ISWC 2020
NOVEMBER 1-6

SemTab 2.0: 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, ICS-FORTH, Greece
Vincenzo Cutrona, University of Milano - Bicocca, Italy

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.

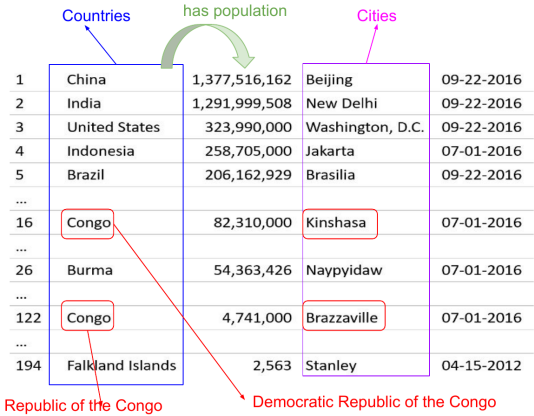
Adding Semantics to Tabular Data: Challenge Tasks

- 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 2020: We relied on **Wikidata KG** as target.

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

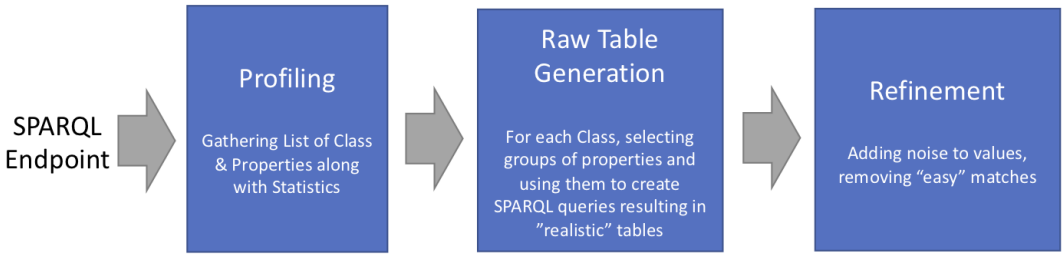
Rounds and Datasets

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

- **Rounds 1-3:**
  - run with the support of AICrowd
  - relied on an automatic dataset generator [1]
- **Round 4:**
  - blind round
  - combination of: (1) an automatically generated (AG) dataset, and (2) the Tough Tables (2T) dataset (CEA and CTA) [2]

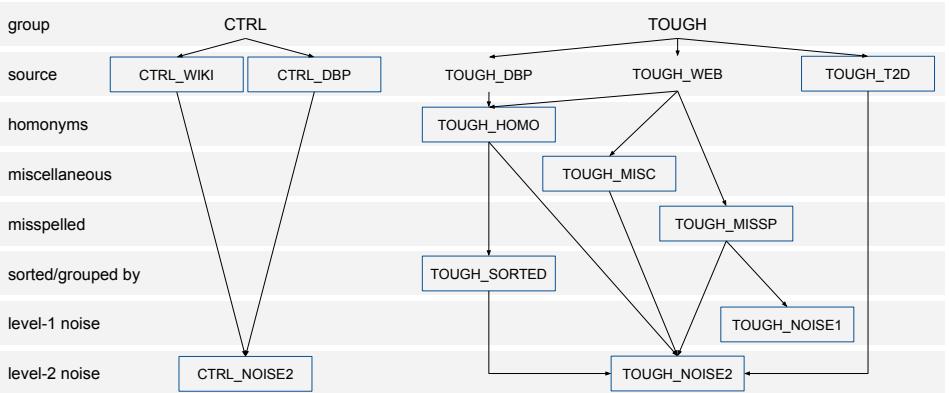
1. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. Extended Semantic Web Conference (ESWC). 2020.
2. Tough Tables: Carefully Evaluating Entity Linking for Tabular Data. International Semantic Web Conference (ISWC). 2020

Automatic Dataset Generator (AG)



Tough Tables (2T) Dataset

Semi-automatically created tables that aim at resembling **real-world scenarios**.



Rounds and Datasets

Stats	Automatically Generated				Tough Tables
	Round 1	Round 2	Round 3	Round 4 (AG)	Round 4 (2T)
Tables	34,295	12,173	62,614	22,207	180
Avg. rows	7.3	6.9	6.3	21	1,080
Avg. cols	4.9	4.6	3.6	3.5	4.5

Tables and ground truth:

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

Participation

– The community seems active and growing

Participants	Round 1	Round 2	Round 3	Round 4
2019	*17*	*11*	*9*	*8*
2020	18	16	18	10
CEA	10	10	9	10(-)
CTA	15	13(*)	16(+)	9(-)
CPA	9	11	8	7

Outliers:
(*) 3 systems with F-score < 0.3
(+) 8 systems with F-score < 0.3
(-) 1 system with F-score < 0.3

Results Overview: Average F1-score

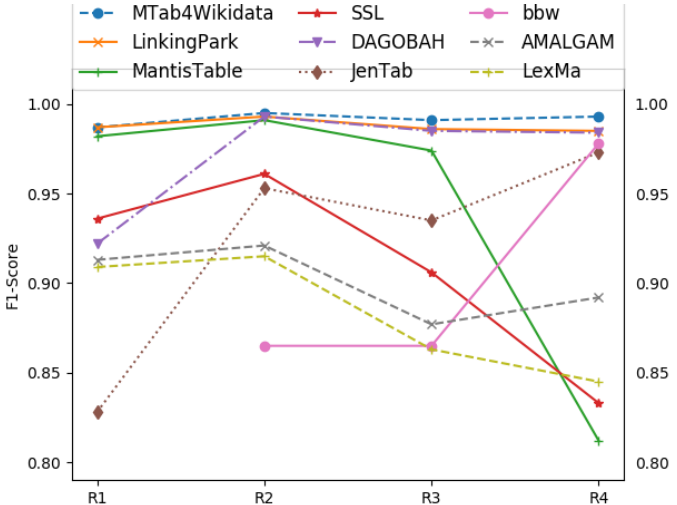
- Noise in synthetic datasets not challenging enough.
- The 2T dataset brings additional complexity.

Task	Automatically Generated				Tough Tables
	Round 1	Round 2	Round 3	Round 4 (AG)	Round 4 (2T)
CEA	0.93	0.95	0.94	0.92	0.54
CTA	0.83	0.93	0.94	0.92	0.59
CPA	0.93	0.97	0.93	0.96	-

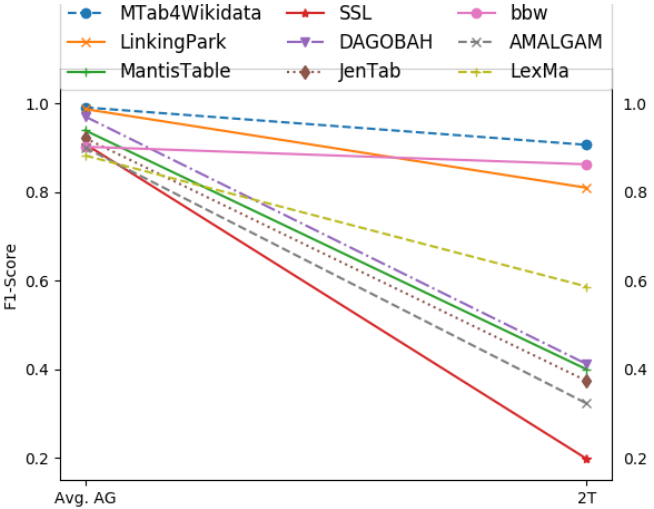
(*) Averages of top-10 without outliers

CEA (Cell-Entity Annotation) Results

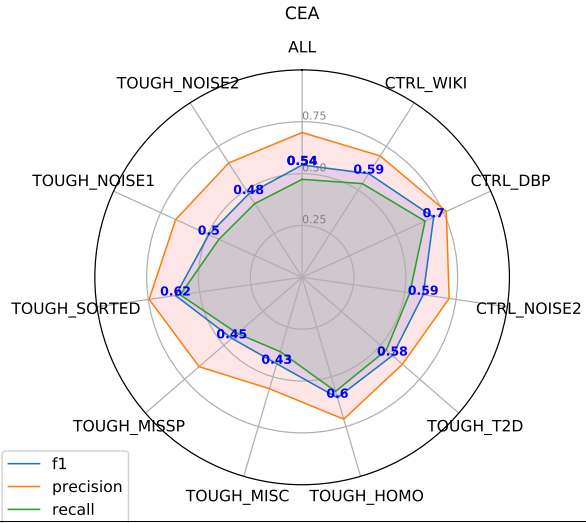
CEA: Automatically generated dataset



CEA: AG vs 2T


MTab4Wikidata
LinkingPark
MantisTable
SSL
DAGOBAH
JenTab
bbw
AMALGAM
LexMa
F1-Score
1.0
0.8
0.6
0.4
0.2
Avg. AG
2T

CEA: 2T results

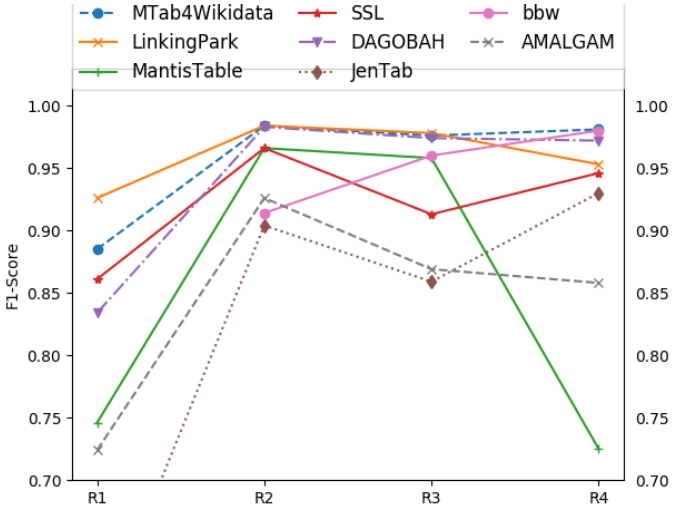


CEA: Knowledge Gap in 2T

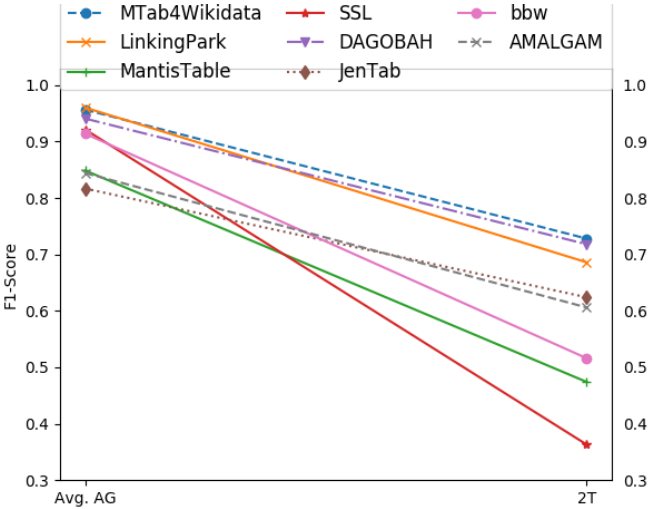
- Cells **without** an expected **KG correspondence**.
- **Synthetic transformations** of cells without a correspondence in Wikidata
- **3,588** out of 666,424 (0.5%) **cells**.
- Participants gave **preference to Recall**. All systems apart from SSL (F1-score 0.198) tried to provide an annotation for these cells.
- **Negative impact** to both CEA and CTA.

CTA (Column-Type Annotation) Results

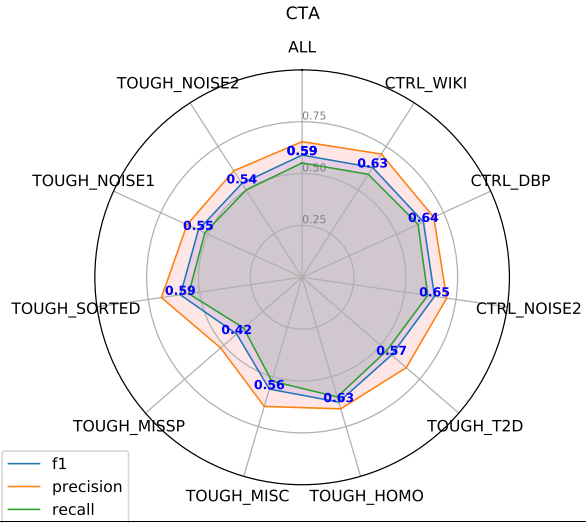
CTA: Automatically generated dataset



CTA: AG vs 2T

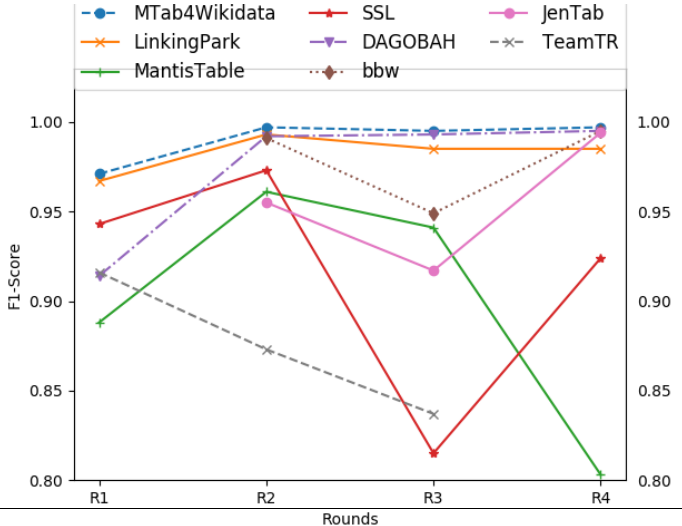

MTab4Wikidata
LinkingPark
MantisTable
SSL
DAGOBAH
JenTab
bbw
AMALGAM
F1-Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Avg. AG
2T

CTA: 2T results



CPA (Columns-Property Annotation) Results

CPA: Automatically generated dataset



Acknowledgements

- All participants
- Challenge organisers and their institutions
- AICrowd
- Our sponsor: IBM Research
- ISWC and OM organisers
- 2T contributors: Federico Bianchi and Matteo Palmonari

Conclusions

- Automatic generator allows to quickly create datasets.
- Automatically introduced noise was not challenging enough.
- 2T brings a very interesting complexity.
- 2T also introduced (synthetic) knowledge gap (*i.e.*, cells without an expected KG correspondence).

- 2021: seed the automatic generator with 2T
- 2021: target both Wikidata and DBpedia

SemTab Challenge Prizes

– Prizes sponsored by **IBM Research**

3rd Prize

3rd Prize

– **DAGOBAH** and **bbw**

2nd Prize

2nd Prize

- **LinkingPark**

1st Prize

1st Prize

- **MTab4Wikidata**

What is coming next?

- **2 sessions** with coffee break in between
  - **Session 7A:**
    - SSL team
    - LinkingPark
  - **Session 8B:**
    - MTab4Wikidata
    - DAGOBAH
    - MantisTable SE
- Q&A and Challenge discussion after the talks in each session
- Use the Q&A facility for questions
- Discussion after sessions in SLACK / REMO / personal ZOOM accounts