Open-Domain Conversational Question Answering with Historical Answers

Hung-Chieh Fang[*] **Kuo-Han Hung**[*] **Chao-Wei Huang** **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
{b09902106,b09902120,f07922069}@csie.ntu.edu.tw
y.v.chen@ieee.org

Abstract

Open-domain conversational question answering can be viewed as two tasks: passage retrieval and conversational question answering, where the former relies on selecting candidate passages from a large corpus and the latter requires better understanding of a question with contexts to predict the answers. This paper proposes **ConvADR-QA** that leverages historical answers to boost retrieval performance and further achieves better answering performance. Our experiments on the benchmark dataset, OR-QuAC, demonstrate that our model outperforms existing baselines in both extractive and generative reader settings, well justifying the effectiveness of historical answers for open-domain conversational question answering.[1]

1 Introduction

Conversational information seeking and conversational question answering (CQA) are fundamental tasks of dialogue systems (Gao et al., 2018). The conversational agents are expected to serve as nature interfaces for users' information need, providing information and answers via multi-turn natural language interactions. The multi-turn natural of CQA makes it challenging as the queries are contextualized, requiring the systems to resolve coreference and ambiguities. With recent advances in language understanding and dialogue modeling, along with the curation of large-scale datasets, e.g., QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019), we have seen substantial progress in CQA.

While the state-of-the-art (SOTA) models have achieved performance comparable or even superior than human performance on QA and CQA datasets, this setting is highly limited as it requires the source document containing evidence to be given, which is unlikely the case in real-world sce-

narios. To address this issue, researchers have expanded the scheme of CQA to an open-domain setting, where the document containing evidence must be retrieved from a large candidate pool (Qu et al., 2020). In the open-domain setting, there are usually millions of candidate documents, making the conventional method which jointly encodes the query and the document infeasible (Chen et al., 2017). The dominant technique to tackle the challenge is dense retrieval (Karpukhin et al., 2020; Qu et al., 2020; Xiong et al., 2021), which encodes a query and documents as dense representations separately and performs nearest neighbor search that is efficient and scalable to millions of documents. It has been shown to outperform traditional sparse retrieval methods on multiple QA benchmarks.

However, applying dense retrieval for conversations may need to consider the dialogue context and structure, which is not trivial. Qu et al. (2020) proposed ORConvQA to include previous questions in the same dialogue. where the context-dependent nature of questions is shown useful. ConvDR (Yu et al., 2021) further improved the retrieval performance by knowledge distillation on reformulated questions with an ad-hoc teacher model. Nevertheless, simply concatenating historical questions is suboptimal. Our hypothesis is that rather than relying on the model to infer helpful knowledge from historical questions, we provide direct signals by adding historical answers to the input. Hence, we propose **ConvADR-QA** (**Conv**ersational **A**nswer-aware **D**ense **R**etrieval) to leverage historical answers for better retrieval and then answering performance for open-domain CQA.

2 Related Work

CQA A unique challenge to CQA is that the questions are context-dependent. Hence, most prior work focused on various history modeling techniques (Huang et al., 2018; Yeh and Chen, 2019; Qu et al., 2019b; Chen et al., 2020). Choi et al.

---

[*]Equal contribution.
[1]The source code is available at https://github.com/MiuLab/ConvADR-QA.

(2018) proposed to mark the previous answers in the passage by adding an answer embedding to the input embeddings. Qu et al. (2019a) extended this method to the large pre-trained language models. However, Chiang et al. (2020) showed that prior conversational models do not fully understand the content, implying that CQA still needs further investigation. While our method also leverages historical answers as additional input signal, our major contribution is that we apply this technique to dense retrieval instead of question answering for better practicality in a open-domain setting.

Open-Domain QA Without a given target passage, most work for this task was built upon the dense retrieval framework for retrieving relevant passages for QA. DPR (Karpukhin et al., 2020) first showed that dense retrieval outperforms sparse retrieval methods. GAR (Mao et al., 2021a) introduced pseudo relevance feedback by augmenting queries with generated texts. RIDER (Mao et al., 2021b) proposed a simple passage reranking method which promotes the passages containing the predicted answers. While these methods consider the predicted answers, they aim at improving single-turn question answering. We instead focus on enhancing model's ability on handling multi-turn conversational questions.

Open-Domain CQA Researchers have put increasing attention on open-domain CQA with the TREC Conversational Assistance Track (Dalton et al., 2020, 2021). However, these datasets have limited supervision, making dense retrieval hardly applicable due to its data-hungry nature. Qu et al. (2020) introduced the first large-scale open-domain CQA data, OR-QuAC, by extending QuAC (Choi et al., 2018) to a open-domain setting. They also proposed ORConvQA, a pipeline system with a DPR retriever and an extractive reader, as a baseline system. ConvDR (Yu et al., 2021) proposed to reformulate questions into their context-independent rewrites with the CANARD dataset (Elgohary et al., 2019), then applied knowledge distillation using a ad-hoc teacher model. Our method is built upon these two methods by incorporating historical answers to aid the retriever. Li et al. (2021) proposed a graph-guided retrieval method which constructs a graph using passages with historical answers and potential answers. Our work does not introduce extra parameters and complex modeling, and we demonstrate that we can achieve better results with a simpler design for better practicality. The All-
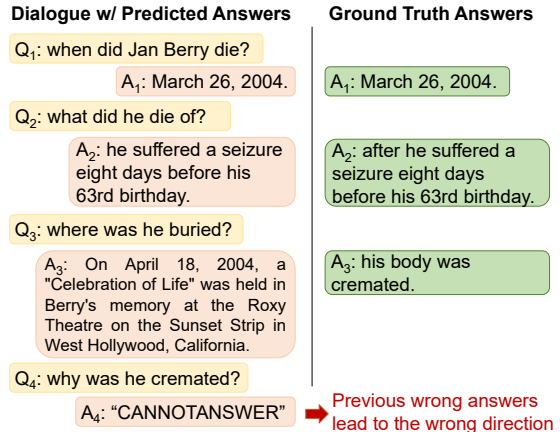

Figure 1: Demonstration of how previous answers affect the quality of an answer.

History strategy from TopiOCQA (Adlakha et al., 2022) is very similar to ours. However, their experimental setting is not realistic as they used ground truth answers as historical answers, which is corresponding to our oracle setting.

3 ConvADR-QA

Let C denote the passage collections with N passages $\{p_i\}_{i=1}^N$, where p_i can be viewed as a sequence of tokens $p_i^1, \ldots, p_i^l$. Given the t-th question q_t and all historical questions $\{q_i\}_{i=1}^{t-1}$ in a conversation, the task of open-domain CQA is to predict a_t from C. In an extractive setting, a_t is a span $p_i^s, \ldots, p_i^e$ from a passage p_i.

The difficulty of open-domain CQA is that the current question usually requires context information from previous turns, which makes it harder for the system to capture the latent information compared with the open-domain QA task. Previous work on open-domain conversational search addressed the problem by concatenating the current and historical questions *without* answers (Qu et al., 2020; Yu et al., 2021). Our motivation is that historical answers can also provide the important signal for the current question to obtain the answers illustrated in Figure 1.

To better leverage the historical answers for open-domain CQA, we propose ConvADR-QA illustrated in Figure 2, which includes a retriever for obtaining relevant passages from a large collection and a reader for CQA.

3.1 Retriever

Following the prior work (Karpukhin et al., 2020; Xiong et al., 2021), we apply a dense retrieval method, which has shown dominant performance
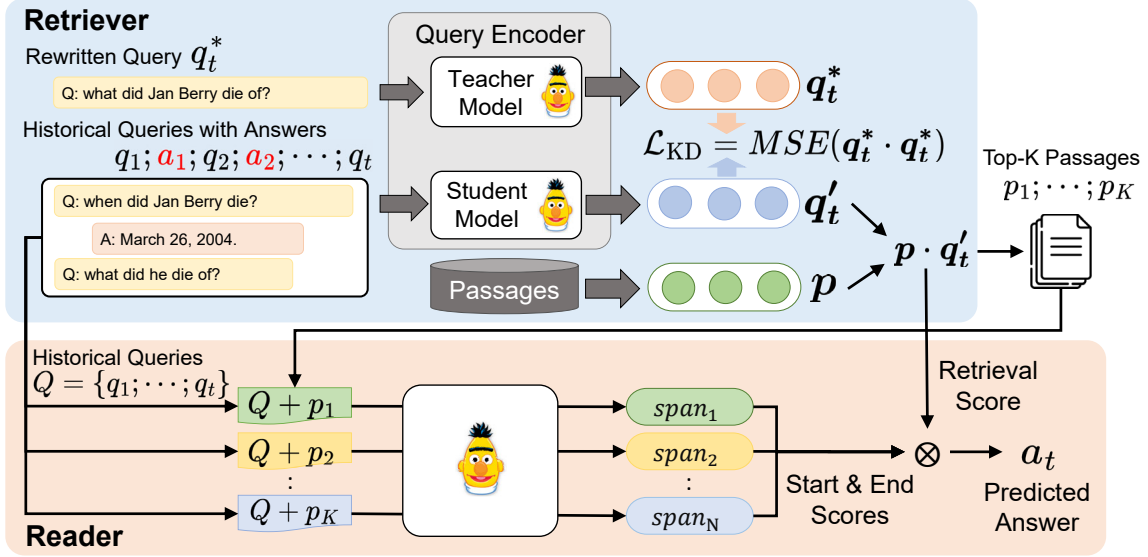
Figure 2: Illustration of our proposed ConvADR-QA model.

over sparse ones. Specifically, the model uses a dual-encoder architecture to map passages and questions to the same embedding space. The input of our question encoder is the concatenation of historical questions and answers:

$$\boldsymbol{p} = E_P(p), \boldsymbol{q}'_{\boldsymbol{k}} = E_Q(\{q_i, a_i\}_{i=1}^{k-1}; q_k).$$

The retrieval score is then defined as the dot product of the passage embedding and the question embedding:

$$S_{\text{rt}}(q_k, p) = \boldsymbol{p} \cdot \boldsymbol{q}'_{\boldsymbol{k}}.$$

In the training process, each question contains one gold passage p^+ and a set of negative passages P^-, ConvADR-QA is then optimized using the negative log likelihood loss:

$$\mathcal{L}_{\text{NLL}} = -\log \frac{e^{S_{\text{rt}}(q_k, p^+)}}{e^{S_{\text{rt}}(q_k, p^+)} + \sum_{p^- \in P^-} e^{S_{\text{rt}}(q_k, p^-)}}.$$

3.2 Knowledge Distillation

In conversational search, dense retrieval is challenging since the current question requires information from previous turns, which aggravate the discrepancy between question embeddings and passage embeddings. Yu et al. (2021) recently addressed the problem using a teacher-student framework to distill knowledge from an ad-hoc teacher model. The input of the teacher model is a manually-rewritten context-independent query $q_k^\star$, and the knowledge distillation (KD) loss is defined as the mean square error (MSE) loss between the teacher's and the student's question embeddings:

$$\boldsymbol{p} = E'_P(p), \boldsymbol{q}_{\boldsymbol{k}}^\star = E'_Q(q_k^\star),$$

$$\mathcal{L}_{\text{KD}} = \text{MSE}(\boldsymbol{q}_{\boldsymbol{k}}^\star, \boldsymbol{q}'_{\boldsymbol{k}}).$$

The retrieval loss of our multi-task learning setting is the sum of NLL loss and KD loss:

$$\mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{KD}}.$$

3.3 Reader

The task of the reader is to extract a span from passages as the final answer. We use a standard BERT model for the machine comprehension task (Devlin et al., 2019). Given the t-th question q_t and top-K candidate passages $\{p_i\}_{i=1}^K$ retrieved by our retriever, the reader first extracts a span for each passage by choosing the highest score of start and end tokens. The score of the m-th token is defined as follows:

$$S_{\text{start}}^{[m]}(q_t; p) = W_{\text{start}}\text{BERT}(\{q_i\}_{i=1}^t; p)[m],$$

$$S_{\text{end}}^{[m]}(q_t; p) = W_{\text{end}}\text{BERT}(\{q_i\}_{i=1}^t; p)[m],$$

$$S_{\text{rd}}(q_t; p) = \max_{m1, m2}[S_{\text{start}}^{[m1]}(q_t; p) + S_{\text{end}}^{[m2]}(q_t; p)].$$

We choose the final answer by multiplying the retriever score S_{rt} and the sum of start/end token score as the reader score S_{rd}:

$$S(q_t, p) = S_{\text{rt}}(q_t, p) \cdot S_{\text{rd}}(q_t; p).$$

4 Experiments

We conduct the experiments on an open-domain CQA benchmark: OR-QuAC (Qu et al., 2020). OR-QuAC is an open-domain conversational retrieval dataset that aggregates three existing datasets: (1)

	Method	Historical Answers	Retrieval			Answering		
			MRR@5	R@5	MAP@10	HEQ-Q	HEQ-D	F1
Extractive	ORConvQA	✗	31.3	31.4	-	24.10	0.60	29.4
	Graph-Guided	predicted	35.1	36.7	-	30.30	1.00	33.4
	ConvDR→Reader	✗	61.6	75.0	60.7	29.92	0.78	36.2
	ConvADR-QA (Reader)	predicted	**66.8**	**77.9**	**64.6**	**32.11**	**1.16**	**38.4**
	ConvADR-QA (Reader)	gold	*74.5*	*82.5*	*71.7*	*35.69*	*1.03*	*42.3*
Generative	RAG	✗	29.9	30.8	28.5	21.98	0.25	26.1
	ConvDR→FiD	✗	**61.6**	75.0	60.7	27.21	0.86	31.5
	ConvADR-QA (FiD)	predicted	60.9	**76.2**	**62.9**	**28.76**	**1.04**	**33.6**
	ConvADR-QA (FiD)	gold	*74.5*	*82.5*	*71.7*	*30.83*	*0.91*	*35.1*

Table 1: Performance on OR-QuAC (%). Best results are marked in **bold**. Oracle results are in *italic*.

the QuAC dataset (Choi et al., 2018) which contains 14K information-seeking QA dialogs, (2) the CANARD dataset (Elgohary et al., 2019) which rewrites context-dependent queries to self-contained questions based on QuAC, and (3) the Wikipedia corpus dump from 10/20/2019 which extends QuAC to the open-domain setting. The experimental setting is detailed in Appendix A.

Following Yu et al. (2021), we use three commonly used metrics, MRR@5, Recall@5 and MAP@10, to evaluate the retrieval performance. In addition, we use word-level F1 and human equivalence score (HEQ) provided by the QuAC challenge to evaluate the overall performance of our system. The definitions of above metrics are detailed in Appendix B.

4.1 Baselines

We compare our model with recently proposed baselines for open-domain CQA, ORConvQA (Qu et al., 2020), Graph-Guided (Li et al., 2021), and RAG (Lewis et al., 2020) for both extractive and generative settings.

- **ORConvQA**: It is an end-to-end system for the open-domain CQA task, which includes a retriever, a reranker and a reader. The retriever use the dense retrieval method where the input of the query encoder is the concatenation of the current and historical questions.

- **Graph-Guided**: Li et al. (2021) proposed a graph-guided retrieval method that models the relations among answers across conversational turns, which is the first work attempting at utilizing historical answers for open-domain CQA. This model utilizes a graph built from the hyperlink-connected passages

containing historical answers to better retrieve relevant passages.

- **RAG**: It is a generation model that can access to pre-trained parametric memory and non-parametric memory like wikipedia. It has shown good performance on the open-domain QA task, we further adapt it to the open-domain CQA task by doing the following modifications: (1) finetuning the base model, where the input of the question encoder is the concatenation of the current and historical questions, (2) using passages from OR-QuAC as our knowledge source (non-parametric memory).

In addition to the existing open-domain CQA approaches, we further implement two baseline where we use ConvDR as the retriever model. ConvDR is a conversational dense retriever, which uses the few-shot strategy to mimic the embeddings of manual oracle queries from an ad hoc dense retriever. It is also the current SOTA model in the retrieval stage. We adopt it to open-domain CQA by enabling it with QA capability using two existing models to generate answers: (1) Reader of ORConvQA (Qu et al., 2020), which adapts a BERT-based extractive QA model to a multi-document setting, (2) FiD (Izacard and Grave, 2021), which uses a sequence-to-sequence model to generate the answer given the input is the question and retrieved passages, which has shown great performance at combining evidences from multiple passages.

4.2 Results

Table 1 summarizes our experimental results. It is obvious that our proposed ConvADR-QA outperforms almost all existing baseline models in both re-

trieval and answering stages, achieving new SOTA performance of open-domain CQA. We can observe that in both extractive and generative QA settings, our model which leverages predicted answers achieves better performance over the one without answers. Moreover, the graph-guided approach also utilizes historical answers in a more complex way, but performs worse than our ConvADR-QA, demonstrating that our model leverages answer signal more effectively. We also report the oracle results using gold historical answers. It shows that the model with gold answers outperforms the one with predicted answers in most of the metrics except HEQ-D. Note that the oracle results can be viewed as the upper bound of our method, as the gold answers are not available during inference. The results well justify our hypothesis that historical answers are informative for open-domain CQA.

Notably, we can notice that the quality of predicted answers can significantly affect the retrieval performance. Our experiment shows that MRR@5 drops when using FiD as the reader, demonstrating that a QA model with weak performance could potentially hurt retrieval performance. Our hypothesis is that due to its lower answering quality, the errors would propagate through the conversation and mislead the retriever, indicating that further improvement on reader performance could also improve the retrieval performance of our method. In sum, the experimental results show the effectiveness of our model for open-domain CQA in both extractive and generative settings. An example is presented in Table 3 for qualitative analysis, where it can be shown that the previous answers affect the following prediction results. More analysis can be found in Appendix C.

4.3 Error Propagation Analysis

To inspect the impact of the errors propagated through the conversation and reduce the robustness. we conduct analysis on accuracy against number of turns in Figure 3. It shows that the benefits of adding the answers outweigh the error propagation, where ConvADR-QA outperforms ConvDR in earlier turns, which tends to drop as the dialogue gets too long. It implies that the issue about error propagation still have a large room for improvement.

5 Conclusions

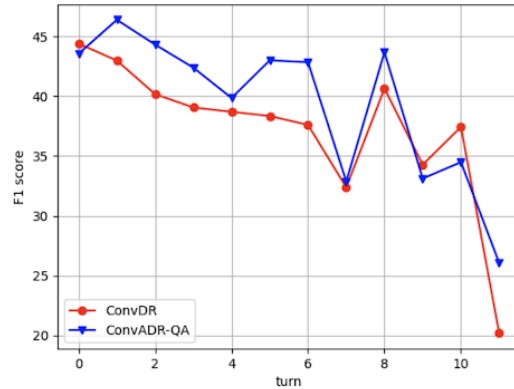This work introduces **ConvADR-QA**, an open-domain CQA model that leverages historical an-

Figure 3: Analysis on accuracy against number of turns.

Q1: What is Roberto Mangabeira Unger,'s programmatic thought?
A1: Key in Unger's thinking is the need to re-imagine social institutions before attempting to revise them.
ConvDR: The beginning of Unger's academic career began with the books Knowledge and Politics and Law in Modern Society,
ConvADR-QA: Key in Unger's thinking is the need to re-imagine social institutions before attempting to revise them.

Q2: Can you explain the mechanism of thinking?
A2: In building this program, however, we must not entertain complete revolutionary overhaul, lest we be plagued by three false assumptions:
ConvDR: CANNOTANSWER
ConvADR-QA: In building this program, however, we must not entertain complete revolutionary overhaul, lest we be plagued by three false assumptions:

Q3: What are the three false assumptions?
A3: Typological Fallacy:
ConvDR: Unger finds three weaknesses that crippled the theory: foremost, the theory claimed that equilibrium would be spontaneously generated in a market economy.
ConvADR-QA: Typological Fallacy: the fallacy that there is closed list of institutional alternatives in history, such as "feudalism" or "capitalism".

Table 2: Qualitative analysis.

swers. The experiments on a benchmark dataset demonstrate that our proposed method outperforms all baselines for both retrieval and answering performance. Our results justify not only the importance of historical answers in a conversation but also the generalizability to different types of readers.

Acknowledgements

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1230–1236. International Joint Conferences on Artificial Intelligence Organization. Main track.

Ting-Rui Chiang, Hao-Tong Ye, and Yun-Nung Chen. 2020. An empirical study of content understanding in conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7578–7585.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. Cast 2020: The conversational assistance track overview. Technical report, Technical report.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7, Melbourne, Australia. Association for Computational Linguistics.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick 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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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*.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021a. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021b. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350, Online. Association for Computational Linguistics.

C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer. 2019a. Bert with history answer embedding for conversational question answering. In *SIGIR '19*.

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, pages 539–548.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019b. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1133–1136, New York, NY, USA. Association for Computing Machinery.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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 *International Conference on Learning Representations*.

Yi-Ting Yeh and Yun-Nung Chen. 2019. FlowDelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90, Hong Kong, China. Association for Computational Linguistics.

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*, pages 829–838.

A Reproducibility

Our source code and the trained model was published at GitHub together with running scripts for better reproducibility. All models are trained with 2 Nvidia Quadro P6000. For the retriever, we set the training batch size to 4, the number of epochs to 3, and the learning rate to 1e-5. For the reader, we set the training batch to 2, the number of epochs to 3, the max sequence length to 512, the max question length to 125 and the learning rate to 3e-5.

B Evaluation Metrics

We use following metrics for evaluating our proposed model.

- **MRR**: Reciprocal Rank (RR) calculates the reciprocal of the rank where the first relevant passage was retrieved, and MRR averages the reciprocal rank across all questions.

- **Recall**: The proportion of the questions that the answer is in the retrieved passages.

- **MAP**: Mean Average Precision is the mean of the average precision scores for each question.

- **F1**: It measures the overlap of the predicted answer span and the ground truth answer span at the word level.

- **HEQ**: HEQ-Q measures the percentage of questions where the system matches or surpasses human performance in terms of F1 score. HEQ-D measures the percentage of dialogues in which all questions have an HEQ-Q = 1.

C Qualitative Study

The results generated by ConvDR and ConvADR-QA are presented in Table 3, where the underline texts indicate the spans appearing in the ground truth answers. It can be found that historical answers can be an important signal to obtain the answers.

	Example 1
Q1	What was unique about Joe Greene's attitude?
A1	ferocious and intimidating.
ConvDR	Greene's nickname remained popular due to his exploits on the playing field, where he was described as ferocious and intimidating.
ConvADR-QA	Greene's nickname remained popular due to his exploits on the playing field, where he was described as ferocious and intimidating.
Q2	Did he intimidate other players?
A2	He instilled fear in opponents with the intensity of his play.
ConvDR	Lynn Swann, a wide receiver, considered Greene a mentor.
ConvADR-QA	He instilled fear in opponents with the intensity of his play.
Q3	Who did he play for?
A3	Pittsburgh
ConvDR	Houston Oilers
ConvADR-QA	Houston Oilers
Q4	Did his teams do well?
A4	CANNOTANSWER
ConvDR	Greene was named the captain of the defense in 1977.
ConvADR-QA	CANNOTANSWER
Q5	What else should I know about his playing style?
A5	In his early years with the Steelers, Greene was at times uncontrollable, and often let his temper get the best of him.
ConvDR	Greene was named the captain of the defense in 1977.
ConvADR-QA	In his early years with the Steelers, Greene was at times uncontrollable, and often let his temper get the best of him.
	Example 2
Q1	What is Roberto Mangabeira Unger,'s programmatic thought?
A1	Key in Unger's thinking is the need to re-imagine social institutions before attempting to revise them.
ConvDR	The beginning of Unger's academic career began with the books Knowledge and Politics and Law in Modern Society,
ConvADR-QA	Key in Unger's thinking is the need to re-imagine social institutions before attempting to revise them.
Q2	Can you explain the mechanism of thinking?
A2	In building this program, however, we must not entertain complete revolutionary overhaul, lest we be plagued by three false assumptions:
ConvDR	CANNOTANSWER
ConvADR-QA	In building this program, however, we must not entertain complete revolutionary overhaul, lest we be plagued by three false assumptions:
Q3	What are the three false assumptions?
A3	Typological Fallacy:
ConvDR	Unger finds three weaknesses that crippled the theory: foremost, the theory claimed that equilibrium would be spontaneously generated in a market economy.
ConvADR-QA	Typological Fallacy: the fallacy that there is closed list of institutional alternatives in history, such as "feudalism" or "capitalism".

Table 3: The comparison between ConvDR and ConvADR-QA.