A Retrieval-Augmented Framework for Tabular Interpretation with LLM

Mengyi Yan[1] Weilong Ren*[2] Yaoshu Wang[2] Jianxin Li*[1]

[1]Beihang University [2]Shenzhen Institute of Computing Sciences

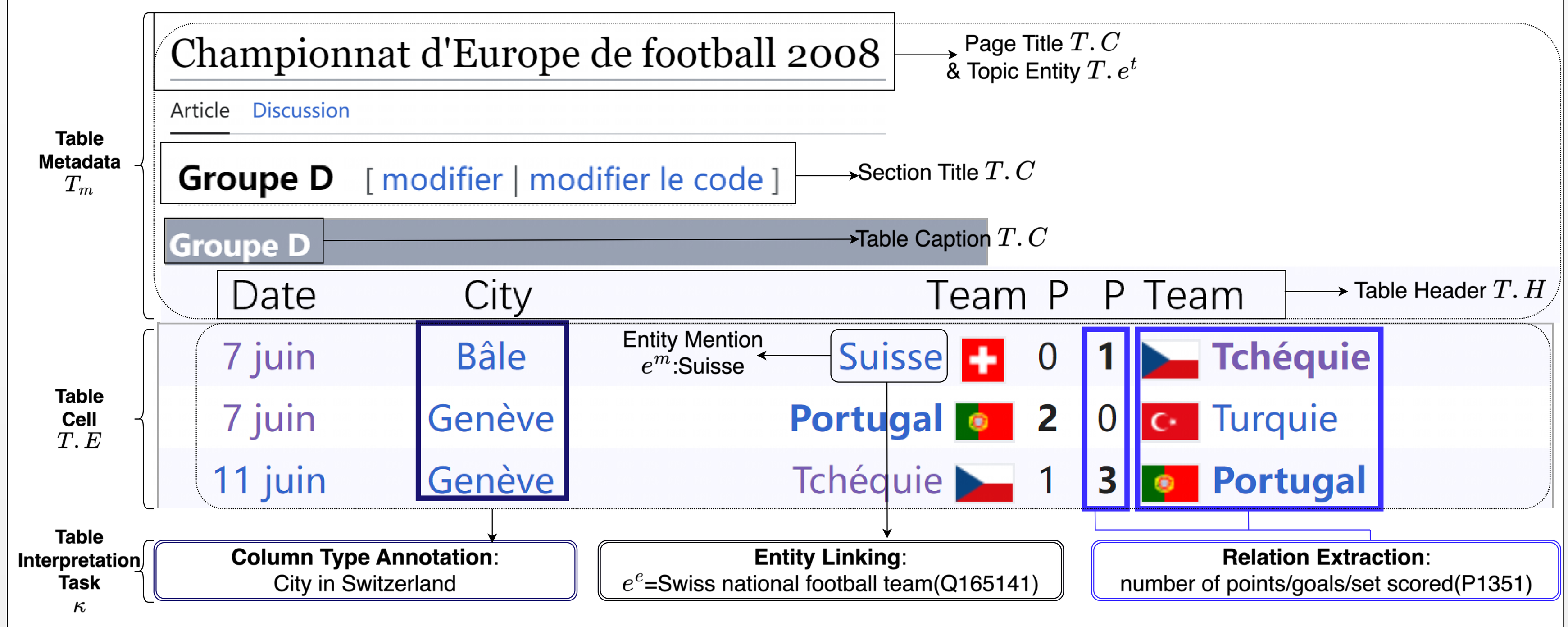
A. Introduction

A Real-World Case for Tabular Interpretation:

As shown in the figure below, the schema-free webtable T contain various metadata, columns and cells with hyperlink.

Column Type Annotation (**CTA**) refers to deciding the column type for column **CITY**;

Entity Linking(**EL**) refer to choosing the KG entity linked with cell **Suisse**. Relation Extraction(**RE**) refer to decide the KG relation for column pair (**Team-P**)



Questions:

- ➢ How can we retrieve related tables from a large amount of web table corpus $\mathcal{T}$?
- ➢ To annotate a cell/column/column-pair, how can we consider both **semantic** and **structural** similarity?
- ➢ How can we teach a LLM to rank and annotate web tables, without hallucination and numerous pre-training data?

B. Motivation

1. Pre-Ranking and Re-Ranking: Weak-to-Strong

Motivated from recommendation system, for a given table T, we apply light-weighted model G, M to retrieve related table set, as well as providing pre-ranking options; next, we apply LLM as a fine-grained selector for re-ranking.

2. Contrastive Learning:

We apply contrastive learning with Sentence-Bert like model, to quickly select top-k most possible options for a variety of schema-free table, consider semantic similarity.

3. Graph Structural Learning(GSL):

We transfer self-annotated tables T to graph G, and apply GCN to learn structural similarity for any given table pair (T_1, T_2).

4. Retrieval-Augmented LLM for re-ranking:

LLM only needs to consider top-k options from pre-ranking phase, and most-related demonstration, retrieved from related table set.

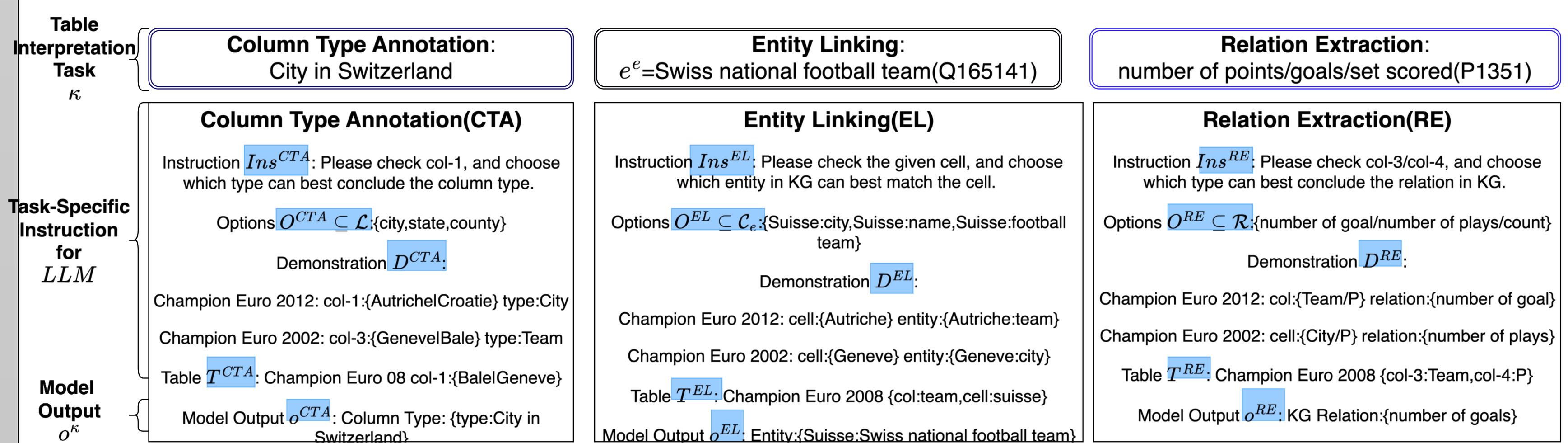
C. Problem Definition

1. Pre-Ranking Model($RAFL_{ret}$)

- **Input:** A schema-free web table $T \in \mathcal{T}$, an annotated training set T_{train}, a knowledge graph $\mathcal{G}$
- **Output**: Related table set $T_{related}$ with self-annotation; self-annotated pre-ranking top-k options O for T

2. Re-Ranking Model($RAFL_{rank}$ with LLM)

- **Input:** Specific task $\kappa \in \{CTA, RE, RL\}$, Instruction Ins^{κ} for task κ, demonstration D^{κ} from $T_{related}$, top-k options O^{κ} for T.
- **Output**: Selection $o^{\kappa} \in O^{\kappa}$ by LLM as re-ranking model.



D. Contribution

- **An unified framework RAFL for tabular interpretation learning**: RAFL handles information retrieval, self-supervised annotation and ranking procedure with state-of-the-art LLM-backboned model in a reliable manner.
- **A graph-enhanced retrieval system:** which can annotate and retrieve related table set, considering both semantic and structural similarity.
- **A two-stage ranking system with LLM:** transfer tabular interpretation task into a ranking problem, and apply RAG paradigm to alleviate LLM hallucination.
- **Comprehensive Experiment**: RAFL has both high precision and few-shot learning capability in various tasks, comparing to non-LLM and LLM solutions.

E. Retrieval System $RAFL_{ret}$

1. Bi-level Ranking Model M:

Given training set T_{train} of annotated tables, M can embed any table $T \in \mathcal{T}$ and task-specific information (e.g. column type $l \in \mathcal{L}$, relation type $r \in \mathcal{R}$) in unified embedding space. Sentence-Bert model M is fine-tuned with contrastive loss.

2. Self-Annotation:

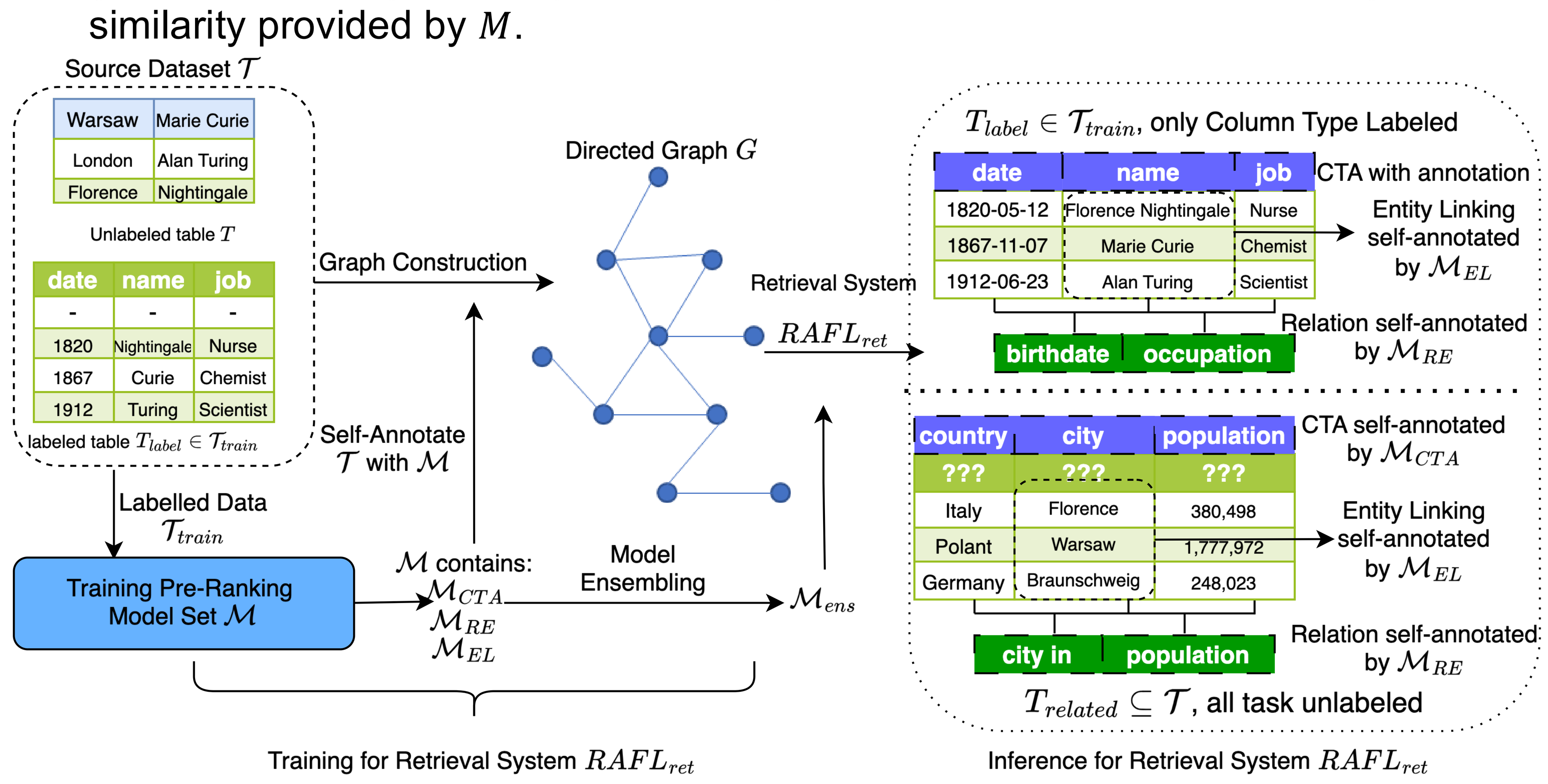
When training is finished, we obtain the task-specific ensembled model set: $M_{ens} = \{M_{CTA}, M_{RE}, M_{EL}\}$. Given T without annotation, we apply the ensembled M_{ens} to predict top-1 annotation for GSL, and top-k annotation for re-ranking.

3. Graph Structural Learning(GSL):

We leverage the annotation result of semantic type by M_{ens} to transfer all $T \in \mathcal{T}$ to a directed graph G. Such procedure refines various headers $T.H \in T$ to a limited pre-defined semantic type set $\mathcal{L}, \mathcal{R} \in \mathcal{G}$. After graph construction, we apply M to initialize the embedding, and apply GCN to further learn the structural information.

4. Similarity Calculation:

related table set $T_{related}$ are firstly selected from filter graph $G_{related}$ to reduce search space; then the similarity score is calculated by ranking the sum of graph embedding similarity provided by G , and the semantic embedding similarity provided by M.



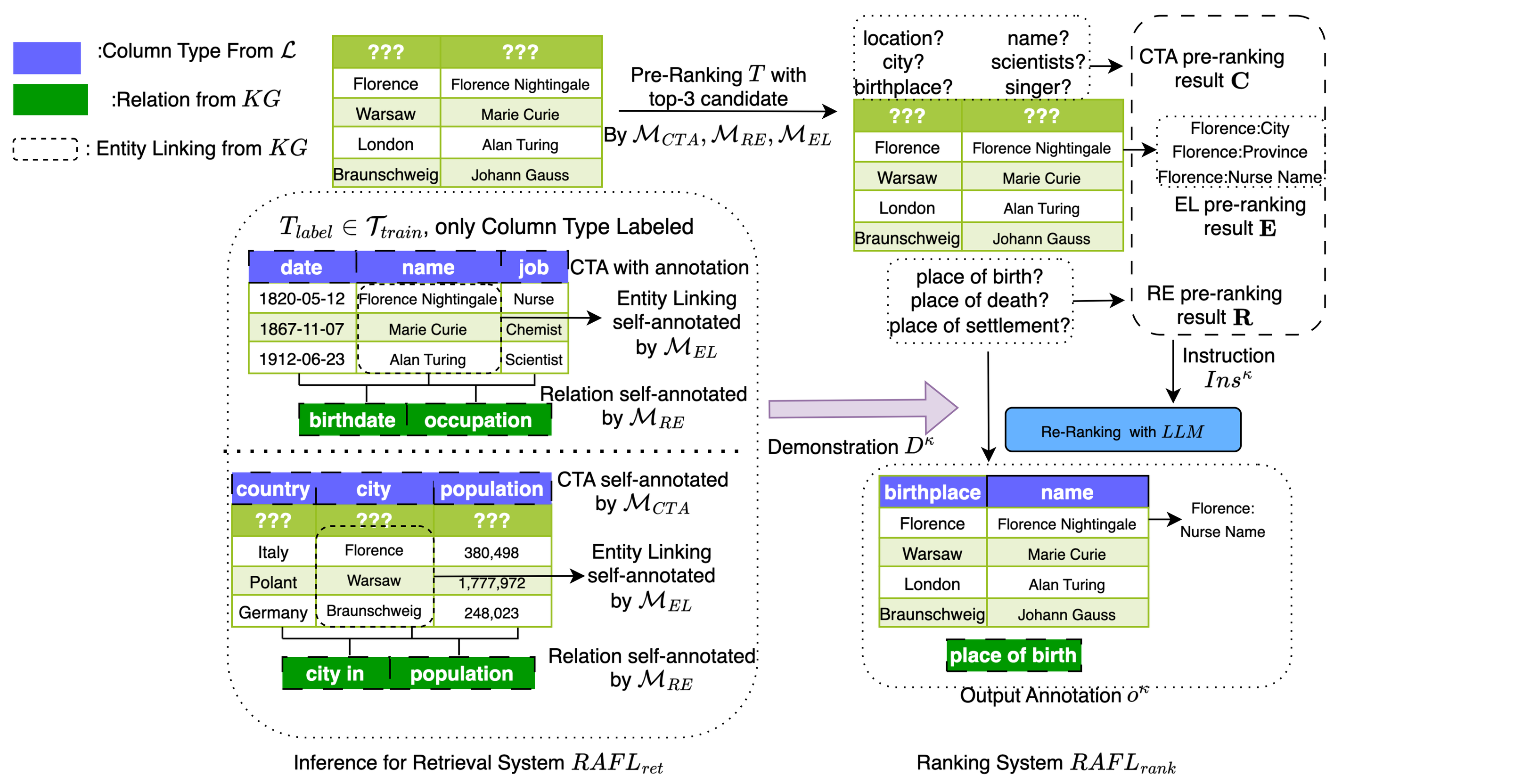
F. Re-Ranking System $RAFL_{rank}$

1. Avoiding Hallucination of LLM:

- LLM cannot select the correct annotation from hundreds of sematic type set $\mathcal{L} \cup \mathcal{R}$. (Limited Input Token Length)
- LLM cannot understand the meaning of each semantic type $l \in \mathcal{L}$ (resp. $r \in \mathcal{R}$) without demonstration.
- **Restrict Selection Domain**: to avoid hallucination, LLM is restricted to select from pre-ranking options O^{κ} from M_{ens}.
- **RAG Paradigm**: LLM is also provided with the most related self-annotated table corpus $T_{related}$ as task-specific demonstration, as illustration

2. LLM Fine-Tuning

To guarantee generation stability, the local LLM is fine-tuned with training data T_{train} with LoRA technique.



G. Experiment

- LLM-backboned model: Mistral-7B, Vicuna-13B; RAG Model: bge-large-en
- ☐ LLM is inherently suitable with few-shot scenario, without feature engineering.
- ☐ RAG significantly alleviate LLM hallucination, output structural prediction.
- ☐ Two-stage ranking strategy compensate the shortage of local LLM ability in understanding long-context multi-table data.

Table 2: Results of task CTA on dataset Semtab2019/WebTables

Model	Semtab2019		WebTables	
	Micro F1	Macro F1	Micro F1	Macro F1
Sherlock (100%)	0.646	0.440	0.844	0.670
TaBERT (100%)	0.768	0.413	0.896	0.650
TABBIE (100%)	0.799	0.607	0.929	0.734
DODUO (100%)	0.820	0.630	0.928	0.742
RECA(25%)	0.697	0.442	0.909	0.680
RAFL (25%)	0.861	0.743	0.963	0.825
RECA(100%)	0.853	0.674	0.937	0.783
RAFL (100%)	**0.875**	**0.766**	**0.967**	**0.834**

Table 4: Results of task RE and EL on dataset WikiGS

Model	WikiGS-RE		WikiGS-EL
	Micro F1	Macro F1	Accuracy
TURL(10%)	0.7350	0.3088	0.6055
RAFL (10%)	0.8930	0.8365	0.8705
TURL(25%)	0.8601	0.6755	0.7394
RAFL (25%)	0.9295	0.8642	0.8861
TURL(100%)	0.9025	0.8016	0.8420
RAFL (100%)	**0.9323**	**0.9153**	**0.9112**
GPT-4	0.5295	0.4326	0.9065

Model	Semtab2019		WebTables		WikiGS-RE	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
TableLLaMA(7B)	0.822	0.559	0.946	0.805	0.658	0.423
RAFL (Mistral-7B)	**0.862**	0.675	0.961	0.791	0.832	0.621
RAFL (Vicuna-13B)	0.861	**0.743**	**0.963**	**0.825**	**0.893**	**0.836**