News (22/11/2020): Datasets are now available in zenodo. Back to other SemTab editions.

SemTab 2020: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

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.

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.

The SemTab challenge aims at benchmarking systems dealing with the tabular data to KG matching problem, so as to facilitate their comparison on the same basis and the reproducibility of the results.

The **2020 edition** of this challenge will be collocated with the 19th International Semantic Web Conference and the 15th International Workshop on Ontology Matching.

We have a discussion group for the challenge where we share the latest news with the participants and we discuss issues risen during the evaluation rounds.

Datasets and Evaluator

The challenge datasets and ground truths are now open:

- Synthetic dataset: DOI 10.5281/zenodo.4282879
- Tough Tables dataset: DOI 10.5281/zenodo.4246370

The codes of the AICrowd evaluator are also available here.

The target Knowledge Graph in SemTab 2020 is Wikidata. Wikidata Truthy Dump (April 24, 2020): DOI 10.5281/zenodo.4282941

Datasets per round:

- **Round 1:** tables, cta targets, cea targets, and cpa targets.
- **Round 2:** tables, cta targets, cea targets, and cpa targets.
- **Round 3:** tables, cta targets, cea targets, and cpa targets.
- **Round 4:** tables, cta targets, cea targets, cpa targets, and gt examples.

Results and Challenge Prizes

Results of all four rounds available here. Summary of SemTab 2020 results.

SemTab-2020 slides presented during the ISWC conference.

Prizes sponsored by IBM Research:

- **1st Prize**: MTab4Wikidata Team.
- **2nd Prize**: LinkingPark Team.
- **3rd Prize**: DAGOBAH Team and bbw Team.

System Papers

Papers published in the Vol-2775 of CEUR Workshop Proceedings.

- Donguk Kim, Heesung Park, Jae Kyu Lee, Wooju Kim. **Generating conceptual subgraph from tabular data for knowledge graph matching** (SSL team). (paper)
- Shuang Chen, Alperen Karaoglu, Carina Negreanu, Tingting Ma, Jin-Ge Yao, Jack Williams, Andy Gordon, Chin-Yew Lin. **LinkingPark: An integrated approach for Semantic Table Interpretation.** (paper)
- Phuc Nguyen, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, Hideaki Takeda. **MTab4Wikidata at the SemTab 2020: Tabular Data Annotation with Wikidata.** (paper)
- Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Thomas Labbé, Pierre Monnin, and Raphaël Troncy. **DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data** (paper)
- Marco Cremaschi, Roberto Avogadro, Andrea Barazzetti, David Chieregato. **MantisTable SE: an enhanced and efficient approach to a complete Semantic Table Interpretation.** (paper)
- Renat Shigapov, Philipp Zumstein, Jan Kamlah, Lars Oberländer, Jörg Mechnich, Irene Schumm. **bbw: Matching CSV to Wikidata via meta-lookup.** (paper)
- Nora Abdelmageed, Sirko Schindler. **JenTab: Matching Tabular Data to Knowledge Graphs.** (paper)
- Gayo Diallo, Rabia Azzi. **AMALGAM: A Matching Approach to fairfy tabular data with knowledge graph model.** (paper)
- Wiem Baazouzi, Marouen Kachroudi, Sami Faiz. **Kepler-aSI : KeplerAs A Semantic Interpreter.** (paper)
- Shalini Tyagi, Ernesto Jiménez-Ruiz. **LexMa: Tabular Data to Knowledge Graph Matching using Lexical Techniques.** (paper)
- Semih Yumusak. **Knowledge graph matching with inter-service information transfer** (TeamTR team). (paper)

ISWC Challenge Presentations

The results of the challenge will be presented on November 5 in two sessions. See full ISWC program here. Five participating teams will also present their systems.

Session 7A (EST: 10:20-11:20. CET: 16:20-17:20. CST: 23:20-00:20):

- Challenge overview & announcement of awards. (slides)
- Generating conceptual subgraph from tabular data for knowledge graph matching (SSL team) by Donguk Kim, Heesung Park, Jae Kyu Lee, Wooju Kim. (video slides)
- LinkingPark: An integrated approach for Semantic Table Interpretation by Shuang Chen, Alperen Karaoglu, Carina Negreanu, Tingting Ma, Jin-Ge Yao, Jack Williams, Andy Gordon, Chin-Yew Lin. (video slides)

Session 8B (EST: 12:00-13:00. CET: 18:00-19:00. CST: 01:00-02:00):

- MTab4Wikidata at the SemTab 2020: Tabular Data Annotation with Wikidata by Phuc Nguyen, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, Hideaki Takeda. (video slides)
- DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data by Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Thomas Labbé, Pierre Monnin, and Raphaël Troncy (video slides)
- MantisTable SE: an enhanced and efficient approach to a complete Semantic Table Interpretation by Marco Cremaschi, Roberto Avogadro, Andrea Barazzetti, David Chieregato (video slides)
- Closing and Discussion

There will also be a slot devoted to SemTab systems during the Ontology Matching workshop on November 2.

- JenTab: Matching Tabular Data to Knowledge Graphs by Nora Abdelmageed, Sirko Schindler (video slides)
- AMALGAM: A Matching Approach to fairfy tabular data with knowledge graph model by Gayo Diallo, Rabia Azzi (video slides)

Participation: forum and registration

We have a discussion group for the challenge where we share the latest news with the participants and we discuss issues risen during the evaluation rounds.

Please register your system using this google form.

Note that participants can join SemTab at any Round for any of the tasks.

Challenge Tasks

The challenge includes the following tasks organised into several evaluation rounds:

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

The challenge will be run with the support of the AICrowd platform.

Round 4 submission: https://tinyurl.com/semtab2020-round4

Important Dates

- **May 26:** Round 1 opens.
- **July 21:** Round 1 closes.
- **July 31:** Round 2 opens.
- **September 08:** Round 2 closes.
- **September 3:** Best participants in Rounds 1 and 2 are invited to present their results during the ISWC conference and the Ontology Matching workshop.
- **September 17:** Round 3 opens.
- **September 27:** Round 3 closes.
- **October 20:** System paper submissions (preliminary version, e.g., system_name_prelim.pdf). Please use this form.
- **October 20:** Round 4 opens.
- **October 27 (23:59 AoE):** Round 4 closes. Submission: https://tinyurl.com/semtab2020-round4.
- **November 2:** Ontology Matching workshop.
- **November 5:** Challenge Presentation and prize announcement.
- **November 15:** System paper submissions (final version, e.g., system_name_final.pdf). Please use this form.

System Papers

We encourage participants to submit a system paper. The paper should be no more than 10 pages long and formatted using the LNCS Style. System papers will be revised by 1-2 challenge organisers. System papers will be published as a volume of CEUR-WS. By submitting a paper, the authors accept the CEUR-WS publishing rules.

Organisation

This challenge is organised by Kavitha Srinivas (IBM Research), Ernesto Jimenez-Ruiz (City, University of London; University of Oslo), Oktie Hassanzadeh (IBM Research), Jiaoyan Chen (University of Oxford), Vasilis Efthymiou (IBM Research), and Vincenzo Cutrona (University of Milano - Bicocca). If you have any problems working with the datasets or any suggestions related to this challenge, do not hesitate to contact us via the discussion group.

Acknowledgements

The challenge is currently supported by the SIRIUS Centre for Research-driven Innovation and IBM Research.