MammoTab: a giant and comprehensive dataset for Semantic Table Interpretation

Mattia Marzocchi[1], Marco Cremaschi[1], Riccardo Pozzi[1], Roberto Avogadro[1] and Matteo Palmonari[1]

[1]*University of Milan - Bicocca, viale Sarca 336, Edificio U14, 20126, Milan, Italy*

Abstract

In this paper, we present MammoTab, a dataset composed of 1M Wikipedia tables extracted from over 20M Wikipedia pages and annotated through Wikidata. The lack of this kind of datasets in the state-of-the-art makes MammoTab a good resource for testing and training Semantic Table Interpretation approaches. The dataset has been designed to cover several key challenges, such as disambiguation, homonymy, and NIL-mentions. The dataset has been evaluated using MTab, one of the best approaches of the SemTab challenge.

Keywords

Semantic Table Interpretation, Tabular Data, SemTab Challenge, Knowledge Graph

1. Introduction

A vast amount of information is provided as structured data on the Web in tables, this quantity grew over the years. This increase can be linked to the uptake of the Open Data movement, whose purpose is to make a large number of tabular data sources freely available, addressing a wide range of domains, such as finance, mobility, tourism, sports, or cultural heritage [1]. The massive availability of tabular data on the Web makes Web tables a valuable source to consider for data miners. For instance, these tables can be employed for data integration tasks or to construct and extend Knowledge Graphs (KGs) [2]. In this field, the table-to-KG matching problem, also referred to as Semantic Table Interpretation (STI), is the process of adding the semantic meaning of a table by mapping its elements (*i.e.*, cells/mentions, columns, rows) to semantic tags (*i.e.*, entities, classes, properties) from KGs (*e.g.*, Wikidata, DBpedia). This process is typically broken down into the following tasks: (i) cell/mentions to Knowledge Graph (KG) entity matching (CEA task), (ii) column to KG class matching (CTA task), and (iii) column pair to KG property matching (CPA task) [3]. In the last decade, the table-to-KG has collected much attention in the research community [3]. This interest is also certified by the introduction of the "SemTab: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching"

SemTab: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 2022

m.marzocchi@campus.unimib.it (M. Marzocchi); marco.cremaschi@unimib.it (M. Cremaschi); riccardo.pozzi@unimib.it (R. Pozzi); roberto.avogadro@unimib.it (R. Avogadro); matteo.palmonari@unimib.it (M. Palmonari)

0000-0003-0855-0245 (M. Marzocchi); 0000-0001-7840-6228 (M. Cremaschi); 0000-0002-4954-3837 (R. Pozzi); 0000-0001-8074-7793 (R. Avogadro); 0000-0002-1801-5118 (M. Palmonari)

international challenge[1], now in its 4th version. The challenge consists of different rounds in which groups of tables with different features and levels of difficulty have to be annotated. The increased interest in the STI has led to the construction of several datasets (gold standards) in the last decade. As it will be better described later, these datasets often include only a part of the characteristics of Web tables (*e.g.*, small tables with few mentions, easy to annotate semantically [4]).

As a consequence, we created a new dataset, MammoTab, composed of 980 254 tables extracted from 21 149 260 Wikipedia pages and annotated through Wikidata. The number and the different features of the tables make MammoTab a good resource for testing and/or training STI approaches. In particular, because of its dimension MammoTab is a useful tool to train data-hungry models, which require a vast amount of data (*e.g.*, entity linking systems based on large language models).

The rest of the paper is organised as follows. Section 2 will present a brief analysis of the state-of-the-art related datasets used in the context of the STI. Subsequently, MammoTab is described (Section 3), its characteristics are listed (Section 3.1), and the pipeline used for its implementation is presented (Section 3.2). The evaluation is depicted in Section 4 through a state-of-the-art STI approach.

2. Datasets

Although several approaches deal with semantic annotations on tabular data, there are limited Gold Standards (GSs) for assessing the quality of these annotations. The main ones are T2Dv2, Limaye, Musicbrainz, IMDB, Taheryan, Tough Table and SemTab. Table 1 shows statistics for the GSs.

Table 1
Statistics for the most common datasets. '-' indicates unknown.

GS		Tables	Columns	Rows	Classes	Entities	Predicates	KG
T2Dv2 [5]		234	1 157	27 996	39	-	154	DBpedia
Limaye [6]		6 522	-	-	747	142 737	90	Wikipedia Yago
LimayeAll [7]		6 310	28 547	135 978	-	227 046	-	Freebase
Limaye200 [7]		200	903	4 144	615	-	361	Freebase
MusicBrainz [7]		1 406	9 842	-	9 842	93 266	7 030	Freebase
IMDB [7]		7 416	7 416	-	7 416	92 321	-	Freebase
Taheriyan [8]		29	2 467	16 006	-	-	-	Schema
Tough Table (2T) [9]		180	194 438	802	540	667 244	0	Wikidata DBpedia
SemTab2019	**R1**	64	320	9 088	120	8 418	116	DBpedia
	R2	11 924	59 620	298 100	14 780	463 796	6 762	
	R3	2 161	10 805	153 431	5 752	406 827	7 575	
	R4	817	3 268	51 471	1 732	107 352	2 747	
SemTab2020	**R1**	34 294	170 068	249 329	135 773	985 109	135 773	Wikidata
	R2	12 173	55 951	84 896	43 752	283 446	43 752	
	R3	62 614	229 321	396 903	166 632	768 324	166 632	
	R4	22 387	79 552	670 335	32 461	1 662 164	56 475	
SemTab2021	**R1**	180	802	194438	539	667243	56475	Wikidata
	R2	1750	5589	29280	2190	47439	3835	
	R3	7207	17902	58949	7206	58948	10694	

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

An excellent STI approach must consider and adequately balance the different features of a table (or a set of tables). The annotation involves several key challenges: i) *disambiguation*: the class of the entities described in a table are not known in advance, and those entities may correspond to more than one class in the KG. ii) *homonymy*: this issue is related to the presence of different entities with the same name and class. iii) *matching*: the mention in the table may be syntactically different from the label of the entity in a KG (*i.e.*, use of acronyms, aliases and typos). iv) *NIL-mentions*: the approach must also consider strings that refer to entities for which a representation has not yet been created within the KG, namely NIL-mentions. v) *literal and named-entity*: in a table, there can be columns that contain named-entity mentions (NE-column) and columns containing strings (L-column). vi) *missing context*: it is often easier to extract the context from textual documents than from tables due to the amount of content to be processed. For instance, the header (*i.e.*, the first row of a table) which usually contains descriptive attributes for the columns, may or may not be present. vii) *amount of data*: the approach must consider large tables with many rows and columns, and tables with very few mentions. viii) *different domains*: the tables within a set can belong to very general or specific domains. MammoTab has been designed to cover all these cases, making it a resource for evaluating or training STI approaches.

The dataset is made up of tables automatically extracted from Wikipedia. Some pipelines for extracting tables from Wikipedia have been presented in the state-of-the-art. Among these, TabEL [10] proposes a STI approach and a dataset (WikiTable corpus) composed of 1.6M tables from the November 2013 XML dump of English Wikipedia. The dataset focuses only on the CEA task using YAGO as reference KG. It should be noted that the WikiTable is outdated and the code has not been made available[2]. Another paper [11] proposes a dataset composed of 670 171 tables extracted from WikiTable corpus to which the Wikipedia page title, section title, and table caption have been added to obtain a more comprehensive description. However, the dataset has not been released, nor has the source code for its generation.

3. The MammoTab Dataset

The annotations inside MammoTab are based on Wikidata v. 20220520 and are provided following the structure used by the SemTab challenge. One table is stored in one CSV file, and each line corresponds to a table row. The target columns for annotation, CTA, and CEA annotations are saved in CSV files. A JSON document has also been created for each Wikipedia page with additional information about the tables extracted by that page (see Listing 1). We released MammoTab in Zenodo[3] following the FAIR Guiding Principles[4]. The dataset is released under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) licence[5]. MammoTab contains a variety of tables that allow to evaluate approaches considering the challenges previously listed: in the dataset there are tables containing i) entities hard to disambiguate (*e.g.*, table id LBJJ1WGD - reactor Clinton)[6], ii) cases of homonymy (*e.g.*, table id MRBWAAOA - soccer player

[2] websail-fe.cs.northwestern.edu/TabEL/

[3] zenodo.org/deposit/7014472

[4] www.nature.com/articles/sdata201618

[5] creativecommons.org/licenses/by-sa/4.0/

[6] en.wikipedia.org/wiki/List_of_cancelled_nuclear_reactors_in_the_United_States

Michael Jordan)[7], iii) aliases (table id JCND1XGG - Tom Riddle alias of Lord Voldemort)[8], and iv) NIL-mentions (*e.g.*, table id MSTBGKPR - KKOP-LP Wildcat Broad. Inc)[9].

3.1. Dataset Profile

The MammoTab tables were extracted from 21 149 260 Wikipedia pages using the XML dump[10]. In these pages 2 803 424 tables were detected. Among these, the tables with at least three links in the same column have been stored for a total amount of 980 254 tables. Some dataset statistics are reported in Table 2.

Table 2
Overall statistics of the MammoTab dataset: the total number of *Tables*, *Columns*, and *Rows*; the minimum, maximum, and the average number of columns and rows per table. *Linked Cells* refers to the number of cells linked to an entity, *Typed Cols* refers to the number of columns with a known class associated, and *NILs* refers to the number of NIL-mentions.

Tables	Columns				Rows				Linked Cells	Typed Cols	NILs
	total	min	max	avg	total	min	max	avg			
980 254	5 638 191	1	500+	5.75	23 376 498	3	14 436	23.85	28 446 720	2 001 902	4 686 457

3.2. Implementation

The dataset was built through a pipeline implemented as a set of Python scripts, which are available in a Git repository[11].

The pipeline consists of 10 steps: i) *dump processing*: each Wikipedia XML dump file (we used the multiple bz2 stream dumps for easier parallelisation) is parsed using BeautifulSoup[12] to extract each page; ii) *tables identification*: pages are scanned to find those that contain at least one table (wikitext syntax: | class=wikitable |); iii) *tables normalisation and cleaning*: each cell is normalised and cleaned up using custom and wikitextparser[13] functions. Elements such as subscripts, superscripts, elements of the wikitext syntax, images, Wikipedia help and project pages links, and links to external pages are removed; iv) *tables analysis*: each table is analysed to check for cells that contain links to Wikipedia pages (wikitext syntax: [[link]]). A cell is considered as a mention of an entity only if the entire cell is a link. The remaining cells, containing multiple links or additional words around the link, may also be mentions of entities, but we consider them uncertain and mark them as UNKNOWN; v) *tables storing*: tables that have at least three fully linked cells in a column are stored; vi) *table header and table caption detection*: table header (wikitext syntax: !header) and table caption, if any, are stored and added to the current table; vii) *column analysis*: each column is analysed and classified into Literal columns (L-column) for datatype values (*e.g.*, strings, numbers, dates, such as 4808, 10/04/1983),

[7] en.wikipedia.org/wiki/USA_Today_All-USA_high_school_football_team
[8] en.wikipedia.org/wiki/Christian_Coulson
[9] en.wikipedia.org/wiki/List_of_radio_stations_in_Nebraska
[10] dumps.wikimedia.org/enwiki/20220720/
[11] bitbucket.org/disco_unimib/mammotab/
[12] www.crummy.com/software/BeautifulSoup/
[13] github.com/5j9/wikitextparser

or Named-Entity columns (NE-column) if it contains links to Wikipedia pages; viii) *entity linking - CEA*: for each Wikipedia link, the related Wikidata entity is extracted; ix) *column annotation - CTA*: column types are set by choosing the most specific entity class (according to Wikidata subclass relationships) that is shared by most of the column rows. For columns with less than 5 rows, all cells must be instances of that class, while, for bigger columns, at least 60% of the cells are instances of that class; x) *NIL-identification*: we mark as NIL the cells containing Wikipedia red links[14], which are those links referring to a page that does not exist.

The Listing 1 shows an example of a JSON document used to manage the results of the process described above on the Wikipedia page about "As Long as You Love Me (Justin Bieber song)"[15].

Listing 1: JSON document with the information relating to a Wikipedia page that contains at least one table.

```
{"wiki_id": '36115735',
 "title": 'As Long as You Love Me (Justin Bieber song)',
 "tables": {
  "XXI7BFMW": {
    "caption": 'Promotional release dates for "As Long as You Love Me"',
    "header": [['Region', 'Date', 'Format', 'Label']],
    "link": [['', '', '', ''],
             ['', '', 'Music_download', 'Island_Records'],
             ['', '', 'Music_download', 'Island_Records'],
             ...],
    "text": [['Region', 'Date', 'Format', 'Label'],
             ['United States', 'June 11, 2012', 'Digital Download', 'Island Records'],
             ['Canada', 'June 11, 2012', 'Digital Download', 'Island Records'],
             ...],
    "target_col": [2],
    "entity": [['', '', '', ''],
               ['', '', 'Q6473564', 'Q190585'],
               ['', '', 'Q6473564', 'Q190585'],
               ...],
    "types": [[[], [], [], []],
              [[], [], ['Q81941037'], ['Q18127']],
              [[], [], ['Q81941037'], ['Q18127']],
              ...],
    "col_types": [[], [], [['Q81941037', 0.8571428571428571]], [['Q18127', 0.2857142857142857]]],
    "col_type_perfect": ['', '', 'Q81941037', '']}}}
```

A re-run of the Python scripts, simply pointing a new Wikipedia XML dump file to process[16], allows to obtain a new version of the dataset.

4. Evaluation

The experiments in this Section aim to demonstrate how the use of MammoTab allows identifying the weaknesses of a STI approach, with particular reference to the key challenges reported in Section 2. The Mtab [12] approach was considered since it won several versions of the SemTab challenge. A sample of 5 000 tables was selected without NIL-mentions due to the limitations of the Mtab[17] free API. Table 3 reports the results obtained by the approach.

The values obtained by MTab versus MammoTab are lower than the other datasets. A substantial decrease in the F1-Score in the CTA can also be noted. The results show that in the current version, the MammoTab tables are a valuable resource for testing STI approaches which must be characterised by sophisticated mechanisms that consider many semantics aspects. However, as done in other datasets [9], it is possible to add some noise (*i.e.*, adding misspelt or fake mentions) to increase the complexity of the annotation task.

[14] en.wikipedia.org/wiki/Wikipedia:Red_link
[15] en.wikipedia.org/wiki/As_Long_as_You_Love_Me_(Justin_Bieber_song)
[16] dumps.wikimedia.org/backup-index.html
[17] mtab.app/mtab/docs

Table 3
Results (F1-Score) obtained by the MTab approach on different datasets.

Mtab [12] on	**CEA**	**CTA**	**CPA**
SemTab2019 R4	0.983	-	0.832
SemTab2020 R4	0.907	0.993	0.997
SemTab2020 2T	0.907	0.728	-
SemTab2021 R3	0.968	0.984	0.993
MammoTab	**0.853**	**0.659**	-

References

[1] S. Neumaier, J. Umbrich, J. X. Parreira, A. Polleres, Multi-level semantic labelling of numerical values, in: The Semantic Web – ISWC 2016, Springer International Publishing, Cham, 2016, pp. 428–445.

[2] M. Kejriwal, C. A. Knoblock, P. Szekely, Knowledge graphs: Fundamentals, techniques, and applications, MIT Press, 2021.

[3] V. Cutrona, J. Chen, V. Efthymiou, O. Hassanzadeh, E. Jimenez-Ruiz, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos, D. Oliveira, C. Pesquita, Results of semtab 2021, in: 20th International Semantic Web Conference, volume 3103, CEUR Workshop Proceedings, 2022, pp. 1–12.

[4] S. Zhang, E. Meij, K. Balog, R. Reinanda, Novel entity discovery from web tables, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1298–1308.

[5] D. Ritze, C. Bizer, Matching web tables to dbpedia - a feature utility study, in: Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017, OpenProceedings, Konstanz, 2017, pp. 210–221.

[6] G. Limaye, S. Sarawagi, S. Chakrabarti, Annotating and searching web tables using entities, types and relationships, Proc. VLDB Endow. 3 (2010) 1338–1347.

[7] Z. Zhang, Effective and efficient semantic table interpretation using tableminer+, Semantic Web 8 (2017) 921–957.

[8] M. Taheriyan, C. A. Knoblock, P. Szekely, J. L. Ambite, Leveraging linked data to discover semantic relations within data sources, in: The Semantic Web – ISWC, 2016, pp. 549–565.

[9] V. Cutrona, F. Bianchi, E. Jimenez-Ruiz, M. Palmonari, Tough tables: Carefully evaluating entity linking for tabular data, in: The Semantic Web - ISWC 2020, Lecture Notes in Computer Science, Springer International Publishing, 2020, pp. 328–343.

[10] C. S. Bhagavatula, T. Noraset, D. Downey, Tabel: Entity linking in web tables, in: M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan, S. Staab (Eds.), The Semantic Web - ISWC 2015, Springer International Publishing, Cham, 2015, pp. 425–441.

[11] X. Deng, H. Sun, A. Lees, Y. Wu, C. Yu, Turl: Table understanding through representation learning, SIGMOD Rec. 51 (2022) 33–40.

[12] P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, H. Takeda, Semtab 2021: Tabular data annotation with mtab tool., in: SemTab@ ISWC, 2021, pp. 92–101.