# JenTab: Do CTA Solutions Affect the Entire Scores?

Nora Abdelmageed[1,2,*], Sirko Schindler[1]

[1]Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena
[2]Michael Stifel Center Jena, Friedrich Schiller University Jena

Abstract
Semantic Table Annotation remains a crucial task to exploit tabular data in knowledge-aware systems. However, in the process, annotation systems have to overcome various issues ranging from mere typos over inconsistent naming conventions to homonymy among values. The Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) continues to provide demanding datasets to evaluate annotation systems and drive their continued development. In this paper, we describe JenTab's adaptations to the 2022 edition of SemTab: In particular, we added an additional preprocessing step to target Tough Tables (2T)'s excessive misspellings and a new pipeline to exploit meaningful header information. In addition, for each round, we execute two different settings of Column Type Annotation (CTA) creation. We report on the impact of these changes on JenTab's results. In 2022, we highlight the effect of the CTA on the overall score per round.
Our GitHub Repository: https://github.com/fusion-jena/JenTab

**Keywords**
Entity Linking, Cell Entity Annotation, Column Type Annotation, Column-Column Property Annotation, Semantic Table Annotation, JenTab, SemTab

## 1. Introduction

Tabular data such as CSV files are a common way to publish data and represent a precious resource. Nevertheless, they are hardly machine-interpretable in their raw form and are thus hidden from many automated processes. The annotation of regular tables with concepts from the Semantic Web faces various challenges, including misspellings, abbreviations, and the general ambiguity of the free text. Over time, different approaches have been developed to cope with these issues and provide a semantic layer on top of common tables [1, 2, 3, 4, 9]. Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)[1] offers a forum for state-of-the-art systems to compare against one another and provides them with various datasets to challenge their capabilities. In its fourth year, it features a series of three rounds. Each round consists of a variety of raw tables. Such tables to be annotated with concepts either from Wikidata [5] or DBpedia [6].

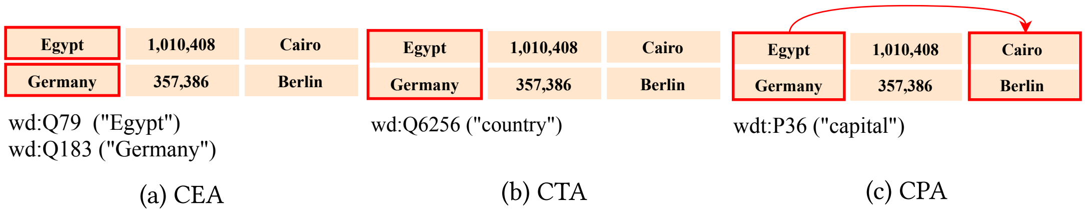
The annotation tasks themselves are called Semantic Table Annotation (STA). Based on the SemTab description of such tasks, the three tasks are namely Cell Entity Annotation (CEA),

*SemTab@ISWC 2022, October 23–27, 2022, Hangzhou, China (Virtual)*

*Corresponding author.
nora.abdelmageed@uni-jena.de (N. Abdelmageed); sirko.schindler@uni-jena.de (S. Schindler)
0000-0002-1405-6860 (N. Abdelmageed); 0000-0002-0964-4457 (S. Schindler)

CEUR Workshop Proceedings (CEUR-WS.org)

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



(a) CEA (b) CTA (c) CPA

**Figure 1:** SemTab tasks summary [7].

CTA, and Column Property Annotation (CPA). Given a data table and a target Knowledge Graph (KG), CEA links a cell to an entity within the KG (cf. Figure 1a). CTA is the task of assigning a semantic type (e.g., a class) to a column (cf. Figure 1b). Finally, CPA assigns a suitable semantic relation (predicate) from the KG to individual column pairs (cf. Figure 1c).

Our previous participation in the SemTab challenge found that the hardest task to solve is the CTA. The challenge call asks for the most precise type to annotate the given column. However, we can consider high-level types as possible to decide on that fine-grained solution. We have investigated the effect of using multiple CTA strategies on the results of STA tasks. In this paper, we focus on analyzing JenTab performance given various strategies for creating and selecting CTA solutions using the provided SemTab 2022 datasets. In addition, we developed a sophisticated cleaning module for the 2T dataset [8] which yielded into significant scores improvement. Finally, we developed a new pipeline configuration that suites datasets with headers.

The remainder of this paper is organized as follows: Section 2 outlines the general approach of JenTab, its pipelines configurations, and CTA creation strategies. Section 3 gives an overview of this year's challenge datasets and requirements. Section 4 highlights the newly developed modules of JenTab during SemTab 2022. Section 5 discusses the given dataset's characteristics of SemTab 2022 and our scores during the rounds under different settings. We conclude and point out future directions in Section 6.

## 2. Background

This section provides an overview of the general approach JenTab follows. Last year, we developed various pipelines based on the given datasets characteristics like `pipeline_full`, `pipeline_no_cpa`, or `pipeline_numeric`. All pipelines follow the Create, Filter, and Select (CFS) pattern developed during SemTab 2020 [4]. The default pipeline, `pipeline_full`, is outlined in Figure 2.

For more details about the CFS pattern, and our various pipelines we refer to our previous publications in 2020 [4], and 2021 [9]. This year, during SemTab 2022, we focus on the `pipeline_full`, which is the most potent pipeline due to its consistent performance on various datasets.

CTA solutions are crucial to solving STA. In 2020, we developed and investigated three strategies to create CTA candidates [7]. We give a brief overview of each strategy in a Wikidata context, as follows:

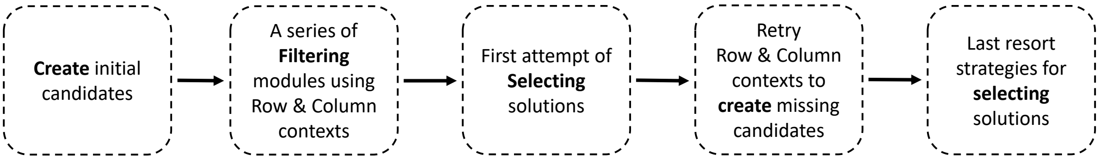

Create initial candidates
A series of Filtering modules using Row & Column contexts
First attempt of Selecting solutions
Retry Row & Column contexts to create missing candidates
Last resort strategies for selecting solutions

**Figure 2:** Abstract view of the `pipeline_full` [9].

- **P31** includes only direct parents using *instance of (P31)* relations. This strategy does not include any further traversal of the class hierarchy.
- **2Hops** extends "P31" with one additional parent (higher level) via *subclass of (P279).*
- **Multi Hops** creates a more general tree of parents following *subclass of (P279)* relations.

From our previous study, Multi Hops gave the lowest scores due to its consideration of very high-level types. Thus, in this year, we focus our experiments on P31 and 2Hops only.

Together with that CTA creation strategies, we have developed two CTA selection methods. On the one hand, we have implemented a "majority vote" technique that can be used with any creation strategies. This technique does not rely on the hierarchical relations among the possible CTA candidates. On the other hand, we have developed a "Least Common Subsumer (LCS)" method that selects the most fine-grained type from the hierarchy of CTA candidates such that this type has the maximum support among column cells.

## 3. SemTab 2022 Datasets & Requirements

In 2022, SemTab consisted of three rounds. Multiple datasets are given for each round. Unlike in previous editions, partial ground truth data is available. Each dataset was divided into two parts: validation and test sets. The validation set is provided with ground truth data and the validator code. This allows self-check on a small portion before the actual system run and the final submission per round. Table 1 shows the given datasets, train/test splits, target KG, and the associated STA tasks per round. The recommended Wikidata dump by the challenge organizers is a custom n-triple dump as of May 21st, 2022, and is hosted on Zenodo [10]. However, using a public API was also recommended since the dump version mentioned is very recent. Table 2 illustrates the characteristics of SemTab 2022 datasets. It shows the number of tables, average rows, columns, and cells. In addition, it shows the number of target annotation for CEA, CTA, and CPA tasks. In this paper, we focus on the test sets since they directly affect the scores. For submission, we were allowed multiple submissions per week, but only the most recent one was evaluated each Friday. This is unlike the previous years when we used to submit our solutions to an AICrowd page.

## 4. What's New in JenTab?

In this section, we discuss our newly developed components. First, we explain the cleaning procedure for one of the provided datasets. Then, we discuss of newly created pipeline that

**Table 1**
Specifications per round of SemTab 2022.

| Round | Dataset | Dev/Test | Target | CEA | CTA | CPA |
|---|---|---|---|---|---|---|
| R1 | HardTables | 200/3,691 | Wikidata | Yes | Yes | Yes |
| R1 | HardTables | 457/4,649 | Wikidata | Yes | Yes | Yes |
| R2 | 2T | 18/144 | Wikidata | Yes | Yes | No |
| R2 | 2T | 18/144 | DBpedia | Yes | Yes | No |
| R3 | GitTables | 4097/4110 | DBpedia & schema.org | No | Yes | No |
| R3 | BiodivTab | 5/45 | DBpedia | Yes | Yes | No |

**Table 2**
SemTab 2022 datasets. Targets created by JenTab are marked with a star (*).

| Round | Dataset | Tables | Avg. Rows #<br>(± Std Dev.) | Avg. Cols #<br>(± Std Dev.) | Avg. Cells #<br>(± Std Dev.) | CEA | CTA | CPA |
|---|---|---|---|---|---|---|---|---|
| R1 | HardTables | $3,691$ | $6 \pm 2$ | $3 \pm 1$ | $14 \pm 6$ | 26,189 | 4,511 | 5,745 |
| R2 | HardTables | $4,649$ | $6 \pm 1$ | $3 \pm 1$ | $14 \pm 5$ | 22,009 | 4,534 | 3,954 |
|  | 2T (WD) | $144$ | $1,181 \pm 2,985$ | $4 \pm 2$ | $4,511 \pm 11,602$ | 586,118 | 443 | 299* |
|  | 2T (DB) | $144$ | $1,008 \pm 2,710$ | $4 \pm 2$ | $3,787 \pm 10,198$ | 486,203 | 429 | 285* |
| R3 | BiodivTab | $45$ | $259 \pm 743$ | $24 \pm 13$ | $4,589 \pm 10,862$ | 33,405 | 569 | NA |

selects CTA solutions based on the header values.

**Tough Tables Cleanup** We have developed a cleaning module for the 2T dataset. This dataset contains a large amount of artificially added misspellings to its tables. Thus, our core idea is to locate the correctly spelled cells and then replace all the artificial occurrences with the correct word. The first step aims to find the correctly spelled words by querying those cells in target KG, Wikidata, Those with exact matches are considered correct words. The second step is to match the remaining values in the tables to the correctly identified values. We converted all the given cells into the embedding space using fasttext [11] to avoid the out-of-vocab (OOV) problem. Then, we applied cosine similarity among those vectors; we picked the final value if the similarity is $\geq 70\%$. We ran this step offline before the actual running of JenTab to solve the STA tasks.

**New Pipeline: `pipeline_headers`** In addition to the previously developed pipelines [9], we added a new one, `pipeline_headers`, during Round 3 of SemTab 2022. It is based on `pipeline_no_cpa`, which contains all modules from the default configuration except the CPA create, filter, and select parts. However, the handling of CTA candidates has changed to accommodate datasets that contain meaningful headers. Already in 2021, BiodivTab [12] was included in SemTab as an example of such datasets. Here, JenTab only achieved rather low scores: $60\%$, and $10\%$ on both CEA, and CTA tasks respectively [9]. Contrary to 2021, BiodivTab in 2022 also asked for DBpedia annotations replacing the previous target KG of

**Table 3**
Generic Lookup: Unique labels and ratio of resolved labels per round.

| Rounds | Dataset | Target | Unique Labels | Unmatched | Matched | Matched (%) |
|---|---|---|---|---|---|---|
| R1 | HardTables | Wikidata | 19,107 | 179 | 18,928 | 99% |
| R2 | HardTables & 2T | Wikidata | 74,177 | 6,191 | 67,986 | 91.6% |
| R2 | 2T | DBpedia | 65,223 | 6,988 | 58,235 | 89.3% |

Wikidata.

## 5. Experimental Results

Spelling mistakes and artificial noise are common challenges across SemTab's datasets. Especially in 2T dataset. We developed the generic lookup as our primary strategy for tackling this crucial issue. Due to the resources required for comparing cell values against all labels (and aliases) within Wikidata or DBpedia, we extracted the unique values from all dataset tables. Then, we matched those against the labels of the respective KG using an optimized Jaro-Winkler Similarity implementation based on [13] and a threshold for minimum similarity of $0.9$. Table 3 illustrates the results of this approach. For the synthetic datasets, HardTables, the matching percentage is high. It reached up to 99% in the first round. This is unlike the case of the 2T dataset; it reached around 89% in the second round, where DBpedia is the target KG. Such lower matching percentage guided us to develop a more sophisticated cleaning step before the actual run, as discussed in Section 4.

Table 4 demonstrates our scores of the three rounds of SemTab 2022 as reported in the results sheet after each round. Each week per round, we have submitted different pipeline setting results. For example, during the first week of Round 1, we submitted the results of the `pipeline_full` combined with the "P31" CTA creation strategy and the majority vote as the selection strategy. In week two of the same round, we tested the same pipeline with "2Hops" CTA creation strategy combined with LCS selection technique instead. From the results, the P31 strategy that is associated with the majority vote selection yielded the best scores on the HardTables dataset in both rounds. However, for the 2T dataset, 2Hops improved the CEA scores significantly compared to the P31 strategy while achieving similar results for CTA. The 2Hops strategy seems better equipped to deal with challenging values like those found in 2T whose values are even hard to annotate for human users [8]. On the other hand, P31 seems a reasonable choice for comparatively more straightforward datasets across all tasks. Omitting higher levels in the hierarchy, P31 is also computationally less expensive and can thus be run faster.

We highlight the impact of the sophisticated cleaning we applied on the 2T dataset. This additional step yielded substantially improved results over past years' attempts: During 2020 our initial pipeline only achieved an F1-score of $10\%$ [4]. In 2021, applying the 2Hops strategy improved this result to an F1-score of $45\%$ [9]. This year, we surpassed the previous scores using the lightweight P31 and the 2Hops strategies by achieving $75.1\%$ and $80.2\%$, respectively.

**Table 4**
JenTab scores during SemTab 2022. F1 - F1 Score, Pr - Precision, AF1 - Average F1 Score, and APr - Average Precision.

| | | | | CEA | | | CTA | | | CPA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rounds | Dataset | Target | Setting | F1 | Pr | R | AF1 | APr | AR | F1 | Pr | R |
| R1 | HardTables | WD | P31 | **0.945** | **0.946** | **0.944** | **0.938** | **0.940** | **0.936** | 0.974 | 0.985 | 0.964 |
| | | | 2Hops | 0.936 | 0.936 | 0.935 | 0.871 | 0.871 | 0.871 | **0.975** | **0.986** | **0.965** |
| R2 | HardTables | WD | P31 | **0.751** | **0.758** | **0.743** | **0.836** | **0.881** | **0.795** | **0.872** | **0.921** | **0.828** |
| | | | 2Hops | 0.713 | 0.720 | 0.707 | 0.72 | 0.752 | 0.691 | 0.862 | 0.913 | 0.816 |
| | 2T | WD | P31 | 0.773 | 0.774 | 0.772 | **0.357** | **0.362** | 0.352 | NA | NA | NA |
| | | | 2Hops | **0.802** | **0.807** | **0.796** | 0.346 | 0.357 | **0.337** | NA | NA | NA |
| R3 | BiodivTab | DBP | P31 | **0.550** | **0.605** | **0.505** | **0.414** | **0.421** | 0.407 | NA | NA | NA |
| | | DBP | 2Hops | 0.547 | 0.601 | 0.502 | 0.408 | 0.410 | 0.407 | NA | NA | NA |

**Table 5**
Generic Lookup effect. Primary, secondary scores, and Ranks for JenTab. F1 - F1 Score, Pr - Precision, AF1 - Average F1 Score, and APr - Average Precision.

| | | | CEA | | | CTA | | | CPA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Target | Generic Lookup | F1 | Pr | R | AF1 | APr | AR | F1 | Pr | R |
| HardTables | WD | Yes | **0.945** | **0.946** | **0.944** | **0.938** | **0.940** | **0.936** | 0.974 | 0.985 | 0.964 |
| | | No | 0.655 | 0.672 | 0.638 | 0.626 | 0.657 | 0.599 | 0.804 | 0.881 | 0.740 |
| 2T | WD | Yes | **0.773** | 0.774 | **0.772** | **0.357** | **0.362** | **0.352** | NA | NA | NA |
| | | No | 0.726 | **0.803** | 0.663 | 0.323 | **0.362** | 0.291 | NA | NA | NA |

Moreover, we investigated the impact of Generic Lookup, shown in Table 5, on both HardTables and 2T datasets during Round 2. We have selected the P31 strategy to perform this experiment since it had the overall best performance across all STA tasks. The absence of Generic Lookup yielded lower scores in general except for the precision of CEA task. This indeed reflects the importance of this module in the JenTab system.

Our solution strategy for BiodivTab in Round 3 differs from our traditional way. Initially, we ran both `pipeline_no_cpa` and `pipeline_header` directly against DBpedia Proxy. However, the scores were deficient, reaching only 20% and 5% for both CEA and CTA tasks, respectively. Thus, we run the same pipelines against Wikidata Proxy; this fetches solutions for the dataset from Wikidata. After a complete run of the dataset, we retrieved `owl:sameAs` mappings that translate the Wikidata annotations to DBpedia resources for both tasks.

## 6. Conclusions & Future Work

In this paper, we have reported on our participation and JenTab's continuous developments as a part of the 2022 edition of Semantic Web Challenge on Tabular Data to Knowledge Graph Matching challenge. We introduced a cleaning module for the Tough Tables (2T) dataset that significantly impacted our results. In addition, we have developed a new pipeline that leverages information from the table header. We used this pipeline during Round 3 for the BiodivTab dataset. JenTab remains a top participant of the SemTab during its third participation and remains without any complex requirements. Our code is publicly available [14]. Moreover, our precomputed generic lookup [15] and solution files [16] for each round of SemTab are also publicly available.

We see various areas for further improvement. First, the binary decision of whether to keep candidates or remove them should be replaced by a scoring system that emphasizes well-supported candidates but maintains other options. In addition, the new pipeline that uses the header candidates as direct CTA solutions also needs a more intelligent mechanism. For instance, we can apply a weighting technique that controls such decisions. Further, we see the need to apply a more detailed investigation on the impact of individual modules within the pipelines. This applies to both the content level (are we removing correct solutions by accident?) as well as on the performance level (can we exclude more candidates earlier in the pipeline?).

## Acknowledgment

The authors thank the Carl Zeiss Foundation for the financial support of the project "A Virtual Werkstatt for Digitization in the Sciences (P5)" within the scope of the program line "Breakthroughs: Exploring Intelligent Systems" for "Digitization - explore the basics, use applications". We would thank Birgitta König-Ries for her guidance and continuous feedback.

## References

[1] P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, H. Takeda, Semtab 2021: Tabular data annotation with mtab tool, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 27, 2021, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 92–101.

[2] V. Huynh, J. Liu, Y. Chabot, F. Deuzé, T. Labbé, P. Monnin, R. Troncy, DAGOBAH: table and graph contexts for efficient semantic annotation of tabular data, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 27, 2021, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 19–31.

[3] R. Shigapov, P. Zumstein, J. Kamlah, L. Oberländer, J. Mechnich, I. Schumm, bbw: Matching CSV to wikidata via meta-lookup, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 5, 2020, volume 2775 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 17–26.

[4] N. Abdelmageed, S. Schindler, Jentab: Matching tabular data to knowledge graphs, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 5, 2020, volume 2775 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 40–49.

[5] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85. doi:10.1145/2629489.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A nucleus for a web of open data, in: The Semantic Web, Springer Berlin Heidelberg, 2007, pp. 722–735. doi:10.1007/978-3-540-76298-0_52.

[7] N. Abdelmageed, S. Schindler, Jentab: A toolkit for semantic table annotations, in: Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021), Online, June 6, 2021, volume 2873 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[8] V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, M. Palmonari, Tough Tables: Carefully Evaluating Entity Linking for Tabular Data, 2020. doi:10.5281/zenodo.4246370.

[9] N. Abdelmageed, S. Schindler, Jentab meets semtab 2021's new challenges, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 27, 2021, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 42–53.

[10] O. Hassanzadeh, Wikidata Truthy Dump from May 21, 2022, 2022. URL: https://doi.org/10.5281/zenodo.6643443. doi:10.5281/zenodo.6643443.

[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. doi:10.1162/tacl_a_00051.

[12] N. Abdelmageed, S. Schindler, B. König-Ries, Biodivtab: A table annotation benchmark based on biodiversity research data, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 27, 2021, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 13–18.

[13] J. M. Keil, Efficient bounded Jaro-Winkler Similarity based search, BTW 2019 (2019). doi:10.18420/BTW2019-13.

[14] N. Abdelmageed, fusion-jena/jentab: Jentab code for semtab 2022, 2022. URL: https://doi.org/10.5281/zenodo.7229238. doi:10.5281/zenodo.7229238.

[15] N. Abdelmageed, fusion-jena/jentab_precomputed_lookup: Semtab2022, 2022. URL: https://doi.org/10.5281/zenodo.7229246. doi:10.5281/zenodo.7229246.

[16] N. Abdelmageed, fusion-jena/jentab_solution_files: Semtab2022, 2022. URL: https://doi.org/10.5281/zenodo.7229243. doi:10.5281/zenodo.7229243.