Information Processing and Management 60 (2023) 103475

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Generative retrieval for conversational question answering

Yongqi Li [a,*], Nan Yang [b], Liang Wang [b], Furu Wei [b], Wenjie Li [a]

[a] The Hong Kong Polytechnic University, Hong Kong
[b] Microsoft, United States of America

ARTICLE INFO

Keywords:
Conversational question answering
Generative retrieval

ABSTRACT

Effective passage retrieval is crucial for conversation question answering (QA) but challenging due to the ambiguity of questions. Current methods rely on the dual-encoder architecture to embed contextualized vectors of questions in conversations. However, this architecture is limited in the embedding bottleneck and the dot-product operation. To alleviate these limitations, we propose generative retrieval for conversational QA (GCoQA). GCoQA assigns distinctive identifiers for passages and retrieves passages by generating their identifiers token-by-token via the encoder–decoder architecture. In this generative way, GCoQA eliminates the need for a vector-style index and could attend to crucial tokens of the conversation context at every decoding step. We conduct experiments on three public datasets over a corpus containing about twenty million passages. The results show GCoQA achieves relative improvements of +13.6% in passage retrieval and +42.9% in document retrieval. GCoQA is also efficient in terms of memory usage and inference speed, which only consumes 1/10 of the memory and takes in less than 33% of the time. The code and data are released at https://github.com/liyongqi67/GCoQA.

1. Introduction

The increasing popularity of conversational agents, such as Alexa,[1] Siri,[2] and Xiaodu,[3] has led to a shift towards dialogue-based interfaces for information-seeking activities. This has spurred the development of conversational question answering (QA) systems. Conversational QA systems over a stable passage (Choi et al., 2018; Reddy, Chen, & Manning, 2019) or a knowledge base (Christmann, Saha Roy, Abujabal, Singh, & Weikum, 2019; Zhang, Dai, Balog, & Callan, 2020) have been successful, but these systems are incapable in open-domain information-seeking (Qu et al., 2020). This motivates the conversational open-domain QA task, which involves retrieving relevant passages from a Web corpus, such as Wikipedia, and providing an answer to the question based on retrieved passages, as shown in Fig. 1.

Effective passage retrieval is crucial for conversational open-domain QA, but it can be challenging due to the ambiguous nature of questions with the conversation context. As depicted in Fig. 1, the term "the first one" corresponds to the previously mentioned "Peddie School" within the conversation context. Due to the presence of references and omissions in conversations, questions need to be interpreted accurately within the context of preceding conversation turns. To enhance question understanding in conversations, some researchers (Anantha et al., 2021; Vakulenko, Longpre, Tu, & Anantha, 2021) adopt a question rewriting model to refine ambiguous questions and a dense retriever to retrieve passages as the single-turn retrieval. Another line considers a single-stage

* Corresponding author.
E-mail addresses: liyongqi0@gmail.com (Y. Li), nanya@microsoft.com (N. Yang), wangliang@microsoft.com (L. Wang), fuwei@microsoft.com (F. Wei), cswjli@comp.polyu.edu.hk (W. Li).

[1] https://www.alexa.com/
[2] https://www.apple.com/siri/
[3] https://dueros.baidu.com/

https://doi.org/10.1016/j.ipm.2023.103475
Received 4 May 2023; Received in revised form 26 July 2023; Accepted 1 August 2023

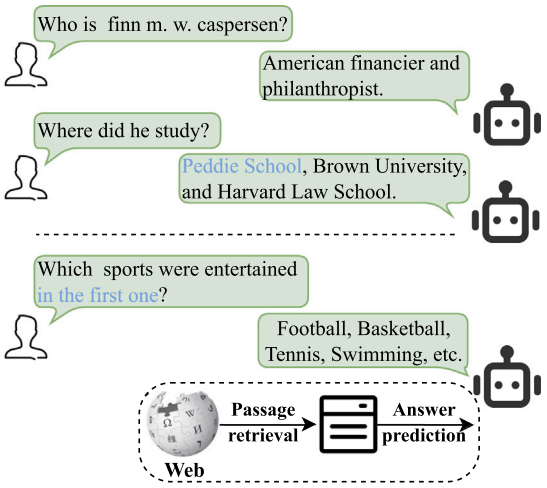

Fig. 1. An example of conversational question answering with Web knowledge. Upon the conversation context, it is required to first retrieve relevant passages from the Web and then give accurate answers.

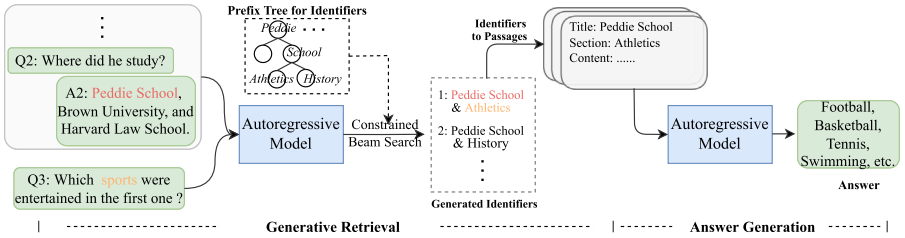

Fig. 2. Our proposed GCoQA method completes the whole QA process with autoregressive language models. GCoQA retrieves passages by generating their identifiers. In this example, GCoQA could correspondingly attend the vital information in the context, like "Peddie School" and "sports", when generating the target identifier "Peddie School & Athletics" step by step.

design (Li, Li, & Nie, 2022a; Qu et al., 2020), where the concatenation of the question and the conversation context is directly input into an encoder, with the belief that the contextualized embeddings can be well learned after training. Basically, current approaches rely on the dual-encoder architecture to retrieve passages, which encode questions (along with the conversation context) and documents into single embedding vectors, and match them via the dot-product operation (Lin, Yang, & Lin, 2021; Qu et al., 2020).

Such dual-encoder based approaches are limited in the conversational setting. (1) The ability of a single embedding vector to capture all the semantics of a question is limited, known as the "embedding bottleneck". This limitation becomes even more significant in conversational QA, where previous questions and answers are constantly added to the current question as the conversation goes. (2) The dot-product operation is not effective in capturing fine-grained interactions, which can prevent important tokens in the conversation context, like the aforementioned "Peddie School", from being properly attended to.

To alleviate the aforementioned limitations, we propose generative retrieval for conversational question answering, called GCoQA. GCoQA uses autoregressive language models to complete the entire QA process, as shown in Fig. 2. We utilize identifier strings, i.e., the page tiles plus section titles, to represent passages in the corpus. As such we train the autoregressive model to generate the identifier of the target passage for a given question. For inference, the identifiers are processed and stored in a prefix tree. As such, GCoQA generates valid identifiers token-by-token via constrained generation upon the prefix tree, and the generated identifiers are correspondingly mapped into relevant passages. Retrieved passages are input into another autoregressive model to generate the answer.

GCoQA does not require encoding context into single vectors and thus could overcome the embedding bottleneck. And benefiting from the cross-interaction of the decoder module, it could attend to the token-level information of the conversation context at every decoding step. For example, as shown in Fig. 2, GCoQA could correspondingly attend the vital information in the context, like "Peddie School" and "sports", when generating the target identifier "Peddie School & Athletics" token by token. The effectiveness of GCoQA is demonstrated through experiments on three public datasets, which show that it outperforms all baseline approaches in terms of both passage-level and document-level retrieval and subsequent answer generation. On average, GCoQA achieves a 13.6% and 42.9% relative improvement in passage and document retrieval compared to the best baseline. Additionally, GCoQA is efficient in terms of memory usage and inference speed.

The rest of this paper is organized as follows: We present a literature review of related studies in Section 2. The studied problem and our GCoQA are presented in Section 3. We conduct experiments and analyze the experiment results in Section 4. Finally, Section 5 concludes this paper and discusses our future work.

2. Related work

Question answering (Neshati, Fallahnejad, & Beigy, 2017; Ni, Lu, Quan, Wenyin, & Hua, 2012; Noraset, Lowphansirikul, & Tuarob, 2021; Ryu, Jang, & Kim, 2014) systems are one of the crucial topics in information processing and management. Our work is related to conversation QA and open-domain QA.

2.1. Conversational question answering

Conversational question answering (QA) methods (Zaib, Zhang, Sheng, Mahmood, & Zhang, 2022) can be classified into four categories based on the knowledge source: knowledge-based conversational QA (Conversational KBQA) (Christmann et al., 2019; Kaiser, Saha Roy, & Weikum, 2021; Lan & Jiang, 2021; Li, Li, & Nie, 2022b; Shen et al., 2019), conversational machine reading comprehension (Choi et al., 2018; Reddy et al., 2019), conversational open-domain QA (Al-Thani, Jansen, & Elsayed, 2023; Li et al., 2022a; Qu et al., 2020), and conversational QA over heterogeneous sources (Christmann, Saha Roy, & Weikum, 2022). A common challenge for all conversational QA methods is understanding users' questions in the context of the conversation. To address this challenge, various methods have been developed, including rewriting (Al-Thani, Elsayed, & Jansen, 2022; Elgohary, Peskov, & Boyd-Graber, 2019; Vakulenko et al., 2021), term classification (Voskarides, Li, Ren, Kanoulas, & de Rijke, 2020), soft-attention (Qu et al., 2019). In the case of conversational open-domain QA, existing methods often directly append the questions with their conversation context (Qu et al., 2020) or use a question rewriting model (Anantha et al., 2021) to refine incomplete questions. As for the conversational open-domain QA, the task bears similarities to open-domain QA, requiring the integration of a passage retriever component and an answer generation component. Notably, in Zhu, Zhang, Zhai, and Liu (2023) presented a method that combines large language models with pseudo-relevance feedback to enhance retrieval efficiency. On a similar note, Wang, Macdonald, Tonellotto, and Ounis (2023) leveraged pseudo-relevance feedback to improve multi-vector retrieval and introduced the ColBERT-PRF method. In this paper, we introduce generative retrieval to conversational QA and find it effective in modeling the conversation context via the encoder–decoder architecture.

2.2. Passage retrieval in question answering

Open-domain question answering (QA) methods (Chen, Fisch, Weston, & Bordes, 2017; Sun et al., 2015; Wang, Yu, Guo, et al., 2018; Wang, Yu, Jiang, et al., 2018) often use a combination of a retriever and a reading comprehension model to find relevant passages and extract the answer to a question. Initially, passage retrieval systems relied on term-based methods such as Term Frequency–Inverse Document Frequency (TF–IDF) and BM25 (Robertson, Zaragoza, et al., 2009). With the advancement of pre-trained language models like BERT (Devlin, Chang, Lee, & Toutanova, 2019), dense retrievers (Karpukhin et al., 2020; Lee, Chang, & Toutanova, 2019) have become more prevalent. Dense retrievers are trained in a contrastive way and require a large number of high-quality negative samples to be effective. However, it can be challenging to find effective negative samples for use in conversational QA. Besides, the few query-passage interactions limit its ability to accurate context modeling in conversational QA. There are also some works that focus on multi-vector representations to improve dense retrievers. In 2020, Omar and Matei (Khattab & Zaharia, 2020) introduced ColBERT, which independently encodes the query and the document using BERT and employs a cheap yet powerful interaction step that models their similarity. In Zhang, Liang, Gong, Jiang, and Duan (2022) proposed better representing a passage with multi vectors from different views.

2.3. Autoregressive models for text retrieval

Autoregressive models have been explored in a variety of tasks in information retrieval, including query expansion, query generation, and question rewriting (Ishii, Wilie, Xu, Cahyawijaya, & Fung, 2022; Ling, Cai, Liu, Chen, & de Rijke, 2023). The authors in Nogueira, Lin, and Epistemic (2019) showed that the generated queries could enhance the document content and bring substantial improvements over BM25. Dai et al. (2022) utilized the autoregressive models to generate a sequence of questions over a document. The authors in Chen, Zhang, Guo, Fan, and Cheng (2022) also explored the generative evidence retrieval for fact verification. In conversational question answering, autoregressive models have often been used to generate self-contained questions, but this can lead to error propagation through multiple stages. More recently, autoregressive models have also been explored for use in entity and document retrieval (Bevilacqua et al., 2022; De Cao, Izacard, Riedel, & Petroni, 2020; Tay et al., 2022), where a language model is used to map queries to target document identifiers. In 2023, Li et al. unified different generative retrieval methods in the proposed multiview identifiers framework (Li, Yang, Wang, Wei, & Li, 2023a) and further proposed a learning-to-rank scheme in generative retrieval (Li, Yang, Wang, Wei, & Li, 2023b), which bridges the novel generative retrieval approach with the typical learning-to-rank paradigm. In this work, we verify that autoregressive models could implement passage-level retrieval with more specific and meaningful identifiers. Besides, we investigate that generative retrieval is more effective in conversational passage retrieval due to its ability to context modeling.

3. Method

3.1. Task definition

Assume the current turn question is q_i, and its conversation context is the set of previous questions and answers, denoted as $C_i = \{q_1, a_1 ..., q_{i-1}, a_{i-1}\}$. Our research objective is to train an effective conversational open-domain QA system to retrieve the target passage p_i from the corpus and generate the answer a_i to the question.

3.2. Overview

The overview of our proposed GCoQA is illustrated in Fig. 2. We detail the generative retrieval and answer generation in Section 3.3 and Section 3.4, respectively. Besides, we give an algorithm analysis to dissect the generative retrieval from different views.

3.3. Generative retrieval

Identifiers. For passages in the corpus, we first assign them with identifiers. An identifier refers to a unique string that could well represent the passage's content. Considering these characteristics, we choose the page title plus the section title as the identifier for a passage. We address that page titles and section titles in Wikipedia are natural identifier strings and contain much semantic information. Take the third question in Fig. 2 as an example. Its target passage is on the page titled "Peddie School", and the evidence to answer this question is concentrated in the section titled "Athletics". "Peddie School & Athletics" could well illustrate that the content of the passage is about the sports of Peddie School.

Denote the identifier string of a passage p_i as $I_i = \{w_1, w_2, \dots, w_l\}$. Given the current question q_i and its conversation context C_i, generative retrieval aims to generate the identifier string I_i via an autoregressive model step by step as follows,

$$p(I_i|q_i; C_i; \theta) = \prod_{j=1}^{l} p(w_j|q_i; C_i; w_{<j}; \theta), \tag{1}$$

where θ is the parameter of the autoregressive model for retrieval. Specifically, the question q_i and C_i are first concatenated with special split tokens, and are input into a Transformer-Encoder to obtain the sequence of token-level representations, formulated as

$$\mathcal{T} = TE([q_1; ...; q_{i-1}; a_{i-1}; q_i]). \tag{2}$$

In the decoding phase, the shifted right identifier tokens and question representations $\mathcal{T}$ are forwarded into a Transformer-Decoder. Benefiting from the multi-head attention of the decoder module, the identifier tokens could perform cross-interactions with the token-level representations $\mathcal{T}$. At different generate steps, the different tokens of the input are expected to be correspondingly attended. At step j, the hidden vector $\mathbf{h}_j$ is used to calculate the probabilities as,

$$p(w_j) = Softmax(\mathbf{h}_j \mathbf{W}), \tag{3}$$

where $W \in R^{d \times v}$ and v is the vocabulary size. In this way, we obtain a predicted identifier. Since each passage is assigned a unique identifier, we could conveniently map an identifier to the corresponding passage.

Constrained beam search. We train the retriever using the standard seq2seq objective, i.e., maximizing the output sequence likelihood with teacher forcing. During the inference phase, we implement constrained decoding (De Cao et al., 2020) using the Trie data structure. The Trie, which is a type of k-ary search tree, is employed to efficiently locate specific keys within a set. In our case, we process and store all passage identifiers from the corpus within the Trie data structure. Given a prefix string, the Trie structure supports providing the possible tokens occurring in identifiers. In this way, we could guarantee that the generated identifiers must be an exact identifier of an existing passage. Additionally, we utilize beam search (Sutskever, Vinyals, & Le, 2014), a commonly-used technique, to generate multiple identifiers instead of just one. Each generated identifier is assigned a language model score, enabling us to obtain a ranking list of generated identifiers based on these scores. The ranking identifiers could naturally correspond to a ranking list of passages, denoted as $\mathcal{P}$.

Algorithm analysis. By discretizing the representations of passages as identifiers rather than vectors, GCoQA transforms the search over passages into a multi-step search over the vocabulary. Denote the identifier length as l and vocabulary size as v. For the beam search size b, the search scope for GCoQA is $O(blv)$, which is usually much smaller than the number of passages. Benefiting from this, GCoQA could perform cross-interactions while keeping retrieval efficient. From another view, GCoQA generates the page title and section title in sequence. This process could also be regarded as a hierarchical search, which first locates a page and then searches a specific passage within the page.

3.4. Answer generation

We use another autoregressive model to generate the answer upon the retrieved passages $\mathcal{P}$. Following the Fusion-in-Decoder (FiD) model (Izacard & Grave, 2021), we encode all retrieved passages independently and jointly attend over all of them in the

Table 1
Statistics for the three datasets. # Avg.Q, # Min.Q, and # Max.Q denote the average, minimum, and maximum number of questions in a conversation. # Avg.P and # Avg.D is the number of passages and documents involved in a conversation on average. Gold Q means whether the dataset contains the self-contained question annotation or not.

Items	TOPIOCQA	QRECC	OR-QuAC
# Dialogs	3714	10,629	4314
# Questions	47,963	56,636	30,992
# Avg.Q	12.9	5.3	7.2
# Min.Q	5	2	4
# Max.Q	25	12	12
# Avg.P	9.0	1.6	1.0
# Avg.D	3.9	1.1	1.0
Gold Q	✗	✔	✔
Answer	abstractive	abstractive	extractive

decoder to generate the answer. Considering that the concatenation of the question q_i and conversation context C_i has been a little long, we split the retrieved passages $\mathcal{P}$ into smaller chunks, $\{c_1, \ldots, c_m\}$.

Each chunk c_j is concatenated with the question q_i and conversation context C_i with special tokens, like $question: q_i; c_j; context: C_i$. The concatenated text is input into the transformer encoder to obtain token-level representations. In the decoding phase, all token-level representations for chunks c_1 to c_m are concatenated and forwarded into the decoder module to generate the answer text.

We train the reader model using the standard seq2seq loss. Since the reader model is trained upon retrieved passages $\mathcal{P}$, we train it after the retriever model training.

4. Experiments

4.1. Benchmark

QA pairs. We conducted experiments on three conversational open-domain QA datasets: OR-QuAC (Qu et al., 2020), QRECC (Anantha et al., 2021), and TOPIOCQA (Adlakha, Dhuliawala, Suleman, de Vries, & Reddy, 2022). To facilitate future research in this area, we unified the three datasets into a benchmark with the same corpus, as DPR (Karpukhin et al., 2020) did.

OR-QuAC (Qu et al., 2020) is transformed from the conversational MRC dataset QuAC (Choi et al., 2018). In OR-QuAC, the questions in a conversation are sourced from the same section in Wikipedia, and the answers are extractive, i.e., exact text spans in passages. Therefore, as the conversation goes, the previous answers leak the gold passage more and more. To avoid this data leak, we followed the previous setting (Qu et al., 2020) that only questions are reserved in the conversation context.

QRECC (Anantha et al., 2021) is conducted specifically for conversational open-domain QA. Most of the questions were sourced from the QuAC (Choi et al., 2018) and CAsT (Dalton, Xiong, Kumar, & Callan, 2020) datasets, and annotators were required to give answers using a web search engine. Besides, the precious gold question (the self-contained one for the incomplete question) label is annotated in this dataset.

TOPIOCQA (Adlakha et al., 2022) is a relatively new published dataset featured with topic switching. As the conversation progresses, the topics may switch to related topics, a phenomenon commonly observed in information-seeking search sessions. Therefore, TOPIOCQA requires the ability of accurate context modeling more, since the previous questions or answers may not be directly related to the current-turn passage retrieval.

Corpus processing. We followed the same setting in TOPIOCQA and used the Wikipedia dump from 10/20/2020, which consists of 5.9 million documents. Wikiextractor[4] is used to clean the text from the document dump. Considering that the retrieval label in QRECC and OR-QuAC concretes to the section-level on a page, we split each Wikipedia document into multiple passages while preserving section boundaries. Finally, we obtained 17,517,456 passages in the end. A handful of questions in QRECC and OR-QuAC are removed because of the transformation of the corpus.

Statistics. The statistics of the three datasets are summarized in Table 1. TOPIOCQA has the longest conversation and the average turn is 12.9. QRECC and OR-QuAC contain the "Gold Q" annotation, but TOPIOCQA does not. Another important difference between the three datasets is the average number of passages per conversation. In TOPIOCQA, the number is 9.0, while they are 1.6 and 1.0 in QRECC and OR-QuAC, respectively. It reveals that in QRECC and OR-QuAC, most conversations only involve one passage. We will discuss its influence on the experimental results in Section 4.3.

[4] https://attardi.github.io/wikiextractor/

Table 2
Retrieval performance on three public datasets. We use Accuracy, Recall@5, MRR@5 to evaluate the retrieval performance. MRR is calculated according to the position of the first passage that contains evidence to answer the question. The "Question Rep" indicates the question representation way of each method. Inapplicable results are marked by "-". The best results in each group are marked in Bold, while the second-best ones are underlined. %improve represents the relative improvement achieved by GCoQA over the best results of the baselines.

Methods	Question Rep	TOPIOCQA			QRECC			OR-QuAC		
		Acc	Recall	MRR@5	Acc	Recall	MRR@5	Acc	Recall	MRR@5
		Passage-level Retrieval								
BM25	Original	1.32	4.10	2.29	1.98	5.27	3.16	4.81	11.46	7.17
GCoQA	Original	7.31	11.79	8.96	4.15	7.38	5.37	9.21	13.98	10.96
CQR(BM25) (Vakulenko et al., 2021)	QR	–	–	–	17.03	36.10	24.02	20.75	44.21	29.03
BM25	All	8.41	18.37	12.11	**38.41**	59.36	46.41	25.37	45.62	33.23
ORConvQA (Qu et al., 2020)	All	9.71	20.54	13.72	22.14	47.82	31.60	50.15	67.92	56.98
DPR (Karpukhin et al., 2020)	Original	2.05	5.34	3.25	1.37	3.69	2.16	3.63	7.91	5.08
CQR(DPR) (Vakulenko et al., 2021)	QR	–	–	–	14.06	34.89	21.46	31.60	55.71	40.54
CQE (Lin et al., 2021)	All	36.06	63.07	46.46	30.12	57.17	40.14	59.11	76.07	65.92
ColBERT (Khattab & Zaharia, 2020)	QR	–	–	–	18.24	37.79	26.63	35.45	61.08	47.82
ColBERT (Khattab & Zaharia, 2020)	All	33.15	61.13	42.25	27.75	55.32	37.28	54.28	73.19	63.32
GCoQA	QR	–	–	–	27.19	53.33	37.24	32.44	59.76	42.12
GCoQA	All	**48.93**	**73.46**	**58.91**	36.98	**69.67**	**50.15**	**64.73**	**77.91**	**69.89**
% improve		35.69%	16.47%	26.80%	−3.87%	17.37%	8.06%	9.05%	2.42%	6.02%
		Document-level Retrieval								
BM25	Original	3.46	7.35	4.88	4.72	9.24	6.36	8.60	15.59	11.19
GCoQA	Original	9.19	14.74	11.27	7.75	10.95	9.01	16.35	19.60	17.59
CQR(BM25)(Karpukhin et al., 2020)	QR	–	–	–	36.51	63.57	46.75	40.05	66.68	50.47
BM25	All	23.03	37.61	28.48	61.72	80.96	69.59	27.21	45.55	34.39
ORConvQA (Qu et al., 2020)	All	10.79	23.75	15.51	26.15	52.12	36.02	43.23	61.44	50.54
DPR (Karpukhin et al., 2020)	Original	1.20	4.49	2.42	1.69	3.81	2.51	3.94	9.51	5.96
CQR(DPR) (Vakulenko et al., 2021)	QR	–	–	–	33.26	59.50	43.33	37.56	64.99	48.67
CQE (Lin et al., 2021)	All	51.28	79.02	62.49	37.07	65.80	48.15	50.82	77.15	61.37
ColBERT (Khattab & Zaharia, 2020)	QR	–	–	–	38.37	68.42	50.13	48.25	73.46	54.19
ColBERT (Khattab & Zaharia, 2020)	All	44.25	72.18	58.61	42.15	72.38	52.38	54.48	80.25	63.39
GCoQA	QR	–	–	–	66.86	73.21	69.57	75.43	78.45	76.69
GCoQA	All	**76.03**	**86.79**	**80.76**	**74.14**	**82.33**	**77.78**	**81.47**	**84.45**	**82.78**
% improve		48.26%	9.83%	29.24%	20.12%	1.69%	11.77%	60.31%	5.23%	34.89%

4.2. Setup

Baselines. We included the widely-used BM25 and DPR (Karpukhin et al., 2020) as basic baselines. Besides, we compared GCoQA with the conversational QA methods, CQR (Vakulenko et al., 2021), ORConvQA (Qu et al., 2020), and CQE (Lin et al., 2021). We also included ColBERT (Khattab & Zaharia, 2020) as a baseline in our experiments due to its potential in addressing the bottleneck issue faced by dense retrievers. To achieve this, we followed the original approach, which employs a two-stage procedure for retrieving the most relevant passages from a large corpus. Additionally, we applied a question rewriting module or included the conversation context to adapt ColBERT to the conversational setting. Some methods (Li et al., 2022a; Yu, Liu, Xiong, Feng, & Liu, 2021) used extra data, like graph structure and pretrained ad-hoc model, are not considered. It is worth noting that the recent and widely-used ChatGPT models[5] cannot be considered as baselines in our experiments. This is primarily due to the fact that ChatGPT does not possess a built-in mechanism for retrieving relevant passages from a large corpus. As a result, it cannot directly serve as a comparison point for our proposed retrieval methods. We considered three different question representations, "All history", "QR", and "Original", for GCoQA. In particular, "All history" refers to appending the entire conversation context to the current question, while "QR" means rewriting the current question with the conversation context. And the "Original" setting only inputs the current question into the model. For all baselines, we adopted the same FiD model as ours to generate the answer for a fair comparison.

Implementation Details. We used Lucene BM25 with default values of k1 = 1.2 and b = 0.75. For baselines, we used the hyperparameters suggested in their codebase. The Question Rewriting model and our generative retriever are initialized from the T5-large model. To train the generative retriever, we adopted Adam with a learning rate of 1e-5, warming up for 4 epochs, and training for 40 epochs. GCoQA adopted the page title as the identifier for document-level retrieval and used the page title plus with section title for passage-level retrieval. The reported results of our model are mean values from several runs. The training is conducted on 8 × 32 GB NVIDIA V100 GPUs.

Evaluation. We evaluated the conversational open-domain QA system from the aspects of passage/document retrieval and answer prediction. To evaluate the retrieval performance, we applied the metrics of Accuracy (Recall@1), Recall, and MRR. For answer prediction, we employed the word-level F1 metric, exact match (EM), and BLEU-4.

[5] https://openai.com/blog/chatgpt

Table 3
Answer prediction performance based on top-5 passages on three public datasets. "-P" and "-D" denote passage-level and document-level, respectively. The best results in each group are marked in Bold.

Methods	TOPIOCQA			QRECC			OR-QuAC		
	F1	EM	BLEU	F1	EM	BLEU	F1	EM	BLEU
BM25(All History)-D	25.05	9.82	5.37	31.44	10.17	11.11	20.98	14.88	2.95
BM25(All History)-P	24.29	9.51	3.21	32.86	11.00	11.78	24.45	15.99	5.33
CQE-D	36.02	15.99	10.41	29.88	10.31	9.33	20.91	14.19	3.24
CQE-P	35.84	15.68	10.86	32.48	11.39	11.41	25.81	16.12	6.89
GCoQA-D	37.93	17.98	11.57	31.27	10.84	10.94	21.16	14.76	3.27
GCoQA-P	**39.58**	**18.66**	**11.87**	**32.95**	**11.39**	**12.28**	**27.51**	**16.60**	**7.89**

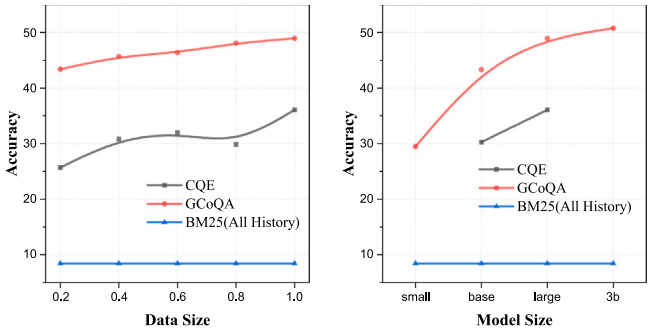

Fig. 3. Retrieval performance on the TOPIOCQA dataset versus the data sizes and model sizes.

4.3. Overall results

The retrieval performance on the three datasets of all approaches is summarized in Table 2, and we reported the answer generation performance in Table 3. By jointly analyzing the results, we gained the following findings.

(1) The CQE model outperforms the BM25 model in terms of its ability to match questions and passages in a deep semantic space and to attend to important information in the conversation context. However, it still performs worse than the GCoQA model, which benefits from fine-grained cross-interactions between the conversation context and identifier strings for the visualization of cross-interactions. In terms of input, the "All History" input performs the best, while the "Original" input performs the worst due to its lack of necessary information. The "Rewriting" input performs better than the "Original" input but still worse than "All History". On the one hand, the question rewriting step may introduce more noise if the rewritten questions are incorrect; on the other hand, the QRECC and OR-QuAC datasets only involve one passage or document in a conversation. Therefore, most of the content of the conversation context is profitable, and rewriting may lose this useful information. (2) Comparing the three datasets, we observed that GCoQA performs particularly well on TOPIOCQA, which involves topic switching and multiple passages in a conversation, as it excels in context modeling. BM25 performs better on QRECC, where there is only one passage in a conversation and the previous answers are included in the context. This is likely because QRECC includes abstractive answers that contain information about the target passage, which benefits term-based methods like BM25. About 10% answers could be directly contained in the passage.

(3) ColBERT (QR) significantly outperforms the CQR(DPR) method. This finding illustrates that the utilization of multi-vector techniques can alleviate the problem of the embedding bottleneck commonly encountered by dense retrievers. ColBERT (All history) not only achieves comparable results with CQE but also surpasses it in certain scenarios. However, despite its improvements, ColBERT still falls short compared to GCoQA in both passage-level and document-level retrieval. This disparity can be attributed to GCoQA's ability to effectively attend to the conversation context through its fine-grained cross-interaction mechanism within the decoder module.

(4) In terms of answer prediction, the performance of all approaches is consistent with the retrieval results because they all use the same FiD reader to generate the answer. The results suggest that providing the same number of passages to the reader is more beneficial than providing documents, possibly because long documents may contain irrelevant information that interferes with answer prediction.

4.4. Comparison

Since GCoQA is a new technical route to the retrieval of passage, we thoroughly compared it with dense retrievers from the aspects of data and model size, memory consumption, and inference efficiency.

Data size and model size. Since GCoQA and CQE (the best baseline in our experiment) are data-driven methods, we run experiments to evaluate their performance with various data sizes. We randomly selected subsets with varying proportions of the training set to train GCoQA and CQE, and the results are summarized in Fig. 3. The results show that GCoQA outperforms CQE on all

Table 4
Retrieval performance of GCoQA on TOPIOCQA with beam size values in {5, 10, 20, 50}.

	BS	Acc	Recall	MRR
Passage level	5	48.93	73.46	58.91
	10	48.93	72.52	58.62
	20	48.85	72.39	58.51
	50	48.76	72.09	58.36
Document level	5	76.03	86.79	80.76
	10	75.94	86.71	80.67
	20	75.81	86.58	80.44
	50	75.56	86.20	80.06

Table 5
Retrieval accuracy of GCoQA on TOPIOCQA with different identifiers.

Identifiers	Accuracy
First Sentence	0.47
Random Sentence	0.17
Page Title & Section Title	48.93

different data sizes. Besides, we reported the results of GCoQA with varying model capacities, which are identical to the small, base, large, and 3b settings of standard T5. As a comparison, the results of CQE with BERT-base and BERT-large are also illustrated. We observed that with the increase in model size, the retrieval performance grows. This implies that the model capacity has a critical impact on retrieval performance. Compared with CQE using the encoders of pre-trained language models, GCoQA could take full advantage of the powerful generative language models, such as T5 (Raffel et al., 2020) and GPT3 (Brown et al., 2020). This may be another potential advantage of generative retrieval against dense retrievers that only use the encoders.

Memory consumption. Memory consumption is another important issue to be addressed for large-scale retrieval. Dense retrievers encode all passages/documents into vectors as indexes and store them for real-time retrieval. In practice, dense retrievers (CQE, ORConvQA, and DPR) consume about 57 GB of CPU memory for 5 million documents and 150 GB for 15 million passages. As a comparison, GCoQA only occupies 5.6 GB and 23 GB for documents and passages, respectively, to store the prefix tree. As the number of passages increases, the memory consumption for dense retrievers lineally grows, while GCoQA will consume less memory. This demonstrates GCoQA's superiority in memory consumption.

Inference time. Efficient passage retrieval is crucial for answering users' questions in real time, particularly for conversational QA. We profiled the passage retrieval speed on a server with Intel Xeon CPU E5-2673 v4 @ 2.30 GHz, 512 GB memory, and one NVIDIA V100 GPU. With the help of FAISS[6] library and using the batch size of 256, CQE processes 2514 questions for about 25 min to give the top-5 results. GCoQA takes about 17 min to retrieve passages for these questions, with a batch size of 2 and beam search size of 5. This is because GCoQA reduces the search scope via beam search as analyzed in Section 3.2, and thus keeps highly efficient. It is noticed the inference time for GCoQA will increase along with a larger beam size.

4.5. In-depth analysis

To further analyze GCoQA, we designed experiments on beam size, identifiers, performance versus conversation turns, as well as visualization.

Beam size. As introduced before, GCoQA relies on beam search to decrease the search scope and give a ranking list. It is essential to study the influence of beam size on retrieval performance. Analyzing the results in Table 4, we found as the beam size increases, the performance declines on both the passage-level and document-level. The results suggest selecting a beam size of k for a top-k ranking. In GCoQA, although we could give a ranking list via beam search decoding, this ranking may not be the same as the retrieval ranking we required.

Identifiers. In this work, we choose page titles plus section titles as identifiers for passage-level retrieval. For comparison, we selected the first or a random sentence in a passage as the identifier, and the results are summarized in Table 5. The results show that whether the first sentence or a random sentence cannot work well as titles and achieves a very poor performance. This results from the fact that a sentence in a passage only covers partial content, and thus this sentence may not be semantically related to the question. Besides, not all words in a sentence are meaningful for retrieval. GCoQA generates an identifier in an autoregressive way, and thus those irrelated words impede the correct generation.

Corpus transition. To fully evaluate GCoQA, we reported results on the original TOPIOCQA (Adlakha et al., 2022) dataset, where the corpus consists of passages with no more than 100 words. We design a simple transition algorithm to transform the retrieved sections by GCoQA into a sequence of fine-grained passages of the original corpus. Specifically, we obtain a new rank list

[6] https://github.com/facebookresearch/faiss

Table 6
Retrieval performance on the original TOPIOCQA (Adlakha et al., 2022) corpus. The results of baselines are officially reported (Adlakha et al., 2022) and released at https://github.com/McGill-NLP/topiocqa.

Model	Question Rep	Recall@5	Recall@20
BM25	Original	2.9	5.2
	All History	14.2	21.1
	Rewrites	19.25	31.9
DPR	Original	5.97	7.3
	All History	51.71	67.0
	Rewrites	34.49	48.6
GCoQA	All History	**59.79**	**71.8**

Turns	Question	Answer	Topic Switch	Identifier	Predictiom
1	When will the new dunkirk film be released on dvd?	18 December 2017	Dunkirk	Dunkirk (2017 film) & Release	Dunkirk (2017 film) & Introduction
2	What is this film about?	Dunkirk evacuation of World War II	Dunkirk	Dunkirk (2017 film) & Introduction	Dunkirk (2017 film) & Introduction
3	Can you mention a few members of the cast?	Fionn Whitehead, Tom Glynn-Carney, ..., Harry Styles	Dunkirk	Dunkirk (2017 film) & Introduction	Dunkirk (2017 film) & Introduction
4	Where was it shot?	Dunkirk and Urk, Netherlands	Dunkirk	Dunkirk (2017 film) & Filming	Dunkirk (2017 film) & Filming
5	Can you tell me anything about the latter place's past?	Until 1475 the High and Low Lordship of Urk and Emmeloord was in the ...	Urk	Urk & History / Lordship of Urk en Emmeloord	Urk & Introduction
6	What does the economy of this place depend upon?	Fishery	Urk	Urk & Economy	Urk & Economy

Fig. 4. A conversation (the first one in the test set) from the TOPIOCQA dataset. We show the questions, answers, topics, and identifiers of the target passage, as well as the model predictions (top-1 prediction). For a better illustration, the coherence of the conversation is colored in green, and the semantic correlation between the conversation and the identifiers is colored in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of passages according to their positions in the retrieved sections. The results in Table 6 show GCoQA still outperforms the officially reported baselines. However, it is impossible to measure the "Recall@100" metric for GCoQA because it cannot recall large passages. We discuss this limitation of GCoQA further in the "Limitations" section.

Case study. To qualitatively illustrate why GCoQA works, we analyzed the prediction results of a conversation in Fig. 4. The example shows that the input text (questions and the conversation context) and target text (identifier strings) are semantically related. For example, the answer to the 4-th question "where was it shot" is in the passage titled "Dunkirk (2017 film) & Filming", where "shot" and "Filming" are semantically corresponding. This case demonstrates well that performing passage retrieval in a generative way is viable based on the semantic correlation. Another observation is that when the topic switches, GCoQA still keeps effective. In the 5-th and 6-th turns, the topic of the conversation has changed from "Dunkirk" to "Urk", and GCoQA correctly predicts the page titles of the two questions. Some error cases are discussed further in the "Limitations" section.

4.6. Limitations

Although GCoQA has been introduced as a potential solution to address the limitations of previous methods, it also brings new limitations: (1) GCoQA recalls passages via beam search, and thus it has difficulty in recalling large-scale passages due to the beam size constraint. Additionally, as the beam size increases, the inference speed decreases. (2) Previous generative retrieval methods primarily focused on entity and document retrieval. Although GCoQA made some progress in passage retrieval by using section titles, it is still limited out of the Wikipedia corpus where titles may be missing. (3) GCoQA lacks the ability to "see" the entire content of a passage, making it difficult to distinguish similar passages. For example, GCoQA incorrectly generates the identifiers for 1-st and 5-st questions in Fig. 4. Although GCoQA has generated the correct page titles for the two questions, it fails to predict the correct section titles. This is also demonstrated by its worse performance at passage-level retrieval compared to document-level retrieval, as shown in Table 2. (4) The generalizability of GCoQA is a legitimate concern. GCoQA heavily relies on the semantic relationship between the question and the passage identifiers for retrieving relevant passages. While GCoQA has been evaluated using three academic datasets, its effectiveness in real-world scenarios, where questions are often ambiguous and challenging to match with the identifiers, remains uncertain and requires further investigation.

Fortunately, several of the aforementioned limitations of GCoQA hold the potential for future improvement. For example, the pseudo-queries generated based on a passage's content could serve as complementary identifiers alongside the current titles. In this way, GCoQA could work in scenarios where the titles are missing and take full advantage of the passages' entire semantics. Overall,

we introduce a new approach to conversational QA and also leave much room for further improvement. Overall, our introduction of a new approach to conversational QA opens up promising possibilities for advancing the field. However, it is important to note that there is still ample room for further refinement and enhancement. Continued research and development efforts are required to overcome the remaining limitations and maximize the potential of GCoQA, ultimately paving the way for more effective and robust conversational QA systems.

4.7. Discussion and practical implications

The results show that our proposed method is effective in retrieving passages in conversations. Our method applies a different architecture, the encoder–decoder way, from previous dual-encoder ones. Benefiting from the encoder–decoder architecture, our method could attend to the crucial information of the conversation context. In this way, GCoQA could retrieve passages and give answers to users more accurately.

Besides, we evaluate our method in terms of memory consumption and inference speed. GCoQA only consumes 1/10 of the memory and takes in less than 33% of the time, compared with the baselines. Therefore, it becomes more convenient and efficient to apply our method in practice.

5. Conclusion and future work

The core problem in conversation QA is accurately identifying and attending to crucial information in the conversation context. We propose the GCoQA, which uses an autoregressive language model to retrieve relevant passages. Benefiting from fine-grained cross-interactions in the decoder module, GCoQA could attend to the conversation context more effectively. Additionally, GCoQA has lower memory consumption and higher inference efficiency in practice.

This work makes the initial attempt to explore generative retrieval for conversational QA, leaving many promising directions for future work: (1) investigating the use of generative retrieval in more general Web search scenarios where identifiers are not directly available from titles; and (2) examining the integration of passage retrieval and answer prediction within a single, generative model in order to better understand their internal relationships.

CRediT authorship contribution statement

Yongqi Li: Conceptualization, Methodology, Experiments, Validation, Formal analysis, Investigation, Manuscript writing. **Nan Yang:** Conceptualization, Formal analysis, Review, Editing, Supervision. **Liang Wang:** Conceptualization, Formal analysis, Review, Editing, Supervision. **Furu Wei:** Conceptualization, Formal analysis, Review, Editing, Supervision. **Wenjie Li:** Conceptualization, Formal analysis, Review, Editing, Supervision.

Data availability

Data will be made available on request.

Acknowledgments

The work described in this paper was supported by Research Grants Council of Hong Kong (PolyU/5210919, PolyU/15207821 and PolyU/15207122), National Natural Science Foundation of China (62076212) and PolyU internal grants (ZVQ0).

References

Adlakha, V., Dhuliawala, S., Suleman, K., de Vries, H., & Reddy, S. (2022). TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, *10*, 468–483.

Al-Thani, H., Elsayed, T., & Jansen, B. J. (2022). Improving conversational search with query reformulation using selective contextual history. *Data and Information Management*, Article 100025.

Al-Thani, H., Jansen, B. J., & Elsayed, T. (2023). ECAsT: A large dataset for conversational search and an evaluation of metric robustness. *PeerJ Computer Science*, *9*, Article e1328.

Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., & Chappidi, S. (2021). Open-domain question answering goes conversational via question rewriting. In *Proceedings of the international conference of the north american chapter of the association for computational linguistics* (pp. 520–534). ACL.

Bevilacqua, M., Ottaviano, G., Lewis, P., Yih, S., Riedel, S., & Petroni, F. (2022). Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, *35*, 31668–31683.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 1870–1879). ACL.

Chen, J., Zhang, R., Guo, J., Fan, Y., & Cheng, X. (2022). GERE: Generative evidence retrieval for fact verification. In *Proceedings of international conference on research and development in information retrieval* (pp. 2184–2189). ACM.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., et al. (2018). QuAC: Question answering in context. In *Proceedings of the international conference on empirical methods in natural language processing* (pp. 2174–2184). ACL.

Christmann, P., Saha Roy, R., Abujabal, A., Singh, J., & Weikum, G. (2019). Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the international conference on information and knowledge management* (pp. 729–738). ACM.

Christmann, P., Saha Roy, R., & Weikum, G. (2022). Conversational question answering on heterogeneous sources. In *Proceedings of the international conference on research and development in information retrieval* (pp. 144–154). ACM.

Dai, Z., Chaganty, A. T., Zhao, V. Y., Amini, A., Rashid, Q. M., Green, M., et al. (2022). Dialog inpainting: Turning documents into dialogs. In *International conference on machine learning* (pp. 4558–4586). MIT Press, PMLR.

Dalton, J., Xiong, C., Kumar, V., & Callan, J. (2020). Cast-19: A dataset for conversational information seeking. In *Proceedings of the international conference on research and development in information retrieval* (pp. 1985–1988). ACM.

De Cao, N., Izacard, G., Riedel, S., & Petroni, F. (2020). Autoregressive entity retrieval. In *International conference on learning representations*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the international conference of the North American chapter of the association for computational linguistics*. Minneapolis, Minnesota: ACL.

Elgohary, A., Peskov, D., & Boyd-Graber, J. (2019). Can you unpack that? Learning to rewrite questions-in-context. In *Proceedings of the international conference on empirical methods in natural language processing* (pp. 5918–5924). ACL.

Ishii, E., Wilie, B., Xu, Y., Cahyawijaya, S., & Fung, P. (2022). Integrating question rewrites in conversational question answering: A reinforcement learning approach. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 55–66). ACL.

Izacard, G., & Grave, É. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the international conference of the European chapter of the association for computational linguistics* (pp. 874–880). ACL.

Kaiser, M., Saha Roy, R., & Weikum, G. (2021). Reinforcement learning from reformulations in conversational question answering over knowledge graphs. In *Proceedings of the international conference on research and development in information retrieval* (pp. 459–469). ACM.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the international conference on empirical methods in natural language processing* (pp. 6769–6781). ACL.

Khattab, O., & Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the international conference on research and development in information retrieval* (pp. 39–48).

Lan, Y., & Jiang, J. (2021). Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 3288–3297). ACL.

Lee, K., Chang, M.-W., & Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 6086–6096). ACL.

Li, Y., Li, W., & Nie, L. (2022a). Dynamic graph reasoning for conversational open-domain question answering. *ACM Transactions on Information Systems*, *40*(4), 1–24.

Li, Y., Li, W., & Nie, L. (2022b). MMCoQA: Conversational question answering over text, tables, and images. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 4220–4231). ACL.

Li, Y., Yang, N., Wang, L., Wei, F., & Li, W. (2023a). Multiview identifiers enhanced generative retrieval. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 6636–6648).

Li, Y., Yang, N., Wang, L., Wei, F., & Li, W. (2023b). Learning to rank in generative retrieval. arXiv preprint arXiv:2306.15222.

Lin, S.-C., Yang, J.-H., & Lin, J. (2021). Contextualized query embeddings for conversational search. In *Proceedings of the international conference on empirical methods in natural language processing* (pp. 1004–1015). ACL.

Ling, Y., Cai, F., Liu, J., Chen, H., & de Rijke, M. (2023). Generating relevant and informative questions for open-domain conversations. *ACM Transactions on Information Systems*, *41*(1), 1–30.

Neshati, M., Fallahnejad, Z., & Beigy, H. (2017). On dynamicity of expert finding in community question answering. *Information Processing & Management*, *53*(5), 1026–1042.

Ni, X., Lu, Y., Quan, X., Wenyin, L., & Hua, B. (2012). User interest modeling and its application for question recommendation in user-interactive question answering systems. *Information Processing & Management*, *48*(2), 218–233.

Nogueira, R., Lin, J., & Epistemic, A. (2019). From doc2query to docTTTTTquery. Online preprint, 6.

Noraset, T., Lowphansirikul, L., & Tuarob, S. (2021). Wabiqa: A wikipedia-based thai question-answering system. *Information processing & management*, *58*(1), Article 102431.

Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W. B., & Iyyer, M. (2020). Open-retrieval conversational question answering. In *Proceedings of the international conference on research and development in information retrieval* (pp. 539–548). ACM.

Qu, C., Yang, L., Qiu, M., Zhang, Y., Chen, C., Croft, W. B., et al. (2019). Attentive history selection for conversational question answering. In *Proceedings of the international conference on information and knowledge management* (pp. 1391–1400). ACM.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Reddy, S., Chen, D., & Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, *7*, 249–266.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389.

Ryu, P.-M., Jang, M.-G., & Kim, H.-K. (2014). Open domain question answering using wikipedia-based knowledge model. *Information Processing & Management*, *50*(5), 683–692.

Shen, T., Geng, X., Qin, T., Guo, D., Tang, D., Duan, N., et al. (2019). Multi-task learning for conversational question answering over a large-scale knowledge base. In *Proceedings of the international conference on empirical methods in natural language processing* (pp. 2442–2451). ACL.

Sun, H., Ma, H., Yih, W.-t., Tsai, C.-T., Liu, J., & Chang, M.-W. (2015). Open domain question answering via semantic enrichment. In *Proceedings of the international conference on world wide web* (pp. 1045–1055). ACM.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, *27*.

Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., et al. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831–21843.

Vakulenko, S., Longpre, S., Tu, Z., & Anantha, R. (2021). Question rewriting for conversational question answering. In *Proceedings of the international conference on web search and data mining* (pp. 355–363). ACM.

Voskarides, N., Li, D., Ren, P., Kanoulas, E., & de Rijke, M. (2020). Query resolution for conversational search with limited supervision. In *Proceedings of the international conference on research and development in information retrieval* (pp. 921–930). ACM.

Wang, X., Macdonald, C., Tonellotto, N., & Ounis, I. (2023). ColBERT-PRF: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, *17*(1), 1–39.

Wang, S., Yu, M., Guo, X., Wang, Z., Klinger, T., Zhang, W., et al. (2018). R3: Reinforced ranker-reader for open-domain question answering. In *AAAI conference on artificial intelligence* (pp. 1–8). AAAI.

Wang, S., Yu, M., Jiang, J., Zhang, W., Guo, X., Chang, S., et al. (2018). Evidence aggregation for answer re-ranking in open-domain question answering. In *International conference on learning representations*.

Yu, S., Liu, Z., Xiong, C., Feng, T., & Liu, Z. (2021). Few-shot conversational dense retrieval. In *Proceedings of the international conference on research and development in information retrieval* (pp. 829–838). ACM.

Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. *Knowledge and Information Systems*, *64*(12), 3151–3195.

Zhang, S., Dai, Z., Balog, K., & Callan, J. (2020). Summarizing and exploring tabular data in conversational search. In *Proceedings of the international conference on research and development in information retrieval* (pp. 1537–1540). ACM.

Zhang, S., Liang, Y., Gong, M., Jiang, D., & Duan, N. (2022). Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 5990–6000).

Zhu, W., Zhang, X., Zhai, Q., & Liu, C. (2023). A hybrid text generation-based query expansion method for open-domain question answering. *Future Internet*, *15*(5), 180.