Auckland 2019
ISWC

Semantic Web Challenge on Tabular Data to KG Matching

Kavitha Srinivas, IBM Research, USA
Ernesto Jiménez-Ruiz, City, University of London, UK
Oktie Hassanzadeh, IBM Research, USA
Jiaoyan Chen, University of Oxford, UK
Vasilis Efthymiou, IBM Research, USA

Introduction

- Special OAEI track / ISWC challenge
- **Tabular data** in the form of **CSV files** is the common input format in a data analytics pipeline.
- **Tables on the Web** may also be the source of highly valuable data for web searches, question answering, and knowledge base (KB) construction.

Motivation

- The **lack of semantics and context in datasets** hinders their application.
- Gaining **semantic understanding** will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks.
- Understanding what the data is can help assess what sorts of **transformation are appropriate on the data**.

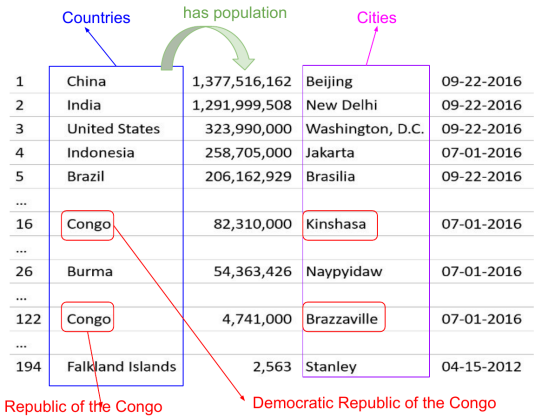
Adding Semantics to Tabular Data: Challenge Tasks

- Assigning a semantic type (e.g., a KG class) to an (entity) column (**CTA task**)
- Matching a cell to a KG entity (**CEA task**)
- Assigning a KG property to the relationship between two columns (**CPA task**)

(*) We assume the existence of a (possibly incomplete) ***Knowledge Graph (KG)*** *relevant to the domain.*

(**) We relied on DBpedia KG.

Adding Semantics to Tabular Data: Example



1	China	1,377,516,162	Beijing	09-22-2016
2	India	1,291,999,508	New Delhi	09-22-2016
3	United States	323,990,000	Washington, D.C.	09-22-2016
4	Indonesia	258,705,000	Jakarta	07-01-2016
5	Brazil	206,162,929	Brasilia	09-22-2016
...				
16	Congo	82,310,000	Kinshasa	07-01-2016
...				
26	Burma	54,363,426	Naypyidaw	07-01-2016
...				
122	Congo	4,741,000	Brazzaville	07-01-2016
...				
194	Falkland Islands	2,563	Stanley	04-15-2012

(*) Adapted from Efthymiou et al. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. ISWC 2017

Challenge Dates and Evaluation Rounds

- **Round 1**
  - April 15: opens / June 30: closes.
  - Best participants are invited to present during ISWC and OM.
- **Round 2**
  - July 17: opens / September 22: closes.
- **Round 3**
  - September 23: opens / October 14: closes.
- **Round 4**
  - October 15: opens / October 21: closes.

Evaluation Platform: AICrowd

The challenge run with the support of the **AICrowd platform**.
(Why not SEALS or HOBBIT?)

- ✓ Testing new platform
- ✓ Registration of participants
- ✓ Flexibility in the submission process
- ✓ Online leaderboards
- × Communication with participants
- × Deployment and problem-solving required AICrowd support

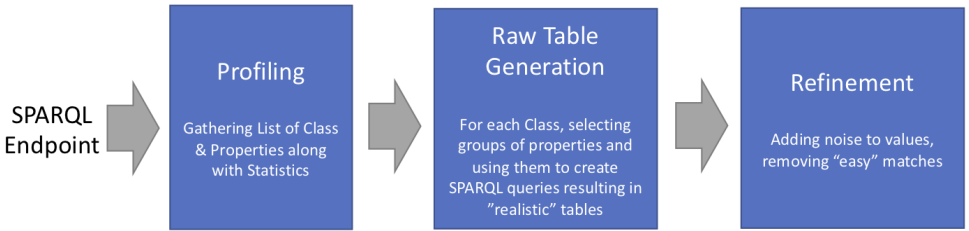
Datasets

- Round 1 (sandbox): extended **T2Dv2 dataset**
- Round 2 (fine-tuning): **Wikipedia tables** dataset + **automatically generated** dataset
- Round 3 (limited tests): **automatically generated** dataset
- Round 4 (limited tests): **automatically generated** dataset with only **hard cases**

Tables and ground truth for all rounds are made publicly available at:

https://doi.org/10.5281/zenodo.3518539

Automatic Dataset Generator



Automatic Dataset Generator - Issues

- **Profiling**
  - Detailed statistics can help create a more diverse corpus (e.g., fair coverage of classes with various levels of popularity)
  - Profiling within SPARQL could be hard to scale
- **Raw Table Generation**
  - The goal is creating SPARQL queries that produce ”realistic” looking tables.
  - There needs to be restrictions on the number of columns, number of rows, number of tables for a given class/property, etc.
- **Refinement**
  - Some instance values can be replaced in a rule-based fashion. E.g., first names of person entities can be abbreviated, synonyms can be used, the precision of numerical values can be adjusted, full dates can be replaced with months/years
  - Tables or rows/columns too “easy” for annotation (e.g., through exact match) can be dropped

Automatic Dataset Generator - Details

- **Profiling**
  - So far only getting a list of classes, properties, and the number of instances for each. Properties with a small number of instances are dropped
- **Raw Table Generation**
  - Each table has between 3-7 columns and 10-200 rows
  - There won't be more than 5 tables with the same set of properties
  - Header row is $(col1, \cdots, coln)$ i.e., property labels are *not* used as headers
- **Refinement**
  - Value refinement: only person name labels are adjusted
  - For Round 4: Subset of the dataset for which the simple lookup method of [1] returned low F-1 scores for the CEA task.
- **RDF Dataset for OM/OAEI**: Generated by [2] with an additional look-up extension

1. Efthymiou, Hassanzadeh, Rodriquez-Muro, Christophides. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. ISWC 2017

2. Efthymiou, Hassanzadeh, Sadoghi, Rodriquez-Muro. Annotating Web tables through ontology matching. OM 2016

Participation

- 7 systems stable across tasks and rounds
- Good starting to create community

#	Round 1	Round 2	Round 3	Round 4
Participants	17	11	9	8
CTA	13	9	8	7
CEA	11	10	8	8
CPA	5	7	7	7

Results Overview: Max Scores

- Standard **F1-score** for CEA, CPA and CTA (Round 1).
- CTA (Rounds 2-4) uses a score to take into account **approximate hits** of the (perfect) semantic type.

#	Round 1	Round 2	Round 3	Round 4
CTA	1.0	1.4	1.96	2.01
CEA	1.0	0.91	0.97	0.98
CPA	0.99	0.88	0.84	0.83

ISWC Challenge Presentation and Prizes

- ISWC challenge presentation on **Wednesday (11:40-12:40)**
- Prizes sponsored by **IBM Research** and **SIRIUS** (Norwegian Center for Scalable Data Access): `http://www.sirius-labs.no/`

IBM **Research**

SIRIUS

Proceedings

- **CEUR-WS**: ISWC Post-event proceedings.
- **November 10**: Final system paper submissions
- Papers:
  - Daniela Oliveira and Mathieu d'Aquin. **ADOG - Anotating Data with Ontologies and Graphs**.
  - Phuc Nguyen et al. **MTab: Matching Tabular Data to Knowledge Graph using Probability Models.**
  - Marco Cremaschi et al. **MantisTable: an automatic approach for the Semantic Table Interpretation.** (Team STI)
  - Avijit Thawani et al. **Entity Linking to Knowledge Graphs to Infer Column Types and Properties.** (Tabularisi)
  - Gilles Vandewiele et al. **ISWC Challenge: Transforming Tabular Data into Semantic Knowledge.** (IDLab)
  - Yoan Chabot et al. **DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System.**
  - Hiroaki Morikawa et al. **Semantic Table Interpretation using LOD4ALL.**

Challenge Talks

Challenge Presentation at ISWC:

- MTab
- Tabularisi
- Team STI
- Team DAGOBAH

Challenge Presentations at OM:

- Tabularisi
- IDLab

Problems, Feedback and Next Steps

- To be discussed during **OM panel session**
- Problems with dbpedia wikiredirects
- Encoding problems
- Errors in datasets (e.g., unexpected relationships, geonames)
- Maximum number of submissions per day
- Availability of GT
- AICrowd as platform
- RDF datasets

Acknowledgements

- All participants
- Challenge organisers and their institutions
- AICrowd and Arjun Nemani
- Our sponsors: IBM Research and SIRIUS
- ISWC and OM organisers