Kepler-aSI at SemTab 2023*

Wiem Baazouzi[1], Marouen Kachroudi[2] and Sami Faiz[3]

[1]Université de Manouba, Ecole Nationale des sciences de l'informatique, Laboratoire de Recherche en génie logiciel, Application distribuées, Manouba 2010, Tunis, Tunisie. wiem.baazouzi@ensi-uma.tn

[2]Université de Tunis El Manar, Faculté des Sciences de Tunis, Informatique Programmation Algorithmique et Heuristique, LR11ES14, 2092, Tunis, Tunisie marouen.kachroudi@fst.rnu.tn

[3]Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, Laboratoire de Télédétection et Systèmes d'Information à Référence Spatiale, 99/UR/11-11, 2092, Tunis, Tunisie sami.faiz@insat.rnu.tn

Abstract

In this article, we present our system Kepler-aSI, for the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2023). This is the fourth participation in the challenge. Kepler-aSI is intended to participate in three content tasks: column type annotation (CTA), cell entity annotation (CEA), and column property annotation (CPA) . Kepler-aSI still relies on SPARQL query to semantically annotate tables in Knowledge Graphs (KG), to solve critical task matching issues. The results obtained during the evaluation phase are encouraging and show the strengths of the proposed system.

Keywords

Tabular Data, Knowledge Graph, Kepler-aSI, Semantic Web Challenge

1. Introduction

It is evident that the World Wide Web encompasses and conveys very large volumes of textual information, in several forms: unstructured text, semi-structured model-based web pages (which represent data in the form widely recognized by key-value notation and lists), and of course arrays. In this broad context, the methods aiming to extract information from these resources to convert them in a structured form have been the subject of several works [1, 2]. As an observation, it is evident that there is a lack of understanding of the semantic structure which can hamper the process of data analysis. Indeed, acquiring this semantic reconciliation will therefore be very useful for data integration, data cleansing, data mining, machine learning and knowledge discovery tasks. For example, understanding the data can help assess the appropriate types of transformation.

SemTab'23: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 2023, co-located with the 22nd International Semantic Web Conference (ISWC), November 6-10, 2023, Athens, Greece

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

2. Presentation of the system Kepler-aSI

Kepler-aSI, as a table annotation system, can efficiently process three tabular data to knowledge graph matching tasks: column type annotation (CTA), entity annotation of cell (CEA) and column property annotation (CPA). Our system makes full use of the structure of tabular data and the information provided by Knowledge Graphs (KG). The datasets demonstrate that Kepler-aSI has disambiguation capability and achieves performance with less query time. Kepler-aSI, is implemented as a set of tools, used in order, to provide three main functionalities.

1. Spelling correction for natural language processing (NLP) which is deployed in multiple applications such as information retrieval, information retrieval and search engines.

2. The annotation of tabular data is a central phase of the system. It allows us to extract candidate annotations from Knowledge Graphs (such as Wikidata) using parameterized SPARQL queries.

3. Disambiguation It is important to recognize that an entity within Knowledge Graphs can be associated with multiple classes. The presence of several classes for an entity enriches its representation and allows a more complete understanding of its semantic context within Knowledge Graphs.

   To provide these functionalities, Kepler-aSI is structured as follows. The preprocessing modules perform table cleanup and a first high-level characterization of their form and content. After performing the various pre-processing treatments, the tabular data annotation phase can be triggered and After having the candidate annotations, we move on to filtering through the Disambiguation phase.

2.1. System description

In order to address the above mentioned SemTab challenge tasks, Kepler-aSI is designed according to the workflow depicted by Figure 1. There are three major complementary modules which consist in, respectively, Preprocessing, Annotation context and Tabular data to KG matching. The aforementioned steps are the same for each round, but the changes remain minimal depending on the variations observed in each case. As shown in Figure 1 Preprocessing aims to prepare the data inside the considered table. While Annotation Context, seeks to create a list of terms denoting the same context.

□ **Preprocessing :**

The content of each table can vary significantly in terms of types and formats, such as numeric, character strings, binary data, date/time, boolean, addresses, and more. This diversity of data types makes the pre-processing step particularly crucial. The main objective of preprocessing is to ensure that the processing of each table can proceed smoothly without encountering any errors. This becomes especially challenging when dealing with data that contains spelling errors. Consequently, resolving these issues becomes a prerequisite before applying any further approach. In order to well carry out this step, we used several

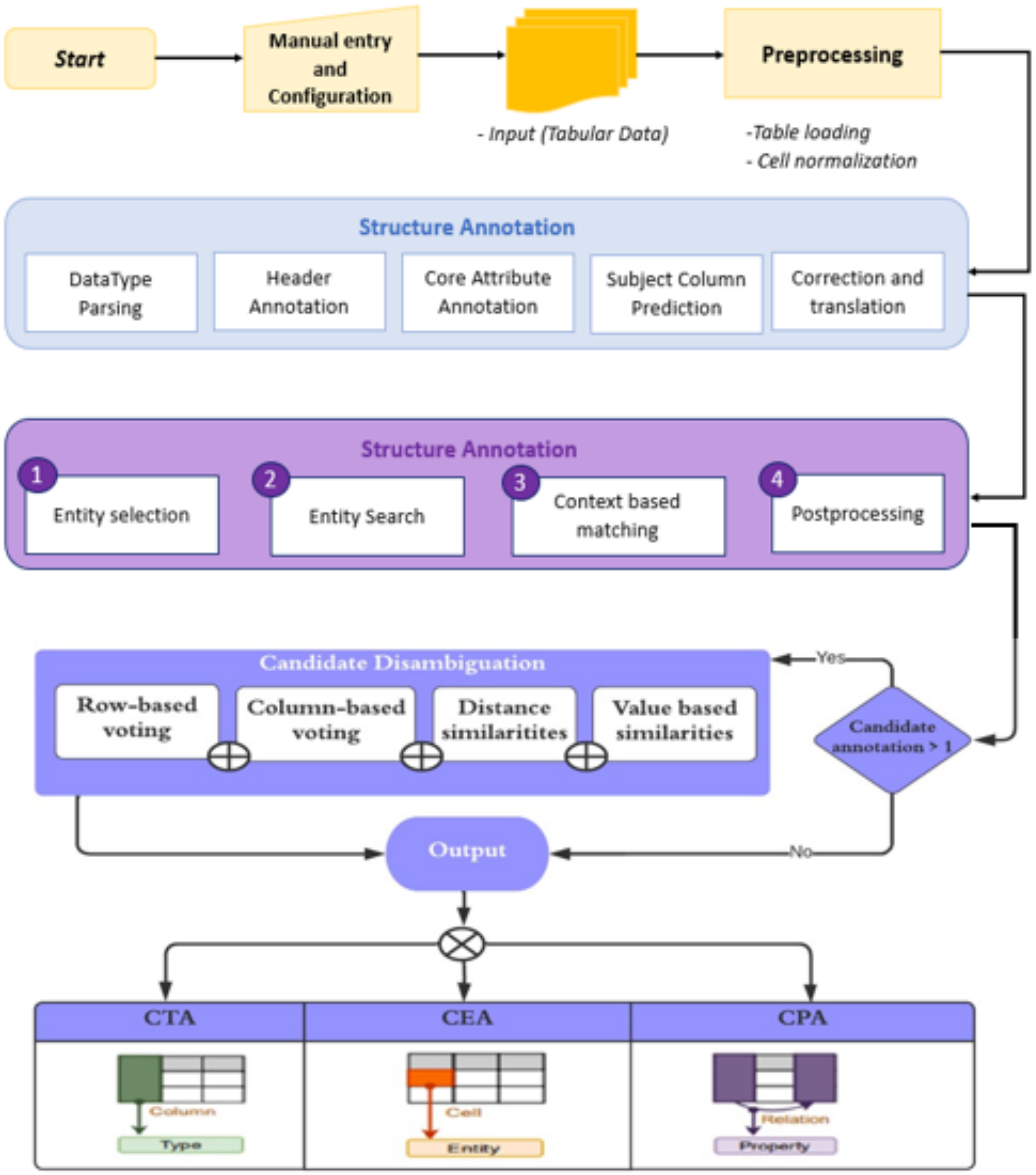

Figure 1: Overview of our approach

techniques and libraries such as (Textblob[1], Pyspellchecker[2], etc.) to rectify and correct all the noisy textual data in the considered tables. As an example, we detect punctuation,

[1]https://textblob.readthedocs.io/en/dev/
[2]https://pypi.org/project/pyspellchecker/

parentheses, hyphen and apostrophe, and also stop words by using the `Pandas`[3] library to remove them. Like a classic treatment in this register, we ended this phase by transforming all the upper case letters into lower case.

☐ **Annotation context :**

During this phase, candidates for the annotation process are explicitly extracted. This extraction is achieved through an analysis of the processing columns, which aims to comprehend and define a set of regular expressions encompassing various units. These units include area, currency, density, electric current, energy, flow rate, force, frequency, energy efficiency, information unit, length, mass, numbers, population density, power, pressure, speed, temperature, time, torque, voltage, and volume. By identifying and delimiting these regular expressions, the relevant units are isolated and prepared for further annotation. This step allows to identify multiple Regextypes using regular expressions (*e.g.* numbers, geographic coordinates, address, code, color, URL). Since all values of type text are selected, preprocessing for natural languages was performed using the `langrid`[4] library to detect 26 languages in our data. By the way, it's a novelty for this year's SemTab campaign, *i.e.*, which makes the task more difficult with the introduction of natural language barriers. The `langrid` library serves as a standalone language identification tool and currently supports a wide range of languages (97 in total). It efficiently handles correction, data type identification, and language detection. By employing this library, the need for repetitive treatments in each subtask and for every cell in the tables is significantly reduced. This reduction in effort and cost can be substantial, as it allows us to avoid the massive repetition of these processes for all table cells.

☐ **Assigning a semantic type to a column (CTA) :**

The task is to annotate each entity column with elements from Wikidata (or possibly Dbpedia) as its type identified during the preprocessing phase. Each item is marked with the tag in Wikidata or Dbpedia. This treatment allows semantics identification. The CTA task can be performed based on Wikidata or Dbpedia APIs which allows us to search for an item according to its description. The main information collected about a given entity and used in our approach are: a list of instances (expressed by the `instanceOf` primitive and accessible by the P31 code), the subclass of (expressed by the `subclassOf` primitive and accessible by code P279) and overlaps (expressed by the `partOf` primitive and accessible by code P361). At this point, we are able to process the CTA task using a SPARQL query. The SPARQL query is our interrogation mean fed by the main information of the entity which governs the choice of each data type, since they are a list of instances (P31), of subclasses (P279) or a part of a class (P361). The result of the SPARQL query may return a single type but for some cases the result is more than one type, so in this case no annotation is produced for the CTA task.

☐ **Matching a cell to a KG entity (CEA)** :

[3]https://pandas.pydata.org
[4]https://github.com/openlangrid

The CEA task involves annotating the cells of a given table with specific entities from Wikidata or Dbpedia. Similar to the CTA task, our approach follows the same principles. We leverage the results obtained from the CTA task process and make necessary modifications to the SPARQL query for the CEA task. In cases where the operation returns more than one annotation, we address the ambiguity problem by examining the context of the column in question, relative to the results obtained from the CTA task. This contextual analysis helps us overcome any ambiguities and refine the annotations for the table cells.

□ **Matching a property to a KG entity (CPA)**

Once the cell values and their respective entity types have been annotated, the next step is to identify relationships between two cells that appear in the same row using a property through a SPARQL query. This task is known as the CPA task, which aims to annotate the relationship between two cells in a row via a property. Similar to the CTA and CEA tasks, the CPA task can be performed using a similar approach. However, in the CPA task, the SPARQL query needs to select both the entity and its corresponding attributes. The properties for establishing relationships are relatively easy to match since we have already determined them during the CEA and CTA task processing. In summary, the CPA task involves finding relationships between cell values using properties through SPARQL queries, and it complements the CTA and CEA tasks in annotating and linking the information in the given table.

□ **Disambiguation** It is important to recognize that an entity in Knowledge Graphs can be associated with multiple classes. This indicates that entities can have different classifications in Knowledge Graphs, reflecting the different aspects, roles, and characteristics associated with them. The presence of several classes for a enriches its representation and provides a more complete understanding of its semantic context within Knowledge Graphs. (For more details [3, 4, 5, 6, 7])

3. Kepler-aSI performance and results

In this section we will present the results of Kepler-aSI for the different matching tasks in the 2 rounds of SemTab 2023 (Table 1). We would like to report that the results are presented according to two scenarios, *i.e.*, before deadline and after deadline (since the organizers allow participants a period of 1 month before freezing the values). These results highlight the strengths of Kepler-aSI with its encouraging performance despite the multiplicity of issues. To measure the effectiveness of the data repair and data augmentation features of the Kepler-aSI.R.A process, we based on the results of Kepler-aSI, we used the following metrics proposed in [8] : Precision (P) , Recall (R) and F-measure (F1). (P), (R) and (F1) of the mapping between the datasets, the Wikidata KG and the DBpedia KG are calculated using the following formula: where a perfect annotation refers to the annotation returned by our approach, which corresponds to ground truth annotations, a submitted annotation refers to the annotation returned by our approach and a ground truth of the annotations corresponds to the number of annotations in the target tables. We combined the predefined measurements, which represent

the harmonic mean between P and R to calculate F1.

Table 1
Results for Kepler-aSI 2023

Task	**F1 Score**	**Precision**	**Rank**
Round1			
tFood-horizontal-TD	0.513	0.513	1
tFood-horizontal-CEA	-	-	
tFood-horizontal-CTA	0.105	0.105	1
tFood-horizontal-CPA	0.000	0.000	
tFood-entity-TD	0.000	0.000	
tFood-entity-CEA	-	-	
WikidataTablesR1-CEA	0.006	0.959	
WikidataTablesR1-CTA	0.739	0.739	3
WikidataTablesR1-CPA	0.777	0.777	3
R1-SOTAB-CTA-SCH	0.364	0.343	1
R1-SOTAB-CPA-SCH	0.235	0.230	1
Round2			
R2-SOTAB-CTA-SCH	0.3372	0.3601	
R2-SOTAB-CPA-SCH	-	-	
R2-SOTAB-CTA-DBP	0.4611	0.4983	
R2-SOTAB-CPA-DBP	0.1316	0.132	
R2-QA	-	-	

The SemTab 2023 challenge[5] has been organized into two different courses: **the Precision course**, which is the standard course offered in previous editions, which focuses on applications in real-world contexts where the output of matching systems can contribute. In this section we will present the results of Kepler-aSI for the different matching tasks in the two rounds of SemTab 2023. These results highlight the strengths of Kepler-aSI with its encouraging performance despite the multiplicity of issues.

4. Conclusion & Future Work

In this paper, we presented our contribution to the SemTab2023 challenge, Kepler-aSI. We tackled the several proposed tasks. Our solution is based on a generic SPARQL query using the cell contents as a description of a given item. In each round, despite the time allocated by the organizers running out, we continued the work and the improvements, having the conviction that each effort counts and brings us closer to the good control of the studied field.Kepler-aSI is a promising approach but which will be further improved: First, we will apply several methods yet to correct spelling mistakes and other typos in the source data. Finally, we will try to develop

[5]https://sem-tab-challenge.github.io/2023/

our system by integrating new data processing techniques (some Big Data oriented paradigms). Indeed, the parallel implementation will allow us to circumvent the data size problem, which is the major gap for our current machines. Eventually, the idea of moving to a data representation using indexes would be a good track to investigate in order to master the search space, formed by the considered tabular data.

References

[1] J. Chen, E. Jiménez-Ruiz, I. Horrocks, C. Sutton, Colnet: Embedding the semantics of web tables for column type prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 29–36.

[2] S. Malyshev, M. Krötzsch, L. González, J. Gonsior, A. Bielefeldt, Getting the most out of wikidata: Semantic technology usage in wikipedia's knowledge graph, in: International Semantic Web Conference, Springer, 2018, pp. 376–394.

[3] W. Baazouzi, M. Kachroudi, S. Faiz, A matching approach to confer semantics over tabular data based on knowledge graphs, in: Model and Data Engineering: 11th International Conference, MEDI 2022, Cairo, Egypt, November 21–24, 2022, Proceedings, Springer, 2022, pp. 236–249.

[4] W. Baazouzi, M. Kachroudi, S. Faiz, Towards an efficient fairification approach of tabular data with knowledge graph models, Procedia Computer Science 207 (2022) 2727–2736.

[5] W. Baazouzi, M. Kachroudi, S. Faiz, Kepler-asi: Kepler as a semantic interpreter., in: SemTab@ ISWC, 2020, pp. 50–58.

[6] W. Baazouzi, M. Kachroudi, S. Faiz, Kepler-asi at semtab 2021., in: SemTab@ ISWC, 2021, pp. 54–67.

[7] W. Baazouzi, M. Kachroudi, S. Faiz, Yet another milestone for kepler-asi at semtab 2022, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), CEUR-WS. org (2022).

[8] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, V. Cutrona, Results of semtab 2020, in: CEUR Workshop Proceedings, volume 2775, 2020, pp. 1–8.