RINK: Reader-Inherited Evidence Reranker for Table-and-Text Open Domain Question Answering

Eunhwan Park[1], Sung-Min Lee[1], Daeryong Seo[3], Seonhoon Kim[2*], Inho Kang[3], Seung-Hoon Na[1†]

[1] Jeonbuk National University
[2] Coupang
[3] Naver Corporation
{judepark, cap1232, nash}@jbnu.ac.kr, sekim625@coupang.com, {daeryong.seo, once.ihkang}@navercorp.com

Abstract

Most approaches used in open-domain question answering on hybrid data that comprises both tabular-and-textual contents are based on a *Retrieval-Reader* pipeline in which the retrieval module finds relevant "heterogenous" evidence for a given question and the reader module generates an answer from the retrieved evidence. In this paper, we present a *Retriever-Reranker-Reader* framework by newly proposing a Reader-INherited evidence reranKer (RINK) where a reranker module is designed by finetuning the reader's neural architecture based on a simple prompting method. Our underlying assumption of reusing the reader's module for the reranker is that the reader's ability to generating an answer from evidence contains the knowledge required for the reranking, because the reranker needs to "read" in-depth a question and evidences more carefully and elaborately than a baseline retriever. Furthermore, we present a simple and effective pretraining method by extensively deploying the commonly used data augmentation methods of *cell corruption* and *cell reordering* based on the pretraining tasks – *tabular-and-textual entailment* and *cross-modal masked language modeling*. Experimental results on OTT-QA, a large-scale table-and-text open-domain question answering dataset, show that the proposed RINK armed with our pretraining procedure makes improvements over the baseline reranking method and leads to state-of-the-art performance.

Introduction

Open-domain question answering (ODQA) is a task that aims to find an answer to a given question, which explicitly requires the "retrieval" step for searching relevant knowledge. Going beyond the classical "textual" ODQA, recent studies have addressed multi-modal ODQA that additionally retrieves and reads other types of knowledge such as images and tabular contents, raising new challenges due to its heterogeneous type of data. Among the variants of multi-modal ODQA tasks, this paper addresses *table-and-text* ODQA, which uses heterogeneous data of tabular and textual contents as real-world knowledge for ODQA, as introduced in the OTT-QA benchmark of (Chen et al. 2021).

The existing approaches for table-and-text ODQA are primarily based on a *Retrieval-Reader* pipeline, in which the retrieval module first finds a set of relevant "heterogenous" evidences and the reader module then generates an answer using a decoder layer that includes the retrieved evidences as an additional input. In contrast to textual ODQA, the *multimodality* is a challenging issue in table-and-text ODQA, requiring the retrieval and reasoning components to be involved with a strong interaction across different types of evidence (i.e., textual, or tabular content). For the retrieval step in table-and-text ODQA, *early fusion* has been widely used to efficiently handle multimodality (Chen et al. 2021), by defining a *fusion block* as the basic retrieval unit, which is obtained by combining table cells with their related passages and differs from the retrieval based on a *single block* that searches tabular or textual blocks separately. Despite its effectiveness over the use of single blocks, the collection size of fusion blocks becomes readily *huge* because of its combinatorial nature, being considerably larger than the separate sizes of table cells and passages, rendering the retrieval process non-trivial[1].

To improve the retrieval performance for table-and-text ODQA, this paper extensively incorporates a "reranker" module for fusion block retrieval, thereby initially exploring the framework of *Retriever-Reranker-Reader* for table-and-text ODQA, described as follows:

1. **Reader-INherited evidence reranKer (RINK)**: In particular, we propose a novel type of reranker, referred to as RINK, in which the reranker module is inherited from the reader module with the same architecture and performs a binary classification based on a *prompting* method, enabling us to maximally share the reader's parameters for the reranker and to apply multi-task learning or transfer learning. The key property of RINK is that the reranker is based on a *set*-level reranking, in which the reranker is first applied to sampled sets of blocks, and then the resulting *set*-level evidences are *aggregated* to compute

[*] Work done at NAVER.
[†] Corresponding author.

[1] The initial retrieval performance reported in the OTT-QA dataset is only about only about 52.4% in terms of HITS@4K (Chen et al. 2021), which is considerably smaller than those of textual ODQA reported in the work of (Asai et al. 2020), which shows about 93% in terms of Paragraph Recall, although the evaluation metrics are not the same.

the relevance score of an individual block. This is different from the standard reranker module which is based on an *instance*-level reranking that directly estimates the relevance score of each individual block, usually via a cross-encoder (Soleimani, Monz, and Worring 2019; Kruengkrai, Yamagishi, and Wang 2021) on the concatenated sequence of a question and an individual retrieval block. The underlying assumption of the reranker's inherited design from the reader is that the reranker requires to "read' a question and a retrieval block more *in-depth* than the retriever, and this requirement for *in-depth reading* is likely presented to the reader even more strongly.

2. **Retriever based on pretrained encoders**: To improve the retriever, we present a simple "pretraining" method of the retriever's encoder to enhance representation by strengthening the multimodal interaction between tabular and textual contents to obtain enhanced multimodal representations based on self-supervised tasks – *tabular-and-textual entailment* and *cross-modal masked language modeling*. To pretrain the retriever's encoder, we deploy a synthetic dataset constructed automatically using two types of common data augmentation methods: *cell corruption* and *cell reordering*.

Experiment results on OTT-QA, a large-scale table-and-text ODQA dataset, show that the proposed RINK with the pretrained retriever outperforms the baseline reranking method and achieves state-of-the-art performance with 35.5 EM in the blind test set. Furthermore, our analysis presents that incorporating a base reranker and decoding-based evidence reranking system boosts retrieval performance and positively affects the performance of generative question answering.

The contribution of this paper is summarized as follows: 1) To the best of our knowledge, our study is the first to apply the "reranker" to improve the retrieval performance in table-and-text ODQA tasks; 2) we newly propose RINK, which is a set-level reranker that is induced by maximally reusing the reader module; 3) we present an effective pretraining method to improve the retriever's encoder using synthetic datasets obtained by simple data augmentation methods; 4) our proposed RINK leads to achieving state-of-the-art performance in OTT-QA.

Related Work

Open-Domain Question Answering

ODQA (Chen et al. 2017) is a task that finds the answer to a given question from a large set of data in multiple domains. The retrieval components for ODQA, which improves the retrieval performances of TF-IDF, are primarily based on dense retrieval based on the encoder of pretrained language models (Karpukhin et al. 2020). Reader components have been explored by either machine reading comprehension (MRC)-based extractive models or generation-based models (Izacard and Grave 2021b; Lewis et al. 2020a; Raffel et al. 2020; Lewis et al. 2020b). Multi-hop reasoning for QA has been studied in HotpotQA (Yang et al. 2018) which requires the model to read and understand multiple documents simultaneously to answer a given question.

Table-and-Text Question Answering

Recent studies on MRC and ODQA have been conducted on hybrid knowledge consisting of both tabular and textual contents. HybridQA (Chen et al. 2020b) is an MRC-style dataset for table-and-text QA and TAT-QA (Zhu et al. 2021) is an MRC task with financial table-and-text contents.

OTT-QA (Chen et al. 2021) is a large-scale table-and-text ODQA benchmark that requires systems to retrieve relevant evidences and to perform reasoning over heterogeneous knowledge of tabular and textual contents. (Chen et al. 2021) proposed "early fusion" mechanism, which defines a 'fusion block' as the basic retrieval unit that combines a table segment and its related passages.

DUREPA (Li et al. 2021) proposed a hybrid framework that uses a dual-reader to jointly encode tabular and textual contents, and generates either an SQL query or an answer on table-and-text contents.

Similar to the reasoning path in (Asai et al. 2020) on textual ODQA, CARP (Zhong et al. 2022) suggested the use of a 'hybrid chain' defined as a sequence of nodes from a heterogeneous graph, where nodes include a question, table cells, and sentences in passages. Furthermore, CARP proposed chain-centric pretraining to improve the domain-specific chain extractor based on data augmentation, which synthesizes a hybrid chain and reversely generates its corresponding question.

UnifiedSKG (Xie et al. 2022) standardizes datasets and models in a single framework to cover 21 knowledge-grounding datasets, including semantic parsing tasks such as Spider (Yu et al. 2018) and GrailQA (Gu et al. 2021), QA tasks such as WikiSQL (Zhong, Xiong, and Socher 2017), WikiTQ (Pasupat and Liang 2015), HybridQA, MultiModalQA (Talmor et al. 2021), and FeTaQA (Nan et al. 2022), as well as other tasks such as ToTTo (Parikh et al. 2020), KVRET (Eric et al. 2017), SQA (Iyyer, Yih, and Chang 2017), TabFact (Chen et al. 2020a), and FEVEROUS (Aly et al. 2021).

Dense Retrieval and Reranking

Dense passage retrieval (DPR) (Karpukhin et al. 2020) usually adopts a bi-encoder architecture (Humeau et al. 2020; Lee et al. 2021; Li et al. 2020) that encodes the question and passage separately. The encoded representations are then used to measure the similarity between the question and passage through the maximum inner-product search.

Related to our work, (Mao et al. 2021) proposed RIDER, the reader-guided reranker, solely based on the top predictions of the reader, by checking each passage from the initial retrieval result and adding it to a reranked list if it contains any reader prediction. In contrast to RIDER, which is an "instance"-level reranker based on the reader's results without requiring any training, our proposed RINK is the "set"-level reranker with an additional finetuning stage.

Another related work (Izacard and Grave 2021a) uses the cross-attention scores obtained from the Fusion-in-Decoder (FiD)-based reader as the relevance scores of passages.

However, the study of (Izacard and Grave 2021a) focuses on learning the bi-encoder DPR module using the reader's attention scores as the teacher's supervision, rather than improving the reranker like ours. Furthermore, in contrast to our proposed RINK that uses "multiple" sets of passages and aggregates the set-level relevance scores, (Izacard and Grave 2021a) derived an instance-level score using only a "single" set of passages without any aggregation. Because of this difference between (Izacard and Grave 2021a)'s work and ours, the method proposed by (Izacard and Grave 2021a) is complementary to our proposed RINK; for example, RINK can be generalized using its cross-attention scores as alternative relevance signals.

Background

Suppose that the task aims to find an answer to a question q and a set of fusion blocks $\mathcal{B}$, each fusion block b is presented by a table segment block b_T and its associated list of passages $b_P^1, \cdots, b_P^L$ as defined in (Chen et al. 2021) where L is the number of associated passages, BERT refers to the encoder of a pretrained language model such as (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), and T5 refers to the encoder-decoder language model such as T5 (Raffel et al. 2020) and BART (Lewis et al. 2020a). Unless otherwise specified, BERT is used by the retriever and T5 is adopted by the reranker and reader.

Retriever Using Bi-encoder

For the Retriever-Reranker-Reader framework, the retriever is based on a bi-encoder (Karpukhin et al. 2020; Humeau et al. 2020; Lee et al. 2021), which uses BERT to encode a question q and a fusion block b, and computes the similarity between them as follows:

$$\begin{aligned} \mathsf{Emb}_q(q) &= \mathbf{W}_q \mathsf{BERT}_{[\mathsf{CLS}]}(q) \\ \mathsf{Emb}_{block}(b) &= \mathbf{W}_{block} \mathsf{BERT}_{[\mathsf{CLS}]}(b) \\ sim(q,b) &= \mathsf{Emb}_q(q)^T \mathsf{Emb}_{block}(q) \end{aligned} \quad (1)$$

where $\mathsf{BERT}_{[\mathsf{CLS}]}(s)$ indicates the BERT's last encoded representation at [CLS] token on the sequence s, $\mathbf{W}_q, \mathbf{W}_{block} \in \mathbb{R}^{d_{model} \times d_{model}}$ are projection matrices, and d_{model} is the dimensionality of BERT. We employ FAISS (Johnson, Douze, and Jégou 2021) to efficiently obtain a set of top-N fusion blocks $\mathcal{B}_{init}$ based on $sim(q,b)$ in $\mathcal{B}$.

(Baseline) Reranker Using Cross-Encoder

The baseline instance-level reranker is based on the cross-encoder that feeds q [SEP] b, the concatenated input of a question q and a fusion block b, to BERT and performs the binary classification of whether the given block is relevant, formulated as follows:

$$\begin{aligned} \mathsf{Emb}_{cross}(q,b) &= \mathsf{BERT}_{[\mathsf{CLS}]}(q\ [\mathsf{SEP}]\ b) \\ rel_{rerank}(q,b) &= \log \sigma\left(\mathsf{Linear}\left(\mathsf{Emb}_{cross}(q,b)\right)\right) \end{aligned} \quad (2)$$

where $\mathsf{Linear}: \mathbb{R}^{d_{model}} \rightarrow \mathbb{R}^1$ is a linear layer using a reranker-specific projection matrix and σ is the sigmoid function. The reranker takes $\mathcal{B}_{init}$, the top N initial retrieval results, and produces the top M reranked results, denoted as $\mathcal{B}_{top}$.

Reader Using Fusion-in-Decoder (FiD)

Given the set of top-retrieved blocks $\mathcal{B}_{top}$, our reader module is based on FiD (Izacard and Grave 2021b); for each fusion block, its concatenated input with question q is fed into T5's encoder to produce its contextualized representation, and all the contextualized block representations are concatenated and then fed into T5's decoder. The FiD-based reader is described as follows:

$$\begin{aligned} \mathbf{C}_i &= \mathsf{T5\text{-}enc}(q\ [\mathsf{SEP}]\ b_i) \\ decode(q, \mathcal{B}_{top}) &= \mathsf{T5\text{-}dec}([\mathbf{C}_1; \cdots; \mathbf{C}_M], [\mathsf{PAD}]) \end{aligned} \quad (3)$$

where $b_i \in \mathcal{B}_{top}$ is the i-th fusion block, ';' is the concatenation operator, T5-enc indicates the T5's encoder, $\mathsf{T5\text{-}dec}(x,y)$ represents the decoder function where x is the input sequence and y is the prefix sequence for the decoder, and [PAD] is the padding token used for the start token of the decoder[2].

Proposed Approach

Figure 1 shows the overall neural architecture of the reranker using RINK under the framework of the Retriever-Reranker-Reader, that is, the retriever with the bi-encoder, the baseline reranker with the cross-encoder, and the proposed RINK using the FiD architecture. The background section presents a retriever and baseline reranker, including the FiD-based reader. This section presents the details of the proposed RINK and describes the proposed pretraining for improving the retriever.

Reader-Inherited Reranker (RINK)

The proposed RINK directly employs the reader module but is fine-tuned based on a *prompting* method to determine whether a given set of blocks is relevant. Here, a set of blocks is relevant if at least one element block is relevant. When multiple sets of blocks and their relevance labels are provided, RINK derives the relevance score of an individual block b by computing the degree to which the sets are likely to be relevant while containing the given block b, which is proportional to the ratio of the size of relevant sets to that of sets containing b.

Set-Level Relevance Classification Suppose that $\mathcal{V}$ is a set of vocabulary and $\mathsf{T5\text{-}dec}_{\mathsf{token}}(x,y)$ is the autoregressive language model of T5 that produces a probability vector over $\mathcal{V}$ for the next token, when x is the input sequence and y is the prefix sequence. Given a set of blocks $\mathcal{B}$ and $b_i \in \mathcal{B}$, we obtain $\mathbf{p}(q,\mathcal{B}) \in \mathbb{R}^{|\mathcal{V}|}$, a probability vector over $\mathcal{V}$ based on T5's decoder, as follows:

$$\begin{aligned} \mathbf{C}'_i &= \mathsf{T5\text{-}enc}(\text{“query:”}\ q\ \text{“block:”}\ b_i\ \text{“relevant:”}) \\ \mathbf{p}(q,\mathcal{B}) &= \mathsf{T5\text{-}dec}_{\mathsf{token}}([\mathbf{C}'_1; \cdots; \mathbf{C}'_M], [\mathsf{PAD}])) \end{aligned} \quad (4)$$

where "query:", "block:", and "relevant:" are *prompt tokens*, being designed by mostly following (Hu et al. 2022)[3].

[2]With abuse of notation, it is assumed that $\mathsf{T5\text{-}dec}(x,y)$ can take the pre-encoded contextualized representations for x.

[3]Although we use the same notation of T5-enc as the reader's one, we fine-tune T5 separately for each module and keep different

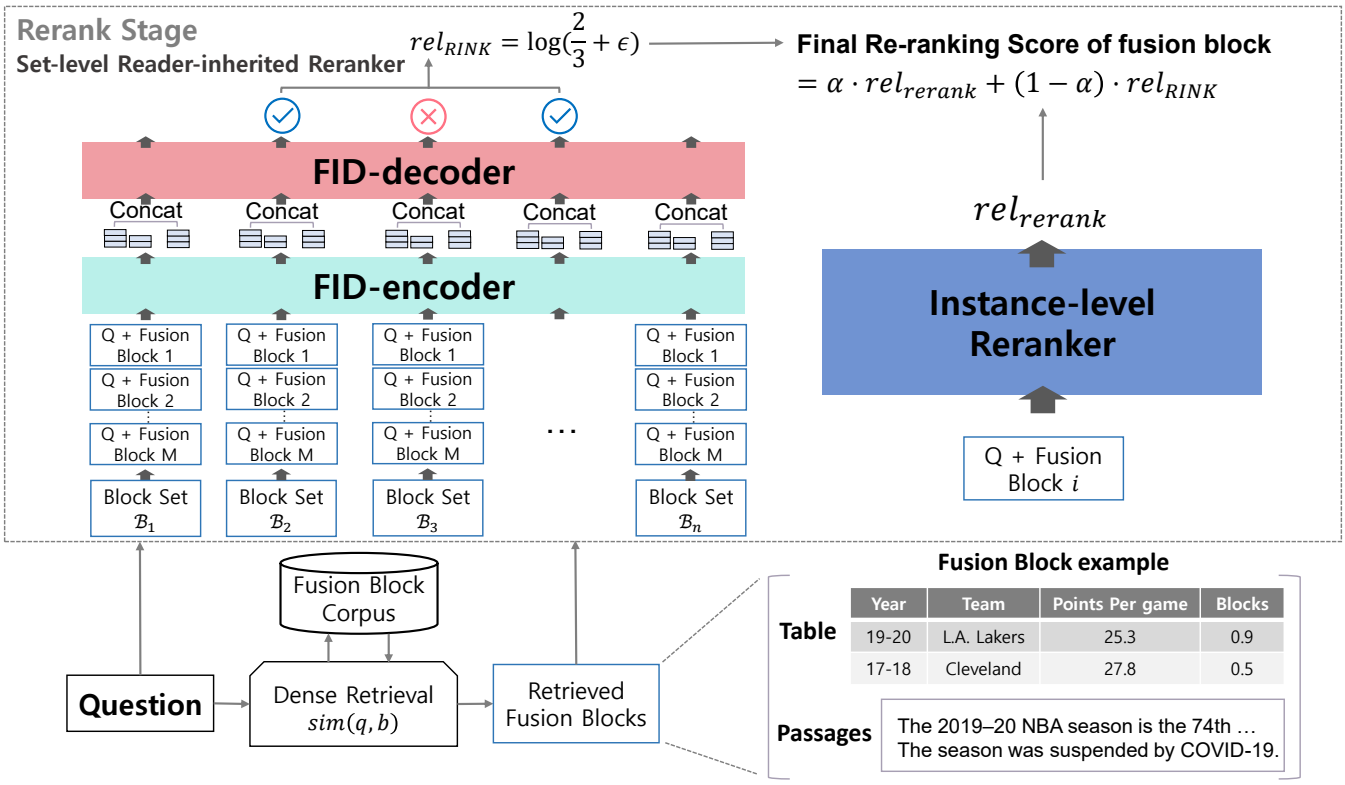

Figure 1: Neural architecture of our proposed reranker using RINK under the framework of Retriever-Reranker-Reader: 1) given a question q and a fusion block b, the dense retrieval computes the similarity $sim(q, b)$ based on bi-encoder using Eq. (1) and returns the initial retrieved results $\mathcal{B}_{init}$; 2) the proposed RINK applies the set-level relevance classification of using Eq. (4)-(5) on a collection of block sets $\mathcal{S} = \{\mathcal{B}\}_{i=1}^{n}$, where $\mathcal{B}_i \subseteq \mathcal{B}_{init}$, and aggregate these set-level relevance scores using Eq. (7) to derive the instance-level score $rel_{RINK}(q, b)$ for the block b; 3) the baseline reranker computes the relevance score $rel_{rerank}(q, b)$ based on cross-encoder using Eq. (2); 4) the proposed set-level RINK and baseline reranker are combined using a linear combination using Eq. (8) to finally produce the relevance score $rel_{combined}(q, b)$.

For prompting-based classification, let $v: \mathcal{Y} \rightarrow \mathcal{V}$ be the verbalizer that converts a label into individual words, where $\mathcal{Y} = \{\text{Nonrel}, \text{Rel}\}$. Here, $v(\text{Nonrel}) =$ "false" and $v(\text{Rel}) =$ "true", which refer to the nonrelevance and relevance labels, respectively.

Given a pair of a question and a set of blocks $(q, \mathcal{B})$, the probability of the relevance label $\text{Rel} \in \mathcal{Y}$ for $\mathcal{B}$ is computed as follows:

$$\text{p}(\text{Rel}|q, \mathcal{B}) = \mathbf{p}(q, \mathcal{B})[v(\text{Rel})] / \sum_{w \in \mathcal{Y}} \mathbf{p}(q, \mathcal{B})[v(w)] \quad (5)$$

where $\mathbf{p}(q, \mathcal{B})[w]$ is an element of the probability vector $\mathbf{p}(q, \mathcal{B})$ corresponding to token w.

Aggregation over Set-Level Relevance Results We first prepare multiple set-level relevance results by constructing a collection of block sets $\mathcal{S} = \{\mathcal{B}_1, \cdots, \mathcal{B}_n\}$ by randomly sampling $B_i \subset \mathcal{B}_{init}$, with the constraint that the number of sets containing each block is the same. Let $\mathcal{S}(b) \subseteq \mathcal{S}$ be the collection of sets containing b, defined as follows:

$$\mathcal{S}(b) = \{\mathcal{B} | b \in \mathcal{B} \text{ and } \mathcal{B} \in \mathcal{S}\} \quad (6)$$

The constraint can be restated as that we need to create a collection $\mathcal{S}$, such that $|\mathcal{S}(b)| = K$ for any block b. For consistency with the setting of the reader, we maintain the size of the set as the same as the reader, as possible, i.e., $|\mathcal{B}| = M$ for any sampled set $\mathcal{B} \in \mathcal{S}$. Given $|\mathcal{B}_{init}| = N$ and $|\mathcal{B}_i| = M$, the number of set samples is $n = N \times K/M$[4]; in our experiment setting where $N = 100$ and $K = 3k$, $n = 20k$ for $M = 15$ and $n = 30k$ for $M = 10$.

We apply the set-level classification using Eq. (4)-5 for all block sets in $\mathcal{S} = \{\mathcal{B}\}_{i=1}^{n}$. From these set-level classification results, we derive the instance-level score of block b based

parameters between the reranker and the reader for the encoder and decoder of T5.

[4]More generally, $n = \lceil N \times K/M \rceil$ to cover the case that $N \times K \bmod M \neq 0$. In this case, a simple solution is to allow an exceptional case of a block $|\mathcal{B}| < M$ or to fill the remaining blocks with the empty block.

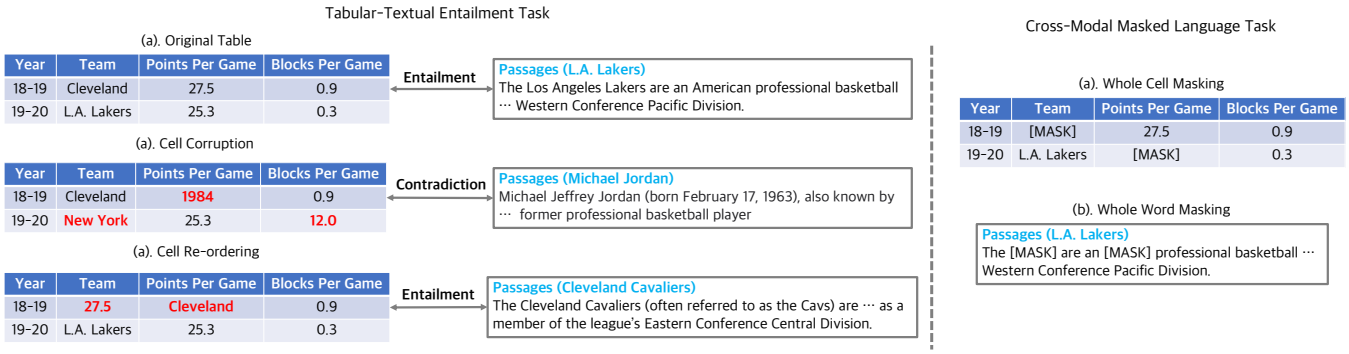

Figure 2: Pretraining tasks for the retriever's encoder. Left: Tabular-and-textual entailment tasks uses the data augmentation of table perturbation methods to generate positive and negative examples – a) the original table; b) the "cell corruption", that randomly selects a cell and replaces its original value with another one; c) The "cell reordering", which randomly changes the order of cells in the same row. Right: Cross-modal masked language modeling task applies the whole-cell/word masking, accepts a "masked" concatenated sequence of tabular and textual contents, and then performs the masked token prediction.

on the following aggregation function:

$$rel_{RINK}(q,b) = \log\left(\frac{\sum_{\mathcal{B}\in\mathcal{S}(b)}\mathcal{I}(\mathsf{p}(\mathsf{Rel}|q,\mathcal{B})>0.5)}{K}+\epsilon\right) \quad (7)$$

where ϵ is the small constant value for smoothing the probability and $\mathcal{I}(e)$ is the indicator function that is one if e is true and zero otherwise.

Combining with the Baseline Instance-Level Reranker It is expected that the baseline instance-level reranker of Eq. (2), and set-level RINK of Eq. (7) capture their specialized features for computing the relevance score of a block. Thus, we combine these two types of reranker using a simple linear function, as follows:

$$rel_{combined}(q,b) = \alpha \cdot rel_{rerank} + (1-\alpha)\cdot rel_{RINK} \quad (8)$$

where α is an interpolation parameter, which is tuned on the development set[5].

Training To train RINK of Eq. (4)-(5), we prepare a collection of positive and negative training examples by randomly creating subsets of blocks from $\mathcal{B}_{init}$. As a positive example, we randomly add one to five "gold" fusion blocks and fill the remaining blocks with nonrelevant blocks. A negative set is constructed by randomly sampling M blocks among the nonrelevant blocks in $\mathcal{B}_{init}$. We use a 1:9 ratio of positive to negative training samples. The T5's parameters of RINK are first initialized by the reader's ones and fine-tuned using the cross-entropy loss on training examples constructed in an aformentioned manner.

A similar training procedure is applied to train the baseline reranker in Eq. (2). A gold fusion block in $\mathcal{B}_{init}$ is used as a positive example, and any nonrelevant block as a negative one.

[5]It is noted that the initial retrieval score Eq. (1) is not used for the combination of Eq. (8).

Pretraining the Encoder of Retriever

Data Augmentation for Pretraining Inspired by (Iida et al. 2021), we construct a large-scale pre-training corpus using the parsed Wikipedia corpus consisting of over 200K tables and 3M hyperlinked passages. To this end, we employ two types of common data augmentation methods of *cell corruption* and *cell reordering*, which are introduced in the work of (Iida et al. 2021). As illustrated in Figure 2, cell corruption is a cell perturbation method that randomly selects one cell and replaces its original value with another. Cell reordering randomly changes the order of cells in the same row[6].

Pretraining Tasks: Tabular-and-Textual Entailment and Cross-modal Masked Language Modeling To pretrain the retriever's encoder BERT, we first introduce the *tabular-and-textual entailment* task to strengthen the interaction between tabular and textual contents, where a pair of a table and a passage is assumed to be an "entailment" class when there is a table cell that is hyperlinked to the passage, and other pairs are merely regarded as "contradiction" class. The ratio of positive to negative examples for the entailment task is set to 1:1.

In addition, we use the *cross-modal masked language modeling task*, where a "masked" concatenated sequence of tabular and textual contents is provided as an input to the retriever's encoder and the masked token prediction is performed. We apply whole-cell masking for table contents and BERT's whole-word masking for textual ones, as proposed by (Herzig et al. 2020).

Experiments

Experimental Setup

Table 1 shows the detailed statistics of OTT-QA (Chen et al. 2021). By employing the data augmentation described in

[6]The cell corruption and the cell reordering correspond to the cases of "sample cells from other tables" and "swap cells on the same row" of Figure 3 in the work of (Iida et al. 2021), respectively.

Category	Number
Fusion Block Number	5,411,408
Passage Number	6,342,314
Table Number	410,740
Train Dataset	41,469
Developement Dataset	2,214
Test Dataset	2,158
Pretraining Dataset	400,000

Table 1: Statistics of the OTT-QA benchmark dataset and automatically constructed pretraining dataset: "Pretraining Dataset" indicates the number of pairs of tables and texts constructed via the data augmentation of Figure 2. All the other statistics derive from the OTT-QA dataset.

Model	Dev		Test	
	EM	F1	EM	F1
HYBRIDER	10.3	13.0	9.7	12.8
IR + SBR	7.9	11.1	9.6	12.8
FR + SBR	13.8	17.2	13.4	16.9
IR + CBR	14.4	18.5	16.9	20.9
FR + CBR	28.1	32.5	27.2	31.5
DUREPA	15.8	–	–	–
CARP	33.2	38.6	32.5	38.5
Instance-level Reranker ($M = 15$)	35.2	41.2	34.5	40.4
RINK ($\alpha = 0, M = 15, K = 30$)	34.6	40.3	32.7	38.6
RINK ($\alpha = 0, M = 10, K = 30$)	35.2	40.8	33.3	39.3
RINK ($\alpha = 0.7, M = 15, K = 30$)	36.2	42.1	35.0	41.0
RINK ($\alpha = 0.7, M = 10, K = 30$)	**36.7**	**42.4**	**35.5**	**41.5**

Table 2: QA performances on the dev and blind test sets of OTT-QA datasets. The proposed RINK with $K = 30$ is compared with various baseline models for $M = 10$ and $M = 15$. "Instance-level reranker" refers to the baseline reranker of using Eq. (2), corresponding to RINK ($\alpha = 1$). The best performance is shown in **bold** text.

the previous section, we automatically constructed about 400,000 pairs of tables and texts as an additional pretraining dataset from the parsed Wikipedia corpus that consists of about 200K tables and 3M hyperlinked passages. We used the **RoBERTa-base** [7] (Liu et al. 2019) model for the BERT (i.e., the encoder used in the retriever and the baseline reranker.). We used the **T5-base** [8] (Raffel et al. 2020) model for T5 and the FiD decoder (Izacard and Grave 2021b) (i.,e., the encoder-decoder used in the RINK and the reader). As mentioned previously, the RINK parameters are initialized by the reader's fine-tuned weights on the OTT-QA dataset. We used a batch size of 16 and learning rates of 5×10^{-5} and 1×10^{-4} to train the BERT and T5, respectively. The AdamW optimizer was used for training. All experiments were conducted using eight NVIDIA Quadro RTX A6000 GPUs. For the initial retrieval and reranker, $N = |\mathcal{B}_{init}|$ was fixed at 100 and $M = |\mathcal{B}_{top}|$ is either 10 or 15. The combination parameter, α, was tuned on the development set.

[7] https://huggingface.co/roberta-base

[8] https://huggingface.co/t5-base

Model	R@1	R@5	R@10	R@15
Bi-Encoder	–	–	72.9	–
Tri-Encoder	–	–	73.8	–
CARP	49.0	–	74.0	–
Retriever	51.0	66.2	76.6	79.6
RINK ($M = 15, \alpha = 0$)	61.4	77.4	81.6	83.7
RINK ($M = 10, \alpha = 0$)	62.7	78.0	81.9	84.0
Instance-level Reranker ($M = 15$)	65.0	77.4	81.2	83.6
RINK ($M = 15, \alpha = 0.7$)	66.1	79.9	83.3	**85.5**
RINK ($M = 10, \alpha = 0.7$)	**66.9**	**80.6**	**83.5**	85.4

Table 3: Retrieval performance on development set of OTT-QA. RINK using $K = 30$ is compared with CARP and other baselines under the settings of $M = 10$ and $M = 15$: "Retriever" indicates the initial retrieval method of using Eq. (1) without the reranking module; "Instance-level reranker" indicates the baseline reranker of using Eq. (2), which corresponds to the case of RINK ($\alpha = 1$). The best performance is shown in **bold** text and we report the retrieval performance of bi-encoder, tri-Encoder, and CARP, as in (Kostić, Risch, and Möller 2021; Zhong et al. 2022).

Model	K	R@1	R@5	R@10	R@15
RINK ($\alpha = 0$)	6	58.7	75.7	80.6	82.8
	9	59.8	76.5	81.1	83.6
	15	60.8	76.6	81.3	83.6
	24	**62.0**	77.0	81.0	83.0
	30	61.4	**77.4**	**81.6**	**83.7**

Table 4: Retrieval performances of RINK ($M = 15, \alpha = 0$) with varied K values on the development set of OTT-QA datasets.

Baselines

We compared RINK with the following methods:

- **HYBRIDER** (Chen et al. 2020b) employs a sparse retriever (i.e., BM25 and TF-IDF) to retrieve relevant tables and passages, and uses a reasoning model based on ranking, hop, and reading comprehension (RC) models to extract an answer.
- **{Iterative, Fusion}-Retriever / {Single, Cross}-Block Reader** (Chen et al. 2021) proposed the iterative retriever (**IR**), fusion retriever (**FR**), single block reader (**SBR**), and cross block reader (**CBR**). IR uses the iterative retrieval protocol, and FR adopts the "early fusion" strategy to handle the multi-modality issue using a fusion block as the basic retrieval unit. The SBR feeds the top-k retrieved blocks to the reader one by one, and selects the answer with the highest confidence score. The CBR feeds all concatenated top-k blocks together into the reader.
- **DUREPA** (Li et al. 2021) jointly reads tables and passages using the dual-reader architecture and generates ei-

Model	K	EM	F1
RINK ($\alpha = 0$)	6	33.4	39.3
	9	33.7	39.4
	15	34.1	40.0
	24	34.4	40.2
	30	**34.6**	**40.3**

Table 5: QA performances of RINK ($M = 15$, $\alpha = 0$) with varied K values on the development set of OTT-QA datasets.

Model	R@1	R@5	R@10	R@15
Retriever (w/ pre-train)	**52.2**	**70.2**	76.6	**79.8**
Retriever (w/o pre-train)	51.0	66.2	76.6	79.6

Table 6: Retrieval performance of baseline retriever on development set with and without pretraining.

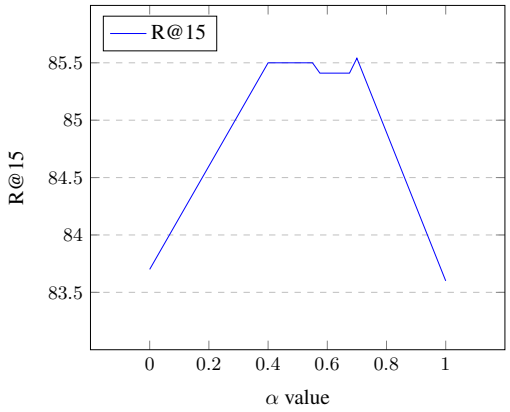

Figure 3: Performance curve of the reranking by RINK with $M = 15$ in terms of R@15, varying α in Eq. (8) on the development set of OTT-QA.

ther an answer or an executable SQL query to derive the answer.

- **CARP** (Zhong et al. 2022) proposed the use of a hybrid chain defined as a sequence of nodes from a heterogeneous graph, whose nodes include a question, table cells, and sentences in passages to promote multimodal reasoning ability.

Main Results

Tables 2 and 3 show the performance of QA and retrieval, respectively. As shown in Table 2, the proposed RINK with $M = 10$ outperforms CARP, the best baseline model, by increasing EM by 3.5 and 3.0 points on the development and blind test sets, respectively, and achieves state-of-the-art performance. In particular, it is interesting to note that RINK using only the set-level reranker without including the baseline reranker (i.e., $\alpha = 0$) further improves the CARP.

In Table 3, it can be observed that the use of the reranker shows substantial improvements in the retrieval performance where "Retriever" refers to the initial retrieval method of using Eq. (1) without the reranking module and "Instance-level reranker" indicates the baseline reranker of using Eq. (2), which corresponds to RINK ($\alpha = 1$). While the retrieval performance of "Retriever" is similar to those CARP results, the reranking either using "Instance-level reranker" or RINK leads to significant improvements, with increases of more than 10% at R@1 and R@5. The proposed set-level RINK shows further improvements over the "Instance-level reranker," particularly showing an increase of approximately 2% at R@5, R@10, and R@15, when $M = 15$. Given this improvement in the retrieval by RINK, it is shown from Table 3 that the resulting QA performances of using RINK lead to further improvements over those of "Instance-level reranker," both on development and blind test sets, increasing EM and F1 by approximately $0.5 \sim 1\%$, for $M = 15$.

Overall, the results consistently suggest us that the Retriever-Reranker-Reader is a promising approach in the table-and-text ODQA literature, and the "reranker" module is one of the key components for achieving state-of-the-art performance, given that our other components, such as the retriever and reader are similar to or simpler than CARP.

Ablation Study

Effect of Varying K for RINK To examine the effect of K under RINK($\alpha = 0$), Tables 4 and 5 present the retrieval and QA performances of RINK with varying K values, respectively. Intuitively, the larger the value of K, the more effective the reranking, because the instance-level aggregated score of Eq. (7) is estimated more accurately for larger values of K. The results in Tables 4 and 5 confirm our expectation that when K is larger, the retrieval and QA performances gradually increase.

Effect of Varying α for RINK Figure 3 shows the curve of the retrieval performance by RINK in terms of R@15, with varying α in Eq. (8) on the development set with an increment of 0.1. It is shown that the performances are relatively high in the range of $[0.4, 0.7]$ for α, where $\alpha = 0.7$ shows the best performance.

Effect of Pretraining via Data Augmentation To examine the effect of pretraining the retriever's encoder, Table 6 lists the initial retrieval performances of the "Retriever" with and without pretraining. Although the improvements are marginal in most cases, the pretraining results in an increase of approximately 4% at R@5 compared to the case without pretraining. Table 7 further presents QA performances of FiD and RINK with $M = 15$ with and without pretraining on blind test set. It is shown that the effect of the pretraining on QA performance is more dominant than that on retrieval performance, leading to increases of about 1.5% over non-pretraining runs both at EM and F1, under both of FiD and RINK.

Considering that the size of the pretraining dataset is 400,000, which is relatively small compared to other pretraining literature, we believe that the effect of pretraining becomes stronger when using synthetic datasets of larger sizes.

(a)

Q: This 70 's Kishore Kumar song was in a film produced by Alankar Chitra and directed by Shanker Mukherjee ?
B1 (Kishore_Kumar_0): [TITLE] Kishore Kumar [SECTITLE] Awards [TABLE] Year is 1970. Song is Roop Tera Mastana. Film is Aradhana. Music director is Sachin Dev Burman. Lyricist is Anand Bakshi. ⋯
B2 (Kishore_Kumar_1): [TITLE] Kishore Kumar [SECTITLE] Awards [TABLE] Year is 1975. Song is Main Pyaasa Tum. Film is Faraar. Music Director is Kalyanji Anandji. Lyricist is Rajendra Krishan. [PASSAGE] ⋯ The film is produced by Alankar Chitra and directed by Shanker Mukherjee.
Gold Table Id: Kishore_Kumar_1
Generated Answer: Main Pyaasa Tum
Reference Answer: Main Pyaasa Tum

(b)

Q: What position does 2009-10 season Vancouver Canucks player Rob Davison currently hold with the Toronto Marlies ?
B1 (2009-10_Vancouver_Canucks_season_14): [TITLE] 2009-10 Vancouver Canucks season [SECTITLE] Free agents lost [TABLE] Player is Rob Davison. New team is New Jersey Devils. Contract terms is . [PASSAGE] ⋯ He is currently serving as assistant coach of the Toronto Marlies ⋯
B2 (2009-10_Vancouver_Canucks_season_15): [TITLE] 2009-10 Vancouver Canucks season [SECTITLE] Draft picks [TABLE] Round is 2. # is 53. Player is Anton Rodin (RW). Nationality is Sweden. College/junior/club team (League) is Brynas IF Jr. (J20 SuperElit) ⋯
Gold Table Id: 2009-10_Vancouver_Canucks_season_15
Generated Answer: assistant coach
Reference Answer: assistant coach

(c)

Q: What is Spain 's oldest sporting club solely devoted to football with a 2014-15 Fenerbahçe S.K . season result F-A of 0-2 ?
B1 (2014-15_Fenerbahçe_S.K._season_7): [TITLE] 2014-15 Fenerbahçe S.K. season [SECTITLE] Pre-season friendlies [TABLE] Date is 16 August 2014. Opponents is Olympiacos. Stadium is Şükrü Saracoğlu Stadium. Result F-A is 2-1. ⋯
B2 (2014-15_Fenerbahçe_S.K._season_7): [TITLE] 2014-15 Fenerbahçe S.K. season [SECTITLE] Pre-season friendlies [TABLE] Date is 8 August 2014. Opponents is Chelsea. Stadium is Şükrü Saracoğlu Stadium. Result F-A is 0-2. Attendance is 21,300. ⋯
B6 (2014-15_FC_Barcelona_season_8): ⋯ with Athletic Bilbao and Barcelona .
Gold Table Id: 2014-15_Fenerbahçe_S.K._season_7
Generated Answer: Barcelona
Reference Answer: Sevilla Fútbol Club

(d)

Q: Is 20th Century Fox or Walt Disney Pictures the creative studio behind the most highest grossing film of Singapore ?
B1 (List_of_highest-grossing_films_in_Singapore_4): [TITLE] List of highest-grossing films in Singapore [SECTITLE] Top-grossing films in 2006 [TABLE] Rank is 1. Title is X-Men : The Last Stand. Studio is 20th Century Fox. Gross is $ 4.8m. ⋯
B14 (List_of_highest-grossing_films_in_Singapore_8): [TITLE] List of highest-grossing films in Singapore [SECTITLE] Top-grossing films of all time [TABLE] Rank is 2. Title is Avengers : Infinity War. Studio is Walt Disney Pictures. Lifetime Gross is S $ 16.22mil. Year is 2018.
Gold Table Id: List_of_highest-grossing_films_in_Singapore_16
Generated Answer: 20th Century Fox
Reference Answer: Walt Disney Pictures

Figure 4: Case studies of QA results by RINK that show samples in the top M retrieved fusion blocks obtained by the proposed reranker of Eq. (8) and generated answers by the reader, where Bi indicates the top i-th retrieved block and the *gold* table segments are presented in Figure 5. (a)-(b): the correct cases where their reasoning types are text → table in B2 (in (a)) and table → text in B1 (in (b)). (c): the case with the *retrieval error* where the gold table segment (i.e., Figure 5-(c)) is not appear in the top M retrieved blocks, while its reasoning path seems to be ended with B6 after trials of question matching. (d): the case with the *numerical reasoning error* where both relevant blocks B1 and B14 are successfully retrieved, while the answer extraction is failed to precisely perform the numerical reasoning that selects the cell with the highest gloss among B1 and B14.

(a) Kishore_Kumar_1

Year	Song	Flim	Music Director	Lyricist
1975	Main Pyaasa Tum	Faraar	Kalyanji-Anandji	Rajendra Krishan

(b) 2009-10_Vancouver_Canucks_season_15

Player	New team	Contract Terms
Rob Davison	New Jersey Devils	N/A

(c) 2014-15_Fenerbahçe_S.K._season_7

Date	Opponents	Stadium	Result F-A	Attendance
2 August 2014	Sevilla Fútbol Club	Brøndby Stadium	2-0	N/A

(d) List_of_highest-grossing_films_in_Singapore_16

Rank	Title	Studio	gross	Year
1	X-Men: The Last Stand	20th Century Fox	$4.8m	2018
2	Avengers: Infinity War	Walt Disney Pictures	$16.22m	2018

Figure 5: The gold table segment blocks manually annotated corresponding to the examples in Figure 4 with their table ids.

Case Studies Figure 4-5 show four illustrating examples of QA results obtained by RINK and their *gold* table segments. Figure 4 (a)-(b) present the correct cases, where relevant fusion blocks are successfully retrieved by the reranker and the reasoning for matching types of the text → table and table → text, which is required to generate a correct answer, is also performed well.

On the other hand, Figure 4 (c)-(d) demonstrate the incorrect cases, due to the *retrieval* and the *numerical reasoning errors*, respectively. In Figure 4-(c), without referring to the gold table segment of Figure 5-(c), the reader unnecessarily performs the reasoning of table → text across B2 and B16.

In Figure 4-(d), while the gold relevant blocks B1 and B14 are successfully retrieved, the reader is failed to perform the necessary numerical reasoning that selects the table row with a higher gloss among two table rows B1 and B14.

Conclusion

This work is an initial exploration of the Retriever-Reranker-Reader framework for table-and-text ODQA, emphasizing the importance of the retrieval step due to the multimodality issue. We proposed RINK, a novel set-level reranking method that reuses the reader's module and its finetuned parameters for reranking, and presented a prompting method

Model	Pretraining	EM	F1
FiD	✗	28.5	34.4
	✓	**30.1**	**36.0**
RINK ($M = 10, \alpha = 0.7$)	✗	33.8	40.0
	✓	**35.5**	**41.5**

Table 7: QA performances of FiD and RINK on blind test set with and without pretraining.

that performs the binary classification for reranking, without any modification of the reader's module. In addition, we presented pretraining method fpr the retriever's encoder, based on tabular-and-textual entailment and cross-modal masked language modeling tasks, on an additionally constructed dataset deploying two data augmentation methods – cell corruption and cell reordering. The experimental results on OTT-QA showed that the proposed RINK led to state-of-the art performance and consistently confirmed that the retrieval step was the key component for improving the QA performance, thus suggesting us to further investigate the Retriever-Reranker-Reader framework as a promising approach to table-and-text ODQA.

In the future, we would like to extend the set-level RINK by using the cross-attention scores of (Izacard and Grave 2021a) as an additional relevance signal. By extending REALM (Guu et al. 2020) and RAG (Lewis et al. 2020b), we would also establish an end-to-end learning framework of Retriever-Reranker-Retriever for table-and-text ODQA and explore data augmentation methods directly to train all the components in the framework in a joint manner.

Acknowledgements

This work was supported by NAVER Corp. We would like to thank all anonymous reviewers for their valuable comments and suggestions.

References

Aly, R.; Guo, Z.; Schlichtkrull, M. S.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; and Mittal, A. 2021. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 1–13. Dominican Republic: Association for Computational Linguistics.

Asai, A.; Hashimoto, K.; Hajishirzi, H.; Socher, R.; and Xiong, C. 2020. Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering. In *International Conference on Learning Representations*.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 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)*, 1870–1879. Vancouver, Canada: Association for Computational Linguistics.

Chen, W.; Chang, M.; Schlinger, E.; Wang, W. Y.; and Cohen, W. W. 2021. Open Question Answering over Tables and Text. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; and Wang, W. Y. 2020a. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations*.

Chen, W.; Zha, H.; Chen, Z.; Xiong, W.; Wang, H.; and Wang, W. Y. 2020b. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Eric, M.; Krishnan, L.; Charette, F.; and Manning, C. D. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 37–49. Saarbrücken, Germany: Association for Computational Linguistics.

Gu, Y.; Kase, S.; Vanni, M.; Sadler, B.; Liang, P.; Yan, X.; and Su, Y. 2021. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In *Proceedings of the Web Conference 2021*, WWW '21, 3477–3488. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval Augmented Language Model Pre-Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 3929–3938. PMLR.

Herzig, J.; Nowak, P. K.; Müller, T.; Piccinno, F.; and Eisenschlos, J. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Hu, X.; Yu, S.; Xiong, C.; Liu, Z.; Liu, Z.; and Yu, G. 2022. P3 Ranker: Mitigating the Gaps between Pre-Training and Ranking Fine-Tuning with Prompt-Based Learning and Pre-Finetuning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 1956–1962. New York, NY, USA: Association for Computing Machinery.

Humeau, S.; Shuster, K.; Lachaux, M.; and Weston, J. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Iida, H.; Thai, D.; Manjunatha, V.; and Iyyer, M. 2021. TABBIE: Pretrained Representations of Tabular Data. In

Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 3446–3456. Association for Computational Linguistics.

Iyyer, M.; Yih, W.-t.; and Chang, M.-W. 2017. Search-based Neural Structured Learning for Sequential Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1821–1831. Vancouver, Canada: Association for Computational Linguistics.

Izacard, G.; and Grave, E. 2021a. Distilling Knowledge from Reader to Retriever for Question Answering. In *International Conference on Learning Representations*.

Izacard, G.; and Grave, E. 2021b. 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*.

Johnson, J.; Douze, M.; and Jégou, H. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data*, 7(3): 535–547.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kostić, B.; Risch, J.; and Möller, T. 2021. Multi-modal Retrieval of Tables and Texts Using Tri-encoder Models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, 82–91. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kruengkrai, C.; Yamagishi, J.; and Wang, X. 2021. A Multi-Level Attention Model for Evidence-Based Fact Checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Lee, J.; Sung, M.; Kang, J.; and Chen, D. 2021. Learning Dense Representations of Phrases at Scale. 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)*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020a. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020b. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, A. H.; Ng, P.; Xu, P.; Zhu, H.; Wang, Z.; and Xiang, B. 2021. Dual Reader-Parser on Hybrid Textual and Tabular Evidence 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)*.

Li, B. Z.; Min, S.; Iyer, S.; Mehdad, Y.; and Yih, W.-t. 2020. Efficient One-Pass End-to-End Entity Linking for Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Mao, Y.; He, P.; Liu, X.; Shen, Y.; Gao, J.; Han, J.; and Chen, W. 2021. Reader-Guided Passage Reranking for Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Nan, L.; Hsieh, C.; Mao, Z.; Lin, X. V.; Verma, N.; Zhang, R.; Kryściński, W.; Schoelkopf, H.; Kong, R.; Tang, X.; Mutuma, M.; Rosand, B.; Trindade, I.; Bandaru, R.; Cunningham, J.; Xiong, C.; Radev, D.; and Radev, D. 2022. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics*, 10: 35–49.

Parikh, A.; Wang, X.; Gehrmann, S.; Faruqui, M.; Dhingra, B.; Yang, D.; and Das, D. 2020. ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pasupat, P.; and Liang, P. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1470–1480. Beijing, China: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. volume 21, 140:1–140:67.

Soleimani, A.; Monz, C.; and Worring, M. 2019. BERT for Evidence Retrieval and Claim Verification. volume abs/1910.02655.

Talmor, A.; Yoran, O.; Catav, A.; Lahav, D.; Wang, Y.; Asai, A.; Ilharco, G.; Hajishirzi, H.; and Berant, J. 2021. MultiModal{QA}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C.-S.; Zhong, M.; Yin, P.; Wang, S. I.; Zhong, V.; Wang, B.; Li, C.; Boyle, C.; Ni, A.; Yao, Z.; Radev, D.; Xiong, C.; Kong, L.; Zhang, R.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. *arXiv preprint arXiv:2201.05966*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yu, T.; Zhang, R.; Yang, K.; Yasunaga, M.; Wang, D.; Li, Z.; Ma, J.; Li, I.; Yao, Q.; Roman, S.; Zhang, Z.; and Radev, D. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3911–3921. Brussels, Belgium: Association for Computational Linguistics.

Zhong, V.; Xiong, C.; and Socher, R. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *CoRR*, abs/1709.00103.

Zhong, W.; Huang, J.; Liu, Q.; Zhou, M.; Wang, J.; Yin, J.; and Duan, N. 2022. Reasoning over Hybrid Chain for Table-and-Text Open Domain Question Answering. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 4531–4537. ijcai.org.

Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; and Chua, T.-S. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. 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)*.