LREC-COLING 2024

KET-QA: A Dataset for Knowledge Enhanced Table Question Answering

Mengkang Hu[1], Haoyu Dong[2*], Ping Luo[1], Shi Han[2], Dongmei Zhang[2]

[1] The University of Hong Kong, [2] Microsoft

HKU MMLAB

Outline

Outline

1.1 Motivation

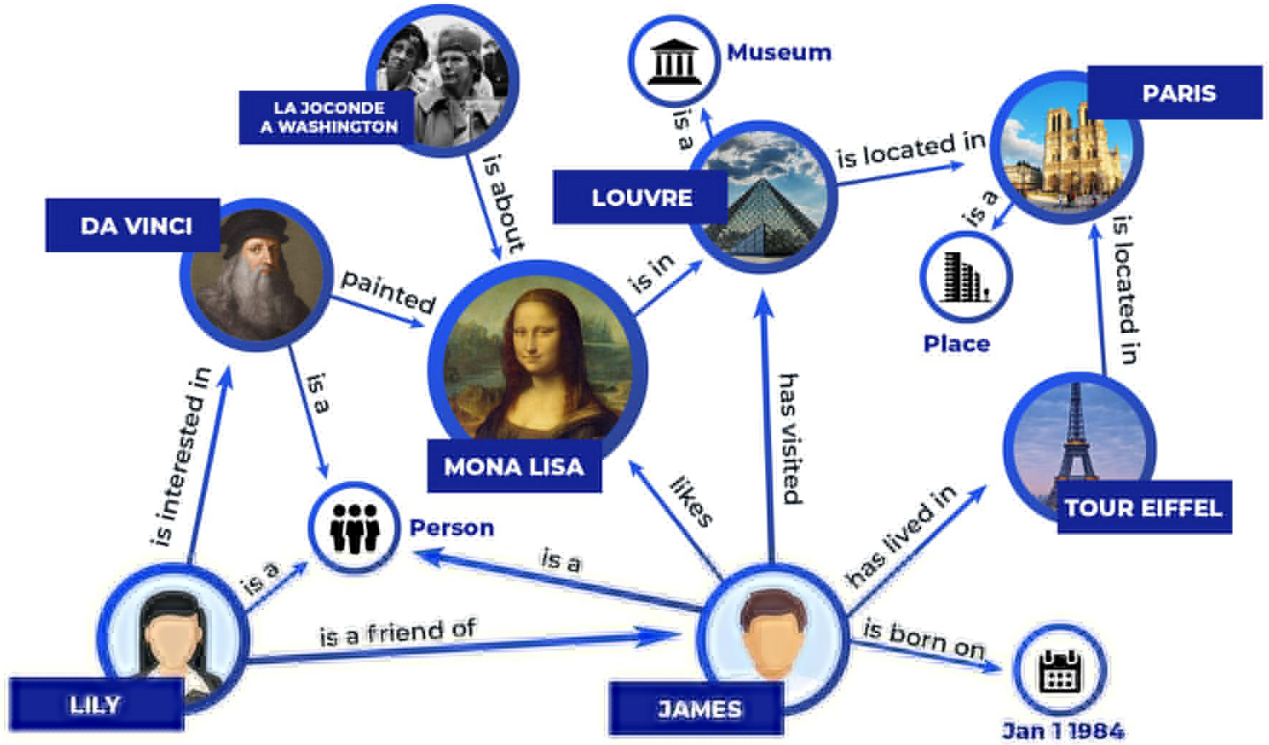
Intuition: Due to the concise and structured nature of tables, the knowledge contained therein may be incomplete or missing, posing a significant challenge for TableQA systems.

Question: What was the release date of the studio album from the artist who signed to the record label GOOD Music ?

Missing in the table.

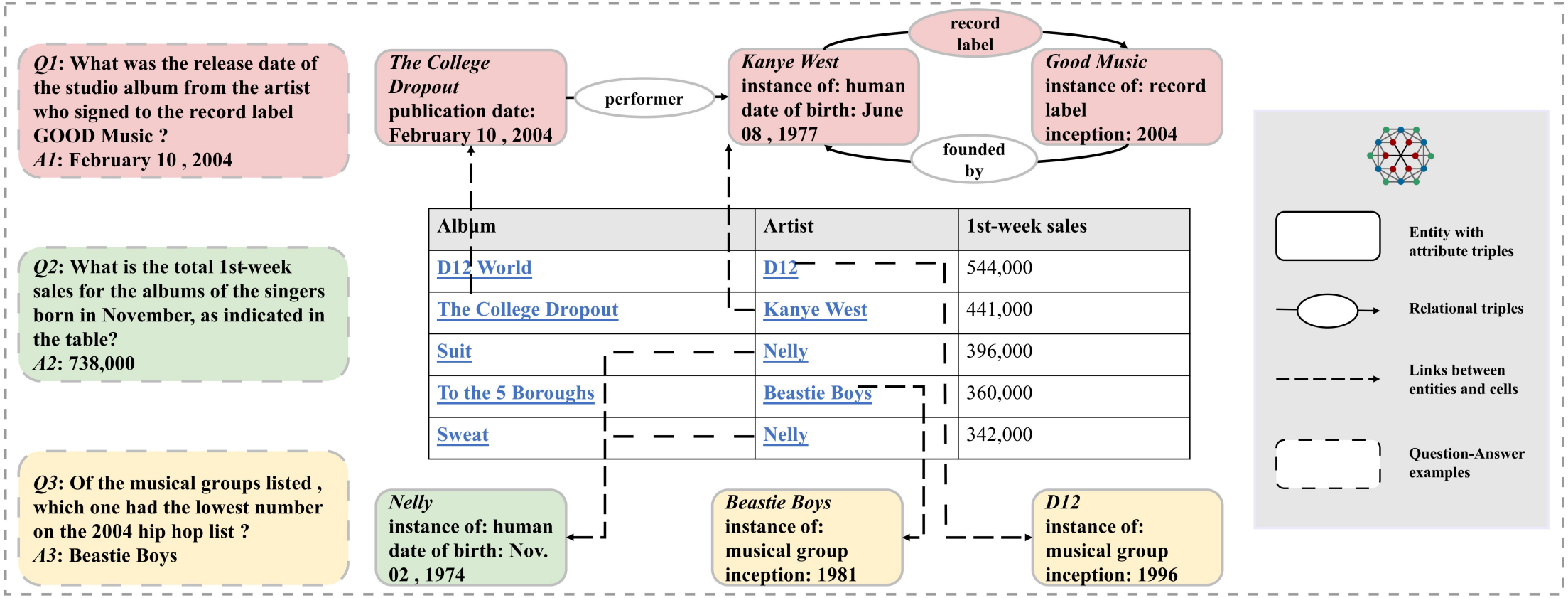
Album	Artist	1st week sales
D12 World	D12	544,000
The College Dropout	Kanye West	441,000
Suit	Nelly	396,000
To the 5 Boroughs	Beastie Boys	360,000
Sweat	Nelly	342,000

Can we utilize the high-quality knowledge in knowledge graphs to alleviate the problem of missing knowledge in tables?



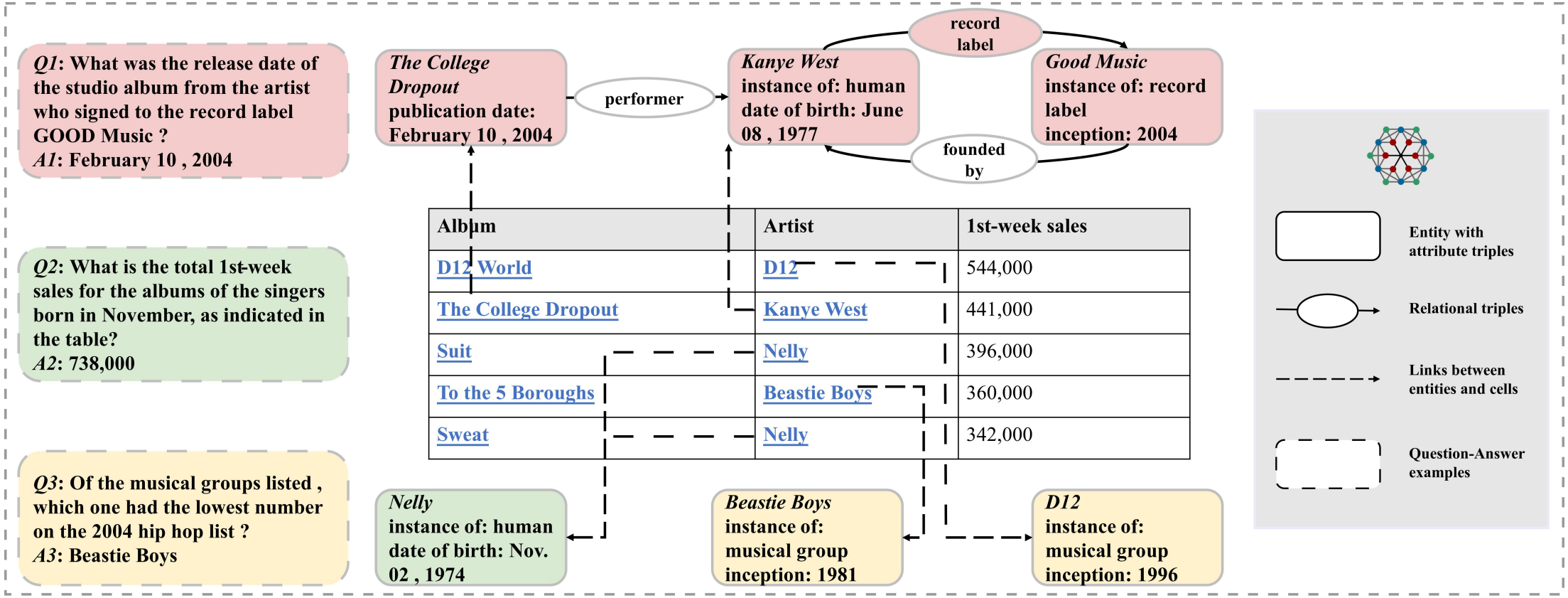
Outline

2.1 Dataset Overview



We also annotated gold evidence for each question. For example, for Q1, the gold evidence for this instance would be *{((2, 1), (Kanye West, record label, Good Music)), ((2, 0), (The College Dropout, publication date, February 10, 2004))}*.

2.1 Task Definition



The process of KET-QA: given a table T, the grounded knowledge sub-graph G, and a natural language question q, output a that answers the question according to the context.

2.2 Statistics

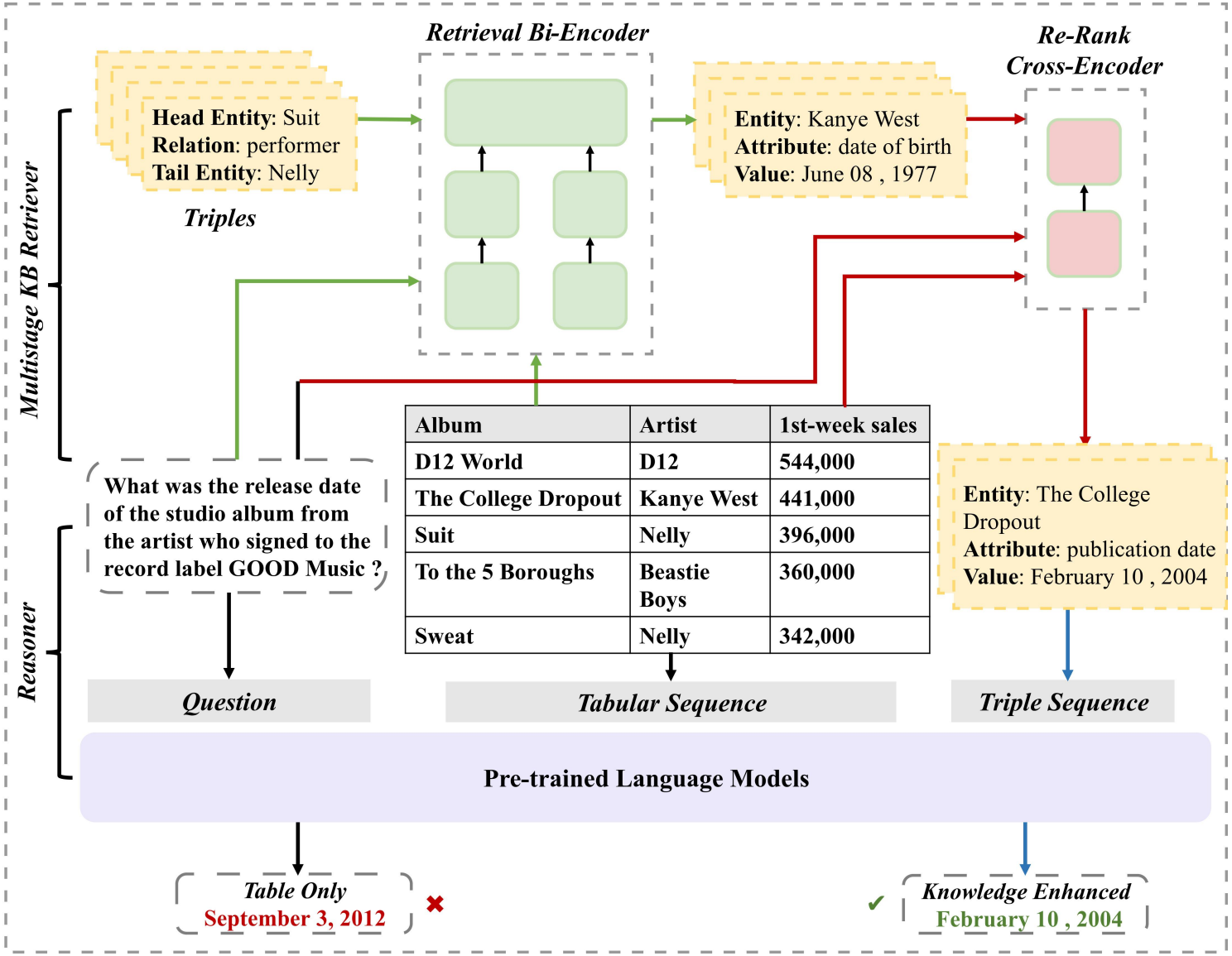
Dataset	Size		External Knowledge		
	#Ques.	#Tables	Type	Source	GE
WTQ	22,033	2,108	-	-	-
WikiSQL	80,654	26,521	-	-	-
Spider	10,181	1,020	-	-	-
HiTab	10,672	3,597	-	-	-
FeTaQA	10,330	10,330	-	-	-
HybridQA	69,611	13,000	Text	Wikipedia	No
TAT-QA	16,552	2,757	Text	Financial reports	Yes
FinQA	8,281	2,776	Text	Financial reports	Yes
KET-QA	9,421	5,721	KB	Wikidata	Yes

#Words/Ques.	#Words/Answer	#Rows/Table
17.2	3.3	15.8
#Columns/Table	#Entities/Table	#Triples/Table
4.5	41.9	1696.7
Answer in KB	Answer in Table	Calculated Answer
5197	4131	93

- KET-QA is the first TableQA dataset that utilizes a knowledge graph as external knowledge source
- KET-QA is the first TableQA dataset that provides fine-grained gold evidence annotations from external knowledge sources.

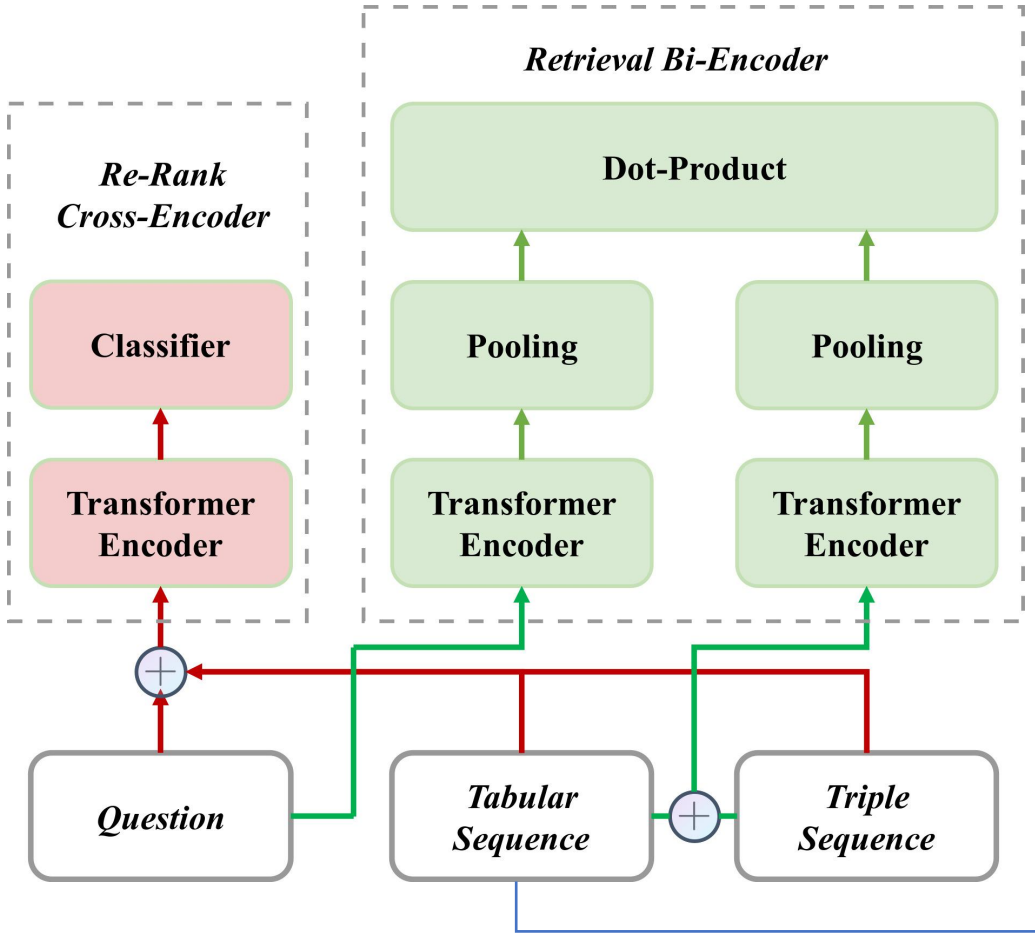
Outline

3.1 Overview



We propose a retriever-reasoner pipeline model to incorporating knowledge from KG to TableQA.

3.2 Multistage Knowledge Base Retriever



Motivation behind 'Multistage Retriever':

Speed of the retrieval operation is important for real-time TableQA scenarios.

Triple-Related Sub-Table: The information within the entire table may be redundant for retrieval and could exceed the length limitations of the transformer model.

$$\mathcal{T} = \{r_i \in T \mid \exists\, c_{ij} \in r_i, e \in f(c_{ij})\}$$

Outline

3.1 Evidence Retrieval - Experimental Setup

Metrics:

We introduce a modified version of Recall@k (R@k) to evaluate the retrieval performance in the context of KET-QA. The purpose of R@k is to measure the percentage of items of gold evidence that are retrieved by the retriever

$$R@k = \frac{1}{N}\sum_{i=1}^{N}\frac{|evidence\ retrieved|_i}{|gold\ evidence|_i}$$

Baseline Methods:
(i) String Match: Triples are retrieved based on whether the label of the triples matches the words in the question;
(ii) Bi-Encoder and Cross-Encoder are used to compare the performance of a single retriever with MKBR.

3.2 Evidence Retrieval - Experimental Results

Method	Top-1	Top-5	Top-20	Top-100
Random	0.05	0.27	2.94	12.49
String Match	5.87	14.65	28.24	43.66
Cross-Encoder	37.83	63.84	82.14	**94.44**
Bi-Encoder	29.17	51.95	72.12	89.62
MKBR	**38.77**	**66.04**	**83.47**	93.51

Table 3: Comparison between retrieval methods on *KET-QA* test set using R@k ($k \in \{1, 5, 20, 100\}$).

Table Rep.	Top-1	Top-5	Top-20	Top-100
Bi-Encoder				
FT	25.81	49.89	71.75	88.91
NT	24.05	48.67	71.89	89.13
TT	**29.17**	**51.95**	**72.12**	**89.62**
Cross-Encoder				
FT	28.59	58.94	78.32	94.42
NT	12.27	34.41	59.78	85.04
TT	**37.83**	**63.84**	**82.14**	**94.44**

Table 4: Ablation study on different table representation methods, which can be chosen from {FT (**F**ull **T**able), NT (**N**o **T**able), TT (**T**riple-Related Sub-**T**able)}.

- in scenarios where k is small (k ≤ 20), MKBR consistently outperforms any single retriever model.
- The proposed Triple-Related Sub-Table is superior to the other two approaches.

Outline

3.1 Question Answering - Experimental Setup

Metrics:

- Exact Match The EM score is a strict all-or-nothing metric, which represents the percentage of predictions that exactly match the ground truth.
- F1 Score measures the token overlap between the predicted answer and ground truth.

Baseline Methods:

- We take table-only models as baselines to explore whether the question can be answered based solely on the table information in the traditional TableQA manner.

3.2 Question Answering - Experimental Results

Model	Table Only				Knowledge Enhanced				Δ			
	Dev		Test		Dev		Test		Dev		Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Fine-Tuning												
TAPEX$_{large}$	14.44	18.52	12.83	17.1	**60.62**	**63.22**	56.63	58.75	46.18	44.7	43.8	41.65
BART$_{large}$	9.34	13.41	8.17	12.17	51.7	54.49	52.81	56.16	42.36	41.08	44.64	43.99
BART$_{base}$	7.64	11.57	8.38	11.56	45.65	48.89	46.87	50.28	38.01	37.32	38.49	38.72
T5$_{base}$	9.77	14.05	9.12	12.97	45.54	48.94	46.02	49	35.77	34.89	36.9	36.03
Zero-Shot												
GPT-3	8.07	17.85	10.07	20.11	33.55	45.04	36.69	47.76	25.48	27.19	26.62	27.65
ChatGPT	3.82	7.65	4.03	7.31	17.73	27.53	15.69	26.79	13.91	19.88	11.66	19.48
Few-Shot												
GPT-3	**33.86**	**39.58**	**31.81**	**37.06**	57.86	63.04	**60.23**	**64.89**	24	23.46	28.42	27.83
ChatGPT	20.7	23.95	19.72	23.92	45.01	49.51	43.26	49.53	24.31	25.56	23.54	25.61

Table 5: Performance of different reasoners on *KET-QA*. Block Δ represents the increase in performance after incorporating knowledge base as an additional source of information. We employ the text-davinci-003 version for GPT-3 and the gpt-3.5-turbo version for ChatGPT.

Outline

5.1 Comparison of Knowledge Sources

Knowledge Source	Dev		Test	
	EM	F1	EM	F1
Table Only	14.44	18.52	12.83	17.1
LLM	29.62	34.23	26.51	30.58
Wikipedia Passages	32.27	36.69	28.31	32.24
Knowledge Base	**60.62**	**63.22**	**56.63**	**58.75**

Table 6: Experimental results with various external knowledge sources. We employed TAPEX$_{large}$ as a representative reasoner.

- LLM-generated Knowledge: We employed the prompt "*Generate some knowledge about the given question and table*" to instruct the Large Language Model to generate knowledge that is beneficial for answering the current question.
- Wikipedia Passage hyperlinked Wikipedia passages can provide additional information that complements the table.

LREC-COLING 2024

Thanks for your attention!

- Project Page: https://ketqa.github.io/

Project Page

Personal Website

YouTube
MMLAB@HKU

HKU MMLAB