Preface SemTab 2019

SemTab 2019 [4, 3] was the first edition of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, successfully collocated with the 18th International Semantic Web Conference (ISWC) and the 14th Ontology Matching (OM) Workshop: `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/`

Description

Tabular data in the form of CSV files is the common input format in a data analytics pipeline. However a lack of understanding of the semantic structure and meaning of the content may hinder the data analytics process. Thus gaining this semantic understanding will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what the data is can help assess what sorts of transformation are appropriate on the data.[1]

Tables on the Web may also be the source of highly valuable data. The addition of semantic information to Web tables may enhance a wide range of applications, such as web search, question answering, and knowledge base (KB) construction.

Tabular data to Knowledge Graph (KG) matching is the process of assigning semantic tags from Knowledge Graphs (e.g., Wikidata or DBpedia) to the elements of the table. This task however is often difficult in practice due to metadata (e.g., table and column names) being missing, incomplete or ambiguous.

Tabular data to KG matching tasks typically include *(i)* cell to KG entity matching, *(ii)* column to KG class matching, and *(iii)* column pair to KG property matching.

There exist several approaches that aim at addressing one or several of above tasks and datasets with ground truths that can serve as benchmarks. Despite this significant amount of work, there was a lack of a common framework to conduct a systematic evaluation of state-of-the-art systems. The creation of SemTab aims at filling this gap and becoming the reference challenge in this community, in the same way the OAEI is for the Ontology Matching community.

The Challenge

The SemTab 2019 challenge started in mid April 2019 and closed in mid October 2019. It was organised into four evaluation rounds where we aimed at testing different

---

[1]AIDA project: `https://www.turing.ac.uk/research/research-projects/artificial-intelligence-data-analytics-aida`

Table 1: Participation in the SemTab 2019 challenge.

	Round 1	Round 2	Round 3	Round 4
Overall	17	11	9	8
CEA task	11	10	8	8
CTA Task	13	9	8	7
CPA task	5	7	7	7

datasets with increasing difficulty. Table 1 shows the participation per round. We had a total of 17 systems participating in Round 1. Round 2 had a reduction of participating systems (from 17 to 11), which helped us identify the core systems and groups actively working in tabular data to KG matching. Round 3 and Round 4 preserved the 7 core participants across rounds and all three tasks. SemTab 2019 core participants: *MTab* [6], *IDLab* [8], *Tabularisi* [9], *ADOG* [7], *DAGOBAH* [1], *Team_sti* [2], and *LOD4ALL* [5].

SemTab 2019 was successful as we managed to *(i)* create a small community around the challenge, *(ii)* advance the state of the art, *(iii)* gather feedback from the evaluation to improve future editions of the challenge, and *(iv)* release 4 generated benchmark datasets with ground truths [3].

Please refer to [4] for a full description of the SemTab 2019 challenge, datasets, evaluation and discussion.

Presentation

The results of the challenge were presented during ISWC 2019. Four participating teams also presented their systems.

- **MTab: Matching Tabular Data to Knowledge Graph using Probability Models** by Phuc Nguyen, Natthawut Kertkeidkachorn, Ryutaro Ichise and Hideaki Takeda.

- **Entity Linking to Knowledge Graphs to Infer Column Types and Properties** *(Tabularisi)* by Avijit Thawani, Minda Hu, Erdong Hu, Husain Zafar, Naren Teja Divvala, Amandeep Singh, Ehsan Qasemi, Pedro Szekely and Jay Pujara.

- **MantisTable: an automatic approach for the Semantic Table Interpretation** *(Team_sti)* by Marco Cremaschi, Roberto Avogadro, and David Chieregato.

- **DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System** by Yoan Chabot, Thomas Labbe, Jixiong Liu and Raphaël Troncy.

We also had a devoted session during the Ontology Matching workshop where we described the challenge and had a system presentation:

- **Transforming tabular data into semantic knowledge** *(IDLab)* by Gilles Vandewiele, Bram Steenwinckel, Filip De Turck, Femke Ongenae.

Slides and photos are available on the challenge website: `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/`

Prizes

SIRIUS[2] and IBM Research[3] sponsored the prizes for the best systems in the challenge. This sponsorship was important not only for the challenge awards, but also because it shows a strong interest from industry.

- **1st Prize** (CTA, CEA and CPA): MTab Team.

- **2nd Prize** (CTA, CEA and CPA): IDLab Team.

- **3rd Prize** (CTA, CEA and CPA): Tabularisi Team.

- **3rd Prize** (CEA): ADOG Team.

- **Outstanding Improvement** (CEA): Team STI.

Organizing committee

Challenge Chairs

- Kavitha Srinivas (IBM Research): `Kavitha.Srinivas@ibm.com`

- Ernesto Jimenez-Ruiz (City, University of London; University of Oslo): `ernesto.jimenez-ruiz@city.ac.uk`

- Jiaoyan Chen (University of Oxford): `jiaoyan.chen@cs.ox.ac.uk`

- Oktie Hassanzadeh (IBM Research): `hassanzadeh@us.ibm.com`,

- Vasilis Efthymiou (IBM Research): `Vasilis.Efthymiou@ibm.com`

Challenge committee members

- Udayan Khurana (IBM Research)

- Erik Bryhn Myklebust (University of Oslo)

- Monika Solanki (Agrimetrics)

- Ole Magnus Holter (University of Oslo)

- Pedro Szekely (University of Southern California)

- Basil Ell (University of Bielefeld; University of Oslo)

- Marco Cremaschi (University of Milano - Bicocca)

- Asan Agibetov (Medical University of Vienna)

---

[2]SIRIUS: Norwegian Centre for Research-driven Innovation: `https://sirius-labs.no`
[3]`https://www.research.ibm.com/`

Support

References

[1] Y. Chabot, T. Labbe, J. Liu, and R. Troncy. DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.

[2] M. Cremaschi, R. Avogadro, and D. Chieregato. MantisTable: an automatic approach for the Semantic Table Interpretation. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.

[3] O. Hassanzadeh, V. Efthymiou, J. Chen, E. Jiménez-Ruiz, and K. Srinivas. SemTab2019: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching - 2019 Data Sets. `https://doi.org/10.5281/zenodo.3518539`, 2019.

[4] E. Jimenez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, and K. Srinivas. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In *The Semantic Web: ESWC 2020*. Springer International Publishing, 2020.

[5] H. Morikawa. Semantic Table Interpretation using LOD4ALL. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.

[6] P. Nguyen, N. Kertkeidkachorn, R. Ichise, and H. Takeda. MTab: Matching Tabular Data to Knowledge Graph using Probability Models. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.

[7] D. Oliveira and M. d'Aquin. ADOG - Anotating Data with Ontologies and Graphs. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.

[8] B. Steenwinckel, G. Vandewiele, F. De Turck, and F. Ongenae. CSV2KG: Transforming Tabular Data into Semantic Knowledge. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.

[9] A. Thawani, M. Hu, E. Hu, H. Zafar, N. T. Divvala, A. Singh, E. Qasemi, P. Szekely, and J. Pujara. Entity Linking to Knowledge Graphs to Infer Column Types and Properties. In *SemTab, ISWC Challenge*, volume 2553. CEUR-WS.org, 2019.