KG2Tables: Your way to generate an STI benchmark for your domain

Nora Abdelmageed[1], Ernesto Jiménez-Ruiz[2,3], Oktie Hassanzadeh[4] and Birgitta König-Ries[1]

[1]Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Jena, Germany
[2]City, University of London, UK
[3]University of Oslo, Norway
[3]IBM Research, USA

Abstract

Tabular data, often found in CSV files, is essential for data analytics workflows. Understanding this data in a semantic context, known as Semantic Table Interpretation (STI), is critical but challenging due to issues like label ambiguity. Consequently, STI has garnered significant attention in recent years. To evaluate STI systems effectively, robust benchmarks are needed. Most existing large-scale benchmarks originate from general domain sources and emphasize ambiguity, whereas domain-specific benchmarks tend to be smaller. This paper presents KG2Tables, a framework designed to create large-scale domain-specific benchmarks from a Knowledge Graph (KG). KG2Tables utilizes the internal hierarchy of relevant KG concepts and their properties. As a proof of concept, we have developed extensive datasets in the food, biodiversity, and biomedical domains. One of these datasets was used in the ISWC 2023 SemTab challenge, and the rest have been integrated into SemTab 2024.

Keywords

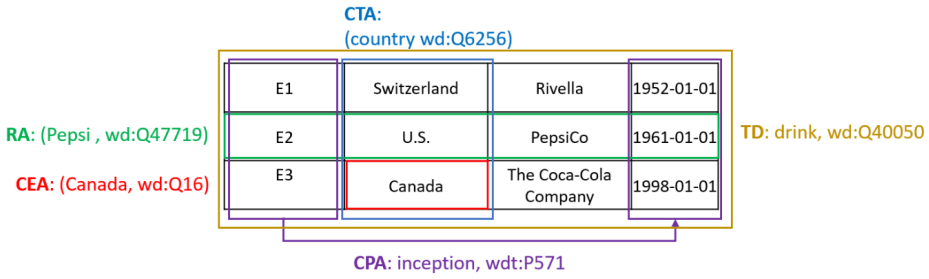
Semantic Table Interpretation (STI), Knowledge Graph (KG), Tabular Data, Benchmark, SemTab

1. Introduction

Semantic Table Interpretation (STI) has recently witnessed increasing attention from the community [1]. The goal of this process is to map individual table components, e.g., columns and cells, to entities and classes from a target Knowledge Graph (KG) such as Wikidata [2], or DBpedia [3]. Since 2019, the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)[1], which is running for the fifth time this year, has aimed at setting a common standard for evaluating STI systems [4, 5, 6, 7, 8]. It poses various challenges and benchmarks every year. Most of the datasets are Automatically Generated (AG) except for the Tough Tables (2T) [9] dataset that has been manually curated. All of these benchmarks, including 2T, are derived from the general domain. BiodivTab [10] is an exception, a manually annotated and biodiversity-specific dataset. The SemTab results show that this dataset was much harder to tackle than the domain-independent ones for state-of-the-art systems participating in the

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA
nora.abdelmageed@uni-jena.de (N. Abdelmageed); ernesto.jimenez-ruiz@city.ac.uk (E. Jiménez-Ruiz); hassanzadeh@us.ibm.com (O. Hassanzadeh); birgitta.koenig-ries@uni-jena.de (B. König-Ries)
0000-0002-1405-6860 (N. Abdelmageed); 0000-0002-9083-4599 (E. Jiménez-Ruiz); 0000-0001-5307-9857 (O. Hassanzadeh); 0000-0002-2382-9722 (B. König-Ries)

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



(a) Horizontal Relational Table.

(b) Entity Table.

Figure 1: A summary of Semantic Table Interpretation (STI) tasks

challenge. We believe this is due to domain-specific challenges that general-purpose systems are ill-equipped to handle or require extensive tuning or training data.

State-of-the-art STI tasks propose ways to annotate tabular data semantically and, thus, facilitate a potential transformation into a KG. We summarize them as follows: 1) Cell Entity Annotation (CEA) links a table cell value to a knowledge graph (KG) entity. 2) Column Type Annotation (CTA) assigns a semantic type to an entire column. 3) Column Property Annotation (CPA) connects a column pair (subject-object) to a semantic property from the KG. 4) Row Annotation (RA) maps an entire row to a KG entity, differing from CEA as the subject column might be missing. 5) Topic Detection (TD) classifies the table into a topic, such as a semantic class. Figure 1 gives an overview of the five most common STI tasks in two table types Horizontal Relation Tables in Figure 1 (a), and Entity Tables in Figure 1 (b). The former includes a set of entities row-wise. The latter represents a single entity with a list of its properties. The solution indicates that Wikidata is the target KG.

In this paper, we introduce `KG2Tables`[2], an STI benchmark generator that constructs both horizontal relational tables and entity tables, given a list of domain-specific concepts from Wikidata. This Zenodo dump[3] shows the code we refer to in this paper and lists six benchmarks we created using `KG2Tables` from three domains: food, biodiversity, and biomedical.

2. Methodology

`KG2Tables` accepts a list of related domain concepts in a CSV file, and constructs a tree structure for these concepts. In Wikidata, domain concepts form a graph structure since the KG allows that. However, `KG2Tables` process each relevant concept only once. We construct the respective

[2]https://github.com/fusion-jena/KG2Tables
[3]https://zenodo.org/records/10285835

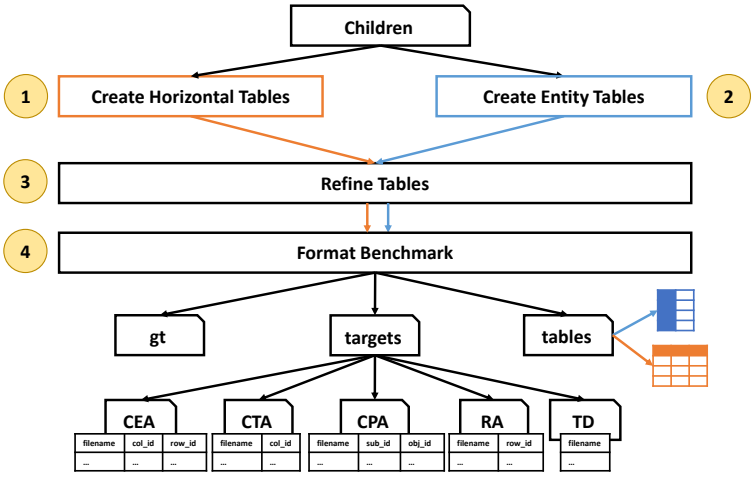


Figure 2: KG2Tables Generator Approach

tree structure using the internal hierarchy of the input concepts. e.g., in Wikidata, we have included all instances and subclasses via `wdt:P31, instance of` and `wdt:P279, subclass of`. We use the term "Children" to generalize related instances or subclasses. This tree structure will be different in DBpedia. In that case, the internal hierarchy is determined via `rdf:type` only. We applied a deduplication step since the overall instances and subclasses may overlap. Such overlap may also occur across different levels of the tree.

Figure 2 depicts the approach we developed to construct domain-specific benchmarks. It starts with the children of domain concepts, i.e., based on the current level of the recursion, and it consists of four steps: **(1) Create Horizontal Tables** and **(2) Create Raw Entity Tables**: we constructed both types of tables based on the properties of the current children; these tables contain the solutions of all STI tasks. We apply several operations for these properties, e.g., union, intersection, or random selection, to create different versions of the same table. **(3) Refine Tables**: we revised the collected data and applied several steps to construct the final tables, i.e., anonymizing column names. **(4) Format Benchmark**: we separated tables from solutions and targets to create a complete set of STI tasks. Targets indicate what to solve regarding column and row IDs, while solutions include the ground truth data of these targets.

3. Conclusions & Remarks

In this paper, we presented KG2Tables, a code generator that creates domain-specific tabular data benchmarks for Semantic Table Interpretation (STI) tasks. It uses the internal hierarchy of related concepts in a target Knowledge Graph (KG) to generate two types of tables: horizontal and entity tables. KG2Tables addresses five common STI tasks and was tested in three domains: Food, Biodiversity, and Biomedical. While our examples use Wikidata, KG2Tables is adaptable to any KG through SPARQL query modifications. KG2Tables accepts and parses a list of given domain concepts, thus validating and ensuring the domain specificity of the resulting dataset using either a data-driven approach or a reuse of existing domain-specific classes.

References

[1] B. Wanders, Repurposing and probabilistic integration of data, SIKS dissertation series, Universiteit Twente, 2016. Isbn:978-90-365-4110-7, number:2016-24.

[2] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. doi:10.1145/2629489.

[3] 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: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings, Springer, 2007, pp. 722–735.

[4] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems, in: The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings, Springer, 2020, pp. 514–530.

[5] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, V. Cutrona, Results of semtab 2020, 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. 1–8.

[6] V. Cutrona, J. Chen, V. Efthymiou, O. Hassanzadeh, E. Jiménez-Ruiz, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos, D. Oliveira, C. Pesquita, Results of SemTab 2021, 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. 1–12.

[7] N. Abdelmageed, J. Chen, V. Cutrona, V. Efthymiou, O. Hassanzadeh, M. Hulsebos, E. Jiménez-Ruiz, J. Sequeda, K. Srinivas, Results of SemTab 2022, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 21st International Semantic Web Conference (ISWC 2022)., volume 3320 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 1–13.

[8] O. Hassanzadeh, N. Abdelmageed, J. Chen, V. Cutrona, V. Efthymiou, M. Hulsebos, E. Jiménez-Ruiz, A. Khatiwada, K. Korini, B. Kruit, J. Sequeda, K. Srinivas, Results of SemTab 2023, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023)., volume 3557 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1–14.

[9] V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, M. Palmonari, Tough Tables: Carefully Evaluating Entity Linking for Tabular Data, in: 19th International Semantic Web Conference (ISWC), 2020, pp. 328–343.

[10] 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.