Mixed-modality Representation Learning and Pre-training for Joint Table-and-Text Retrieval in OpenQA

Anonymous ACL submission

Abstract

Retrieving evidences from tabular and textual resources is essential for open-domain question answering (OpenQA), which provides more comprehensive information. However, training an effective dense table-text retriever is difficult due to the challenges of table-text discrepancy and data sparsity problem. To address the above challenges, we introduce an optimized **O**penQA **T**able-**TE**xt **R**etriever (OTTER) to jointly retrieve tabular and textual evidences. Firstly, we propose to enhance mixed-modality representation learning via two mechanisms: modality-enhanced representation and mixed-modality negative sampling strategy. Secondly, to alleviate data sparsity problem and enhance the general retrieval ability, we conduct retrieval-centric mixed-modality synthetic pre-training. Experimental results demonstrate that OTTER substantially improves the performance of table-and-text retrieval on the OTT-QA dataset. Comprehensive analyses examine the effectiveness of all the proposed mechanisms. Besides, equipped with OTTER, our OpenQA system achieves the state-of-the-art result on the downstream QA task, with 10.1% absolute improvement in terms of the exact match over the previous best system. [1]

1 Introduction

Open-domain question answering (Joshi et al., 2017; Dunn et al., 2017; Lee et al., 2019) aims to answer questions with evidence retrieved from a large-scale corpus. The prevailing solution follows a two-stage framework (Chen et al., 2017), where a *retriever* first retrieves relevant evidences and then a *reader* extracts answers from the evidences. Existing OpenQA systems (Lee et al., 2019; Karpukhin et al., 2020; Mao et al., 2021) have demonstrated great success in retrieving and reading passages. However, most approaches are limited to questions whose answers reside in single modal evidences,

[1] All the code and data will be released upon acceptance.

Question:

What date was the location established where the 1920 Summer Olympics boxing and wrestling events were held?

Retrieved Table:

venues were used in the 1920 Summer Olympics

Venue	Sports	Capacity
Antwerp [4]	Cycling [5] (road)	Not listed
Antwerp Zoo [1]	Boxing [2], Wrestling[2]	Not listed

Retrieved Passages:

[1] *Antwerp Zoo*: Antwerp Zoo is a zoo in the centre of Antwerp, Belgium. It is …, established on 21 July 1843.
[2] *Boxing*: These are the results of the boxing competition at the 1920 Summer Olympics in Antwerp.
[3] *Wrestling*: At the 1920 Summer Olympics, ten wrestling events were contested, for all men. There were five weight classes …
[4] *Antwerp*: … [5] *Cycling*: …

Answer: 21 July 1843

Figure 1: An example of the open question answering over tables and text. Highlighted phrases in the same color indicate evidence pieces related to the question in each single modality. The answer is marked in red.

such as free-form text (Xiong et al., 2021b) or semi-structured tables (Herzig et al., 2021). However, solving many real-world questions requires aggregating heterogeneous knowledge (e.g., tables and passages), because massive amounts of human knowledge are stored in different modalities. As the example shown in Figure 1, the supporting evidence for the given question resides in both the table and related passages. Therefore, retrieving relevant evidence from heterogeneous knowledge resources involving tables and passages is essential for advanced OpenQA, which is also our focus.

There are two major challenges in joint table-and-text retrieval: (1) There exists the discrepancy between table and text, which leads to the difficulty of jointly retrieving heterogeneous knowledge and considering their cross-modality connections; (2) The data sparsity problem is extremely severe because training a joint table-text retriever requires large-scale supervised data to cover all targeted areas, which is labourious and impractical to obtain.

In light of this two challenges, we introduce an optimized **O**penQA **T**able-**TE**xt **R**etriever, dubbed

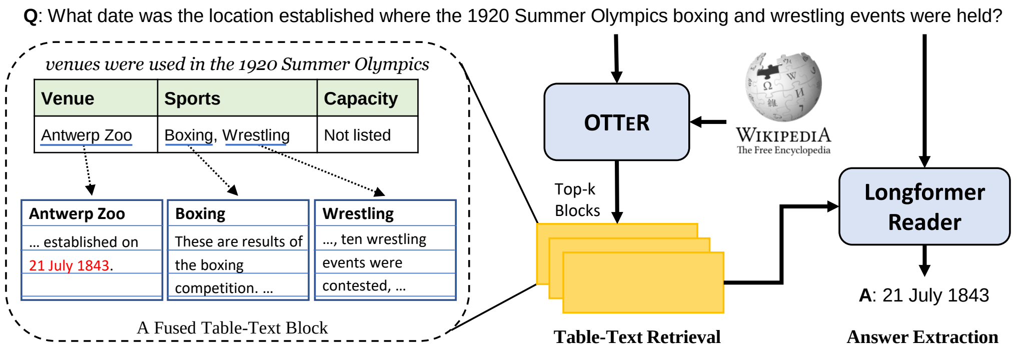

Figure 2: The framework of the overall OpenQA system. It first jointly retrieves top-k table-text blocks with our OTTER. Then it answers the questions from the retrieved evidence with a reader model.

OTTER, which utilizes mixed-modality dense representations to jointly retrieve tables and text. Firstly, to model the interaction between tables and text, we propose to enhance mixed-modality representation learning via two novel mechanisms: modality-enhanced representations (MER) and mixed-modality hard negative sampling (MMHN). MER incorporates fine-grained representations of each modality to enrich the semantics. MMHN utilizes table structures and creates hard negatives by substituting fine-grained key information in two modalities, to encourage better discrimination of relevant evidences. Secondly, to alleviate the data sparsity problem and empower the model with general retrieval ability, we propose a retrieval-centric pre-training task with a large-scale synthesized corpus, which is constructed by automatically synthesizing mixed-modal evidences and reversely generating questions by a BART-based generator.

Our primary contributions are three-fold: (1) We propose three novel mechanisms to improve table-and-text retrieval for OpenQA, namely modality-enhanced representation, mixed-modality hard negative sampling strategy, and mixed-modality synthetic pre-training. (2) Evaluated on OTT-QA, OTTER substantially improves retrieval performance compared with baselines. Extensive experiments and analyses further examine the effectiveness of the above three mechanisms. (3) Equipped with OTTER, our OpenQA system significantly surpasses previous state-of-the-art models with 10.1% absolute improvement in terms of exact match.

2 Background

2.1 Problem Formulation

The task of OpenQA over tables and text is defined as follows. Given two corpus of tables $C_T = \{t_1, ..., t_T\}$ and passages $C_P = \{p_1, ..., p_P\}$, the task aims to answer question q by extracting answer a from the knowledge resources C_P and C_T. The standard system of solving this task involves two components: a *retriever* that first retrieves relevant evidences $c \subset C_T \cup C_P$, and a *reader* to extract a from the retrieved evidence set.

2.2 Table-and-text Retrieval

In this paper, we focus on table-and-text retrieval for OpenQA. To better align the mixed-modality information in table-and-text retrieval, we follow Chen et al. (2020a) and take a table-text block as a basic retrieval unit, which consists of a table segment and relevant passages. Different from retrieving a single table/passage, retrieving table-text blocks could bring more clues for retrievers to utilize since single modal data often contain incomplete context. Figure 2 illustrates table-and-text retrieval and our overall system.

2.2.1 Table-Text Block

Since relevant tables and passages do not necessarily naturally coexist, we need to construct table-text blocks before retrieval. One observation is that tables often hold large quantities of entities and events. Based on this observation and prior work (Chen et al., 2020b), we apply entity linking to group the heterogeneous data. Here we apply BLINK (Ledell et al., 2020) to fuse tables and text, which is an effective entity linker and capable to link against all Wikipedia entities and their corresponding passages. Given a flat table segment, BLINK returns l relevant passages linked to the entities in table. However, as table size and passage quantity grow, the input may become too long for BERT-based encoders (Devlin et al., 2019). Thus, we split a table into several segments that each of them contains only a single row. More details about block constructions and representations can be found in Appendix A.1.

3 Methodology

We present **OTTER**, an **O**penQA **T**able-**T**Ext **R**etriever. We first introduce the basic dual-encoder architecture for dense retrieval (§ 3.1). We then describe three mechanisms to mitigate the table-text discrepancy and data sparsity problems, i.e., modality-enhanced representation (§ 3.2), mixed-modality hard negative sampling (§ 3.3), and mixed-modality synthetic pre-training (§ 3.4).

3.1 The Dual-Encoder Architecture

The prevailing choice for dense retrieval is the dual-encoder method. In this framework, a question q and a table-text block b are separately encoded into two d-dimensional vectors by a neural encoder $E(\cdot)$. Then, the relevance between q and b is measured by dot product over these two vectors:

$$s(q, b) = \mathbf{q}^\top \cdot \mathbf{b} = E(q)^\top \cdot E(b). \qquad (1)$$

The benefit of this method is that all the table-text blocks can be pre-encoded into vectors to support indexed searching during inference time. In this work, we initialize the encoder with a pre-trained RoBERTa (Liu et al., 2019), and take the representation of the first `[CLS]` token as the encoded vector. When an incoming question is encoded, the approximate nearest neighbor search can be leveraged for efficient retrieval (Johnson et al., 2021).

Training The training objective aims to learn representations by maximizing the relevance of the gold table-text block and the question. We follow Karpukhin et al. (2020) to learn the representations. Formally, given a training set of N instances, the i^{th} instance $(q_i, b_i^+, b_{i,1}^-, ..., b_{i,m}^-)$ consists of a positive block b_i^+ and m negative blocks $\{b_{i,j}^-\}_{j=1}^m$, we minimize the cross-entropy loss as follows:

$$L(q_i, b_i^+, \{b_{i,j}^-\}_{j=1}^m) = -\log \frac{e^{s(q_i, b_i^+)}}{e^{s(q_i, b_i^+)} + \sum_{j=1}^m e^{s(q_i, b_{i,j}^-)}}.$$

Negatives are a *hard negative* and $m-1$ *in-batch negatives* from other instances in a mini-batch.

3.2 Modality-enhanced Representation

Most dense retrievers use a coarse-grained single-modal representation from either the representation of the `[CLS]` token or the averaged representations of tokens (Zhan et al., 2020), which is insufficient to represent cross-modal information. To remedy this, we propose to learn modality-enhanced representation (MER) of table-text blocks.

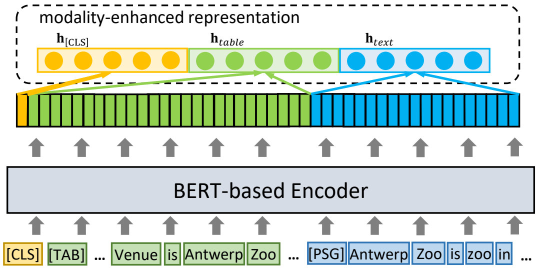

Figure 3: The illustration of modality-enhanced representation in OTTER. Segments in green and blue denote information of tables and passages respectively.

As illustrated in Figure 3, instead of using only the coarse representation $\mathbf{h}_{\texttt{[CLS]}}$ at the `[CLS]` token, MER incorporates tabular and textual representations ($\mathbf{h}_{table}$ and $\mathbf{h}_{text}$) to enhance the semantics of table and text. Thus, the modality-enhanced representation is $\mathbf{b} = [\mathbf{h}_{\texttt{[CLS]}}; \mathbf{h}_{table}; \mathbf{h}_{text}]$, where ; denotes concatenation.

Given the tokens in a tabular/textual modality, we calculate a representation in the following ways: (1) **FIRST**: representations of the beginning token (i.e., `[TAB]` and `[PSG]`); (2) **AVG**: averaged token representations; (3) **MAX**: max pooling over token representations ; (4) **SelfAtt**: weighted average over token representations where weights are computed by a self attention layer. We discuss the impact of different types of MERs in § 5.4. Our best model adopts FIRST as the final setting. To ensure the same vector dimensionality with the enriched representation, we represent the question by replicating the encoded question representation.

3.3 Mixed-modality Hard Negative Sampling

Prior studies (Nogueira and Cho, 2019; Gillick et al., 2019) have found that hard negative sampling is essential in training a dense retriever. These methods take each evidence as a whole and retrieve the most similar irrelevant one as the hard negative. Instead of finding an entire irrelevant block, we propose a *mixed-modality hard negative sampling* mechanism, which constructs more challenging hard negatives by only substituting partial information in the table or text.

Formally, suppose a positive block $b^{j+} = (t^j, p^j)$ is from the j-th row in the table, the answer a resides in either table segment t^j or passages p^j. We decide to replace either the table row or the passage depending on where the answer exists. If a exists in the table row, we construct a hard negative $b^{j-} = (t^k, p^j)$ by replacing t^j with a random row t^k in the same table. Similarly, if a resides in the passages, we create hard negative $b^{j-} = (t^j, p^k)$ by replacing passages with p^k in other blocks.

3.4 Mixed-modality Synthetic Pre-training

To alleviate the issue of data sparsity, we propose a mixed-modality synthetic pre-training (MMSP) task. MMSP enhances the retrieval ability by pre-training on a large-scale synthesized corpus, which involves mixed-modality pseudo training data with *(question, table-text block)* pairs. Here, we introduce a novel way to construct the pseudo training corpus in two steps, including table-text block mining and question back generation.

(1) Mine relevant table-text pairs. One observation is that Wikipedia hyperlinks often link explanatory passages to entities in tables, which provides high-quality relevant table-text pairs. Based on this, we believe Wikipedia is an excellent resource for parsing table-text pairs. Specially, we select a row in a table, and find corresponding passages with the hyperlinks to form a fused table-text block. We only keep the first section in each Wikipedia page as it always contains the most important information about the linked entity. (2) Write pseudo questions for fused blocks. The questions are expected to not only contain the mixed-modality information from the blocks, but also have good fluency and naturalness. Therefore, instead of using template-based synthesizing methods, we use a generation-based method to derive more fluent and diverse questions, which is called *back-generation*. Specially, we use BART$_{base}$ (Lewis et al., 2019) as the backbone of our generator, which is fine-tuned with oracle pairs of *(question, table-text block)* in the OTT-QA training set. The input to the generator is a sequence of the flat table and linked passages, and the output is a mixed-modality question. Finally, we automatically construct a large-scale pre-training corpus. We present some examples of generated pseudo questions in Appendix A.2.

During pre-training, we adopt a similar ranking task where the training objective is the same as described in § 3.1. As for negative sampling, we use in-batch negatives and one hard negative randomly sampled from the same table. Finally, we obtain a synthesized corpus with 3M pairs of table-text blocks and pseudo questions.

4 Experiment Settings

In this section, we describe the experiment settings on the task of open-domain question answering over tables and text, and report the performance of our system on the table-and-text retrieval, and downstream question answering.

4.1 Dataset

Our system is evaluated on the **OTT-QA** dataset (Chen et al., 2020a), which is a large-scale open-domain table-text question answering benchmark. Answering questions in OTT-QA requires aggregating multi-modal information from both tables and text. OTT-QA dataset contains over 40K questions with human annotated answers and ground truth evidences. It also provides a corpus of over 400K tables and 6M passages collected from Wikipedia. Data statistics of OTT-QA dataset and table-text corpus are shown in Table 1.

# questions (train/dev/test)	41,469 / 2,214 / 2,158
# of tables in the corpus	410,740
# of passages in the corpus	6,342,314
# of fused table-text blocks in the corpus	5,409,903
Average tokens in fused blocks	357.5
Average fused table-text blocks in each table	12.9

Table 1: Statistics of OTT-QA and table-text corpus.

4.2 Evaluation Metrics

A well-recognized metric for information retrieval is the recall at top k ranks (Recall@k), which is the proportion of relevant items found in the top-k returned items. In this paper, we use two metrics to evaluate the retrieval system: one is table recall and the other is table-text block recall. Table recall indicates whether the top-k retrieved blocks come from the ground-truth table. However, in table-and-text retrieval, table recall is imperfect as an coarse-grained metric since our basic retrieval unit is a table-text block corresponding to a specific row in the table. Therefore we propose a more fine-grained and challenging metric: table-text block recall at top k ranks, where a fused block is considered as a correct match when it meets two requirements. Firstly, it comes from the ground truth table. Second, it contains the correct answer. On the downstream QA task, we report the exact match (EM) and F1 score (Chen et al., 2020a) to evaluate OpenQA system.

5 Experiments: Table-and-Text Retrieval

In this section, we evaluate the retrieval performance of our OpenQA Table-Text Retriever

(OTTeR). We first compare OTTeR with previous retrieval approaches on OTT-QA. Then we conduct extensive experiments to examine the effectiveness of the three proposed mechanisms.

5.1 Baseline Methods

We compare with the following retrievers. (1) **BM25** (Chen et al., 2020a) is a sparse method to retrieve tabular evidence, where the flat table with metadata (i.e., table title and section title) and content are used for retrieval. (2) **Bi-Encoder** (Kosti'c et al., 2021) is a dense retriever which uses a BERT encoder for questions, and a shared BERT encoder to separately encode tables and text as representations for retrieval. (3) **Tri-Encoder** (Kosti'c et al., 2021) is a dense retriever that uses three individual BERT encoders to separately encode questions, tables and text as representations. (4) **Iterative Retriever** (Chen et al., 2020a) is a dense retriever which iteratively retrieves tables and passages in 3 steps. (5) **Fusion Retriever** (Chen et al., 2020a) is the only existing dense method to retrieve table-text block, which uses a GPT2 (Radford et al., 2019) to link passages and the Inverse Cloze Task (Lee et al., 2019) to pre-train the encoder. We also report results of OTTeR-baseline (removing three proposed strategies) and OTTeR w/o text (removing textual passages during retrieval).

5.2 Implementation Details

We use RoBERTa-base (Liu et al., 2019) as the backbone of our retrievers with a maximum input length of 512 tokens per table-text block and 70 tokens per question. The retrievers are trained using the in-batch negative and one additional hard negative setting for both pre-training and fine-tuning. On the pre-training stage, we pre-train on the synthesized corpus for 5 epochs on 8 Nvidia Tesla V100 32GB GPUs with a batch size of 168. We use AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3e-5, linear scheduling with 5% warm-up. On the fine-tuning stage, we train the retrievers for 20 epochs with a batch size of 64, learning rate of 2e-5 and warm-up ratio of 10 % for all encoders on 8 Nvidia Tesla V100 16GB GPUs.

5.3 Main Results

Table 2 compares different retrievers on OTT-QA dev. set, using the table recall at top k ranks ($k \in \{1, 10, 20, 50, 100\}$) because the results from other papers are mainly reported in table recall. We find that: (1) OTTeR significantly outperforms previous sparse and dense retrievers and the gap is especially large when k is smaller (e.g., 8.2% absolute gain for R@10), which demonstrates the effectiveness of OTTeR; (2) When textual passages are removed during retrieval (OTTeR w/o text), the performance of OTTeR drops dramatically, especially when k is smaller. This phenomenon shows the importance of taking textual information as a complement to tables.

Models	R@1	R@10	R@20	R@50	R@100	Hit@4K
BM25	41.0	68.5	73.7	80.4	-	-
Bi-Encoder	-	72.9	78.0	-	89.4	-
Tri-Encoder	-	73.8	79.7	-	90.1	-
Iterative Retriever	-	-	-	-	-	27.2
Fusion Retriever	-	-	-	-	-	52.4
OTTeR-baseline	46.3	69.4	74.4	80.1	83.9	54.6
OTTeR w/o text	48.7	73.9	79.5	85.8	88.8	34.6
OTTeR	**58.5**	**82.0**	**86.3**	**90.6**	**92.8**	**66.4**

Table 2: Overall retrieval results on OTT-QA dev set. Hit@4K (Chen et al., 2020a) is used to measure whether the answer exists in the retrieved 4096 subword tokens.

5.4 Ablation Study

To examine the effectiveness of the three mechanisms in OTTeR, we conduct extensive ablation studies on OTT-QA and discuss our findings below.

Effect of Modality-enhanced Representation

In this experiment, we explore the effect of modality-enhanced representations (MER) on retrieval performance. Table 3 reports the table recall and block recall of our models with different MER strategies on the OTT-QA dev. set. We also report the result after eliminating MER, i.e., using only the representation of the `[CLS]` token for ranking. We find that integrating modality-enhanced representations improves the retrieval performance significantly. As MER incorporates single-modal representations to enrich the mixed-modal representation, retrievers can easily capture the comprehensive semantics of table-text blocks. In addition, among all the strategies for MER, the FIRST strategy using the representation of the beginning special token of each modality achieves the best performance. This observation verifies the stronger representative ability of the FIRST strategy compared with other pooling strategies.

Effect of Mixed-modality Negative Sampling

To investigate the effectiveness of hard negative sampling on retrieval performance, we evaluate our system under following settings of hard negative sampling on the OTT-QA development set: (1) Mixed-modality hard negative (MMHN) described

Models	Table Recall			Block Recall		
	R@1	R@10	R@100	R@1	R@10	R@100
OTTeR						
MER=FIRST	**58.5**	**82.0**	**92.8**	**30.9**	**66.4**	**87.0**
MER=AVG	57.1	81.2	92.5	29.8	65.3	85.9
MER=MAX	56.7	81.4	92.2	29.0	65.1	86.4
MER=SelfAtt	57.9	81.2	92.6	29.5	65.3	86.0
w/o MER	50.0	76.8	89.9	22.7	55.2	79.3

Table 3: Retrieval performance of OTTeR under different modality-enhanced representations (MER) settings.

Models	Table Recall			Block Recall		
	R@1	R@10	R@100	R@1	R@10	R@100
OTTeR						
PT=BartQ	**58.5**	**82.0**	**92.8**	**30.9**	**66.4**	**87.0**
PT=TitleQ	56.6	79.3	91.8	23.1	60.0	83.1
DA w/o PT	39.3	68.9	73.0	14.8	45.9	74.5
w/o PT	53.1	77.8	91.2	20.5	57.2	81.3

Table 5: Retrieval performance of OTTeR under different settings. PT denotes pre-training.

in § 3.3; (2) BM25: the most similar irrelevant table-text block searched by BM25; (3) Random: a random table-text block in the same table containing no answer.

From the results shown in Table 4, we can observe that training the retriever with MMHN yields the best performance compared with other hard negative sampling strategies. Since mixed-modality hard negatives is constructed by only replacing partial information from the positive block, it is more challenging and it enables the retriever to better distinguish important information in the evidence.

Models	Table Recall			Block Recall		
	R@1	R@10	R@100	R@1	R@10	R@100
OTTeR						
HN=MMHN	**58.5**	**82.0**	**92.8**	**30.9**	**66.4**	**87.0**
HN=BM25	51.4	79.8	92.4	25.8	58.2	81.9
HN=Random	50.3	79.0	92.7	28.4	58.7	80.1

Table 4: Retrieval performance of OTTeR under different hard negative sampling settings. MMHN denotes mixed-modality hard negatives.

Effect of Mixed-modality Synthetic Pre-training We investigate the effectiveness of mixed-modality synthetic pre-training. We first pre-train the retriever and then fine-tune the retriever with OTT-QA training set. The pre-training corpus consisting of 3 millions of *(question, evidence)* pairs, with questions synthesized in the following ways: (1) **BartQ**: the questions are generated by BART as described in § 3.4; (2) **TitleQ**: the questions are constructed from passage titles and table titles. (3) **DA w/o PT**: data augmentation without pre-training, where we integrate the BART synthetic corpus with the oracle data together for fine-tuning. (4)**w/o PT** direct fine-tuning without pre-training.

The retrieval results on the dev. set of OTT-QA are exhibited in Table 5. We can find that: (1) Pre-training brings substantial performance gain to dense retrieval, showing the benefits of automatically synthesizing large-scale pre-training corpus to improve retrievers. (2) synthesizing questions using BART-based generator performs better than using template-based method (TitleQ). We attribute it to more fluent and diverse questions synthesized by generation-based method. (3) Using the synthesized corpus for data augmentation performs much poorer than using it for pre-training, and even worse than directly fine-tuning without pre-training. One explanation is that pre-training targets to help the model in learning a more general retrieving ability beforehand, while fine-tuning aims to learns a more specific and accurate retriever. As the synthesized corpus is more noisy, using it as augmented fine-tuning data may make the training unstable and lead to a performance drop. This observation again verifies the effectiveness of pre-training with mixed-modality synthetic corpus.

5.5 Case Study

Here, we give an example of retrieved evidences to show that OTTeR correctly represents questions and blocks with the proposed three strategies.

As shown in Figure 4, to answer the question, the model should find relevant table-text blocks with two pieces of evidences distributed in tables and passages, including the *"skier who won 6 gold medals at the FIS Nordic Junior World Ski Championships"* and the *"year when the skier started competing"*. As we can see, OTTeR successfully returns a correct table-text block at rank 1, which includes all necessary information. The top-2 retrieved block by OTTeR is also reasonable, since partial evidences like *6 gold medals* and *Ski Championships* are matched. However, OTTeR-baseline (w/o three mechanisms) returns an unsatisfactory block. Though the retriever finds the *Ski Championships* , which is a strong signal to locate the table, it fails to capture fine-grained information like *6 gold medals* and *starting year.*

This case demonstrates that OTTeR can capture the more accurate meanings of fused table-text block, especially when the supported information resides separately. It shows that enhancing cross-modal representations with proposed mechanisms is beneficial to modeling heterogeneous data.

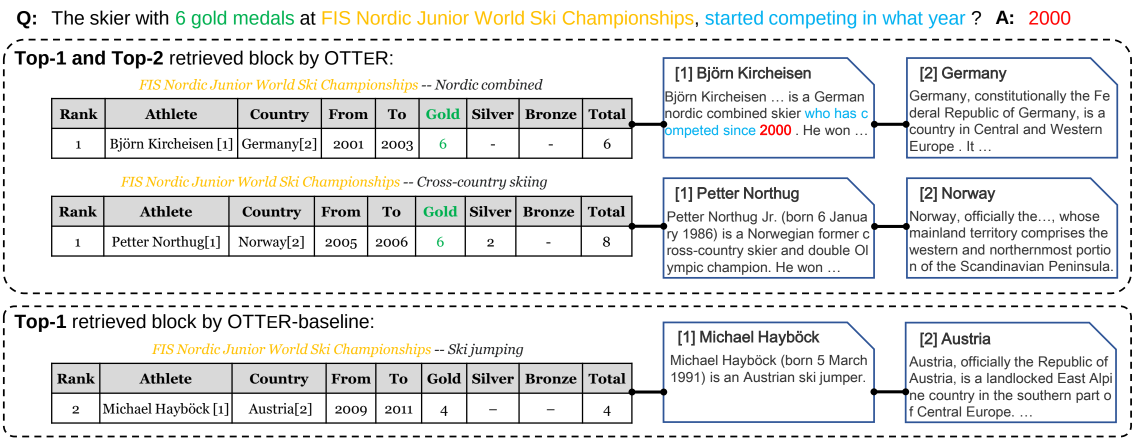

Figure 4: Examples of table-text blocks returned by full OTTER and OTTER without modality-enhanced representations. Words in the retrieved blocks of the same color denote the evidences corresponding to questions.

6 Experiments: Question Answering

In this section, we experiment to show how OTTER affects the downstream QA performance.

6.1 Reader

We implement a two-stage open-domain question answering system, which is equipped with our OTTER as the *retriever* and a *reader* model for extracting the answer from the retrieved evidence. As we mainly focus on improving the retriever in this paper, we use the state-of-the-art reader model to evaluate the downstream QA performance.

Following Chen et al. (2020a), we use the *Cross Block Reader* (CBR) to extract the answer. The CBR jointly reads the concatenated top-k retrieved table-text blocks and outputs a best answer span from these blocks. In contrast to *Single Block Readers* (SBR) that read only one block at a time, CBR is more powerful in utilizing the cross-attention mechanism to model the cross-block dependencies. Here we take the pre-trained Long-Document Transformer (Longformer) (Beltagy et al., 2020) as the backbone of CBR, which applies sparse attention mechanism and accepts longer input sequence of up to 4,096 tokens. For fair comparison with Chen et al. (2020a), we feed top-15 retrieved blocks into the reader model for inference. To balance the distribution of training data and inference data, we also takes k table-text blocks for training, which contains several ground-truth blocks and the rest of retrieved blocks. The training objective is to maximize the marginal log-likelihood of all the correct answer spans in the positive block. The reader is trained with 8 Nvidia V100 GPUs for 5 epochs, using the batch size of 16 and learning rate of 1e-5.

		Dev		Test	
Retriever	Reader	EM	F1	EM	F1
BM25	HYBRIDER (Chen et al., 2020b)	10.3	13.0	9.7	12.8
BM25	DUREPA (Li et al., 2021)	15.8	-	-	-
Iterative Retriever	SBR (Chen et al., 2020a)	7.9	11.1	9.6	13.1
Fusion Retriever	SBR (Chen et al., 2020a)	13.8	17.2	13.4	16.9
Iterative Retriever	CBR (Chen et al., 2020a)	14.4	18.5	16.9	20.9
Fusion Retriever	CBR (Chen et al., 2020a)	28.1	32.5	27.2	31.5
OTTER (ours)	CBR	**37.1**	**42.8**	**37.3**	**43.1**

Table 6: QA Results on the dev. set and blind test set.

6.2 Results

The results are shown in Table 6. We find that OTTER+CBR significantly outperforms existing OpenQA systems, with 10.1% performance gain in terms of EM over the prior state-of-the-art system. The results demonstrate that our approach can retrieve better supported evidences to the question, which leads to further improvement on the downstream QA performance.

To further analyze the effect of different components of OTTER on QA performance, we conduct an ablation study on OTT-QA after eliminating different components. As shown in Figure 5, the OpenQA system with full OTTER achieves the best performance, and removing each component leads to a substantial performance drop. This observation verifies the effectiveness of our proposed three mechanisms, i.e., modality-enhanced representations (MER), mixed-modality hard negatives (MMHN) and mixed-modality synthetic pretraining. We also evaluate the impact of taking different numbers of retrieved blocks as the inputs for inference. As shown in Figure 5, the EM score increases rapidly with k when $k < 20$ but the growth slows down when $k > 20$, which can help to find a better tradeoff between efficiency and performance.

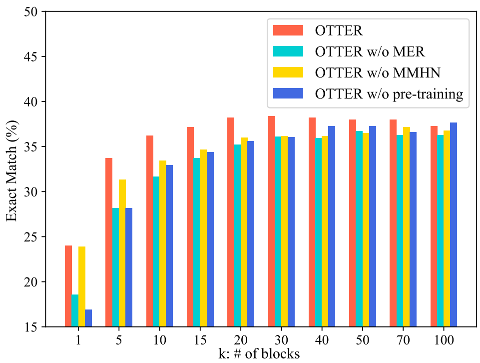

Figure 5: QA Performance on the OTT-QA dev. set with different number of table-text blocks as input.

7 Related Works

In OpenQA (Chen et al., 2017; Joshi et al., 2017; Dunn et al., 2017; Lee et al., 2019), the retriever is an essential component to identify relevant evidences for answer extraction. In contrast to sparse information retrieval methods (Wang et al., 2018; Nogueira and Cho, 2019; Yang et al., 2019), recent OpenQA systems tend to adopt dense retrieval approaches utilizing dense representations learned by pre-trained language models (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020). These methods are powerful in capturing contextual semantics.

The prevailing OpenQA datasets mainly take the unstructured passage as evidence, including Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (Berant et al., 2013), CuratedTREC (Baudis and Sedivý, 2015) and SQuAD (Rajpurkar et al., 2016). Recently, Herzig et al. (2021) study OpenQA in the tabular domain. Chen et al. (2020a) consider a more challenging setting that takes both tabular corpus and textual corpus as the knowledge sources, which is also the setting in this paper.

Our approach differs from existing methods mainly in two aspects: targeted evidence source and mixed-modality learning mechanisms. First of all, we retrieve mixed-modality evidence from both tabular and textual corpus, which is different from text-based retrievers (Karpukhin et al., 2020; Asai et al., 2020; Xiong et al., 2021b) and table-based retrievers (Chen et al., 2020c; Shraga et al., 2020; Pan et al., 2021a). Secondly, our proposed three mixed-modality learning mechanisms also differ from existing methods. As for mixed-modality representation, previous work (Karpukhin et al., 2020) mainly uses the single representation of the special token for ranking. Our method incorporates single modal representation to enrich the mixed modal representation. As for mixed-modality negative sampling, instead of finding an entire negative evidence with either sparse or dense methods (Yang et al., 2021; Luan et al., 2021; Lu et al., 2020; Xiong et al., 2021a; Lu et al., 2021; Zhan et al., 2021), we construct more challenging hard negative by only replacing partial single-modality information at once. As for mixed-modality synthetic pre-training, our pre-training strategy is different in the pre-training task, knowledge source and the method of synthesizing pseudo question. There are also works investigating joint pre-training over tables and text (Herzig et al., 2020; Eisenschlos et al., 2020; Yin et al., 2020; Oğuz et al., 2020). However, these methods mainly take the table metadata as the source of text and do not consider the retrieval task. Instead, we use linked passages as a more reliable knowledge source, and target on retrieval-based pre-training. There are some attempts on incorporating pre-training task to improve retrieval performance (Chang et al., 2020; Sachan et al., 2021; Ouguz et al., 2021), which target on textual-domain retrieval or using template-based method for query construction. Differently, our approach focuses on a more challenging setting that retrieves evidence from tabular and textual corpus and adopts a generation-based query synthetic method. Besides, Pan et al. (2021b) explore to generate multi-hop questions for tables and text, but they focus on an unsupervised manner.

8 Conclusion

In this paper, we propose an optimized dense retriever called OTTER, to retrieve joint table-text evidences for OpenQA. OTTER involves three novel mechanisms to address table-text discrepancy and data sparsity challenges, i.e., modality-enhanced representations, mixed-modality hard negative sampling, and mixed-modality synthetic pre-training. We experiment on OTT-QA dataset and evaluate on two subtasks, including retrieval and QA. Results show that OTTER significantly outperforms other retrieval methods by a large margin, which further leads to a substantial absolute performance gain of 10.1% EM on the downstream QA. Extensive experiments illustrate the effectiveness of all three mechanisms in improving retrieval and QA performance. Further analyses also show the ability of OTTER in retrieving more relevant evidences from heterogeneous knowledge resources.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, R. Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *ICLR*.

P. Baudis and J. Sedivý. 2015. Modeling of the question answering task in the yodaqa system. In *CLEF*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jonathan Berant, Andrew K. Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Wei-Cheng Chang, F. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *ICLR*.

Danqi Chen, Adam Fisch, J. Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *ACL*.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020a. Open question answering over tables and text. In *ICLR*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of EMNLP*.

Zhiyu Chen, Mohamed Trabelsi, J. Heflin, Yinan Xu, and Brian D. Davison. 2020c. Table search using a deep contextualized language model. In *SIGIR*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Julian Martin Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of EMNLP*.

D. Gillick, Sayali Kulkarni, L. Lansing, A. Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *CoNLL*.

Kelvin Guu, Kenton Lee, Z. Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *NAACL*.

Jonathan Herzig, Pawel Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *ACL*.

Jeff Johnson, M. Douze, and H. Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

Bogdan Kosti'c, Julian Risch, and Timo Moller. 2021. Multi-modal retrieval of tables and texts using tri-encoder models. In *MRQA*.

T. Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, D. Epstein, Illia Polosukhin, J. Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Wu Ledell, Petroni Fabio, Josifoski Martin, Riedel Sebastian, and Zettlemoyer Luke. 2020. Zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Alexander Hanbo Li, Patrick Ng, Peng Xu, Henghui Zhu, Zhiguo Wang, and Bing Xiang. 2021. Dual reader-parser on hybrid textual and tabular evidence for open domain question answering. In *ACL/IJCNLP*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

J. Lu, G. Ábrego, Ji Ma, Jianmo Ni, and Yinfei Yang. 2020. Neural passage retrieval with improved negative contrast. *ArXiv*, abs/2010.12523.

Shuqi Lu, Chenyan Xiong, Di He, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul N. Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pre-training a strong siamese encoder using a weak decoder. In *EMNLP*.

Y. Luan, Jacob Eisenstein, Kristina Toutanova, and M. Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *ACL/IJCNLP*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *ArXiv*, abs/1901.04085.

Barlas Ouguz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen tau Yih, Sonal Gupta, and Yashar Mehdad. 2021. Domain-matched pre-training tasks for dense retrieval. *ArXiv*, abs/2107.13602.

Barlas Oğuz, Xilun Chen, Vladimir Karpukhin, Stanislav Peshterliev, Dmytro Okhonko, M. Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2020. Unified open-domain question answering with structured and unstructured knowledge. *ArXiv*, abs/2012.14610.

Feifei Pan, Mustafa Canim, Michael R. Glass, A. Gliozzo, and P. Fox. 2021a. Cltr: An end-to-end, transformer-based system for cell level table retrieval and table question answering. In *ACL/IJCNLP*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021b. Unsupervised multi-hop question answering by question generation. In *NAACL*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Devendra Singh Sachan, M. Patwary, M. Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *ACL/IJCNLP*.

Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Canim. 2020. Web table retrieval using multimodal deep learning. In *SIGIR*.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, G. Tesauro, Bowen Zhou, and Jing Jiang. 2018. R3: Reinforced ranker-reader for open-domain question answering. In *AAAI*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021a. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Ouguz. 2021b. Answering complex open-domain questions with multi-hop dense retrieval. In *ICLR*.

Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xmoco: Cross momentum contrastive learning for open-domain question answering. In *ACL/IJCNLP*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy J. Lin. 2019. End-to-end open-domain question answering with bertserini. In *NAACL*.

Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *ACL*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *SIGIR*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *ArXiv*, abs/2006.15498.

A Method Details

A.1 Table-Text Block Representation

The table-text block representation is illustrated in Figure 6. Following Chen et al. (2020a), we involve the title and section title of a table and prefix them to the table cell. We also flatten the column name and column value with an "is " token to obtain more natural and fluent utterance. In addition, we add different special tokens to separate different segments, including `[TAB]` for table segment, `[PSG]` for passage segment, `[TITLE]` for table title, `[SECTITLE]` for section title, `[DATA]` for table content, and `[SEP]` to separate different passages. Such a flattened block will be used throughout this paper as the input string to the retriever and the reader.

In OTT-QA dataset, long rows frequently appear in tables, which leads to more entities and passages in a single table-text block. To maintain more relevant information in a block, we rank the passages with the TF-IDF score to table schema and table content. Then we remove the tokens when a flattened block is out of the input length limit of the RoBERTa tokenizer.

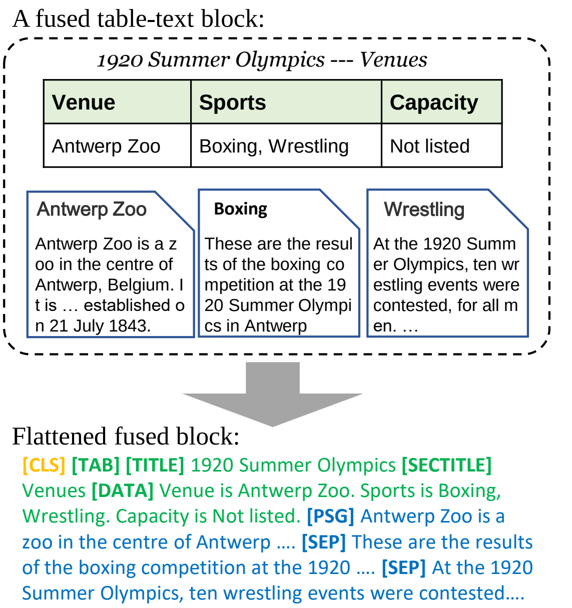

Figure 6: The flattened fused block representation of the each table-text block.

A.2 Examples of Synthesized Corpus

To provide a better understanding of mixed-modality synthetics pre-training, we give some examples of pseudo training data with *(question, table-text block)* pairs in Table 7. As we can see, the generated questions not only are fluent and natural, but also consider mixed-modality information from tables and passages.

B Performance Analysis

B.1 Top-k Retrieval Results

Here, we show the detailed retrieval results of OTTER with different components in Figure 7. The table recall at top-k ranks and block recall at top-k ranks are reported. We can find that full OTTER substantially surpasses the models of other settings in block recall, and in table recall when $k \leq 50$.

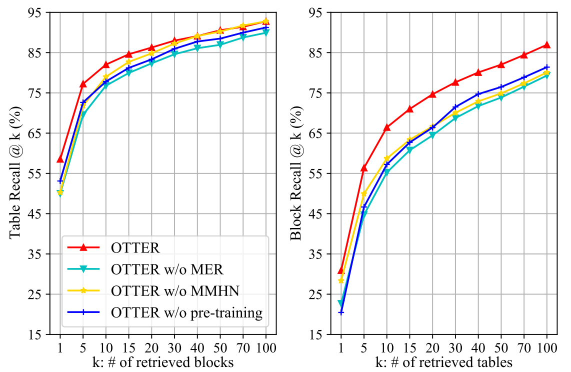

Figure 7: Top-k retrieval performance of retrievers on the dev set of OTT-QA. Full OTTER substantially surpasses the other models in block recall, and in table recall when $k \leq 50$.

B.2 Entity Linking

To understand the effects of entity linking, we evaluate the standalone entity linking accuracy and the retrieval performance. We consider the following linking models: (1) GPT-2 used in Chen et al. (2020a), which first augments the cell value by the context with a GPT-2 (Radford et al., 2019) and then uses BM25 to rank the blocks to the augmented form, (2) BLINK (Ledell et al., 2020) used in OTTER, which applys a bi-encoder ranker and cross-encoder re-ranker to link Wikipedia passages to the entities in flattened tables, (3) Oracle linker, which uses the original linking passages in the table.

We evaluate the entity linking of on the OTTQA dev. set following the settings in Chen et al. (2020a) and report the table-segment-wise F1 score. Table 8 shows the performance. We find that the F1 score of BLINK is higher than GPT-2, which leads to more relevant passages for tables.

Flattened Table Segment	Generated Question
[TAB] [TITLE] J1 League [SECTITLE] History – Timeline [DATA] Year is 2003. Important events is Extra time. # J clubs is 16. Rele . slots is 2. [PSG] The 2003 season was the 11th season since the establishment of the J.League . The league began on March 15 and ended on November 29 .	What is the number of slots in the J1 League for the season that began on March 15 and ended on November 29 ?
[TAB] [TITLE] 2010 President's Cup (tennis) [SECTITLE] ATP entrants – Seeds [DATA] Nationality is KAZ. Player is Mikhail Kukushkin. Ranking is 88. Seeding is 1. [PSG] Mikhail Aleksandrovich Kukushkin (; born 26 December 1987) is a Kazakh professional tennis player of Russian descent .	What is the nationality of the player in the 2010 President 's Cup who was born on 26 December 1987 ?
[TAB] [TITLE] Washington House of Representatives [SECTITLE] Composition – Members (2019-2021 , 66th Legislature) [DATA] District is 7. Position is 1. Representative is Jacquelin Maycumber. Party is Republican. Residence is Republic. Counties Represented is Ferry Okanogan Pend Oreille Spokane Stevens. First elected is 2017. [PSG]	What is the residence of the Washington House of Representatives representative who was first elected in 2017.
[TAB] [TITLE] 1961 NFL expansion draft [SECTITLE] Player selections [DATA] Player is Don Joyce. Position is Defensive end. College Team is Tulane. Original NFL Team is Baltimore Colts. [PSG] Don Joyce (October 8 , 1929 - February 26 , 2012) was an American football defensive end and professional wrestler.	What was the original NFL team of the 1961 NFL expansion draft player who died on February 26 , 2012 ?
[TAB] [TITLE] 2009 Formula One World Championship [SECTITLE] Results and standings – Grands Prix [DATA] Round is 2. Grand Prix is Malaysian Grand Prix. Pole position is Jenson Button. Fastest lap is Jenson Button. Winning driver is Jenson Button. Winning constructor is Brawn Mercedes. Report is Report. [PSG] The Malaysian Grand Prix was an annual auto race held in Malaysia . It was part of the Formula One World Championship from 1999 to 2017 and it was held during these years at the Sepang International Circuit . The first Malaysian Grand Prix was held in 1962 in what is now Singapore .	What is the name of the constructor that won the 2009 Formula One World Championship round that was held at the Sepang International Circuit ?

Table 7: Examples of synthesized corpus for pre-training. The queries are generated by a fine-tuned BART generotor given the input of flattened table segment. The generated questions not only are fluent and natural, but also consider mixed-modality information from tables and passages.

	Linking	Table Recall			Block Recall		
Linker	F1	R@1	R@10	R@100	R@1	R@10	R@100
GPT2	50.4	58.2	81.5	92.5	28.6	64.0	83.7
BLINK	55.9	58.5	82.0	92.8	30.9	66.4	87.0
Oracle	100	60.5	83.5	93.9	35.3	71.5	88.5

Table 8: Entity linking and retrieval results of different linkers.

We further evaluate the retrieval performance with table-text corpus constructed by different entity linkers. Comparing GPT-2 and BLINK, we can find that the retrieval performance improves with the increased linking F1, especially when evaluated in block recall. The result indicates the importance of sufficient context information.

B.3 Embedding Dimension

To maximumly eliminate the impact of embedding dimension in modality-enhanced representation (MER), we add a new ablation by concatenating three [CLS] vectors as block representations, (i.e., $\mathbf{b} = [\mathbf{h}_{\texttt{[CLS]}}; \mathbf{h}_{\texttt{[CLS]}}; \mathbf{h}_{\texttt{[CLS]}}]$), and training in the same way as MER=First (i.e., $\mathbf{b} = [\mathbf{h}_{\texttt{[CLS]}}; \mathbf{h}_{\texttt{[TAB]}}; \mathbf{h}_{\texttt{[PSG]}}]$). The results in Table 9 show that using specific representations of each modality still brings more sufficient information than [CLS] after maximumly eliminating the dimension bias.

	Table Recall			Block Recall		
Model	R@1	R@10	R@100	R@1	R@10	R@100
MER=First	58.5	82.0	92.8	30.9	66.4	87.0
CLS	57.5	80.4	92.5	29.6	64.0	86.5

Table 9: Ablation results on retrieval of MER dimension.