Results of SemTab 2020*

Ernesto Jiménez-Ruiz[1,2], Oktie Hassanzadeh[3], Vasilis Efthymiou[4], Jiaoyan Chen[5], Kavitha Srinivas[3], and Vincenzo Cutrona[6]

[1] City, University of London, UK. ernesto.jimenez-ruiz@city.ac.uk
[2] SIRIUS, University of Oslo, Norway. ernestoj@uio.no
[3] IBM Research, USA. hassanzadeh@us.ibm.com, kavitha.srinivas@ibm.com
[4] ICS-FORTH, Greece. vefthym@ics.forth.gr
[5] University of Oxford, UK. jiaoyan.chen@cs.ox.ac.uk
[6] Università degli Studi di Milano - Bicocca, Italy. vincenzo.cutrona@unimib.it

Abstract. SemTab 2020 was the second edition of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, successfully collocated with the 19th International Semantic Web Conference (ISWC) and the 15th Ontology Matching (OM) Workshop. SemTab provides a common framework to conduct a systematic evaluation of state-of-the-art systems.

Keywords: Tabular data · Knowledge Graphs · Matching · Semantic Table Interpretation

1 Motivation

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 (CEA task), *(ii)* column to KG class matching (CTA task), and *(iii)* column pair to KG property matching (CPA task).

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

Table 1: Statistics of the datasets in each SemTab 2020 round.

	Automatically Generated (AG)				Tough Tables (2T)
	Round 1	**Round 2**	**Round 3**	**Round 4**	**Round 4**
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

There existed several approaches that aim at addressing one or several of above tasks and datasets with ground truths that can serve as benchmarks (*e.g.*, [14, 13]). 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[2] [11] aimed at filling this gap and becoming the reference challenge in this community, in the same way the OAEI[3] is for the Ontology Matching community.[4]

2 The Challenge

The SemTab 2020 challenge started on May 26 and closed on October 27. The target KG in this edition was Wikidata [18]:

- Wikidata Dump (April 24, 2020): `https://doi.org/10.5281/zenodo.4282941`

SemTab 2020 was organised into four evaluation rounds where we aimed at testing different datasets with increasing difficulty. Rounds 1-3 were run with the support of Alcrowd[5], which provided an automatic evaluation of the submitted solutions, and relied on an automatic dataset generator [11]. Round 4 was a blind round (*i.e.*, no evaluation of submissions via Alcrowd) combining: *(i)* an automatically generated (AG) dataset as in previous rounds, and *(ii)* the Tough Tables (2T) dataset for CEA and CTA [7]. Table 1 provides a summary of the statistics of the datasets used in each round. Both the AG datasets and the 2T dataset are available in Zenodo [9, 6]:

- AG datasets: `https://doi.org/10.5281/zenodo.4282879`
- 2T dataset: `https://doi.org/10.5281/zenodo.4246370`

Table 2 shows the participation per round. We had a total of 28 different systems[6] participating across the four rounds; however only 11 systems have produced results in 3 or more rounds, which we identify as the SemTab 2020 core participants:[7] *MTab4Wikidata* [15], *LinkingPark* [3], *DAGOBAH* [10], *bbw* [16], *JenTab* [1], *ManstisTable SE* [4], *AMALGAM* [8], *SSL* [12], *LexMa* [17], *Kepler-aSI* [2], and *TeamTR* [19].

[2] `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/`
[3] `http://oaei.ontologymatching.org/`
[4] `http://ontologymatching.org/`
[5] `https://www.aicrowd.com/`
[6] At least 28 different Alcrowd submissions.
[7] These participants also submitted a system paper to the challenge

Table 2: Participation in the SemTab 2020 challenge. Outliers (F-score < 0.3): † 3 systems, ‡ 8 systems, § 1 system.

	Round 1	Round 2	Round 3	Round 4-AG	Round 4-2T
Total 2019	17	11	9	8	-
Total 2020	18	16	18	10	9
CEA	10	10	9	9	9
CTA	15	13†	16‡	9§	8
CPA	9§	11	8	7	-

2.1 Evaluation measures

Systems were requested to submit a single solution for each of the provided targets in each of the tasks. For example, a target cell in the CEA task is to be annotated with a single entity in the target KG.

The evaluation measures for CEA and CPA are the standard Precision, Recall and F1-score, as defined in Equation 1:

$$P = \frac{|\text{Correct Annotations}|}{|\text{System Annotations}|}, \quad R = \frac{|\text{Correct Annotations}|}{|\text{Target Annotations}|}, \; F1 = \frac{2 \times P \times R}{P + R} \tag{1}$$

where target annotations refer to the target cells for CEA and the target column pairs for CPA. An annotation is regarded to be correct, if it is contained in the ground truth. Note that it is possible that one target cell or column pair may have multiple annotations in the ground truth.

For the evaluation of CTA, we used approximations of Precision and Recall, by adapting their numerators to consider as partially correct annotations, columns annotated with one of the ancestors or descendants of the ground truth (GT) classes. In that sense, we define the correctness score $cscore$ of a CTA annotation α as

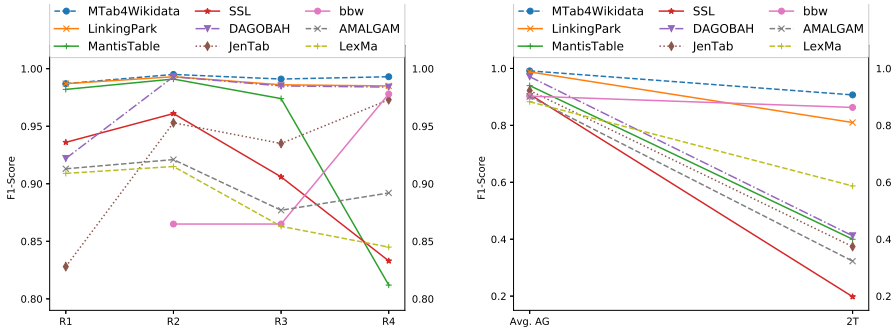
$$\text{cscore}(\alpha) = \begin{cases} 0.8^{d(\alpha)}, & \text{if } \alpha \text{ is in GT, or an ancestor of the GT,} \\ 0.7^{d(\alpha)}, & \text{if } \alpha \text{ is a descendant of the GT,} \\ 0, & \text{otherwise;} \end{cases} \tag{2}$$

where $d(\alpha)$ is the shortest distance to one of the ground truth classes. For example, $d(\alpha) = 0$ if α is a class in the ground truth, and $d(\alpha) = 2$ if α is a grandparent of a class in the ground truth. In the former case, the scoring of α will be 1, while in the later 0.64. Then, our approximations of Precision (AP), Recall (AR), and F1-score (AF1) for the evaluation of CTA are computed as follows:

$$AP = \frac{\sum cscore(\alpha)}{|\text{System Annotations}|}, \quad AR = \frac{\sum cscore(\alpha)}{|\text{Target Annotations}|}, \; AF1 = \frac{2 \times AP \times AR}{AP + AR} \tag{3}$$

Table 3: Average F1-score for the Top-10 systems discarding outliers.

	Automatically Generated (AG)				Tough Tables (2T)
	Round 1	**Round 2**	**Round 3**	**Round 4**	**Round 4**
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	-



(a) AG Datasets

(b) Average AG vs 2T

Fig. 1: Results in the CEA task with the Automatically Generated (AG) and Tough Tables (2T) Datasets.

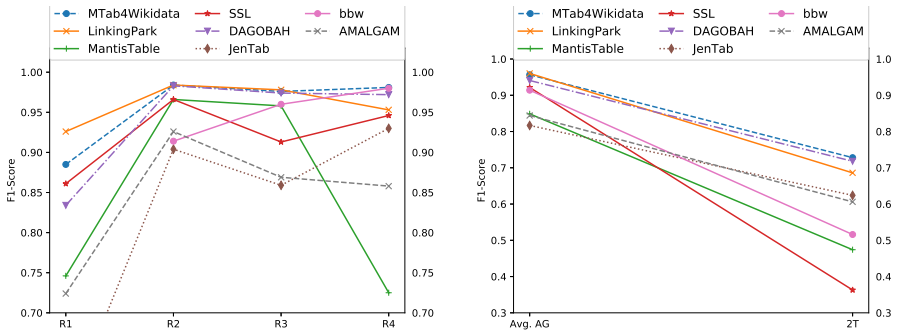
2.2 Results

Table 3 shows the average F1-score for the top-10 systems after discarding outliers (*i.e.*, systems or submissions with very low performance). We can observe that the complexity brought by the Tough Tables dataset was significant with respect to the previous rounds in terms of average results. Note that Round 4 was blind.

CEA task. Figure 1a shows the results for the CEA task over the AG datasets. The dataset in each round aimed at bringing new challenges. This is somehow reflected in MantisTable, LinkingPark, LexMa and SSL. Nevertheless, the overall results for the AG datasets where very positive. In contrast, as shown in Figure 1b, the performance over the 2T dataset is significantly reduced where only three systems (MTab4Wikidata, bbw and Linkingpark) managed to maintain and F1-score over 0.8. It is worth mentioning the performance of the LexMa system in the 2T dataset, where it ranked 4th, while it was among the last ranked systems with the AG dataset.

CTA task. As shown in Figure 2, the results in the CTA tasks are relatively similar to the CEA task, where the average performance against the AG datasets is very good. However, the F1-score with the 2T dataset is dramatically reduced for all the systems (see Figure 2b). It is worth emphasising a general improvement from Round 1 to Round 2 (see Figure 2a).

CPA task. Figure 3 summarises the results for the CPA tasks over the AG datastes. Note that, currently, the 2T dataset is only available for the CEA and CTA tasks. It



(a) AG Datasets (b) Average AG vs 2T

Fig. 2: Results in the CTA task with the Automatically Generated (AG) and Tough Tables (2T) Datasets.

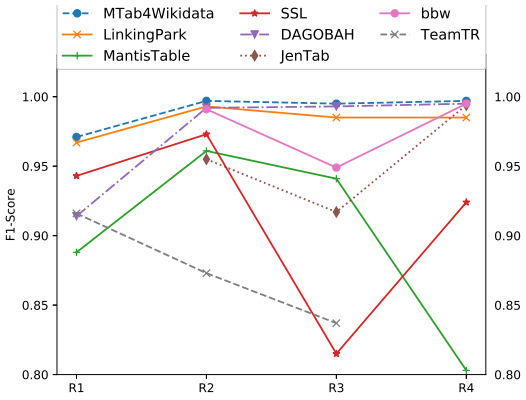


Fig. 3: Results in the CPA task with the Automatically Generated (AG) Datasets.

is worth mentioning the improvement of systems like JenTab and bbw in Round 4. MTab4Wikidata, LinkingPark and DAGOBAH maintained a constant performance during the last rounds.

2.3 Prizes

IBM Research[8] 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**: MTab4Wikidata was the top system in all tasks and the least impacted by the 2T dataset.

[8] https://www.research.ibm.com/

- **2nd Prize**: LinkingPark had a very good and constant performance just below MTab4Wikidata.
- **3rd Prize**: DAGOBAH and bbw. DAGOBAH had overall very positive results, apart from the CEA task in the 2T dataset. On the other hand, bbw had an outstanding performance in the last CEA round.

3 Lessons Learned and Future Work

As in SemTab 2019 [11], the experience of SemTab 2020 has been successful and has served to increase the community interest in the semantic enrichment of tabular data. Both from the organization side and the participation side, we aim at preparing a new edition of the SemTab challenge in 2021. Next, we summarize the ideas for the future editions of the challenge we discussed during the International Semantic Web Conference.

Data shifting. Several participants preferred the options of having a fixed target KG, given as a data dump, instead of using the SPARQL endpoints and related services to access the latest version of the KG. This concerns was specially important as Wikidata is continuously updated. This, however, may be challenging as it may require to "locally" store and process a large amount of data, and it may hinder the participation of systems that rely on online lookup services (*e.g.*, DBpedia Lookup), since the exposed KG may differ from the fixed target KG.

Blind evaluation. Some participants proposed to have both a public and private leaderboard in AIcrowd to cover the rounds with a blind evaluation. It is unclear if AIcrowd provides this service, but systems like STILTool [5] may support this functionality.

Systems as services. To improve reproducibility we are considering to request participants to submit their systems as services, following a pre-defined API, at least in one of the SemTab rounds. This could be achieved by offering the systems as a (Web) service or by submitting a docker image of the system.

The user as a metric. In addition to the standard evaluation measures, it was proposed to include tasks that allow measuring the productivity from a user point of view. For example, a user may be more interested in a system that is easy to setup and run, than in a sophisticated system that provides better results but requires a non-trivial effort for installation or execution.

Complexity of annotations. In future editions, SemTab should also consider datasets that involve more complex annotations to reflect more realistic use cases.

Domain-specific datasets. It was also proposed to use in the future domain-specific datasets and KGs (*e.g.*, biomedical) as targets, in addition to cross-domain KGs like DBpedia and Wikidata and related datasets.

Acknowledgements

We would like to thank the challenge participants, the ISWC & OM organisers, the AIcrowd team, and our sponsors (SIRIUS and IBM Research) that played a key role

in the success of SemTab. Special mention require Federico Bianchi and Matteo Palmonari who contributed to the creation of the Tough Tables dataset. This work was also supported by the AIDA project (Alan Turing Institute), the SIRIUS Centre for Scalable Data Access (Research Council of Norway), Samsung Research UK, Siemens AG, and the EPSRC projects AnaLOG, OASIS and UK FIRES.

References

1. N. Abdelmageed and S. Schindler. JenTab: Matching Tabular Data to Knowledge Graphs. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
2. W. Baazouzi, M. Kachroudi, and S. Faiz. Kepler-aSI : Kepler as a Semantic Interpreter. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
3. S. Chen, A. Karaoglu, C. Negreanu, T. Ma, J.-G. Yao, J. Williams, A. Gordon, and C.-Y. Lin. LinkingPark: An integrated approach for Semantic Table Interpretation. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
4. M. Cremaschi, R. Avogadro, A. Barazzetti, and D. Chieregato. MantisTable SE: an Efficient Approach for the Semantic Table Interpretation. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
5. M. Cremaschi, A. Siano, R. Avogadro, E. Jiménez-Ruiz, and A. Maurino. STILTool: A Semantic Table Interpretation evaLuation Tool. In *ESWC 2020 Satellite Events*, pages 61–66, 2020.
6. V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, and M. Palmonari. Tough Tables: Carefully Benchmarking Semantic Table Annotators [Data set]. `https://doi.org/10.5281/zenodo.3840646`, 2020.
7. V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, and M. Palmonari. Tough tables: Carefully evaluating entity linking for tabular data. In *19th International Semantic Web Conference*, pages 328–343, 2020.
8. G. Diallo and R. Azzi. AMALGAM: making tabular dataset explicit with knowledge graph. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
9. O. Hassanzadeh, V. Efthymiou, J. Chen, E. Jiménez-Ruiz, and K. Srinivas. SemTab 2020: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching Data Sets. `https://doi.org/10.5281/zenodo.4282879`, 2020.
10. V.-P. Huynh, J. Liu, Y. Chabot, T. Labbé, P. Monnin, , and R. Troncy. DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
11. 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.
12. D. Kim, H. Park, J. K. Lee, and W. Kim. Generating conceptual subgraph from tabular data for knowledge graph matching. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
13. O. Lehmberg, D. Ritze, R. Meusel, and C. Bizer. A large public corpus of web tables containing time and context metadata. In *WWW*, 2016.
14. G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. *VLDB Endowment*, 3(1-2):1338–1347, 2010.

15. P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, and H. Takeda. MTab4Wikidata at the SemTab 2020: Tabular Data Annotation with Wikidata. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
16. R. Shigapov, P. Zumstein, J. Kamlah, L. Oberländer, J. Mechnich, and I. Schumm. bbw: Matching CSV to Wikidata via Meta-lookup. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
17. S. Tyagi and E. Jiménez-Ruiz. LexMa: Tabular Data to Knowledge Graph Matching using Lexical Techniques. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.
18. D. Vrandecic and M. Krötzsch. Wikidata: a free collaborative knowledge base. *Commun. ACM*, 57(10):78–85, 2014.
19. S. Yumusak. Knowledge graph matching with inter-service information transfer. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*. CEUR-WS.org, 2020.