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

- **dbp:Democratic_Republic_of_the_Congo**
  - China
  - Indonesia
  - Congo
  - Brazil
  - Kinshasa
  - Brazzaville

- **dbp:DeepMind**
  - OST
  - DeepReason.ai
  - Oxstem
  - Oxbotica
  - DeepMind
  - 2017
  - 2018
  - 2011
  - 2014
  - 2010

- **dbp:Country**
- **dbp:City**
- **dbp:capital**
- **dbp:Republic_of_the_Congo**
- **dbp:Company**
- **xsd:gYear**
- **dbp:foundingYear**

International Semantic Web Conference, Virtual
**SemTab 3.0**: Semantic Web Challenge on Tabular Data to KG Matching
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

---

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
Participation

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

<table>
<thead>
<tr>
<th>Participants</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>17</td>
<td>11</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>2020</td>
<td>18</td>
<td>16</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>2021</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Task</th>
<th>Round 1</th>
<th></th>
<th>Round 2</th>
<th></th>
<th>Round 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2T-DBP</td>
<td>2T-WD</td>
<td>Auto</td>
<td>Bio</td>
<td>Auto</td>
<td>BioDiv</td>
</tr>
<tr>
<td>CEA</td>
<td>0.51</td>
<td>0.52</td>
<td>0.91</td>
<td>0.82</td>
<td>0.89</td>
<td>0.41</td>
</tr>
<tr>
<td>CTA</td>
<td>0.35</td>
<td>0.53</td>
<td>0.91</td>
<td>0.78</td>
<td>0.80</td>
<td>0.23</td>
</tr>
<tr>
<td>CPA</td>
<td>-</td>
<td>-</td>
<td>0.96</td>
<td>0.88</td>
<td>0.95</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Git-DBP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
</tr>
</tbody>
</table>
Results Overview: Top F1-score

- Results for top F1-score per task.

<table>
<thead>
<tr>
<th>Task</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2T-DBP</td>
<td>2T-WD</td>
<td>Auto</td>
</tr>
<tr>
<td>CEA</td>
<td>0.95</td>
<td>0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>CTA</td>
<td>0.46</td>
<td>0.83</td>
<td>0.98</td>
</tr>
<tr>
<td>CPA</td>
<td>-</td>
<td>-</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Results Overview: Top System

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

<table>
<thead>
<tr>
<th>Task</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2T-DBP</td>
<td>2T-WD</td>
<td>Auto</td>
</tr>
<tr>
<td>CEA</td>
<td>DAGOBAH</td>
<td>DAGOBAH</td>
<td>MTab</td>
</tr>
<tr>
<td>CTA</td>
<td>JenTab</td>
<td>MTab</td>
<td></td>
</tr>
<tr>
<td>CPA</td>
<td>DAGOBAH</td>
<td>MTab</td>
<td>DAGOBAH</td>
</tr>
<tr>
<td></td>
<td>MTab</td>
<td>MTab</td>
<td>KEPLER-aSI</td>
</tr>
<tr>
<td></td>
<td>MTab</td>
<td>MTab</td>
<td>DAGOBAH</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MantisTable</td>
</tr>
</tbody>
</table>
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/](https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/)
  - SemTab 2021: to appear
- Results: [http://www.cs.ox.ac.uk/isg/challenges/sem-tab/](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.