Curriculum Contrastive Context Denoising for Few-shot Conversational Dense Retrieval

Kelong Mao
Zhicheng Dou
mkl@ruc.edu.cn
dou@ruc.edu.cn
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China

Hongjin Qian
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
Beijing Key Laboratory of Big Data Management and Analysis Methods
Beijing, China

ABSTRACT

Conversational search is a crucial and promising branch in information retrieval. In this paper, we reveal that not all historical conversational turns are necessary for understanding the intent of the current query. The redundant noisy turns in the context largely hinder the improvement of search performance. However, enhancing the context denoising ability for conversational search is quite challenging due to data scarcity and the steep difficulty for simultaneously learning conversational query encoding and context denoising. To address these issues, in this paper, we present a novel Curriculum cOntrastive conTExt Denoising framework, COTED, towards few-shot conversational dense retrieval. Under a curriculum training order, we progressively endow the model with the capability of context denoising via contrastive learning between noised samples and denoised samples generated by a new conversation data augmentation strategy. Three curriculums tailored to conversational search are exploited in our framework. Extensive experiments on two few-shot conversational search datasets, i.e., CAsT-19 and CAsT-20, validate the effectiveness and superiority of our method compared with the state-of-the-art baselines.

CCS CONCEPTS

• **Information systems** → **Query representation**.

KEYWORDS

Conversational search; contrastive learning; curriculum learning; few-shot learning; conversational dense retrieval

ACM Reference Format:
Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022. Curriculum Contrastive Context Denoising for Few-shot Conversational Dense Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531961

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531961

1 INTRODUCTION

With the rapid development of conversational AI, a new research direction, *Conversational Search (CS)*, has raised more and more attention in the field of information retrieval (IR) in recent years. A conversational search system can interact with users through multiple rounds of dialogues, as shown in Figure 1, to help satisfy users' more complex information-seeking needs [9]. It has been deemed as the next-generation search paradigm [3], especially for facilitating search in more resource-constrained scenarios and promoting social care for people with visual disabilities.

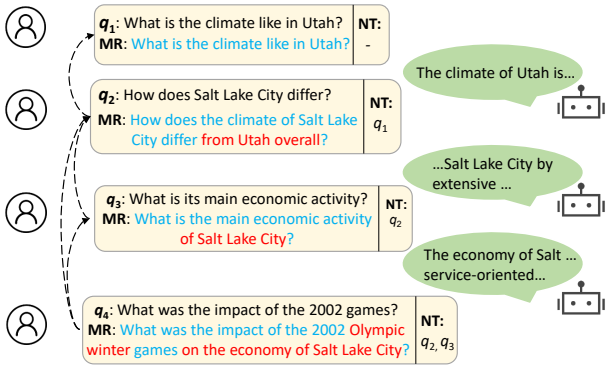


Figure 1: A case of conversational search. MR and NT mean "manual rewrite" and "necessary turns", respectively. The red words should be recovered from the conversation context for accurate conversational search.

Different from ad-hoc search, users will use multi-round natural language-based queries instead of the traditional keyword-based ones to express their complex information needs in conversational search. Such changes in the search form yield big challenges for query understanding since human conversations usually contain more linguistic problems, such as omissions, references, and ambiguities [33]. Thus, recovering the underlying information needs from the conversation context is crucial. To solve this problem, a series of *query reformulation based methods* [21, 36, 39, 43] are first proposed. These methods first train a query rewriting model to create an explicit de-contextualized query and then use this reformulated query to perform standard ad-hoc search. Although such a two-stage way is straightforward and appealing, its drawbacks are also apparent: the query rewriting model is hard to be optimized directly towards retrieval performance and the separate

query reformulation phase further increases search latency [10, 20], resulting in unsatisfactory effectiveness and efficiency.

Recently, *Conversational Dense Retrieval* is proposed to overcome the above limitations. Without the generation of an explicit reformulated query, this integrated approach implicitly represents the conversational query together with its dialog context in a contextualized vector via a conversational query encoder. Different from the previous two-stage approaches, this is an end-to-end solution for conversational search: the learned query vector can be directly optimized for the downstream retrieval. However, it is well-known that dense retrieval needs a large number of labeled query-document pairs for training [**?**], which is contradictory to the reality of conversational search. Unlike the case of ad-hoc search where a large quantity of pseudo query-document relevance signals can be extracted from search click logs, in practice, we cannot get such large-scale search logs for conversational search since nature language based conversational search engines have not been widely deployed [10]. Therefore, the research of conversational search inevitably faces the intractable few-shot learning problem. To the best of our knowledge, there are only a few works [20, 44] focusing on training an effective conversational query encoder under the few-shot scenario. Specifically, Yu et al. [44] developed a few-shot learning framework based on knowledge distillation, where the conversational query encoder learns to mimic the output of a well-trained ad-hoc dense retrieval query encoder [41]. From another perspective, Lin et al. [20] proposed to leverage external datasets to produce more pseudo relevance signals to satisfy the common training requirement of dense retrieval.

The above two pioneering works lay a solid foundation for the research of few-shot conversational dense retrieval. However, their considerations for conversational query encoding are far from optimal. In particular, they simply leverage all historical queries to learn the contextualized representation of the current query turn, but in fact, not all historical conversational turns are necessary for understanding the current turn. In Figure 1, we show necessary turns of each query. For example, for understanding q_4, turns q_2 and q_3 are necessary, while q_1 is not, which is considered as a noisy turn. We argue that including noisy turns will affect the search performance, since they will probably degrade the quality of the representation. For example, they may cause the contextualized query representation to mistakenly include undesirable semantics (e.g., "climate" (in q_1) for q_4), leading to retrieval of wrong documents. Our preliminary studies in Section 3 will show that if we remove noisy turns from the input of the conversational query encoder for each query, we can get better search effectiveness.

Denoising noisy turns is non-trivial for few-shot conversational dense retrieval. First, limited by the amount of training data, we cannot design a new conversational query encoder architecture and train it from scratch to enable context denoising as it is hard to be fully trained to take effect with very little training data. Currently, a common practice is to employ a pre-trained dense retrieval query encoder as a start point and fine-tune it to get the ability of handling conversation context [20, 44]. Second, the multi-task learning of conversational query encoding and context denoising from zero further aggravates learning difficulty for the conversational query encoder, especially under such a few-shot scenario.

To overcome the aforementioned challenges, in this paper, we propose a novel **C**urriculum c**O**ntrastive con**TE**xt **D**enoising framework (**COTED**), towards few-shot conversational dense retrieval. On the whole, we progressively endow the conversational query encoder with the capability of context denoising via a contrastive learning between noised conversational samples and denoised conversational samples, under a tailored curriculum training order. Concretely, for each conversational turn, we assemble its query, its corresponding manual oracle query, its all previous turns, and its necessary turns to be a conversational sample. A conversational sample which does not have noisy turns in its context is a *denoised sample*, otherwise, it is a *noised sample*. In our framework, we first generate much more noised samples from the original training data via a new conversation data augmentation strategy. Then, we develop a contrastive denoising loss by aligning the representations of the same turns in the noised sample to those in its corresponding denoised sample. This denoising loss is optimized together with a knowledge distillation loss through a two-step multi-task learning approach to simultaneously enhance the context denoising and conversational query encoding abilities of the conversational query encoder. To alleviate the difficulty of multi-task learning, we exploit three curriculums tailored to conversational search and form an easy-to-hard training order to further improve the learning process. We conduct extensive experiments on two widely used few-shot conversational search datasets CAsT-19 [5] and CAsT-20 [4]. Experimental results show our proposed method significantly improves search effectiveness over existing state-of-the-art baselines.

In summary, the main contributions of this work are:

(1) We empirically demonstrate that the noisy turns in the conversational context are a critical bottleneck for the improvements of model performance in conversational dense retrieval.

(2) We propose to use contrastive learning to train the conversational query encoder for context denoising, and design a data augmentation strategy to enhance the model learning via generating more noised conversational samples.

(3) We exploit three curriculums tailored to conversational search to further improve the multi-task learning process of conversational query encoding and context denoising.

2 RELATED WORK

To promote the research of conversational search, the TREC Conversational Assistant Track (CAsT) holds an evaluation benchmark [4, 5]. They annually design dozens of artificial conversations to simulate the conversational search process based on the rich voice search experience in Bing. Users express their information needs through multiple turns of dialogue queries, and the task is to retrieve the desired passages for each conversational turn. Qu et al. [30] design an open-domain conversational question answering (QA) task named OR-QuAC, which uses the crowed sourced questions as the query and the target is to retrieve the evidence passages. While the open-domain conversational QA task is very similar to conversational search and can be seen as a sub-domain of conversational search to some extent, some important limitations of OR-QuAC hinder it to be an ideal evaluation benchmark. First, although OR-QuAC provides much more synthetic labels than CAsT datasets, it is originally designed for QA but not for search. The

questions in OR-QuAC are mainly factoid questions while queries in conversational search are much more diverse. Second, in OR-QuAC, there is usually one passage containing evidence for the system to extract the answer, while multiple passages can be regarded as positive to a query in conversational search [17, 30]. Furthermore, all the relevant passages of a dialogue in OR-QuAC reside in the same section of a Wikipedia document due to its synthetic nature [30], which is not the real case of conversational search in practice. Hence, in this work, we focus on more realistic conversational search using two CAsT datasets, facing the intractable few-shot learning problems.

Many studies use query reformulation based methods to build an explicit rewritten query to perform conversational search. Specifically, a few researches study how to select important terms from the previous context turns to expand the current query turn, such as designing rules [21] or training a binary term classifier [39]. Another group of work [23, 32, 36, 37, 43] leverages the powerful pre-trained generative language model to directly generate the reformulated queries. There are also some studies [14, 22] that combine the term selection and query generation methods.

In addition to the query reformulation, some researchers try to boost the retrieval performance by introducing dense retrieval into conversational search or conversational QA. The core of conversational dense retrieval is to train an effective conversational query encoder, which can encode the current query turn and its dialogue history. Qu et al. [30] propose a common conversational query encoding paradigm by extending the pre-trained language model (e.g., BERT) to encode the concatenation of all conversational queries in the context. The architecture of the conversational query encoder is inherited from the pre-trained language model without modification. They show that the conversational query encoder can achieve good performance on the open-retrieval conversational QA task if trained with sufficient query-document relevance labels using a widely used ranking loss [13, 30]. Some other works [16, 17] also demonstrate the effectiveness of conversational dense retrieval for conversational QA if trained with sufficient data.

However, as aforementioned, conversational search currently faces a serious data scarcity problem. To the best of our knowledge, there are only a few works [20, 44] study how to train the conversational query encoder under the few-shot case, towards conversational search. Specifically, Yu et al. [44] propose a knowledge distillation loss with limited data, which forces the conversational query encoder to mimic the output of a well-trained dense retrieval query encoder. From a different perspective, Lin et al. [20] propose to create a large number of pseudo query-document relevance labels using other related datasets and train with the normal ranking loss. Based on these great works, we study how to enhance the context denoising ability of the conversational query encoder with limited available data for conversational search, which is a crucial problem but is neglected by existing studies.

Besides, it is worth emphasizing that the architecture of the conversational query encoder is usually based on a well-trained ad-hoc dense retrieval query encoder without modification [20, 44], because we do not have large-scale high-quality data to train a new sophisticated architecture to take effect. This is also one of the main challenges for our work as we cannot enable context denoising by modifying the model architecture. Therefore, although there exist some complex architectures for context modeling in the related tasks of conversational search, such as HAM [31] for conversational QA and HBA-Transformers [29] for session search, they are not applicable to our work.

3 STUDY OF CONVERSATIONAL TURNS

Before diving into our proposed method, we first conduct preliminary experiments to empirically justify our motivation, i.e., noisy conversational turns will impair the model performance for conversational dense retrieval. We will first introduce the general paradigm of conversational dense retrieval and our experimental model, then elaborate our experimental settings, results, and findings.

3.1 Conversational Dense Retrieval

Conversational dense retrieval is an important research branch of conversational search. Formally, the target of conversational search is to find the relevant document d from a collection of D for each query turn in a multi-round conversation $Q = \{q_k\}_{k=1}^{n}$, where n is the number of turns of the conversation. Different from traditional ad-hoc search, the conversational query q_k itself is usually ambiguous, context-dependent, and requires more sophisticated query understanding approaches to recover its real search intents from the conversation context $Q_{1:k-1}$ [5, 44].

To achieve this goal, the idea of conversational dense retrieval is to directly maps the current query turn together with its context and documents into a unified embedding space to perform dense retrieval, without generation of a new explicit query:

$$\mathbf{q}'_k = \text{CQE}(q_k, Q_{1:k-1}), \quad (1)$$

$$\mathbf{d} = \text{DE}(d), \quad (2)$$

where CQE and DE denote the conversational query encoder and the document encoder, respectively. The retrieval score is computed as the dot product between the contextualized query representation $\mathbf{q}'_k$ and the document representation $\mathbf{d}$, which can be efficiently done with many libraries (e.g., Faiss [12]). As the meaningful information in a document probably has no difference when serving ad-hoc search and conversational search, the document encoder is usually set to the same as in ad-hoc dense retrieval [25, 44] and frozen. Therefore, the core of research in conversational dense retrieval is to study how to get a better conversational query encoder (CQE). However, a unique challenge in this area is that we can hardly design a new CQE architecture and train it from scratch, due to the limited amount of training data. Hence, a common practice is to employ a pre-trained dense retrieval query encoder as a start point and fine-tune it to be an effective CQE [20, 44].

In this section, we choose ConvDR [44] as the experimental model to validate our motivation, because it is the state-of-the-art and the most representative model for few-shot conversational dense retrieval. Specifically, ConvDR adopts a state-of-the-art ad-hoc dense retriever ANCE [41], which is a BERT-Siamese model [15], as its architecture:

$$\mathbf{q}'_k = \text{BERT}([\text{CLS}] \circ q_1 \circ [\text{SEP}] \circ \ldots \circ [\text{CLS}] \circ q_k \circ [\text{SEP}]), \quad (3)$$

$$\mathbf{d} = \text{BERT}([\text{CLS}] \circ d \circ [\text{SEP}]). \quad (4)$$

The input of CQE is the concatenation of all tokens in the conversational queries $Q_{1:k}$, and it uses the BERT first [CLS] embeddings as

the representations of the query turn and the document. Under the few-shot scenario, ConvDR is trained with the knowledge distillation technique. It first uses the same well-trained ANCE [41] as the teacher model to get the representation $\mathbf{q}_k^*$ of the corresponding manual oracle query q_k^*. Then, it is trained with a MSE loss to make the distance between the conversational query q_k and its oracle counterpart q_k^* closer in the embedding space:

$$\mathbf{q}_\mathbf{k}^* = \text{TM}(q_k^*), \quad (5)$$

$$\mathcal{L}_{\text{KD}} = \text{MSE}(\mathbf{q}_\mathbf{k}', \mathbf{q}_\mathbf{k}^*), \quad (6)$$

where TM is the teacher model (i.e., ANCE). The intuition of using knowledge distillation is that the underlying information needs in the manual oracle q_k^* and the conversational query q_k are the same and thus their embeddings should be the same [44].

Note that CQE in ConvDR can also be trained with the ranking loss which is widely used in dense retrieval [**?**]. However, in [44], it has been demonstrated that the ranking loss needs a large number of query-document relevance signals to take effect, which is useless for training ConvDR for few-shot conversational dense retrieval. Therefore, we do not consider the ranking loss and only adopt the KD training strategy for ConvDR in our experiments.

3.2 Necessary Turns Annotation

To validate our motivation, a straightforward method is to compare the performances of ConvDR on queries with noisy turns and without noisy turns. However, the existing conversational search datasets (i.e., CAsT-19 [5] and CAsT-20 [4]) do not provide the explicit information of turn dependency that indicates which previous turns are necessary for understanding the current turn [1]. Therefore, to solve this obstacle, we manually annotated the necessary turns on these two datasets. Specifically, for each conversational turn q_k, we manually select the necessary turns from its actual context queries $Q_{1:k-1}$ by comparing them with the oracle query q_k^*. Necessary turns and noisy turns are defined as follows.

Definition 1. *Necessary turns of a query q_k are the smallest sufficient subset of $Q_{1:k-1}$ which can provide enough information for humans to complete the lost information of q_k compared with q_k^*.*

Definition 2. *Noisy turns of a query q_k refer to the remaining turns of $Q_{1:k-1}$ except the necessary turns.*

Every conversational turn is annotated by at least three information retrieval researchers, and we resolve the inconsistent cases of annotation through discussion and majority rule. The whole process of annotation finally takes us around nine hours.

The statistics of manual annotation results are shown in Table 1. As can be seen, from the human perspective, the average number of necessary turns that a conversational turn depends on is just a little more than 1. On average, necessary turns only make up less than 40% of the actual context of a turn. Therefore, there indeed exists lots of noise in the conversation context, especially for those later conversational turns since they have more preceding queries. We argue that the large proportion of noisy turns may affect the training process and further hurt the model performance.

[1] Although there are part of turn dependency annotations in CAsT 20, we find it is not very accurate and sufficient

Table 1: Statistics of annotated necessary turns on two CAsT datasets. Note that the first turn of each conversation is not considered since it has no context.

Statistics	CAsT-19	CAsT-20
# Conversations	50	25
# Turns (queries)	479	208
# Avg. Question Tokens	6.1	6.8
# Avg. Questions / Conversation	9.6	8.6
# Avg. Necessary Turns / Turn	1.02	1.25
# Avg. Necessary Turns Ratio / Turn	0.31	0.39

3.3 Denoising Control Experiments

Since we have got the annotation of necessary turns, then we conduct a series of control experiments to investigate the impact of noisy turns on ConvDR. For a conversational turn q_k, we denote its actual context $Q_{1:k-1}$ as AC_k, its necessary context as NC_k. Thus the noisy context is $AC_k \setminus NC_k$. Specifically, we design the following four comparative experiments for ConvDR:

- Training with the necessary context NC and test with NC ($NC \rightarrow NC$): For each query turn, we use its necessary context instead of its original actual context as the input of ConvDR (i.e., changing $\mathbf{q}_\mathbf{k}' = \text{CQE}(q_k, AC_k)$ into $\mathbf{q}_\mathbf{k}' = \text{CQE}(q_k, NC_k)$), in both training and testing.
- Training with AC and test with NC ($AC \rightarrow NC$): For each query turn, we only use its necessary context for testing but still use its actual context for training.
- Training with NC and test with AC ($NC \rightarrow AC$): For each query turn, we only use its necessary context for training but still use its actual context for testing.
- Training with AC and test with AC ($AC \rightarrow AC$): For each query turn, we use its actual context for both training and testing, which is the same as the original ConvDR.

The settings and implementations follow the open-source code of ConvDR [2] and the same five-fold cross-validation is used for fair comparisons. Besides, in the original paper of ConvDR, they warm up ConvDR on an external query rewrite dataset CANARD [7] before training on CAsT-20. Since our work targets complete few-shot conversational dense retrieval (i.e., only limited data is available), we remove this warm-up, but we will also show the results of ConvDR with warm-up for reference. Following ConvDR, we use MRR and NDCG@3 as the evaluation metrics.

The comparison results are summarized in Table 2. By analyzing this table, we gain the following observations:

(1) By comparing $NC \rightarrow NC$ and $AC \rightarrow NC$ to $AC \rightarrow AC$, we can find that filtering out noisy turns during both training and testing or only during testing can significantly improve 12.2% and 4.7% model performance w.r.t. NDCG@3 on CAsT-20, respectively.

(2) The improvement of removing noisy turns on CAsT-19 is not as significant as that on CAsT-20. For example, compared with $AC \rightarrow AC$, $NC \rightarrow NC$ just has 0.6% improvements w.r.t. NDCG@3 on CAsT-19. This is because CAsT-19 is an "easier" dataset compared with CAsT-20. Through the comparison with "ANCE with Oracle Query", we can find that the original $AC \rightarrow AC$ (i.e., ConvDR)

[2] https://github.com/thunlp/ConvDR

Table 2: Performance comparison between different ConvDR variants on two few-shot CAsT datasets. *AC→AC* (warmed) is ConvDR warmed up on CANARD. ‡ denotes significant differences ($p < 0.05$) with respect to ConvDR (i.e., *AC→AC*).

Method	CAsT-19		CAsT-20	
	MRR	NDCG@3	MRR	NDCG@3
AC→AC	0.740	0.466	0.476	0.319
AC→NC	0.740	0.467	0.484‡	0.334‡
NC→AC	0.627	0.384	0.396	0.259
NC→NC	**0.743**‡	**0.469**‡	**0.513**‡	**0.358**‡
AC→AC (warmed)	0.746	0.463	0.510‡	0.340‡
ANCE with Oracle Query	0.740	0.461	0.591	0.422

has already reached the performance of using oracle queries on CAsT-19, but still has significant gaps on CAsT-20. In fact, the conversational search cases in CAsT-20 are more complex and realistic than those of CAsT-19 (See Section 5.1 for details). In our annotation process, we also find that it is much easier to recover the conversational queries to the oracle query in CAsT-19 from human views. Therefore, such improvements gaps on the two datasets are reasonable.

(3) When we use the denoised context for training but use the noisy actual context for testing (i.e., *NC→AC*), the model performance degrades significantly on both two datasets. This is probably caused by the large difference between the distribution of test data and training data, but it also indicates that the context denoising ability of the current model is not strong.

In general, the overall results justify our claim that noisy turns can hurt model performance for conversational dense retrieval.

4 OUR METHODOLOGY

Through the preliminary experiments in Section 3, we prove context denoising is beneficial to few-shot conversational dense retrieval. However, since we cannot leak the annotation information in the test phase, how to help CQE learn to automatically denoise becomes a key problem, which faces the following unique challenges:

(1) Although there are many existing studies [27, 45, 46] about context denoising or sequence denoising in various research fields, most of them usually resort to designing a specific trainable denoising module, such as attention mechanism [46]. However, as shown in Table 1, the number of conversation data is extremely small, which is insufficient to train any new parameterized denoising module. Thus, we can only use the existing pre-trained ad-hoc dense retrieval query encoder (e.g., ANCE) as the conversational query encoder and can hardly modify its architecture (as aforementioned in Section 2 and Section 3.1). How to perform context denoising under such a data scarcity challenge has seldom been explored by existing studies.

(2) Simultaneously teaching the ad-hoc dense retrieval query encoder to effectively encode conversational queries (i.e., *Conversational Adaption*) and have the context denoising ability (i.e., *Context Denoising*) further aggravates the learning difficulty, especially under the few-shot learning scenario.

In this section, we elaborate our proposed curriculum contrastive context denoising framework (COTED), which can effectively solve the above challenges for few-shot conversational dense retrieval. Figure 2 shows an overview of our proposed COTED framework. Basically, it consists of three components, namely Conversation Data Augmentation, Curriculum Sampling, and Two-step Multi-task Learning. We will first describe the general training workflow of our framework, and then introduce each component in detail.

4.1 Training Workflow

An illustration of the training workflow of COTED is shown at the bottom of Figure 2. First, we perform a conversation data augmentation on the original dataset to obtain a new augmented dataset containing much more training samples. Then, we use a curriculum sampling strategy to sample a batch of training samples from the new augmented dataset, and finally adopt a two-step multi-task learning method to optimize conversational query encoder for both context denoising and conversational adaption with the samples. The three important components, including conversation data augmentation, curriculum sampling, and two-step multi-task learning, are organically organized together to jointly help CQE achieve better context denoising and generalization abilities.

4.2 Conversation Data Augmentation

Inspired by the idea of contrastive learning [24], we try to teach CQE context denoising through the contrast between noised conversational samples and denoised conversational samples. Formally, in our task, a conversational sample s can be created from a conversational turn and is defined as: $s = (q, q^*, NC, AC)$, where q, q^*, NC, and AC denote the query, the corresponding manual oracle query, the necessary context, and the actual context of the conversational turn, respectively. If $NC = AC$, we call s a *denoised sample*, otherwise, it is a *noised sample*. We can easily get the denoised version of a noised sample by setting its AC to its NC.

For better learning, we develop a new conversation data augmentation method to create more noised samples. Specifically, for a sample s, we first randomly select m turns from its noisy turns (i.e., $AC \setminus NC$), and then combine the selected turns with the necessary turns to form a new noised actual context. Finally, we assemble this new actual context with the query, the oracle query, and the necessary context of the original sample s to be a new noised sample. An example of conversation data augmentation is shown in Figure 2. For example, we can sample q_1 and q_3 ($m = 2$) from the noisy turns (i.e., $\{q_1, q_3, q_4, q_6\}$), and combine them with the necessary context (i.e., $\{q_2, q_5\}$) to be a new actual context $\{q_1, q_2, q_3, q_5\}$.

Theoretically, for an original sample which has M noisy turns, we can produce at most $\sum_{m=1}^{M-1} \binom{M}{m}$ new samples based on it, which is a considerable amount of data augmentation for our few-shot learning task. In practice, to control the number of augmented samples and balance the number of their noisy turns, we set a sampling threshold p. Then, we only perform at most p times non-repetitive sampling for each original sample with m sampling rate.

Our conversation data augmentation provides much more contrastive samples (i.e., noised samples vs. their denoised versions) to facilitate the later learning to denoise for CQE. Besides, similar to data augmentation in other fields [8, 34], it may also help to improve the generalization ability of CQE for conversational search.

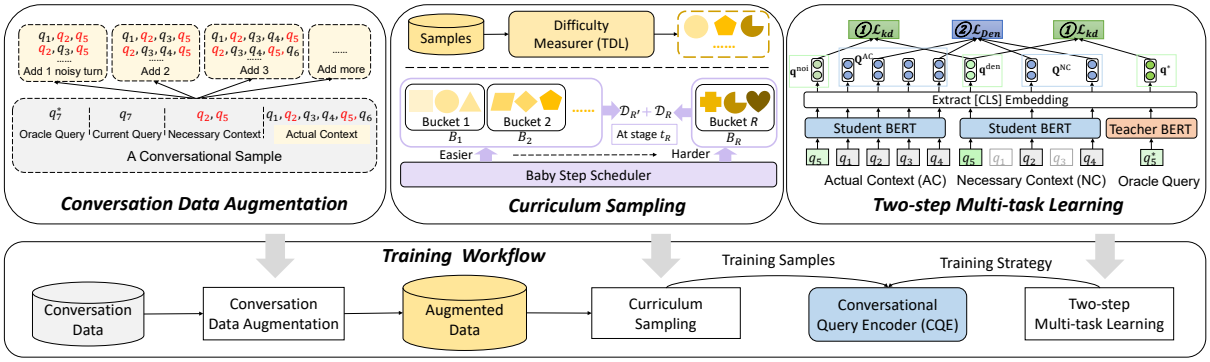


Figure 2: Overview of our proposed COTED.

4.3 Curriculum Sampling

Considering the steep difficulty of simultaneously learning adaption to conversational queries and context denoising for CQE, we adopt an easy-to-hard training strategy inspired by curriculum learning [2, 28, 40, 42] to facilitate the model training. Curriculum learning is to train from easier data to harder data to improve the learning, which mimics the human learning process. The core of curriculum learning is to design an appropriate *Difficulty Measurer* which evaluates the learning difficulty of each sample, and a *Training Scheduler* which decides the sequence of data subsets throughout the training process. Therefore, we elaborate our curriculum sampling strategy from these two aspects.

4.3.1 ***Difficulty Measurer.*** A suitable difficulty measurer is very important for curriculum learning to take effect. However, to the best of our knowledge, it is still unknown what difficulty measurers are effective for conversational search. In this work, we propose to use *Token Difference Length (TDL)* to measure the difficulty of the conversational sample. Specifically, for a conversational sample $s = (q, q^*, NC, AC)$, we first get the two token sets $\text{TS}(q^*)$ and $\text{TS}(AC+q)$, which contains all of tokens in q^* and $AC+q$, respectively. Then, its TDL is defined as:

$$\text{TDL}(s) = |(\text{TS}(q^*) \cup \text{TS}(AC+q)) \setminus (\text{TS}(q^*) \cap \text{TS}(AC+q))|. \quad (7)$$

where $|\cdot|$ denotes the number of tokens in the set. TDL counts how many unique tokens do we need to delete and add to transform $AC+q$ into q^*, which also reflects the semantic difference between $AC+q$ and q^*. Intuitively, a sample with larger TDL tends to be harder to learn to encode for CQE. Thus, TDL is suitable to measure the sample difficulty with respect to the conversational adaption task. Besides, as introduced in Section 4.2, for all new generated samples $S_{\text{aug}}(s)$ based on the same original sample s, they will all share the same necessary context and the same oracle query with s. Thus, in $S_{\text{aug}}(s)$, a sample with larger TDL generally tends to have more noisy turns and be harder to denoise, so TDL is also suitable to measure the sample difficulty with respect to the context denoising. We stipulate that **a larger TDL indicates a harder sample.** Moreover, to investigate the effect of TDL, we also design another two difficulty measurers for comparisons (See Section 5.4.2).

4.3.2 ***Training Scheduler.*** While difficulty measurers vary among different data types and tasks, the choice of the training scheduler is usually data (or task) agnostic [40]. In this work, we choose to study the effectiveness of curriculum learning for conversational search with one of the most popular schedulers, i.e., Baby Step [1, 35]. An illustration is shown in the middle of Figure 2. With all training samples $\mathcal{D} = \{s_1, s_2, ..., s_N\}$ sorted by their difficulty scores, we employ the widely used baby step paradigm to arrange them into a learning curriculum . Specifically, we first averagely distribute $\mathcal{D}$ into R buckets $\{B_1, B_2, ...B_R\}$ from easy to hard, and then train with each bucket one-by-one. When training with the current bucket B_r at stage t_r, for each training step, we randomly select two half batches of samples $\mathcal{D}_r$ and $\mathcal{D}_{r'}$ from B_r and all previous buckets $B_{1:r-1}$, respectively. After all of the buckets are used, we get back to a normal training mode that randomly samples from $\mathcal{D}$ and finish the pre-defined training epochs.

In this way, CQE finally enjoys a better learning curriculum, i.e., first learning how to encode and denoise from easier conversational samples and then gradually learning from harder ones, effectively improving the final model performance (See Section 5.4.2).

4.4 Two-step Multi-task Learning

We design a simple and effective two-step multi-task learning strategy to teach CQE how to perform context denoising as well as encoding conversational queries. An illustration is shown on the right side of Figure 2. Specifically, for a conversational sample $s = (q, q^*, NC, AC)$, we use CQE to encode it from both a noised view and a denoised view:

$$\mathbf{q}^{\text{noi}}, \mathbf{Q}^{\text{AC}} = \text{CQE}(q, AC), \quad (8)$$
$$\mathbf{q}^{\text{den}}, \mathbf{Q}^{\text{NC}} = \text{CQE}(q, NC), \quad (9)$$

where $\mathbf{q}^{\text{noi}}$ and $\mathbf{q}^{\text{den}}$ denote the noised and denoised query representation of this conversational sample, respectively. $\mathbf{Q}^{\text{AC}} = [\mathbf{q}^{\text{AC}}_1, ..., \mathbf{q}^{\text{AC}}_{|\text{AC}|}]$ and $\mathbf{Q}^{\text{NC}} = [\mathbf{q}^{\text{NC}}_1, ..., \mathbf{q}^{\text{NC}}_{|\text{NC}|}]$ are the representation matrices of all turns of the actual context and the necessary context, respectively. |AC| and |NC| are number of turns contained in AC and NC. The architecture of CQE in this work follows ConvDR (Equation 3) except that we move the current query to the first of the input so that the first [CLS] belongs to the current query. We use the [CLS] embedding of the query turn as its representation. But note that our method is a general framework and is not restricted to any specific CQE architecture.

Without introducing any new parameters, we propose a context denoising loss from a contrastive view, by aligning the representations of the same turns in the noised context (i.e., $\mathbf{Q}^{\mathbf{AC}}$) and the denoised context (i.e., $\mathbf{Q}^{\mathbf{NC}}$):

$$\mathcal{L}_{\text{den}} = \frac{1}{C} \sum_{i=1}^{|AC|} \sum_{j=1}^{|NC|} \mathbb{I}_{g(i)=g(j)} \cdot \text{MSE}(\mathbf{q}_\mathbf{i}^{\mathbf{AC}}, \mathbf{q}_\mathbf{j}^{\mathbf{NC}}), \tag{10}$$

where $g(\cdot)$ is an index mapping function to map the index of a turn in the (actual or necessary) context to its index in the all previous turns. Thus, $g(i) = g(j)$ means that these two turns are the same turn. $\mathbb{I}$ is the indicator function. C is the number of the same turns in these two contexts. Note that we detach the gradients of $\mathbf{Q}^{\mathbf{NC}}$, so this is a unidirectional alignment from $\mathbf{Q}^{\mathbf{AC}}$ to $\mathbf{Q}^{\mathbf{NC}}$.

The intuition of the proposed denoising loss is that, if CQE has a good context denoising ability, its output representations under different input contexts should be close to those under the necessary context, since noisy turns are expected to not affect the output of CQE. Besides, it is notable that, similar to BYOL [11], our denoising loss is a special contrastive loss which does not need negative samples. We avoid collapsed solutions through multi-task learning with a knowledge distillation loss (See below).

Similar to ConvDR, we also adopt knowledge distillation to distill knowledge from a dense retriever teacher into CQE to learn how to encode conversational queries:

$$\mathbf{q}^* = \text{TM}(q^*), \tag{11}$$

$$\mathcal{L}_{\text{kd}} = \text{MSE}(\mathbf{q}^{\text{noi}}, \mathbf{q}^*) + \alpha \cdot \text{MSE}(\mathbf{q}^{\text{den}}, \mathbf{q}^*), \tag{12}$$

where α is a hyper-parameter to balance two losses.

Intuitively, if CQE has even not known how to do the basic conversational adaption, it would be harder to learn context denoising since the representations of turns may be meaningless. Therefore, we choose a two-step optimization way for multi-task learning . Specifically, for a batch of training samples, we first train CQE with $\mathcal{L}_{\text{kd}}$, update the model parameters, and then train it with $\mathcal{L}_{\text{den}}$. In practice, we use two optimizers to control the training with these two losses, respectively.

5 EXPERIMENTS

We carried out a series of experiments to justify the effectiveness of our COTED for few-shot conversational dense retrieval, and provide comprehensive analysis for a better understanding of COTED.

5.1 Datasets

We use the only two available datasets for (few-shot) conversational search, i.e., TREC CAsT 2019 [5] and TREC CAsT 2020 [4], in our experiments. Table 1 shows their statistics information.

CAsT-19: The TREC Conversational Assistance Track (CAsT) 2019 benchmark provides 50 test conversations (topics) for conversational search. Each conversation contains an average of 9 to 10 natural language based-queries with common natural language issues, such as coreferences and omissions. The query turns often depend on their previous turns. Besides, each query has a corresponding manual oracle de-contextualized query. Among them, 173 queries in 20 test conversations have relevance judgments. The corpus consists of around 38 million passages from MS MARCO [26] and TREC Complex Answer Retrieval (CAR) [6].

CAsT-20: This is the dataset of the second year of Conversational Assistance Track, which contains 25 conversations. Different from the CAsT-19 dataset, queries in this dataset can refer to previous answers from system responses and are more realistic and complex. All queries have the corresponding manual oracle queries and most of them have relevance judgments. CAsT-20 shares the same passage corpus as CAsT-19.

5.2 Experimental Settings

5.2.1 ***Baselines and Evaluation Protocols.*** We compare the proposed COTED with the following baselines:

Query Reformulation Methods: (1) Transformer++ [36]: It fine-tunes GPT-2 on CANARD [7] dataset, and then generates reformulated queries. (2) QueryRewriter [43]: It fine-tunes GPT-2 with large-scale synthetic conversation data and then generates reformulated queries. (3) QuReTeC [39]: It trains a binary tagger to find relevant terms in the context and then append them to the original query to be a reformulated query.

Conversational Dense Retrieval Methods: (4) ConvDR [44]: A state-of-the-art model for conversational dense retrieval, which is the main direct competitor of our work. We also compare with its warmed-up version. (5) ContQE [20]: Another state-of-the-art model for conversational dense retrieval. Note that it employs a different pre-trained query encoder TCT-ColBERT [19] but not ANCE [41] as the conversational query encoder. It is trained with large-scale pseudo relevance signals.

Reference Methods: (5) Raw: The original context-dependent query. (6) Manual: The manual oracle query. (7) *NC→NC*: We use the necessary context to replace the original actual context as the input of ConvDR in both training and test (See Section 3.3).

For the first three query reformulation methods, we perform both sparse retrieval and dense retrieval with Pyserini BM25 [18] [3] and ANCE [41], respectively, on the reformulated queries for evaluation. The last three baselines involve human intervention (except Raw) and we include them for reference.

Following the existing works [43, 44], we adopt MRR [4] and NDCG@3 as the evaluation metrics, and NDCG@3 is the primary metric as prescribed by TREC CAsT [4, 5].

5.2.2 ***Implementation Details.*** We implement COTED with PyTorch library. The Adam optimizer is employed with a mini-batch size 4 and a learning rate of 5e-6. Most of our settings follow ConvDR [44] for fair comparisons. Concretely, we use the same ANCE checkpoint employed in ConvDR as the teacher and the start point of CQE. All passage embeddings are also encoded by ANCE and fixed. The input of CQE is the concatenation of context queries and the current query, and early tokens will be discarded if the concatenation length exceeds 256 tokens. On CAsT-20, we adopt the automatic canonical setting as ConvDR and also concatenate the last automatic canonical response to the input of CQE. All training and evaluation exactly follow the five-fold cross-validation setting as ConvDR. We tune hyper-parameters with grid search. Finally, we set the sampling threshold p to 2 on CAsT-19 and 3 on CAsT-20, the loss balance weight α to 0.01, the number of buckets R to 5, and train

[3] We use the default setting of LuceneSearcher.

[4] Following the official evaluation setting [4], we use relevance scale ≥ 2 as positive for MRR on CAsT-20

Table 3: Overall results. † and § indicate the model uses ANCE and TCT-ColBERT, respectively. * means the results are quoted from their original papers. ‡ denotes significant differences with respect to all compared baselines.

Search	Method	CAsT-19		CAsT-20	
		MRR	NDCG@3	MRR	NDCG@3
Sparse	Tansformer++	0.557	0.267	0.162	0.100
	QueryRewriter	0.581*	0.277*	0.250*	0.159*
	QuReTeC	0.605	0.338	0.262	0.171
Dense	Transformer++†	0.696	0.441	0.296	0.185
	QueryRewriter†	0.665*	0.409*	0.375*	0.255*
	QuReTeC†	0.709	0.443	0.430	0.287
	ContQE§	-	**0.499***	-	0.312*
	ConvDR†	0.740	0.466	0.476	0.319
	COTED (Ours)†	**0.769**‡	0.478‡	**0.491**‡	**0.342**‡
For Reference					
Sparse	Raw	0.322	0.134	0.160	0.101
	Manual	0.671	0.347	0.445	0.301
Dense	Raw†	0.420	0.247	0.234	0.150
	Manual†	0.740*	0.461*	**0.591***	0.422*
	Manual§	-	**0.507***	-	**0.460***
	ConvDR (warmed)†	**0.746**	0.463	0.510*	0.340*
	NC→*NC*†	0.743	0.469	0.513	0.358

$T = 6$ epochs on both two datasets. We conduct the statistical significance test using the permutation test ($p < 0.05$) between COTED and the compared baselines. Besides, we emphasize that we do not use any external datasets to assist the training since our work targets solving the complete few-shot problem (i.e., only limited data is available). So we do not warm up on CANARD and all experiments are conducted based on the two CAsT datasets. The annotation data and our code are released at https://github.com/kyriemao/COTED.

Same as the experiments in Section 3, the settings and implementations of ConvDR follow their official open-source code. The results of QueryRewriter and ContQE are quoted from the original papers of ConvDR and ContQE, respectively. For Transformer++ and QuReTeC, we use their reformulated queries provided by [38].

5.3 Performance Comparisons

The overall results are listed in Table 3. From the results, we have the following observations:

(1) **Our proposed COTED outperforms the majority of baselines on two CAsT datasets.** In particular, COTED beats its main competitor ConvDR by 3.0% and 7.2% w.r.t. the main metric NDCG@3 on CAsT-19 and CAsT-20, respectively, showing the superiority of our designed training framework. Compared with ConvDR, COTED enjoys more self-augmented data, a more sophisticated training strategy, and a more reasonable training curriculum, finally resulting in better context denoising and generalization abilities for few-shot conversational dense retrieval.

(2) ContQE seems to outperform our COTED on CAsT-19. However, note that ContQE is based on a different pre-trained query encoder (ANCE for COTED while TCT-ColBERT for ContQE). From the comparison of two Manual results, we can find that TCT-ColBERT performs better than ANCE on the two CAsT datasets. **But even in this unfair case, our model can still achieve 9.6% improvements than ContQE w.r.t. NDCG@3 on CAsT-20. Furthermore, on CAsT-19, our model can relatively outperform Manual (0.478 vs. 0.461) while ContQE failed (0.499 vs. 0.507), w.r.t. NDCG@3. Such results prove the advantages of our model.** Besides, since our work targets the complete few-shot scenario (i.e., only limited data is available), we choose to mainly follow the experimental settings of ConvDR for fair comparisons, and leave additional comprehensive comparisons with ContQE, which leverages external datasets, in future work.

Table 4: Performance comparisons of COTED training with different sampling thresholds p with respect to NDCG@3.

Dataset	Sampling Threshold p						
	0	1	2	3	4	6	8
CAsT-19	0.467	0.473	**0.478**	0.476	0.473	0.462	0.459
CAsT-20	0.330	0.334	0.339	**0.342**	0.337	0.331	0.321

Table 5: Performance with different difficulty measurers.

Method	CAsT-19		CAsT-20	
	MRR	NDCG@3	MRR	NDCG@3
COTED (Random)	0.762	0.474	0.482	0.332
COTED (ACL)	**0.769**	**0.480**	0.486	0.337
COTED (MPS)	0.765	0.477	0.485	0.338
COTED (TDL)	**0.769**	0.478	**0.491**	**0.342**

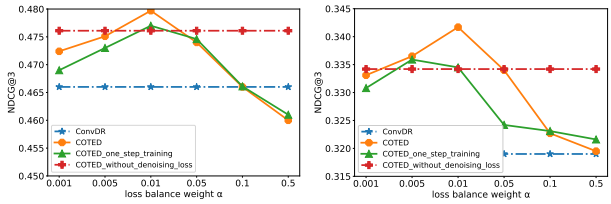
(3) **Surprisingly, our COTED even outperforms *NC*→*NC* on CAsT-19.** As we have manually removed the noisy turns for both training and test in *NC*→*NC*, it is expected to be a performance ceiling from the view of context denoising. But in fact, COTED benefits from three aspects not only from the denoising strategy, and thus achieves such a desirable breakthrough. Besides, compared with ConvDR (warmed), which first warms up ConvDR on the large external CANARD [7] dataset, our COTED can still achieve a slight performance lead w.r.t. the main evaluation metric NDCG@3 on the more difficult CAsT-20 dataset. Such good results demonstrate the effectiveness and superiority of our COTED.

(4) **On CAsT-19, both COTED and ConvDR can outperform ANCE with oracle queries,** indicating that the CQE may be able to surpass its teacher. It is reasonable because the oracle query is not always absolutely perfect. With better training strategies, CQE has the potential to encode users' real information needs from the comprehensive context to achieve more accurate passage retrieval.

5.4 Component Analysis

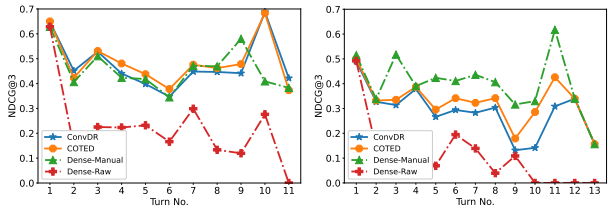
In this section, we further analyze the source of the effectiveness of our COTED. Specifically, we separately explore how the three important components of our framework affect its performance.

5.4.1 ***Effect of Conversation Data Augmentation.*** This component produces much more noised samples for training to enhance the few-shot ability of CQE. We investigate it by conducting experiments with different sampling thresholds $p \in [0, 1, 2, 4, 6, 8]$ on two CAsT datasets. Results are shown in Table 4. In the beginning,



(a) CAsT-19 (b) CAsT-20

Figure 3: Ablation study of the multi-task learning.



(a) CAsT-19 (b) CAsT-20

Figure 4: Turn-level performance comparisons.

the model performance gradually increased as the number of augmented conversational samples continued to increase. However, when there are too many augmented noised samples (i.e., large p), the performance starts to suffer a large negative impact. It is probably because too many similar noised samples lead to overfitting, and it indicates that a proper sampling threshold is important for the conversation data augmentation to take effect.

5.4.2 ***Effect of Curriculum Sampling.*** To investigate the effect of our curriculum sampling strategy, we test the performance of COTED using random sampling. Furthermore, except TDL, we also design another two difficulty measurers, including *Actual Context Length (ACL)* and *Model Prediction Score (MPS)*, to further study the influence of the difficulty measurer. Specifically, for a sample $s = (q, q^*, NC, AC)$, ACL denotes the number of turns in the actual context (i.e., $|\text{AC}|$), and a larger ACL indicates a harder sample. Obviously, ACL is simpler than TDL as it only considers the actual context; Different from TDL and ACL which are heuristically defined by humans, we let the model itself decide the sample difficulty in MPS. Consistent with the golden metric (i.e., NDCG@3) of our target task, we adopt 5-fold cross-validation to get the predicted NDCG@3 for each sample, and use it as the sample difficulty for MPS. A lower NDCG@3 score indicates a harder sample.

We fine-tune to get the best bucket number R for ACL and MPS for fair comparisons, and show results in Table 5. Generally, using our three curriculums gains better performance than using random sampling, verifying the effectiveness of our proposed curriculum sampling for conversational search. In particular, We find the performances of using different difficulty measurers are close on CAsT-19, and the performance gain of curriculum sampling on CAsT-19 is small, compared with that on CAsT-20. It may be because the conversational turns in CAsT-19 are much easier to learn to encode, so the curriculum sampling seems to be little help on CAsT-19. While on the more complex CAsT-20, TDL relatively outperforms the others. On the whole, the more comprehensive heuristic difficulty measurer TDL is a better choice.

5.4.3 ***Effect of Two-step Multi-task Learning.*** The denoising loss $\mathcal{L}_{\text{den}}$ plays an important role in teaching CQE for context denoising. To justify the effectiveness of our designed denoising loss and the two-step learning strategy, we compare COTED with the following two variants:

- COTED without the denoising loss: We remove $\mathcal{L}_{\text{den}}$ from COTED and only optimize it with $\mathcal{L}_{\text{KD}}$.
- COTED with one-step training: We perform a one-step joint optimization of the two tasks by combining the two losses together with proper weights.

Besides, we also investigate the influence of the hyper-parameter α on the performance. Results are shown in Figure 3. We observe that no matter using one-step or two-step optimization, incorporating the denoising loss can achieve better performance than not. Overall, the two-step optimization is more effective, which is probably because the more basic conversation encoding should be learned first to help the learning of context denoising. Besides, we find that too large α will hurt the performance. As one denoised sample corresponds to multiple noised samples after performing conversation data augmentation, too large α will further cause the model to pay too much attention to the optimization of denoised ones.

5.5 Turn-level Performance Comparison

In this section, we compare COTED with ConvDR at a more fine-grained turn-level. Figure 4 shows the results. It is clear that our COTED outperforms ConvDR in most of turns on both two datasets. Particularly, the performance advantages of our COTED are more significant in later turns (e.g., No.6 ∼ No.11 turns on CAsT-20). As the conversation goes on, the context becomes longer and more noise appears. But thanks to the advanced design for context denoising, our COTED is more robust to alleviate the negative impacts of noisy turns and keeps better retrieval performance.

5.6 Case Study

We finally show some typical winning cases in Table 6 to help more intuitively understand how COTED achieves better performance. In the first case, the model is expected to understand "the problem" as "provide a habitat for bees" in the current turn, which only needs the 1st, the 3rd, and the 5th previous turns as the necessary context. From the clue words "native" and "region" in the retrieved passage, we can identify that COTED correctly recovers the lost semantic information. While ConvDR seems to be overwhelmed by the long context and confused about what "the problem" is, as its retrieved passage is not related to "habitat". Similarly, in the second case, ConvDR is not sure "what to get started" ("snowboarding" or "strap-in bindings"), while our COTED accurately recognizes that the omitted object should be "snowboarding". We also find that the learned query representations provide some hints to understand the different behaviors of COTED and ConvDR. Specifically, compared with ConvDR, COTED tends to give less attention to the noisy turns. The larger attention scores that ConvDR pays to the 4th turn in the first case (i.e., 302.1) and to the 3rd and 5th turns (i.e., 317.9 and 315.3) in the second case may illustrate its results.

Table 6: Two typical winning examples of COTED on CAsT-20. The current query is underlined. The blue turns are necessary turns. The first disagreed retrieved passages of COTED, ConvDR and Dense-Manual are shown. For each turn, the two [bracketed] numbers are the dot product similarities between its representation and the representation of the whole conversation in COTED (first) and ConvDR (second), respectively. The red words in passages are clues to help understand the model behaviors.

CAsT-20 Topic-83	**CAsT-20 Topic-94**
1: What are some interesting facts about bees? [300.1] [299.3] 2: Why doesn't it spoil? [280.7] [281.3] 3: Why are so many dying? [295.5] [298.2] 4: What can be done to stop it? [294.5] [302.1] 5: What has happened to their habitat? [316.7] [309.2] 6: What can I do to help with the problem?	1: How did snowboarding begin? [337.0] [332.6] 2: Interesting. That's later than I expected. Who were the winners? [289.2] [296.5] 3: What are strap-in bindings? [304.3] [317.9] 4: What's an alternative to this binding style? [290.1] [296.0] 5: What else do I need for my first time? [292.7] [315.3] 6: How can I teach myself to get started?
Manual Oracle: What can I do to help provide a habitat for bees?	**Manual Oracle**: How can I teach myself to get started snowboarding?
First Disagreed Retrieved Passages	
COTED: Plant native flowers. Native flowers help feed your bees and are uniquely adapted to your region. Select single flower tops such as daisies and marigolds, rather than double flower tops such as double impatiens ...	**COTED**: ... boots should be well-fitted, with toes snug in the end of the boot when standing upright ... To further help avoid injury ... recommended to use the right technique ... one should be taught by a qualified instructor ...
ConvDR: ... commercial beekeepers specialize in minimal care ... hobbyists can keep their bees going with care ... Bees are highly susceptible to insecticides ...	**ConvDR**: Snowboard boots and bindings are normally far simpler than their downhill counterpart ... when the sport was first ... more common to use semi ...
Dense-Manual: ... plant a bee garden and create an oasis for bees and ...	**Dense-Manual**: Learn to snowboard in a day. 1. Being strapped to a board at ...

6 CONCLUSION

In this paper, we empirically identify the negative impacts of noisy turns on the few-shot learning of the conversational query encoder. To tackle it, We present a new framework COTED. The three important components of COTED jointly help to achieve better context denoising and generalization abilities for the model. Extensive experiments and analysis on two CAsT datasets justify the superiority of our method over the state-of-the-art baselines.

Limitations and Future work: Our work shows the importance and feasibility of context denoising for conversational search. Nevertheless, it needs human efforts to annotate the necessary turns of the training conversations, making it quite laboursome to be tested on other large-scale conversational search-related data (e.g., ORQuAC). Designing an automatic annotation method (even coarse-grained) may be a possible solution to complement our method. Due to the lack of data in the current conversational search field, our work presents a solution, i.e., COTED, to improve the performance of the conversational search model in the few-shot scenario. But the idea of COTED is scalable to large-scale scenarios. It is also interesting to explore how the ranking loss can fit into our framework after we can get the necessary turns of large-scale conversations. We leave them to future work.

Developing better model architectures requires a large amount of training data. Now, the development of conversational search is largely hindered by the data scarcity problem. More and richer conversational search-related data, such as search-oriented conversations, conversational query-passage/answer relevance labels, and conversational query rewrites, would all be a huge boost to this promising research field. Therefore, in the future, it is important to study the data augmentation method for conversational search.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. The work was partially done at Key Laboratory of Data Engineering and Knowledge Engineering, MOE. This work was supported by National Natural Science Foundation of China No. 61872370, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, China Unicom Innovation Ecological Cooperation Plan, Beijing Academy of Artificial Intelligence(BAAI), and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. The authors would like to thank the reviewers for their valuable comments and thank Liangcai Su and Dr. Jieming Zhu for their helpful discussion. We also acknowledge the support provided and contribution made by the Public Policy and Decision-making Research Lab of RUC.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[2] Yudong Chen, Xin Wang, Miao Fan, Jizhou Huang, Shengwen Yang, and Wenwu Zhu. 2021. Curriculum meta-learning for next poi recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2692–2702.

[3] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 34–90.

[4] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview.. In *In Proceedings of TREC*.

[5] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. In *In Proceedings of TREC*.

[6] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview.. In *TREC*.

[7] Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 5917–5923.

[8] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *ACL/IJCNLP (Findings) (Findings of ACL)*, Vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 968–988.

[9] Jianfeng Gao, Chenyan Xiong, and Paul Bennett. 2020. Recent Advances in Conversational Information Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2424.

[10] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural Approaches to Conversational Information Retrieval. *arXiv preprint arXiv:2201.05176* (2022).

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos,

and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *NeurIPS*.

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* 7, 3 (2021), 535–547.

[13] 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 *EMNLP (1)*. Association for Computational Linguistics, 6769–6781.

[14] Vaibhav Kumar and Jamie Callan. 2020. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

[15] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6086–6096.

[16] Yongqi Li, Wenjie Li, and Liqiang Nie. 2021. A Graph-guided Multi-round Retrieval Method for Conversational Open-domain Question Answering. *arXiv preprint arXiv:2104.08443* (2021).

[17] Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. Dynamic Graph Reasoning for Conversational Open-Domain Question Answering. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–24.

[18] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.

[19] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386* (2020).

[20] 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)*.

[21] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Query reformulation using query history for passage retrieval in conversational search. *arXiv preprint arXiv:2005.02230* (2020).

[22] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–29.

[23] Hang Liu, Meng Chen, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Conversational Query Rewriting with Self-Supervised Learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7628–7632.

[24] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[25] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Trans. Assoc. Comput. Linguistics* 9 (2021), 329–345.

[26] 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 *CoCo@ NIPS*.

[27] Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. 2018. Perceive your users in depth: Learning universal user representations from multiple e-commerce tasks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 596–605.

[28] Gustavo Penha and Claudia Hauff. 2019. Curriculum learning strategies for ir: An empirical study on conversation response ranking. *arXiv preprint arXiv:1912.08555* (2019).

[29] Chen Qu, Chenyan Xiong, Yizhe Zhang, Corby Rosset, W Bruce Croft, and Paul Bennett. 2020. Contextual Re-Ranking with Behavior Aware Transformers. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1589–1592.

[30] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 539–548.

[31] Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019. Attentive History Selection for Conversational Question Answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. 1391–1400.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[33] 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*. 117–126.

[34] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* 6, 1 (2019), 1–48.

[35] Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 751–759.

[36] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 355–363.

[37] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. In *ECIR (2) (Lecture Notes in Computer Science)*, Vol. 12657. Springer, 418–424.

[38] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. 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*.

[39] 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*. 921–930.

[40] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[41] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. [n.d.]. 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*.

[42] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6095–6104.

[43] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*. 1933–1936.

[44] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*.

[45] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.

[46] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.