ISWC 2021

SemTab 3.0: Semantic Web Challenge on Tabular Data to KG Matching

Oktie Hassanzadeh and **Kavitha Srinivas**, IBM Research, USA
Jiaoyan Chen, University of Oxford, UK
Vasilis Efthymiou, ICS-FORTH, Greece
Ernesto Jiménez-Ruiz, City, University of London, UK and SIRIUS, University of Oslo, Norway
Juan Sequeda, data.world, USA
Vincenzo Cutrona, University of Milano - Bicocca, Italy
Daniela Oliveira and **Catia Pesquita**, Universidade de Lisboa, Portugal
Nora Abdelmageed, University of Jena, Germany
Madelon Hulsebos, University of Amsterdam, Netherlands

Organisers

O. Hassanzadeh

J. Chen

V. Efthymiou

K.Srinivas

E. Jiménez-Ruiz

J. Sequeda

V. Cutrona

D. Oliveira

C. Pesquita

N. Abdelmageed

M. Hulsebos

SemTab @ ISWC 2021

- Challenge overview - 10 min. live.
- Presentations (10 min each, recorded)
  - DAGOBAH
  - MTab
  - JenTab
- QA - 10 min. live
- Announcement of awards and wrap-up - 10 min. live.
- Poster sessions (wonder.me):
  - Today, Oct 27, 18:30-19:10 CET
  - Tomorrow, Oct 28, 15:00-15:30 CET

Motivation

- **Tabular data** in the form of **CSV files** is the common input format in a data analytics pipeline.
- Gaining **semantic understanding** will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks.
- Lack of a **systematic evaluation** framework of Semantic Web solutions.

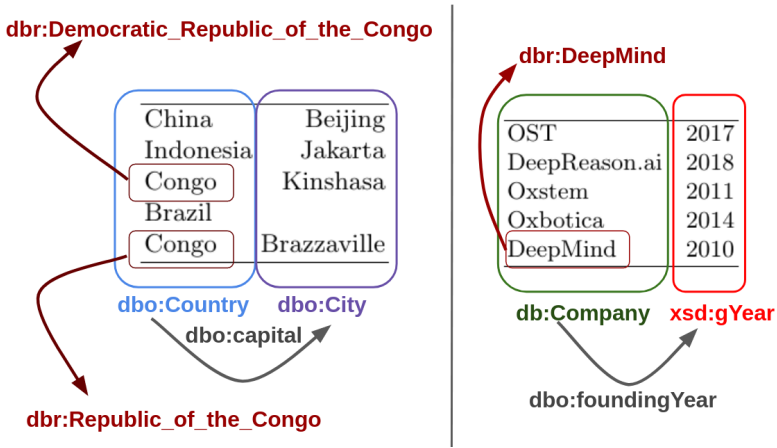
Adding Semantics to Tabular Data: Challenge Tasks

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

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

(**) In SemTab 2021: We relied on **Wikidata**, **DBPedia** and **Schema.org** as target KGs.

Adding Semantics to Tabular Data: Example



Rounds and Datasets (i)

Challenge Web: http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

- **Rounds 1:** CEA and CTA with Wikidata and DBpedia
  - Tough Tables (2T) dataset [1].
- **Round 2:** CEA, CTA and CPA with Wikidata.
  - Automatic generator [2]
  - Biotables

[1] Tough Tables: Carefully Evaluating Entity Linking for Tabular Data. International Semantic Web Conference (ISWC). 2020
[2] SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. Extended Semantic Web Conference (ESWC). 2020.

Rounds and Datasets (ii)

Challenge Web: http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

- **Round 3:**
  - BiodivTab: CEA and CTA with Wikidata [1]
  - GitTables: CTA with DBpedia and Schema.org [2]
  - Automatic generator [3]: CEA, CTA and CPA with Wikidata

[1] BiodivTab: https://github.com/fusion-jena/BiodivTab
[2] GitTables: A Large-Scale Corpus of Relational Tables: https://gittables.github.io/
[3] SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. Extended Semantic Web Conference (ESWC). 2020.

Participation

- The community seems active
- 6 core participants vs 8 in 2020
- More datasets in 2021

Participants	Round 1	Round 2	Round 3	Round 4
2019	*17*	*11*	*9*	*8*
2020	*18*	*16*	*18*	*10*
2021	11	7	6	-

Results Overview: Average F1-score

- Noise in **synthetic datasets** not challenging enough for top systems.
- The **2T** (R1) dataset is still challenging.
- **BiodivTab** (R3) brings an interesting challenge.
- **GitTables** (R3) is very challenging.
  - Brings a new type of task combining CTA and CPA.

Task	Round 1		Round 2		Round 3			
	2T-DBP	2T-WD	Auto	Bio	Auto	BioDiv	Git-DBP	Git-SCH
CEA	0.51	0.52	0.91	0.82	0.89	0.41	-	-
CTA	0.35	0.53	0.91	0.78	0.80	0.23	0.04	0.01
CPA	-	-	0.96	0.88	0.95	-	-	-

Results Overview: Top F1-score

– Results for top F1-score per task.

Task	Round 1		Round 2		Round 3			
	2T-DBP	2T-WD	Auto	Bio	Auto	BioDiv	Git-DBP	Git-SCH
CEA	0.95	0.92	0.99	0.97	0.97	0.60	-	-
CTA	0.46	0.83	0.98	0.96	0.99	0.59	0.1	0.21
CPA	-	-	0.99	0.95	0.99	-	-	-

Results Overview: Top System

- Top System per task:
  - DAGOBAH x 7
  - MTab x 6
  - JenTab x 2
  - KEPLER-aSI x 1
  - MantisTabe x 1

Task	Round 1		Round 2		Round 3			
	2T-DBP	2T-WD	Auto	Bio	Auto	BioDiv	Git-DBP	Git-SCH
CEA	DAGOBAH	DAGOBAH	MTab	DAGOBAH	DAGOBAH	JenTab	-	-
CTA	JenTab	DAGOBAH	MTab	MTab	DAGOBAH	KEPLER-aSI	DAGOBAH	MantisTable
CPA	-	-	MTab	MTab	MTab	-	-	-

Usability “Track" (Prize?)

New this year, addressing a pain point in the community regarding a lack of **publicly available easy-to-use and generic solution**.

Criteria:

- Is the solution open-source?
- Does the solution require specific platform that could affect its use in common settings?
- Does the solution require extensive training and tuning for a new application/domain?
- Is the solution offered as a public service?
- Does the solution include a well-designed user interface?

Results Overview: Usability

Out of the 6 core participants in SemTab 2021:

- 3 are open-source (1 Apache License)
- 1 has open/public service, very well-documented
- Varying complexity of requirements, but reasonable overall
- Most are pre-configured and can be used out of the box
- 4 have GUIs

Results and Proceedings

- CEUR WS proceedings
  - Preliminary version of papers online (prior light review): `https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/`
  - SemTab 2021: to appear
  - SemTab 2020: `http://ceur-ws.org/Vol-2775/`
  - SemTab 2019: `http://ceur-ws.org/Vol-2553/`
- Results: http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

What is coming next?

Recorded presentations:

- DAGOBAH
- MTab
- Jentab

QA after the presentations.

Please feel free to add your questions in chat or the slack channel.