Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

eXascale Infolab

Annotating Web Tables through Knowledge Bases: A Context-Based Approach

Yasamin Eslahi, Akansha Bhardwaj, Paolo Rosso,

Kurt Stockinger, Philippe Cudré-Mauroux

26 June 2020

Agenda

- Introduction
- Methodology
- Experiments
- Conclusions

Introduction

- Web table:
  - Structured source of information with relations and metadata
  - 154 million relational tables on web
  - 1.6 million relational tables on Wikipedia
- Knowledge base:
  - Contains rich information about entities, their semantic classes and their mutual relationships
- Web Table Annotation:
  - Disambiguation and annotation of the named entities in web tables
  - Linking to entities in web knowledge base
- Input for Information retrieval tasks

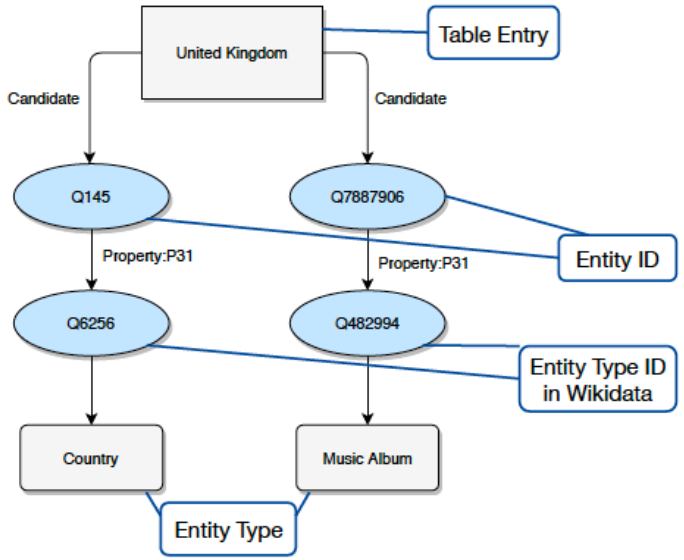
Introduction

- Web table Example:

Label Column

Reference Columns

No.	Country	Capital	Population	Government
1	France	Paris	67,088,000	Unitary semi-presidential republic
2	Germany	Berlin	83,149,300	Federal democratic parliamentary republic
3	India	New Delhi	1,358,036,300	Federal republic
4	Iran	Tehran	83,161,915	Islamic republic
5	Italy	Rome	60,252,824	Unitary parliamentary republic
6	Switzerland	Bern	8,586,550	Federal semi-direct democracy
7	United Kingdom	London	66,435,600	Unitary parliamentary constitutional monarchy



Methodology

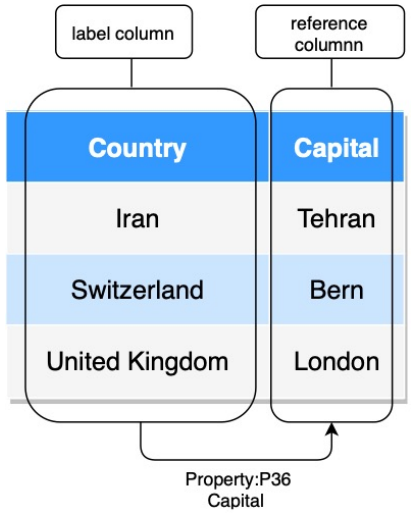
- Context lookup:
  - A systematic majority-based method
  - Uses all candidate entity types from the list of candidates from the surface form
- Looping:
  - An iterative graph-based approach
  - Uses contextual information of the embedding of the entities

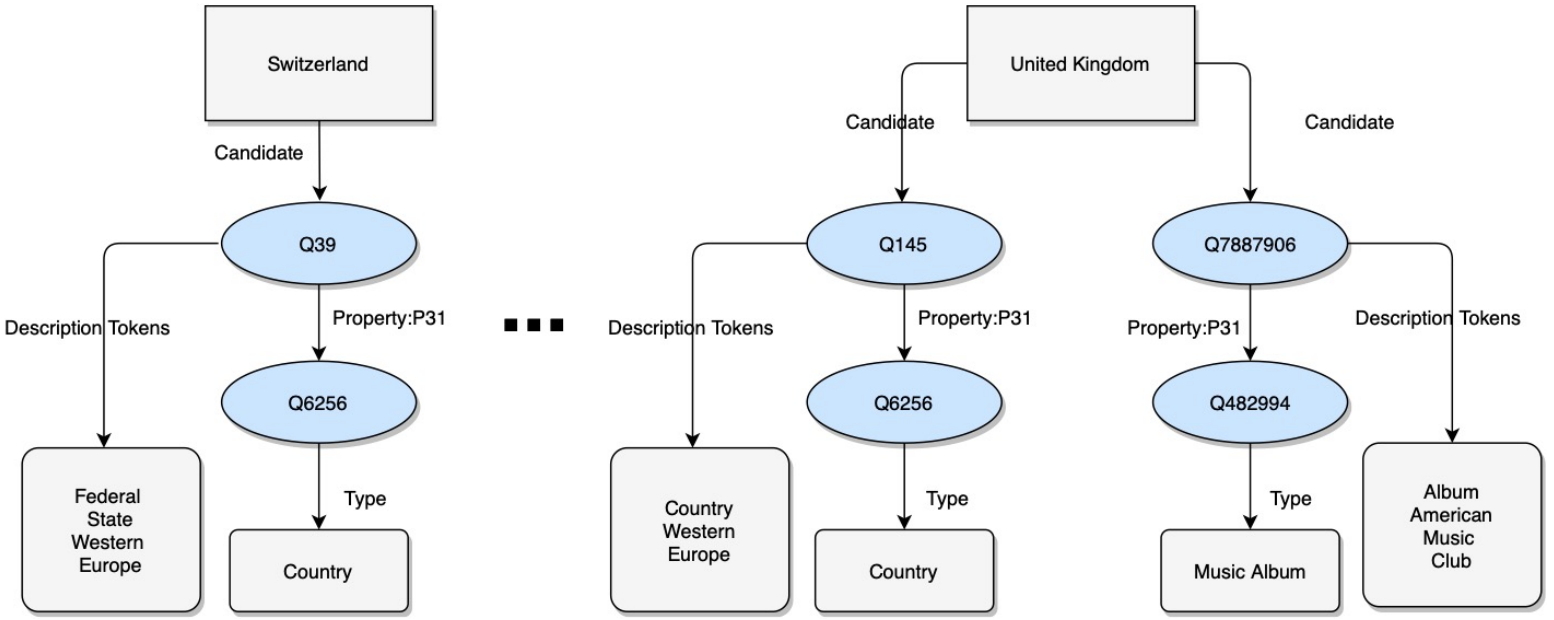
Methodology

- Context lookup:
  - First phase of the lookup-based methods:
    - Preprocessing the table
    - Label Column detection, relation extraction between columns
    - Get the type(s) and description tokens of all possible candidates of the entities
    - Get the top-n frequent types and top-n frequent description words for each column
  - Second phase of the lookup-based methods:
    - If only one result: Correct annotation
    - If more than one result: Repeat the candidate selection with the restriction of type(s) and description tokens
    - If no result: Use of edit distance function and relations

Methodology

No.	Country	Capital	Population
1	Iran	Tehran	83,161,915
2	Switzerland	Bern	8,586,550
3	United Kingdom	London	66,435,600





Most Frequent Types:

Country, City, Capital, ...

Frequent Description Tokens:

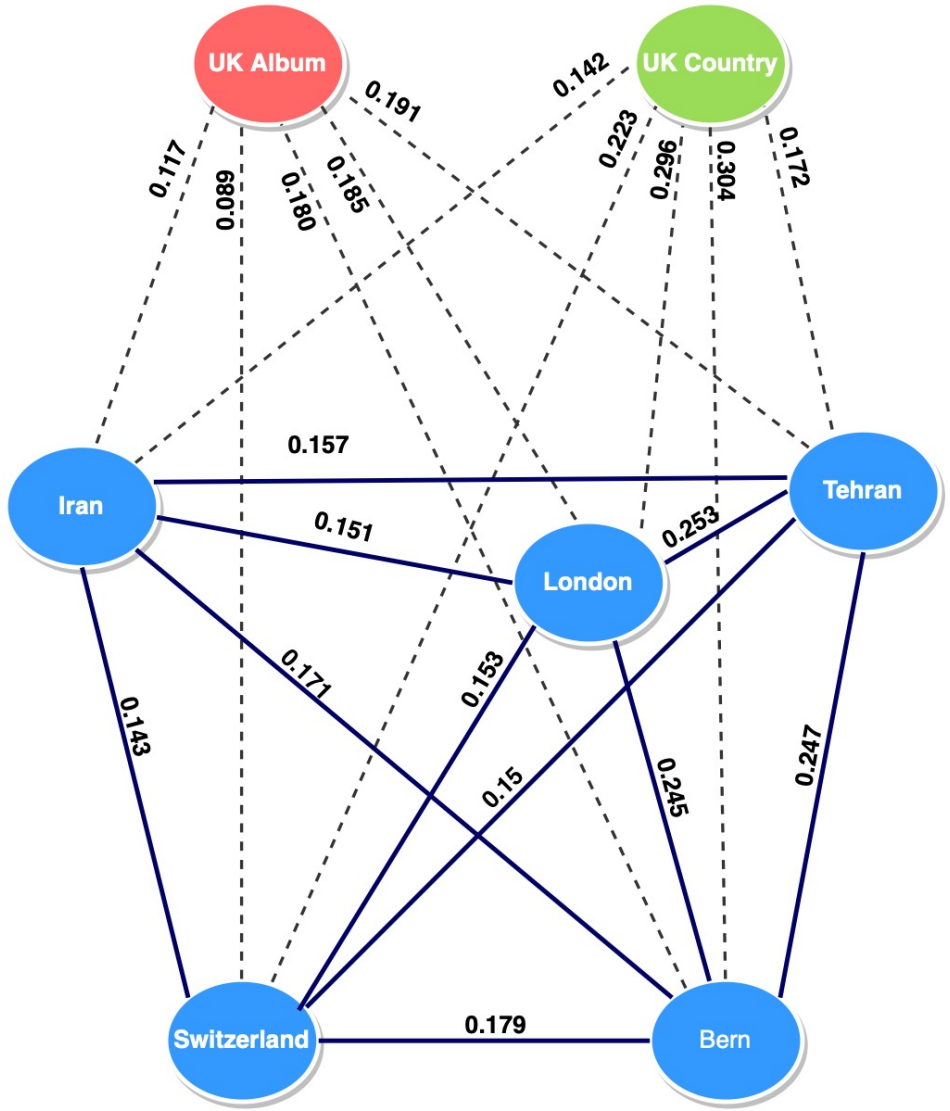
Country, State, City, Capital, Europe, ...

Methodology

- Looping method:
  - Get the table items' candidates
  - Add to the initial Looping graph if only one candidate
  - Keep in a list of ambiguous table items if more than one candidate
  - For all ambiguous items, iterate:
    - Apply ranking algorithm
    - Annotate with the candidate with the highest rank
    - Ensure the highest rank node has one of the top-5 types
    - Add annotation to the Looping graph

Methodology

- A 2-partite graph
- Two possible candidates for the UK
- Weights of the edges: the cosine similarity between vector representation of the nodes
- The nodes shown in blue are already disambiguated.



Datasets

- T2D:
  - 233 tables
  - Structuredness 0.97
  - Tables with relations 46 %
- Refined Limaye:
  - 399 tables
  - ~3,000 manually modified or newly added annotations

Experimental Results – Web Table Annotation

- Baselines:
  - Lookup-Baseline:
    - A text-based approach
    - Use of the types of the top entities and their descriptions in the knowledge base
    - Various search strategies on the basis of the number of candidates for each entity
  - Embedding-Baseline:
    - A graph-based approach
    - Use of contextual information of the embedding of the entities
    - Rank the nodes by ranking algorithm to disambiguate the entities

Rodriguez-Muro M. Christophides V. Efthymiou V., Hassanzadeh O. Matching web tables with knowledge base entities: From entity lookups to entity embeddings. d'Amato C. et al. (eds) The Semantic Web – ISWC 2017.

Experimental Results – Web Table Annotation

Method	T2D			Limaye		
	Pr	*Re*	*F1*	*Pr*	*Re*	*F1*
Lookup-Baseline	0.8784	0.7814	**0.827**	0.788	0.834	0.81
Context-Lookup	0.897	0.72	0.799	0.82	0.82	**0.82**

- Lookup-based methods:
  - T2D dataset:
    - Significant number of entities with unique candidates
  - Limaye dataset:
    - Difficult to find unique candidates
    - Different cell values in table and the entity names in knowledge base

Experimental Results – Web Table Annotation

Method	T2D			Limaye		
	Pr	*Re*	*F1*	*Pr*	*Re*	*F1*
Embedding-Baseline	0.62	0.70	0.66	0.76	0.82	0.79
Looping	0.85	0.82	**0.84**	0.82	0.87	**0.85**

- Semantic embedding methods:
  - Looping focuses on unambiguous entities
  - Leveraging type checking

Experimental Results – Web Table Annotation

Method	T2D		
	Pr	*Re*	*F1*
Looping – Initial Levenshtein	0.67	0.86	0.70
Looping – without Initial Levenshtein	0.84	0.80	**0.82**

- Looping with vs. without initial levenshtein distance:
  - Significant number of erroneous candidates from levenshtein distance search

Conclusions

- Investigating the problem of web table annotation
- Introducing two disambiguation methods for web table annotation, context-lookup and looping
- Context-lookup:
  - Using types and relations in the web table to select the best candidate for annotation
- Looping:
  - Creating weighted graphs to find the best candidate mapping between an entity in the web table and the corresponding entity in the KB
  - Exceeding the state of the art in web table annotation methods by up to 18%
- Refining the Limaye gold standard:
  - Manually correcting the wrong annotations
  - Adding the missing annotations to the dataset

Thank you!
Questions?

github.com/eXascaleInfolab/sds2020_web_table_annotation

www.linkedin.com/in/eslahi-yasamin