CFDA & CLIP at TREC 2022 Conversational Assistance Track (CAsT)

Jia-Huei Ju[*], **Sheng-Chieh Lin**[†], **Li-Young Chang**[‡],
Ming-Feng Tsai[‡] and **Chuan-Ju Wang**[*]
[*] Research Center for Information Technology Innovation, Academia Sinica,
[†] David R. Cheriton School of Computer Science University of Waterloo,
[‡] Department of Computer Science, National Chengchi University

Abstract

In this notebook, we introduce a new pipeline for TREC CAsT 2022. Comparing to the common multistage pipeline for conversational search, we experimented an alternative that does not require conversational query reformulation (CQR). Specifically, our pipeline equipped with conversational dense retriever and conversational passage re-ranker. Our empirical evaluation result on TREC CAsT dataset is also reported in this paper.

1 Introduction

One of the main challenge of conversational search is the ambiguous user's information needs. The latter turn user's information needs in the conversation (i.e., the raw utterances) often leaves out the important context. Some of these context-missing queries may result in the poor effectiveness in the scenario of conversational information seeking. To resolve the problem, the conversational query reformulation methods (CQR) is the crucial component in conversational information seeking systems. The CQR models aim at replenishing the context for current turn of user's utterances from historical context.

For example, Voskarides et al. (2020) used contextualized text features to classify important historical context (words); Lin et al. (2020) used the sequence-to-sequence (T5) model to generate queries with standalone meaning. Both of which augmented the CANARD dataset (Elgohary et al., 2019), which has the aligned pairs of multi-turn raw utterances and manually rewritten query. Recently, some studies have made substantial progress on dense retrieval approaches for conversational search (Qu et al., 2020; Yu et al., 2021; Lin et al., 2021b). They integrate CQR into dense retrieval models; thus, the models can directly retrieve passages in an end-to-end manner similar to standard ad-hoc retrieval.

In this paper, we treat conversational dense retrieval as the first-stage retrieval in our pipeline. As for the second-stage in our pipeline, we follow these works and build up an experimental conversastional passage re-ranking models. Specifically, we fine-tune a conversational passage re-ranking model (ConvRerank), which aims at refining the passage candidates retrieved from contextualized query embedding (CQE) approach (Lin et al., 2021b), one of the conversational dense retrieval methods. For the passage re-ranking model, we used monoT5 (Nogueira et al., 2020) and further fine-tune it with weakly-supervised training data from CQE (Lin et al., 2021b). Finally, by combining the first-stage retrieved top 1000 passage candidates with ConvRerank, we construct a multistage pipeline without CQR, namely CQR-free multistage pipeline.

In the following sections, we introduce the common multistage pipeline approaches in Section 2, including the CQR-driven pipeline (Section 2.1) and the CQR-free pipeline (Section 2.2). Section 3 reports our experimental results on TREC CAsT dataset.

2 The Multistage Pipeline for Conversational Search

In this section, we describe two settings of multistage pipeline for conversational search. First, we recap preliminaries in previous multistage pipeline, which we regarded as the baseline submitted runs. Second, we introduce the CQR-free multistage pipeline.

2.1 Preliminary

Formally, the goal of conversational search is to retrieve the relevant document d from collections D with a sequence set of multi-turn user utterances $U = (u_1, u_2, ..., u_i)$, where u_i indicates user's utterance at the i-th turn. We will introduce the components we used in our baseline submitted ap-

proach, including the following three stages: (1) Conversational query reformulation; (2) First-stage passage retrieval for top 1000 passages; (3) Second-stage passage re-ranking.

Conversational query reformulation. We use a CQR approach followed the previous work (Lin et al., 2020),[1] T5 neural transfer reformulation (T5-NTR), denoted as $\mathcal{F}^{\text{CQR}}$. We can thereby reformulate the raw utterances at i-turn with the previous context as

$$q_i = \mathcal{F}^{\text{CQR}}(u_1, u_2, ...u_{i-3}, r_{i-3}, u_{i-2}, ...; u_i)^2$$

where q_i is the reformulated query and can be regarded as a standalone, omission-free query for the later stages in *decontextualized* manner. r indicates the system response provided in evaluation set and we use at most three responses due to the length limitation of T5 models. Both user's and system's utterances are regarded as the context for T5-NTR to rewrite.

First-stage passage retrieval. In this stage, we adopt the sparse and dense retrieval methods using Pyserini IR toolkit (Lin et al., 2021a) as the first-stage (candidate) passage retrieval. For the dense retrieval, we use the bi-encoders models, TCT-ColBERT (Lin et al., 2021c), as well as the released checkpoint[3]. With the reformulated query q and segmented passage p, we encode dense representations of passages and index via FAISS (Johnson et al., 2017). Finally, we retrieve the top1000 relevant passages as the pool of first-stage retrieved candidates P_i for each reformulated query q_i. As for the sparse retrieval, we use the BM25 to retrieve top1000 relevant documents; subsequently, we segment the documents into passages[4] as another kind of first-stage retrieved passages pool P_i.

Second-stage passage ranking. In the passage re-ranking stage, we use monoT5 (Nogueira et al., 2020) and the released checkpoint.[5] The text-to-text input format of monoT5 is:

$$\text{Query: } q_i \text{ Document: } p \text{ Relevant:}, \quad (1)$$

where $p \in P_i$ indicates the segmented passages in first-stage retrieved passages pool. Afterwards, we follow the "true/false" token logit trick of monoT5 (Nogueira et al., 2020); the relevance of each query-passage pair can be estimated by softmax of "true" logit over "true/false" logits. The probability is regarded as relevance score for final re-ranking results.

2.2 The CQR-free multistage Pipeline

We introduce a multistage pipeline for conversational search without pre-processing the multi-turn query. Our proposed CQR-free multistage pipeline is comprised of conversational dense retrieval and conversational passage re-ranking.

Conversational dense passage retrieval. As the first-stage of CQR-free multistage pipeline, we follow the contextualized query embeddings (CQE) approach (Lin et al., 2021b). We use the same CQE's conversational query encoder and the released checkpoint[6], which is basically a fine-tuned conversational query encoder, denoted as $\mathcal{F}_q^{\text{ConvDPR}}$. For a certain user's raw utterance u_i, we append its previous context $U_{<i}$ and encoded into a single dense representation as follow

$$\begin{aligned} E_q^i &= \mathcal{F}_q^{\text{ConvDPR}}(U_{<i}; u_i), \\ E_p &= \mathcal{F}_d^{\text{ConvDPR}}(p), \end{aligned} \quad (2)$$

where E_q^i is an aggregated embedding by adopting average pooling over all tokens's last hidden layers, excluding the BERT's special tokens as same as the original work. E_p is the passage representation encoded by $\mathcal{F}_d^{\text{ConvDPR}}$. In CQE paper, the document encoder is identical to TCT-ColBERT fine-tuned on MS MARCO (Bajaj et al., 2016). Similar to dense retrieval described in Section 2.1, we used FAISS-supported Pyserini toolkit. Finally, we acquire the top1000 retrieved passage candidates pool P_i for each raw utterance u_i without query reformulation.

Conversational passage re-ranking. In this stage, we introduce a CQR-free re-ranking model (ConvRerank). We adopted monoT5 model checkpoints (T5-large, fine-tuned on MS MARCO for 100K steps) and further fine-tuned with conversational query $U_{<i}; u_i$ for 20K steps; the training data is similar to the pseudo-labeled dataset in CQE. To

[1] https://huggingface.co/castorini/t5-base-canard.

[2] The parathesis indicates the sequence follows the temporal order in a dialogue.

[3] https://huggingface.co/castorini/tct_colbert-v2-msmarco.

[4] The official segmentation tools: https://github.com/grill-lab/trec-cast-tools/.

[5] https://huggingface.co/castorini/monot5-large-msmarco

[6] https://huggingface.co/castorini/tct_colbert-v2-msmarco-cqe

avoid truncation of T5 tokenization (maximum sequence length is 512), we recast the input of ConvRerank as the following text-to-text format:

$$\text{Query: } u_i \mid \mathcal{J}(U_{<i}) \text{ Document: } p \text{ Relevant:}$$

where $\mathcal{J}$ indicate the joining function that concatenate each elements in the sequence with vertical bars " | " as boundaries. Once the ConvRerank is fine-tuned, we can estimate the relevance scores of each conversational query-passage pairs similar to monoT5 described in Section 2.1.

3 Experiments and Results

Evaluation dataset. To validate the effectiveness of our proposed methods, we use the TREC CAsT evaluation set released in 2020. The dataset has 208 evaluation queries with human-judged relevance scores of passages from 0-4. The relevant passages are from TREC CAR (Nanni et al., 2017) and MS MARCO (v1) passage ranking dataset (Bajaj et al., 2016).

Compared Methods We compare few settings of CQR-driven multistage pipeline as our baselins. Our baselines are all equipped with conversational query reformulation (CQR), and using the T5 rewritten query (See Section 2.1). BM25 and TCT-ColBERT are two of our baseline retriever; monoT5 is our baseline passage re-ranking model.

As for the CQR-free multistage pipeline, we use CQE and ConvRerank (see Section 2.2) for the automatic session. In TREC CAsT 2022 submitted runs, we only submitted the results using BM25 and CQE as the first-stage retrieval.

Results on CAsT 2020. In Table 1, we report the full ranking resuls of TREC CAsT 2020 evaluation topics with the nDCG cut-off at 3, 5, 500 and 1000 on the columns. The first two blocks in the Table are our baseline multistage pipeline, and the pipeine with asterisk marks in last two blocks are our proposed CQR-free multistage pipeline.

Generally, we observed that our proposed CQR-free multistage pipeline achieve the higher effectiveness of nDCG in shallower depth. Specifically, the CQE approach is superior than the BM25 and TCT-ColBERT (with reformulated query) in terms of nDCG cutoff at 3 and 5. Moreover, ConvRerank in the last two blocks approach outperform the baseline T5 reranking models (monoT5) in almost all judgement settings. We conclude that our proposed CQR-free multistage pipeline with conversationally encoded query can provide more fine-grained historical context in the latent space compared to performing CQR in advance.

Table 1: The full-ranking results of CQR-free pipeline, with nDCG judgements cut-off at 3, 5, 500 and 1000.

Pipeline	CQR	nDCG			
		3	5	500	1000
BM25	✓	0.1464	0.1432	0.2582	0.2824
+ monoT5	✓	0.3701	0.3613	0.4067	0.4089
TCT-ColBERT	✓	0.3381	0.3271	0.4349	0.4520
+ monoT5	✓	0.3819	0.3786	0.4801	0.4888
CQE	✗	0.3416	0.3288	0.4335	0.4532
+ monoT5	✓	0.3987	0.3876	0.4838	0.4946
+ ConvRerank*	✗	0.4026	0.3973	0.4818	0.4977
CQE-hybrid	✗	0.3676	0.3506	0.4752	0.4954
+ monoT5	✓	0.3939	0.3857	0.5051	0.5196
+ ConvRerank*	✗	0.4087	0.3993	0.5097	0.5273

Table 2: The full-ranking results of our submitted approaches using baseline multistage pipeline and CQR-free multistage pipeline.

Pipeline (Run)	CQR	nDCG@3	nDCG	Recall
Type: Automatic				
BM25 + monoT5 (CNC_AS)	✓	0.235	0.369	0.515
BM25 + ConvRerank (CNC_AS-C)	✗	0.377	0.411	0.527
CQE + monoT5 (CNC_AD)	✓	0.334	0.286	0.320
CQE + ConvRerank (CNC_AD-C)	✗	0.347	0.294	0.320
Type: Manual				
BM25 + ConvRerank (CNC_AD-C)	-	0.397	0.537	0.702
TCT-ColBERT + ConvRerank (CNC_MD-C)	-	0.512	0.350	0.339

Evaluation on CAsT 2022. In Table 2, we reported our submitted runs and evaluation results on TREC CAsT 2022. Two types of CAsT 2022 tasks, automatic and manual are in the first and second blocks in the table, respectively. As expected, the CQR-free multistage pipeline outperformed our baseline multistage pipeline (i.e, the pipeline with query rewriting). Particularly, ConvRerank showed the better re-ranking effectiveness under different first-stage retrieval setting regardless of BM25 and CQE (the run CNC_AS-C and CNC_AD-C). Note that all our submitted runs using dense retrieval (e.g. CQE, TCT-ColBERT), we only use the first 4 segmented passages for each document in provided collections. Therefore, some results of dense retrieval may be inferior to the pipeline using BM25.

4 Conclusion

According to the evaluation, the CQR-free multistage pipeline seems to be better than traditional multistage pipeline. Particularly, we hypothesize

the dependencies of turns (of utterances) may affect the relevance of passage. For example, the passages which previous user's information needs had been satisfied shall be considered as less relevant. To attest our hypothesis, we leave the rationalizing CQE or ConvRerank as our future works.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021c. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online. Association for Computational Linguistics.

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

Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. 2017. Benchmark for complex answer retrieval.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

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

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *SIGIR 2020: 43rd international ACM SIGIR conference on Research and Development in Information Retrieval*. ACM.

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 '21, page 829–838, New York, NY, USA. Association for Computing Machinery.