DRAMA: Dynamic Multi-Granularity Graph Estimate Retrieval over Tabular and Textual Question Answering

Ruize Yuan[1,2,3], **Xiang Ao**[1,2,3,4,*], **Li Zeng**[5] , **Qing He**[1,2,3]
[1] Key Laboratory of AI Safety of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China.
[2] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China.
[3] University of Chinese Academy of Sciences, Beijing 100049, China.
[4] Institute of Intelligent Computing Technology, Suzhou, CAS.
[5] Shenzhen Stock Exchange, Shenzhen 518038, China.
{yuanruize21s, aoxiang, heqing}@ict.ac.cn
lzeng@szse.cn

Abstract

The TableTextQA task requires finding the answer to the question from a combination of tabular and textual data, which has been gaining increasing attention. The row-based approaches have demonstrated remarkable effectiveness. However, they suffer from the following limitations: (1) a lack of interaction between rows; (2) excessively long input lengths; and (3) question attention shifts in the multi-hop QA task. To this end, we propose a novel method: Dynamic Multi-Granularity Graph Estimate Retrieval - **DRAMA**. Our method incorporates an interaction mechanism among multiple rows. Specifically, we utilize a memory bank to store the features of each row, thereby facilitating the construction of a heterogeneous graph with multi-row information. Besides, a Dynamic Graph Attention Network (DGAT) module is engaged to gauge the attention shift in the multi-hop question and eliminate the noise information dynamically. Empirical results on the widely used HybridQA and TabFact datasets demonstrate that the proposed model is effective.

Keywords: TableTextQA, Structural Data Understanding, Graph Neural Network

1. Introduction

The TableTextQA task, which requires finding the answer to the question from a combination of tabular and textual data, has been gaining increasing attention. Each data type offers unique benefits: tables are an exceptional tool for comparing statistical data, while textual information continues to be a cornerstone in daily communication.

Considering complex questions in real-world applications require the combination of data from multiple resources, TableTextQA tasks are burgeoning in domains such as finance, science, and medicine (Chen et al., 2020b, 2021; Ulmer et al., 2020; Zhu et al., 2021; Talmor et al., 2021).

Prevailing methods for addressing TableTextQA tasks can generally be classified into three categories. Knowledge-based methods: These methods enhance the overall understanding of data by incorporating additional information about entities through knowledge injection (Liu et al., 2020; Karpukhin et al., 2020). Pre-training methods: By enhancing Pretrained Language Models (PLMs) or designing new pre-training tasks specifically for tables, these methods enhance the understanding of structural data (Eisenschlos et al., 2021; Pi et al., 2022). Lightweight local retrieval methods: These methods conduct searches based on specific parts of the table content, such as row-based (Kumar et al., 2021; Lee et al., 2023; Huang et al., 2022) or cell-based approaches (Zhong et al., 2022).

Notably, the row-based approach, which employs a specific framework for managing data from hybrid tabular and textual sources, has exhibited remarkable effectiveness.

Despite their success, existing solutions suffer from the following limitations: Firstly, the lack of interaction between rows precludes a direct comparison of information across different rows in the table; secondly, the incorporation of information from multiple rows can lead to excessively long input lengths, posing an additional challenge; thirdly, in multi-hop question-answering tasks where question attention spans multiple steps, static analyses run the risk of overlooking part of the target information.

To address these problems, we propose a novel method: Dynamic Multi-Granularity Graph Estimate Retrieval - **DRAMA** over tabular and textual question answering. Based on the (Kumar et al., 2021) framework, we incorporate an interaction mechanism among multiple rows to address commonly encountered comparison-type questions. Specifically, we first utilize a memory bank to store the feature vectors of each row. During the evidence retrieval process, we employ these stored vectors to construct a heterogeneous

[*] Corresponding author.

LREC-COLING 2024, pages 5365–5375
20-25 May, 2024.

graph, supplementing the comparative information of other rows often missed in row-based methods. Between each layer of the graph model, the Dynamic Graph Attention Network (DGAT) module is engaged to gauge the attention shift in multi-hop questions dynamically. Empirical results on the widely used HybridQA dataset demonstrate that the proposed model is effective, achieving state-of-the-art performance.

Accordingly, our main contributions are summarized as follows:

- We propose a multi-granularity retrieval architecture based on heterogeneous graphs, enhancing the comprehension of hybrid tabular and textual data.
- With the momentum mechanism, we are able to process the long input, addressing the multi-row table retrieval issue while enhancing the cross-row retrieval capability.
- We propose the Dynamic Graph Attention Network (DGAT) module to estimate attention shifts in the complex question, effectively eliminating the noise and unrelated information introduced during multi-row retrieval.

2. Preliminary

2.1. Task Definition

Question answering over hybrid tabular and textual data requires extracting the answer from heterogeneous information. The problem's input includes a question q, a table t, and linked passages P. Specifically, the cells in the ith row, denoted as row_i, may link to a subset of referring passages $\{p_{i1}, p_{i2}, ...\} \subset P$. The table is comprised of m rows and n columns of cells denoted as $\{c_{ij}\}_{i=1}^{m}{}_{j=1}^{n}$, along with a $header$. Each $Col_j \in header$ defines the description for the cells within that respective column. Given a triplet of a question, a table, and several linked passages (q, t, P), the goal is to retrieve the answer A corresponding to the question.

2.2. Basic Framework

Answering based on the entire table may seem like a straightforward approach, but it's not practical for many tables on Wikipedia, which have a large number of rows and columns. This approach results in long encoded token lengths, including huge noise that cannot simply fit into the input length of pre-trained language models (PLMs).

However, concatenating all the linked passages with the table cells can result in a high calculation complexity when processing over-long sequences with existing language models. Additionally, as the table size increases, the model's scalability decreases, making it more challenging to handle large-scale tables.

Prevailing row-based methods adopt a two-stage architecture that consists of a retrieval component and a passage reader.

In the retrieval stage, constrained by the excessive number of tokens contained in the full table, retrieval utilizes partial table information as the basic unit for searching (Li et al., 2021; Kumar et al., 2021; Lee et al., 2023). Heterogeneous data is typically encoded on the row level. Each row serves as the smallest unit to preserve the complete structured attribute information of the table. This method allows for the preservation of structured information within the table while simultaneously controlling the length of the input sequence.

Moreover, the Pretrained Language Model (PLM) is used to extract features. A classifier is trained to select the row that is most likely to contain retrieval evidence, and this evidence is subsequently consolidated.

For the Reader stage, the evidence obtained from the retrieval part and the target question are used as input. A reader model is trained to analyze and generate the final answer (Chen et al., 2022, 2020b; Kumar et al., 2021).

3. Methodology

3.1. Overview

The overall DRAMA architecture is shown in Figure 1. DRAMA consists of a pretrained language model-based extractor for features, a Dynamic Graph Attention Network (DGAT) module with cross-row interaction capabilities, and a passage reader for answer generation.

We introduce an innovative Dynamic Graph Attention Network (DGAT) module in the middle of the framework. This module comprises two key elements: a dynamic estimator designed to evaluate edge correlations and a memory bank constructed for node feature storage and retrieval. Initially, a memory bank stores the features inferred by the extractor. Subsequently, a heterogeneous graph model is constructed based on the current row's relationship and the table's remaining instances. Features from non-current rows are fetched via the memory bank. Then, the dynamic estimator evaluates the shift of question attention between each layer of the graph network and adjusts the edge weights in the graph structure, which helps to eliminate the noise connection introduced in the multi-row retrieval. Finally, the answer is obtained through a passage reader. The fundamental framework for the retrieval and the passage reader

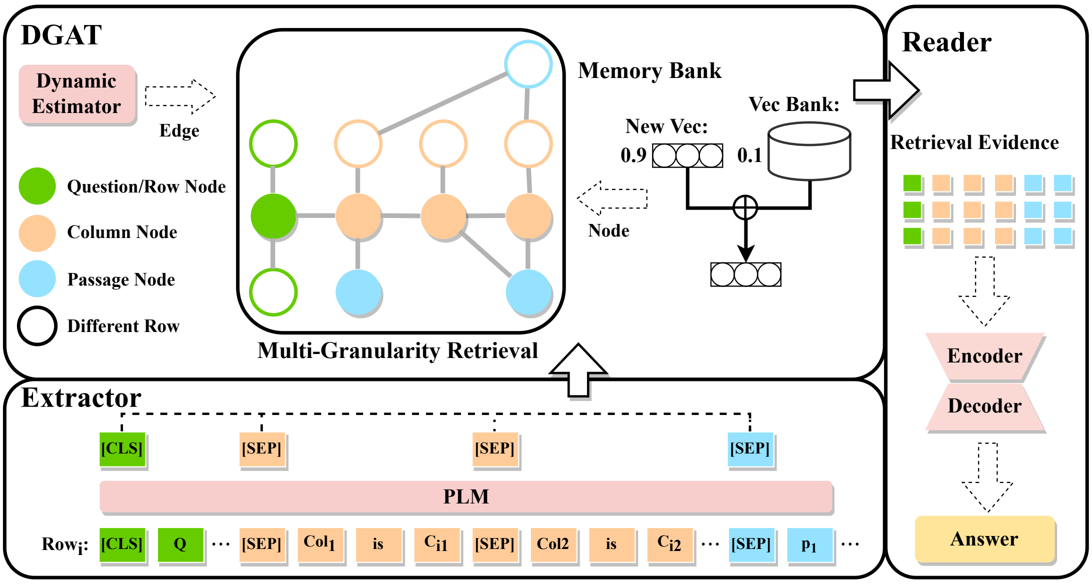


Figure 1: The overall DRAMA model architecture. The extractor generates the textual features, and then DGAT refines the structural information from heterogeneous sources. Depending on the evidence scores retrieved, the reader component generates the final answer.

stages are delineated in Section 2.2. In the following, we describe the DGAT module within the DRAMA model in more detail and then explain its learning process.

3.2. Multi-Granularity Retrieval

For multi-granularity retrieval, we incorporate the features from different granularity units, like rows, columns, and passages, in the graph $\mathcal{G}_t$=<$\mathcal{V}, \mathcal{E}$>, that is $\mathcal{V}$=$\{q\} \cup \{t\} \cup P$.

$$Seq_i = \underbrace{[\text{CLS}]\ q\ ;}_{question} \underbrace{\bigcup_{j=1}^{M} [\text{SEP}]\ j \text{ column is } Col_j\ ;}_{column}$$

$$\underbrace{[\text{SEP}] \text{ row } i\ ;}_{row_i} \bigcup_{j=1}^{M} \underbrace{[\text{SEP}]\ Col_j \text{ is } c_{ij}}_{cell_{ij}}\ ; \qquad (1)$$

$$\underbrace{[\text{SEP}]\ p_i}_{passage}$$

Specifically, given the question q, the column name, the-ith-row from the table $\{header, row_i\} \in t$, and passages from the link set $p_i \in P$, we concatenate them with the [SEP] and [CLS] tokens into an encoded sequence Seq_i. We use the hidden states from the last layer of PLM as the node features for $\mathcal{V} = \{h_{type}\}, type \in \{question, row, column, cell, passage\}$. The hidden states of the [CLS] position encapsulate question-related information, while other hidden states of [SEP] positions represent the fundamental details of different granularities, like row, column, cell, and passage.

Based on the edge ablation study in Section 4.4, we set the edge connections between different granularities. We set multi-granularity connections as $\mathcal{E} = \mathcal{E}_{ques-cell} \cup \mathcal{E}_{cell-col} \cup \mathcal{E}_{cell-passage} \cup \mathcal{E}_{ques-row_{other}} \cup \mathcal{E}_{cell-cell_{other}}$. The sketch map for part of the connections is shown in Figure 1. Additionally, to integrate the heterogeneous structural information between rows, cells, and passages, we establish their relationship using the relational graph attention transformers (Wang et al., 2020). Different types of edges are used to represent the connections between various granularities.

3.3. Memory Bank

Encoding each table row with an encoder during multi-row retrieval will lead to considerable computational and memory expenses. It may be unavailable due to the limitation of the input length of the language model.

To avoid redundantly encoding the table context, we preprocess the semantic information of each row and each granularity level in the table and store the semantic information of the entire table in a memory bank. We only encode the context of the current row during training. In contrast, the information from other rows is sampled from a memory bank.

However, there is instability during the training process due to the inconsistency between the PLM encoder and the memory bank embeddings. If a database is simply used to store all the preprocessed row text information, the semantic space of the new encoded vectors will gradually diverge from the preprocessed ones during the training process. Otherwise, updating all the vectors in the memory bank instantly during training may also lead to significant semantic discrepancies among the vectors. Inspired by the momentum approach introduced in the MoCo (He et al., 2020) paper on computer vision, we introduce the momentum mechanism to smooth the semantic discrepancies between the memory bank and the rapidly updating encoder. The memory bank is updated with the momentum mechanism:

$$\mathcal{E}_x^{(i+1)} = m \cdot \mathcal{E}_x^{(i)} + (1-m) \cdot f_{encoder}(x) \tag{2}$$

The parameter m denotes the momentum coefficient, and $\mathcal{E}_x$ is the representation from the memory bank.

3.4. Dynamic Retrieval Relevance Estimator

In the HybridQA task, question-answer retrieval often involves a multi-hop reading process. At each retrieval step, new evidence is introduced to shift the focus of the target question continually.

We introduce a dynamic estimator to estimate the target evidence connection in the graph network. As shown in Figure 2, we assess the directed focus attributed to the question $h_q^{(t)}$ within individual retrieval iterations during the question-answering process. Based on the structural information within the table, a heterogeneous graph $\mathcal{G}_t$ is constructed between different rows, columns, and granularities. However, the original heterogeneous graph often contains redundant nodes and edges that bring noise information unrelated to the target question, which can affect the acquisition of valid information. We further introduce the dynamic estimation strategy, adjusting the weights of the edges dynamically in each layer of the graph network based on the attention to the target question.

Specifically, due to the utilization of the entire table cells and linked passages feature in retrieval, a considerable amount of irrelevant text information is introduced while providing evidence pertinent to the question. Our method computes the relevance score between the target task and the candidate text based on the target question and the evidential information aggregated within the graph network. This, in turn, aids in diminishing the noise information in the graph that is irrelevant to the target question.

$$\begin{aligned} h_q^{(t)}, h_t^{(t)} &= \mathrm{GRU}(h_q^{(t-1)}, h_t^{(t-1)}) \\ \hat{h}_q^{(t)} &= h_q^{(t)} + h_q^{(t-1)} \\ c_{qi}^{(t)} &= \hat{h}_q^{(t)} \odot h_i^{(t)} \\ d_{ij}^{(t)} &= \sqrt{c_{qi}^{(t)} \odot c_{qj}^{(t)}} \end{aligned} \tag{3}$$

where h_q is the question embedding between each graph layer and h_t is the hidden states initialized from zero vector. We use a gated recurrent unit (GRU) (Dey and Salem, 2017) to estimate the question attention and refine the relationship embedding into d_{ij}. $\hat{h}_q^{(t)}$ is the aggregation of the target question and the evidence retrieved from the $t-1$ graph layer. Finally, we obtain $d_{ij} \in [0,1]$ for the dynamic retrieval correlation coefficient between neighbouring nodes.

$$\begin{aligned} e_{ij}^{(t)} &= \frac{h_i^{(t)} W_q (h_j^{(t)} W_k + r_{ij} \odot d_{ij}^{(t)})^{\mathrm{T}}}{\sqrt{d_z/H}} \\ \alpha_{ij}^{(t)} &= \mathrm{softmax}(e_{ij}^{(t)}) \\ h_i^{(t+1)} &= \sum_{j=1}^{n} \alpha_{ij}^{(t)} (h_j^{(t)} W_v + r_{ij} \odot d_{ij}^{(t)}) \end{aligned} \tag{4}$$

By employing dynamic retrieval relevance estimators, the connections between nodes with weak relevance to the question are diminished. Similarly, in the node representation aggregation step, the connection weights of weak relevance are weakened. That is, among the various layers of the graph network, the weights of the connecting edges within the graph network are dynamically adjusted according to the attention shift towards the question, thereby reducing the noise brought about by irrelevant information.

3.5. Multi-granularity Training

The loss $\mathcal{L}$ for the retrieval part consists of cross-entropy losses for all granularities. The labels are determined based on whether the answer is a substring of the instance.

$$\begin{aligned} \mathcal{L}_R &= -\frac{1}{N} \sum_{i=1}^{N} y_i^R \cdot \log(\hat{y}_i^R) \\ \mathcal{L}_C &= -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n} \sum_{j=1}^{n} y_{ij}^C \cdot \log(\hat{y}_{ij}^C) \\ \mathcal{L}_P &= -\frac{1}{N} \sum_{i=1}^{N} y_i^P \cdot \log(\hat{y}_i^P) \end{aligned} \tag{5}$$

where N denotes the number of training instances, and $\mathcal{L}_R$, $\mathcal{L}_C$, and $\mathcal{L}_P$ respectively signify the loss functions for the current row, each column in the row, and the linked passages. The overall loss is the weighted sum of the losses at each granularity:

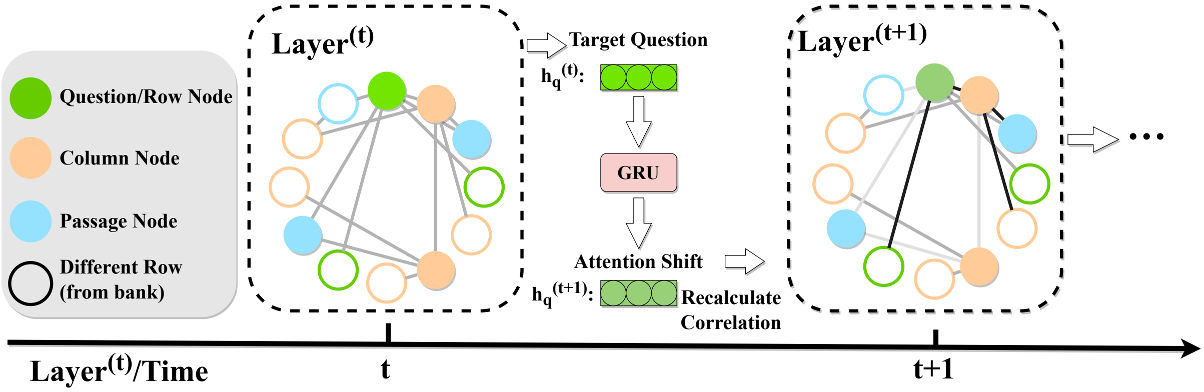


Figure 2: Dynamic Graph Attention Network (DGAT). The basic part of the dynamic retrieval relevance estimation strategy. Different color shades represent the various edge weights. In each layer of the graph network, the structure is dynamically adjusted.

$$\mathcal{L}_{retrieval} = \alpha \cdot \mathcal{L}_R + \beta \cdot \mathcal{L}_C + \gamma \cdot \mathcal{L}_P \tag{6}$$

where α, β, γ are the task coefficients.

As for inference, we first calculate the probabilities of each granularity. Then, we select the row with the highest probability, as well as the two highest-ranked granularities at both column and passage levels, apart from the current row. These selected pieces of evidence are then concatenated for inference.

For the Reader section, we train a generative model, BART, optimizing the parameters of the Reader by taking the product of the probabilities of the output sequences a_1, a_2, ..., a_n.

$$\mathcal{L}_{reader} = -\sum_{i=1}^{n} \log(P(a_i|a_{<i})) \tag{7}$$

4. Experiment

4.1. Experimental Setup

Setup

Datasets.

HybridQA (Chen et al., 2020b) comprises 69,611 problems, each containing a context with a table and multiple linked textual paragraphs to the several table cells. It is a widely used dataset of multi-hop question-answering over tabular and textual data. In terms of the evaluation data split, 'In-table' implies that the answer is retrieved directly from a table cell value, whereas 'In-Passage' indicates that the answer is retrieved from the linked passages.

TabFact (Chen et al., 2019) consists of 16,573 tables. It is a dataset for verifying language understanding on tabular data. This task aims to judge whether a target statement is entailed or denied based on the given table. It is used to evaluate hybrid reasoning skills in symbols and language on structured data.

Evaluation Metrics.

The evaluation of the datasets is based on their respective metrics: Exact Match (EM) and numeracy-focused F1 (Dua et al., 2019) calculated across different data types for HybridQA and accuracy for TabFact.

Implementation Details.

We use the *bert-large-uncased* (Devlin et al., 2018) as a pre-trained language model for the text information encoding and evidence retrieval stage. The Pretrained Language Model (PLM) and Dynamic Graph Attention Network (DGAT) are trained concurrently, with DGAT using an 8-layer network. The task coefficients $\{m, \alpha, \beta, \gamma\}$ is set to $\{0.9, 1.0, 1.0, 1.0\}$. The model is trained with AdamW optimizer (Loshchilov and Hutter), with a learning rate of 1e-5, across 5 epochs. The batch size is set at 24 with 3 gradient accumulation steps. This part of the training requires 2 days on an NVIDIA Tesla V100 GPU.

For the Reader stage, we use the *bart-large-cnn* (Lewis et al., 2020) to generate answers based on the consolidated evidence for each given question. We utilize the AdamW optimizer with a learning rate of 1e-5, training over 10 epochs with a batch size of 4.

4.2. Baselines

In the experiment, we compare DRAMA with other methods of HybridQA as follows:

- **Hybrider**: (Chen et al., 2020b) leverages a two-stage model in their approach. The first phase employs a sparse passage retriever to identify pertinent cells and their corresponding textual data. Then, the target answer is extracted.

	In-Table				In-Passage				Total			
	Dev		Test		Dev		Test		Dev		Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Hybrider	54.3	61.4	56.2	63.3	39.1	45.7	37.5	44.4	44.0	50.7	43.8	50.6
DocHopper	-	-	-	-	-	-	-	-	47.7	55.0	46.3	53.3
MATE	68.6	74.2	66.9	72.3	62.8	71.9	62.8	71.9	63.4	71.0	62.8	70.2
MuGER^2	58.2	66.1	56.7	64.0	52.9	64.6	52.3	63.9	53.7	63.6	52.8	62.5
MITQA	68.1	73.3	68.5	74.4	66.7	75.6	64.3	73.3	65.5	72.7	64.3	71.9
MAFiD	69.4	75.2	68.5	74.9	66.5	75.5	65.7	75.3	66.2	74.1	65.4	73.6
DRAMA	**71.4**$_{\pm0.1}$	**77.6**$_{\pm0.2}$	**69.5**$_{\pm0.1}$	**76.1**$_{\pm0.1}$	**67.4**$_{\pm0.1}$	**76.3**$_{\pm0.2}$	**66.4**$_{\pm0.2}$	**76.2**$_{\pm0.3}$	**67.3**$_{\pm0.2}$	**75.3**$_{\pm0.2}$	**66.2**$_{\pm0.1}$	**74.5**$_{\pm0.2}$
Human	-	-	-	-	-	-	-	-	-	-	88.2	93.5

Table 1: EM and F1 results of our model and related work on the HybridQA dataset. The experimental results are obtained by averaging the performance on the test set across three different random seeds.

- **DocHopper**: (Sun et al., 2021) proposes an approach where the table, along with its hyperlinked passages, is encoded into an extensive document. Subsequently, elements such as column headers, cell text, and linked passages are concatenated to form a comprehensive paragraph.

- **MATE**: (Eisenschlos et al., 2021) applies sparse attention to the rows and columns within the table, a strategy aimed at curtailing computational complexity. Employing the PointR module, 'golden cells' are selected from which corresponding answers are subsequently retrieved.

- **MuGER**^{2}: (Wang et al., 2022) evaluates various instances such as cells, rows, columns, and linked passages at different granularities, subsequently assigning respective scores. Based on these evaluated scores, the reader module is employed to aggregate the final answer.

- **MITQA**: (Kumar et al., 2021) employs a multi-instance, distance-based training approach, designed specifically to mitigate the impact of noise originating from examples with multiple answer spans.

- **MAFiD**: (Lee et al., 2023) leverages a fusion-in-decoder approach, which amalgamates diverse encoding representations. Subsequently, a generative model is deployed to facilitate the production of answers.

In the sentence entailment task, we further evaluate the DGAT on different types of pre-trained models:

- **TAPAS**: (Herzig et al., 2020) proposes a table-based BERT model, which has been pre-trained and weakly supervised and fine-tuned on a large amount of structured tabular data, exhibiting better understanding and representation capabilities for structured data.

Model	Method	Acc
BERT-large	FT	65.1$_{\pm0.3}$
BERT-large+GAT	FT	74.4$_{\pm0.2}$
BERT-large+DGAT	FT	**78.1**$_{\pm0.4}$
TAPAS	PT+FT	81.0$_{\pm0.1}$
TAPAS+GAT	PT+FT	81.7$_{\pm0.2}$
TAPAS+DGAT	PT+FT	**82.9**$_{\pm0.2}$

Table 2: The accuracy of retrieval achieved by DGAT on TabFact, based on different pre-trained models.

4.3. Main Results

As summarized in Table 1, DRAMA shows the state-of-art performance by achieving EM and F1 by 66.2 and 74.5 on the blind test set of HybridQA. It is observed that DRAMA outperforms MATE (Eisenschlos et al., 2021), a pre-training model that utilizes sequential encoding for tabular data. This suggests that effectively processing the structural information is crucial for interpreting heterogeneous source data and for selecting the potential evidence accurately. Additionally, DRAMA displays enhancements over the multi-granularity baseline. This underscores the salience of the interconnectedness across data of varying granularities. Rather than examining the features of each granularity in isolation, integrating information across multiple granularities is instrumental for a comprehensive understanding of hybrid table-text data.

As shown in Table 2, we further analyze the retrieval ability of DGAT on the TabFact dataset. The experiment compares the static graph method GAT with the proposed dynamic graph method DGAT. The node features in the graph network are based on pre-trained language models on different data domains. A direct representation of the structure is introduced into the graph network, contributing to a 9.3% improvement in the accuracy of the BERT-large+GAT model. Utilizing dynamic graph estimation methods helps further enhance the understanding capability for complex reasoning

Model	Dev EM	Dev F1
DRAMA	**67.3**	**75.3**
w/o multi-row	64.4	71.4
w/o multi-granularity	64.9	71.9
w/o dynamic estimate	66.0	73.9
w/o momentum	66.3	74.6

Table 3: Ablation study on different modules of DRAMA.

Edge Type	Dev EM	Dev F1
No Graph Edge	64.9	71.9
+ Ques — Cell	65.1	72.2
+ Cell — Passage & Col	65.3	72.4
+ Ques — Row_{other}	65.5	72.6
+ Cell — Cell_{other}	**66.0**	**73.9**
+ Ques — Cell_{other}	65.6	73.1
+ Row_{other} — Cell_{other}	65.5	72.5

Table 4: Ablation study of graph edge construction.

tasks. Employing a pre-trained model with TAPAS on structured data for feature representation improves the understanding of tabular data. The experimental results on different PLMs demonstrate that the introduced module possesses a certain level of universality.

4.4. Ablation Studies

Since the test set is blind, we can't analyze the detailed results on it. Consequently, We conduct ablation studies on the dev set.

Effect of multi-granularity retrieval. As shown in Table 3, we conduct ablation studies on the modules of DRAMA. When only using the sequential encoding (w/o multi-row), the EM score declined by 3.9 points. However, with the introduction of the memory bank and multi-row access (w/o multi-granularity), there is a slight improvement in the model's performance. The introduction of a heterogeneous graph leads to a significant improvement in model performance (w/o dynamic estimate). Besides, these results underscore the ability of the Dynamic Graph Attention Network (DGAT) to locate inferential evidence through its computation of question attention more accurately. Further incorporation of the momentum (w/o momentum) updating strategy during training leads to more stable model training and further improvement in performance.

Effect of graph construction. As shown in Table 4, we conduct ablation studies on different types of edges in the heterogeneous graph. For a more intuitive comparison of the impact of structural information on evidence retrieval, we do not incorporate dynamic estimation of question attention in these experiments. We use the sequential encoding method as the baseline method. From top to bottom, we incrementally add different types of edges to the graph. The subscript of $other$ represents the nodes of other rows from the memory bank. It shows that the addition of all *cell-to-cell* structural information in the table leads to a significant improvement in model performance, validating the efficacy of direct encoding of structural information. Based on the experimental results, we select the crucial edge combinations that play a crucial role in this task as the final approach.

m	0.0	0.5	0.9	0.99	1.0
Dev EM	-	65.9	**67.3**	67.0	66.3

Table 5: Ablation study of momentum update.

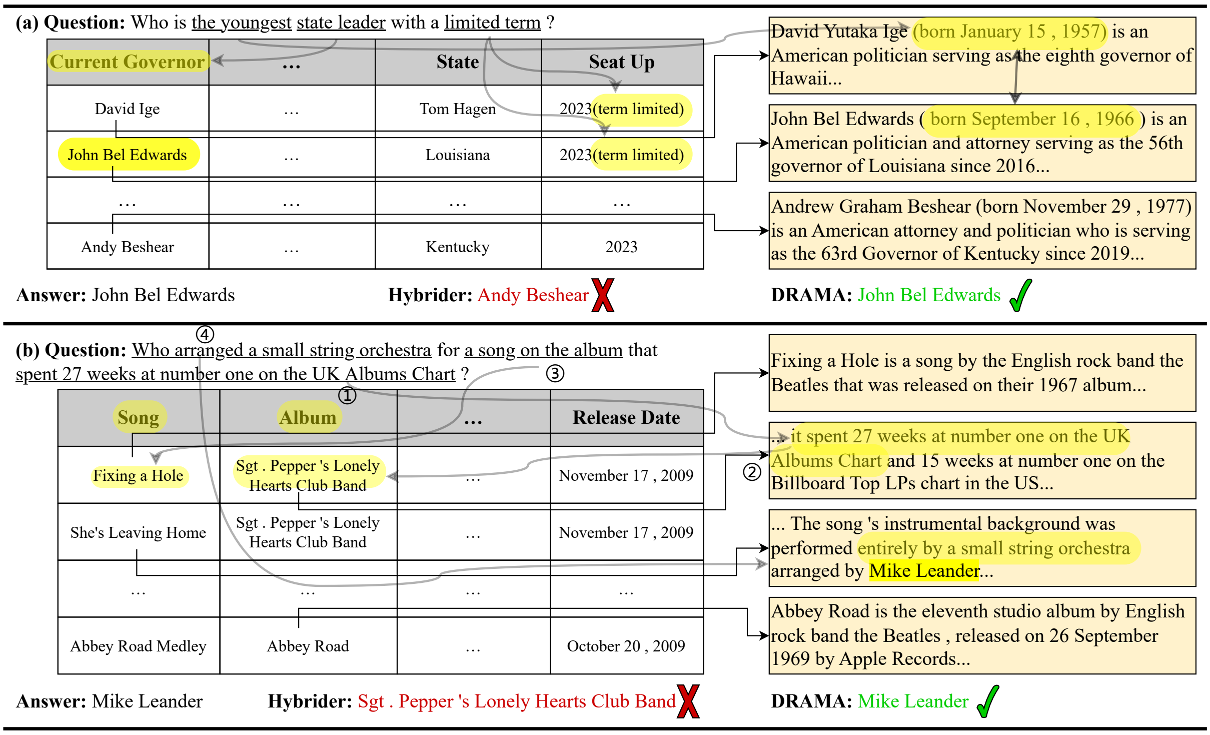
Effect of momentum vector update. As shown in Table 5, we analyze the impact of the momentum updating method for vectors on model training. Here, m represents the momentum parameter in equation 2. When $m = 0$, indicating real-time updating of vectors in the memory bank, it introduces a substantial computational overhead, which is impractical in the experimental setting. Conversely, when $m = 1$, meaning the vectors in the memory bank are not updated and inference is performed on all vectors only at the onset of training, the EM score decreases by 1.0. We ultimately select $m = 0.9$ as the optimal momentum coefficient.

4.5. Case Studies

As shown in Figure 3, we analyze the performance of the DRAMA and Hybrider in several cases.

In case (a), the question involves comparisons between multiple rows in the given table. The Hybrider baseline, which only considers single-row information retrieval, does not have a sufficient inferential basis to answer the question. However, in the DRAMA model, we introduce features of multiple rows, allowing the model to compare the ages of the current governors in different rows. Concurrently, the heterogeneous graph integrates information from different modalities, allowing the model to evaluate cell information in the table while comparing linked passages. The model correctly determines that the answer originates from the row corresponding to *term limited* in the *Seat Up* column.

In case (b), the question involves multi-hop retrieval of the table and linked text information, and the large amount of repetitive information in the table can introduce noise. With the phrase *spent*



(a) Question: Who is the youngest state leader with a limited term ?

Current Governor	...	State	Seat Up
David Ige	...	Tom Hagen	2023(term limited)
John Bel Edwards	...	Louisiana	2023(term limited)
...	...	...	...
Andy Beshear	...	Kentucky	2023

David Yutaka Ige (born January 15 , 1957) is an American politician serving as the eighth governor of Hawaii...

John Bel Edwards (born September 16 , 1966) is an American politician and attorney serving as the 56th governor of Louisiana since 2016...

Andrew Graham Beshear (born November 29 , 1977) is an American attorney and politician who is serving as the 63rd Governor of Kentucky since 2019...

Answer: John Bel Edwards **Hybrider:** Andy Beshear ✗ **DRAMA:** John Bel Edwards ✓

(b) Question: Who arranged a small string orchestra for a song on the album that spent 27 weeks at number one on the UK Albums Chart ?

Song	Album	...	Release Date
Fixing a Hole	Sgt . Pepper 's Lonely Hearts Club Band	...	November 17 , 2009
She's Leaving Home	Sgt . Pepper 's Lonely Hearts Club Band	...	November 17 , 2009
...	...	...	...
Abbey Road Medley	Abbey Road	...	October 20 , 2009

Fixing a Hole is a song by the English rock band the Beatles that was released on their 1967 album...

... it spent 27 weeks at number one on the UK Albums Chart and 15 weeks at number one on the Billboard Top LPs chart in the US...

... The song 's instrumental background was performed entirely by a small string orchestra arranged by Mike Leander...

Abbey Road is the eleventh studio album by English rock band the Beatles , released on 26 September 1969 by Apple Records...

Answer: Mike Leander **Hybrider:** Sgt . Pepper 's Lonely Hearts Club Band ✗ **DRAMA:** Mike Leander ✓

Figure 3: Case studies of Hybrider and DRAMA on the HybridQA dataset.

27 weeks at number one in the question, all the models precisely locate the *Album: Sgt. Pepper's Lonely Hearts Club Band*. The DGAT module in DRAMA targets the question by dynamically estimating the attention of the question, thus eliminating noisy connections and correctly identifying the pertinent row. However, Hybrider, being affected by fields that repeatedly appear in the table and are somewhat relevant to the question, fails to locate the ultimate target of the question.

5. Related Work

In recent years, traditional question-answering tasks focused solely on textual or tabular data have been systematically researched (Herzig et al., 2020; Sun et al., 2019; Zhou et al., 2022). Each type of data has its own characteristics; textual data can be obtained in large quantities from various sources, while tabular data is beneficial for presenting comparative information between similarly structured data. There has been an increasing trend in practical applications towards integrating unstructured text information with structured table information, a technique known as Hybrid Question Answering (HQA) (Nakamura et al., 2022). To bridge this gap, (Chen et al., 2020b) proposed the HybridQA dataset, which has been widely used to study heterogeneous QA tasks. Each question in the HybridQA dataset includes a WiKiTable and its corresponding linked Wikipedia passages as sources of evidence for answer retrieval. Furthermore, (Chen et al., 2020a) proposed an open-domain hybrid question-answering dataset (OTT-QA) based on the Wikipedia dataset. More specifically, (Zhu et al., 2021) and (Chen et al., 2021) proposed TAT-QA and FinQA for numerical calculation question-answering on hybrid data in the financial domain.

In terms of heterogeneous data representation, traditional methods transform structured tables into sequential text forms through specific encoding strategies. Hybrider (Chen et al., 2020b) proposes a two-stage framework that divides tables into cells and concatenates them with linked text. The method then utilizes a reading comprehension (RC) model to extract answers. DocHopper (Sun et al., 2021) uses an end-to-end method to retrieve and locate passage sentences or cells as inference evidence. However, converting both types of data into sequential encoding directly can result in excessively long encoding lengths. MITQA (Kumar et al., 2021) conducts retrieval based on rows as units. MATE (Eisenschlos et al., 2021) modifies the attention layer of the transformer network into sparse attention, reducing the computational complexity and thus supporting longer sequence inputs.

6. Conclusion

In this paper, we propose a Dynamic Multi-granularity Graph Estimate Retrieval method, **DRAMA**, to address multi-hop TableTextQA problems. When answering questions from given structured tables and unstructured passages, DRAMA employs a heterogeneous graph to create various types of nodes and edges, representing data from diverse sources and their associated interconnections. This approach effectively mitigates the problem of structural information loss when combining heterogeneous data. Due to the large volume of text input, a memory bank is used to store some of the feature information from the table, and a momentum update method is used during the training process to smooth out differences in the feature space within the memory bank. To enhance the capability to evaluate the question attention shift in multi-hop QA problems, we design a Dynamic Graph Attention Network, which dynamically estimates question attention between each layer of the graph network and recalculates the correlation between evidence and the question to eliminate the noise. DRAMA has demonstrated state-of-the-art performance on the widely used HybridQA benchmark.

7. Limitations

Since the TableTextQA task has only one publicly available dataset, HybridQA, for heterogeneous data, on which we conduct our experiments exclusively. This may result in a certain lack of generalizability of our model. To further verify the effectiveness of our proposed DGAT module, we also conduct experimental validation on the purely structured dataset TabFact. However, this dataset may not reflect the retrieval capability at the passage level.

For count-based statistical questions, the target answer cannot be directly extracted from the evidence. Therefore, the method of obtaining answers varies between these types of questions. However, such questions are less frequent and are not the primary focus of our model. Since the dataset does not explicitly provide the type for each question, we may consider classifying questions using methods such as rule matching in the future.

8. Acknowledgements

This research work is supported by National Key R&D Plan No. 2022YFC3303303, the National Natural Science Foundation of China under Grant No.61976204, and the CAAI Huawei MindSpore Open Fund. Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS, Beijing Nova Program 20230484430.

9. Bibliographical References

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020a. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. Hitab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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)*, pages 2368–2378.

Julian Eisenschlos, Maharshi Gor, Thomas Mueller, and William Cohen. 2021. Mate: Multi-view attention for table transformer efficiency. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7606–7619.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333.

Junjie Huang, Wanjun Zhong, Qian Liu, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Mixed-modality representation learning and pre-training for joint table-and-text retrieval in openqa. *arXiv preprint arXiv:2210.05197*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Vishwajeet Kumar, Saneem Chemmengath, Yash Gupta, Jaydeep Sen, Samarth Bharadwaj, and Soumen Chakrabarti. 2021. Multi-instance training for question answering across table and linked text. *arXiv preprint arXiv:2112.07337*.

Sung-Min Lee, Eunhwan Park, Daeryong Seo, Donghyeon Jeon, Inho Kang, and Seung-Hoon Na. 2023. Mafid: Moving average equipped fusion-in-decoder for question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2292–2299.

Fangyu Lei, Shizhu He, Xiang Li, Jun Zhao, and Kang Liu. 2022. Answering numerical reasoning questions in table-text hybrid contents with graph-based encoder and tree-based decoder. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1379–1390.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4078–4088.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. In *International Conference on Learning Representations*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492.

Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Qiang Fu, Yan Gao, Jian-Guang Lou, and Weizhu Chen. 2022. Reasoning like program executors. *arXiv preprint arXiv:2201.11473*.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2380–2390.

Haitian Sun, WilliamW. Cohen, and Ruslan Salakhutdinov. 2021. End-to-end multihop retrieval for compositional question answering over long documents. *arXiv: Computation and Language*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. 2020. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *Machine Learning for Health*, pages 341–354. PMLR.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578.

Yingyao Wang, Junwei Bao, Chaoqun Duan, Youzheng Wu, Xiaodong He, and Tiejun Zhao. 2022. MuGER2: Multi-granularity evidence retrieval and reasoning for hybrid question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6687–6697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600.

Wanjun Zhong, Junjie Huang, Qian Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. Reasoning over hybrid chain for table-and-text open domain question answering. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4531–4537.

Yongwei Zhou, Junwei Bao, Chaoqun Duan, Youzheng Wu, Xiaodong He, and Tiejun Zhao. 2022. Unirpg: Unified discrete reasoning over table and text as program generation. *arXiv preprint arXiv:2210.08249*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 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)*, pages 3277–3287, Online. Association for Computational Linguistics.