JenTab Meets SemTab 2021’s New Challenges

Nora Abdelmageed, Sirko Schindler

Friedrich Schiller University Jena, Germany

ISWC 2021
### Semantic Table Annotation Tasks

<table>
<thead>
<tr>
<th>Country</th>
<th>Area</th>
<th>Capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt</td>
<td>1,010,408</td>
<td>Cairo</td>
</tr>
<tr>
<td>Germany</td>
<td>357,386</td>
<td>Berlin</td>
</tr>
</tbody>
</table>

- [https://www.wikidata.org/wiki/Q79](https://www.wikidata.org/wiki/Q79)
- [https://www.wikidata.org/wiki/Q183](https://www.wikidata.org/wiki/Q183)

<table>
<thead>
<tr>
<th>Country</th>
<th>Area</th>
<th>Capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt</td>
<td>1,010,408</td>
<td>Cairo</td>
</tr>
<tr>
<td>Germany</td>
<td>357,386</td>
<td>Berlin</td>
</tr>
</tbody>
</table>

- [https://www.wikidata.org/wiki/Q6256](https://www.wikidata.org/wiki/Q6256)

<table>
<thead>
<tr>
<th>Country</th>
<th>Area</th>
<th>Capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt</td>
<td>1,010,408</td>
<td>Cairo</td>
</tr>
<tr>
<td>Germany</td>
<td>357,386</td>
<td>Berlin</td>
</tr>
</tbody>
</table>

- [https://www.wikidata.org/wiki/Q5119](https://www.wikidata.org/wiki/Q5119)

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

- JenTab uses **only** the public available **lookup API** and **SPARQL query endpoints** of the target knowledge graph

- Follows (CFS Pattern):
  - **C**reate as many as candidates per task
  - **F**ilter these candidates by using semantic context
  - **S**elect the most appropriate candidate from the remaining

- Uses 4 contexts:
Background – Default Pipeline

- Abstract View of pipeline_full
Outlook

1. Semantic Table Annotation Tasks
2. Background
3. New Design
   • System Architecture
   • Various Pipelines
4. Evolution
5. Experiences & Results
6. Conclusions & Future work
System Architecture

- **Manager**
  - Work load balance

- **Solver**
  - Performs pre-processing
  - Executes the actual pipeline

- **Runner**
  - Mediator between Manager & Solver

- **Wikidata & DBpedia Proxies**
  - Encapsulate the publicly available API and SPARQL endpoint for Wikidata and DBpedia

- **Generic Lookup**
  - Holds a precomputed mapping for the unique cell-values of the entire dataset using Jaro Winkler distance*

- **Caching Server**
  - Central caching server for the last three services

---

* W. E. Winkler, String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, 1990
Highlighting New Changes

**Solver + Clean Cells + Type Prediction**
- Less data transfers
- Faster execution

**DBpedia Proxy**
- New requirement by SemTab 2021

**Caching Server**
- Reduced the redundancy made by individual caches

**Generic Lookup**
- Optimized, it is our primary auto-correction strategy
Various Pipelines

- **pipeline_essential**
  - Core parts of the pipeline_full (no re-executions, faster)

- **pipeline_no_cpa**
  - Omits the CPA components (Create, Filter and Select)

- **pipeline_keyTables**
  - Key-value pair grouping of the dataset
  - Key → clean table, without artificial noise
  - Value → noisy tables derived from that clean
  - Keys are solved by pipeline_full
  - Values obtain their solutions via broadcasting
Various Pipelines

- **pipeline_numeric**
  - Designed for tables with single object column (the subject) with a rest of numerical columns
  - Emphasizes the Row Context
    - Highest priority goes to candidates with the most support by their row

- **pipeline_conditional**
  - Combines `pipeline_numeric` and `pipeline_full`
  - If the table structure meets (like above)
    - Solve using `pipeline_numeric`
    - If coverage >= 80%, return solutions
  - Otherwise, Solve using `pipeline_full`
JenTab Evolution – SemTab 2021 Round 1

- **Givens:**
  - 180 tables, KG: Wikidata, Tasks: CEA & CTA
  - 180 tables, KG: DBpedia, Tasks: CEA & CTA

- **Configuration:**
  - DBpedia Proxy $\rightarrow$ Lookup = Lookup$^1$ + spotlight$^2$
  - `pipeline_no_cpa` $\rightarrow$ No CPA required
  - `pipeline_keyTables` $\rightarrow$ Massive amount of noise, manual clean up and grouping data

- **Findings:**
  - DBpedia spotlight $\rightarrow$ False positives
  - CTA might not be enough semantic context to filter candidates
    - CPA increases scores, precision.

---

$^1$ [http://lookup.dbpedia.org/api/search/PrefixSearch](http://lookup.dbpedia.org/api/search/PrefixSearch)
JenTab Evolution – SemTab 2021 Round 2

- **Givens:**
  - HardTables, 1750 tables, KG: Wikidata, Tasks: CEA, CTA, CPA
  - BioTables, 110 tables, KG: Wikidata, Tasks: CEA, CTA, CPA

- **Configuration:**
  - `pipeline_essential` -> BioTables, wide tables with long strings, timeouts.
  - `pipeline_full` -> HardTables, ambiguity.

- **Findings:**
  - `pipeline_full` is optimized and managed to run over BioTables
  - Best Results for both datasets
JenTab Evolution – SemTab 2021 Round 3

• Givens:
  • HardTables, 7207 tables. KG: Wikidata. Tasks: CEA, CTA, CPA
  • BiodivTab, 50 tables. KG: Wikidata. Tasks: CEA, CTA
  • GitTables, 1101 tables. KG: DBpedia & schema.org. Tasks: CTA

• Configuration:
  • pipeline_numeric & pipeline_conditional

• Findings:
  • Most of the given tables follows
    • Subject_col & [num1, num2, num3]
JenTab Evolution – SemTab 2021 Round 3

• Givens:
  • HardTables, 7207 tables. KG: Wikidata. Tasks: CEA, CTA, CPA
  • BiodivTab, 50 tables. KG: Wikidata. Tasks: CEA, CTA
  • GitTables, 1101 tables. KG: DBpedia & schema.org. Tasks: CTA

• Configuration:
  • pipeline_no_cpa → We couldn’t relate the given object column to properties.
  • Dictionary based lookup for taxons, to solve abbreviations
    • Canna glauca → (C.gluca , ca.gluca)
  • Split nested entities and solve the first part only
    • David Eichenberg (University of Halle-Wittenberg) → David Eichenberg (University of Halle-Wittenberg)

• Findings:
  • Our nested entity solution is not optimal
    • The first entity may not be the target entity
  • Abbreviated taxons are ambiguous
    • C.gluca matches 48 plant species in Wikidata 14 Oct 2021
JenTab Evolution – SemTab 2021 Round 3

• Givens:
  • HardTables, 7207 tables. KG: Wikidata. Tasks: CEA, CTA, CPA
  • BiodivTab, 50 tables. KG: Wikidata. Tasks: CEA, CTA
  • GitTables, 1101 tables. KG: DBpedia & schema.org. Tasks: CTA

• Configuration:
  • GitTables CTA targets structure mixes between column types (CTA) and properties (CPA)
    • We have split that into normal CTA and CPA targets
  • CTA/CPA depends on CEA
    • For our generated CTA, CEA is created.
  • Ontology is changed from dbpedia.org to schema.org in DBpedia Proxy
  • pipeline_essential.

• Findings:
  • pipeline_full suffers from timeout here
  • DBpedia Proxy managed to retrieve schema.org types but failed for properties.
Experiences and Results

- SemTab2021 datasets. KGs: DBpedia (DPB), Wikidata (WD), and schema.org (SCH).

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2T DBP</td>
<td>2T WD</td>
<td>BioTables WD</td>
</tr>
<tr>
<td>Tables</td>
<td>180</td>
<td>180</td>
<td>110</td>
</tr>
<tr>
<td>Avg. Rows # (± Std Dev.)</td>
<td>1,080 ± 2,798</td>
<td>1,080 ± 2,798</td>
<td>2,448 ± 193</td>
</tr>
<tr>
<td>Avg. Cols # (± Std Dev.)</td>
<td>5 ± 2</td>
<td>4 ± 2</td>
<td>6 ± 1</td>
</tr>
<tr>
<td>Avg. Cells # (± Std Dev.)</td>
<td>4125 ± 10947</td>
<td>3952 ± 10129</td>
<td>14605 ± 2338</td>
</tr>
<tr>
<td>CEA #</td>
<td>663,655</td>
<td>636,185</td>
<td>1,391,324</td>
</tr>
<tr>
<td>CTA #</td>
<td>539</td>
<td>535</td>
<td>656</td>
</tr>
<tr>
<td>CPA #</td>
<td>359*</td>
<td>355*</td>
<td>546</td>
</tr>
</tbody>
</table>
Experiences and Results

- Generic Lookup: Unique labels and ratio of resolved labels per round.

<table>
<thead>
<tr>
<th>Rounds</th>
<th>Dataset</th>
<th>Target</th>
<th>Unique Labels</th>
<th>Unmatched</th>
<th>Matched</th>
<th>Matched (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>2T</td>
<td>Wikidata</td>
<td>69,980</td>
<td>7,072</td>
<td>62,908</td>
<td>89.89%</td>
</tr>
<tr>
<td>R1</td>
<td>2T</td>
<td>DBpedia</td>
<td>66,340</td>
<td>7,172</td>
<td>59,168</td>
<td>89.19%</td>
</tr>
<tr>
<td>R2</td>
<td>HardTables</td>
<td>Wikidata</td>
<td>249,625</td>
<td>600</td>
<td>249,025</td>
<td>99.76%</td>
</tr>
<tr>
<td>R3</td>
<td>HardTables</td>
<td>Wikidata</td>
<td>47,809</td>
<td>944</td>
<td>46,865</td>
<td>98.03%</td>
</tr>
<tr>
<td>R3</td>
<td>GitTables</td>
<td>DBpedia</td>
<td>37,780</td>
<td>21,253</td>
<td>16,527</td>
<td>43.75%</td>
</tr>
</tbody>
</table>
Experiences and Results

- Primary, secondary scores, and Ranks for JenTab.

<table>
<thead>
<tr>
<th>Rounds</th>
<th>Dataset</th>
<th>Target</th>
<th>CEA</th>
<th></th>
<th>CTA</th>
<th></th>
<th>CPA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>F1</td>
<td>Pr</td>
<td>Rank</td>
<td>AF1</td>
<td>APr</td>
<td>Rank</td>
</tr>
<tr>
<td>Round 1</td>
<td>2T</td>
<td>DBpedia</td>
<td>0.607</td>
<td>0.669</td>
<td>3rd</td>
<td>0.460</td>
<td>0.468</td>
<td>1st</td>
</tr>
<tr>
<td>Round 1</td>
<td>2T</td>
<td>Wikidata</td>
<td>0.457</td>
<td>0.520</td>
<td>3rd</td>
<td>0.697</td>
<td>0.697</td>
<td>2nd</td>
</tr>
<tr>
<td>Round 2</td>
<td>HardTables</td>
<td>Wikidata</td>
<td>0.966</td>
<td>0.967</td>
<td>4th</td>
<td>0.914</td>
<td>0.917</td>
<td>4th</td>
</tr>
<tr>
<td>Round 2</td>
<td>BioTables</td>
<td>Wikidata</td>
<td>0.857</td>
<td>0.858</td>
<td>4th</td>
<td>0.835</td>
<td>0.843</td>
<td>5th</td>
</tr>
</tbody>
</table>

F1 - F1 Score, Pr - Precision, AF1 - Average F1 Score, and APr - Average Precision.
Conclusions

• JenTab matches table components to Wikidata/DBpedia KGs
• Updates of JenTab as a participant of SemTab2021
  • System redesign
  • A variety of pipelines based on dataset characteristics
• Our code, generic lookup & solution files are publicly available*

* https://github.com/fusion-jena/JenTab
Future Work

• Change binary decision keep/remove candidate to a scoring mechanism
• Address is the lack of targets for specific tasks
  • GitTables
• Continuously improve the performance
  • Timeouts received from the public endpoints
Acknowledgement

Sarah Böning
Muhammad Abbady

Prof. Dr. Birgitta König-Ries

The support is from the Carl Zeiss Foundation
Thank You!

Nora Abdelmageed  
nora.abdelmageed@uni-jena.de  
@NoraYoussef

Sirko Schindler  
sirko.schindler@uni-jena.de