954	4.25
Conv	Conv-GTR	42.2	28.4	46.4	-	-
	ConvInv	41.9(-0.3)	28.2(-0.2)	41.7(-4.7)	0.664	2.80
	Conv-ANCE	41.1	25.2	42.1	-	-
	ConvInv	30.1(-11)	16.9(-8.3)	31.2(-10.9)	0.973	2.73
	Conv-BGE	48.4	32.8	51.1	-	-
	ConvInv	50.5(+2.1)	32.4(-0.4)	50.5(-0.6)	0.740	3.07

Table 11: Retrieval performance and interpretability of generated transformed text based on different ad-hoc retrievers on CAsT-21 Dataset. The best performance is bold.