STILTool: a Semantic Table Interpretation evaLuation Tool

Marco Cremaschi[1], Alessandra Siano[1], Roberto Avogadro[1], Ernesto Jimenez-Ruiz[2,3], and Andrea Maurino[1]

[1] University of Milan - Bicocca, Milan, Italy
{marco.cremaschi,alessandra.siano,roberto.avogadro, andrea.maurino}@unimib.it
[2] City, University of London, London, United Kingdom
ernesto.jimenez-ruiz@city.ac.uk
[3] University of Oslo, Oslo, Norway

Abstract. This paper describes STILTool, an open-source tool for the automatic evaluation of the quality of semantic annotations computed by semantic table interpretation approaches. STILTool provides a graphical interface allowing users to analyse the correctness of the annotations of tabular data. The tool also provides a set of statistics in order to identify the most common error patterns.

Keywords: Semantic Web · Ontology · Linked Data · Knowledge Graph · Semantic Table Interpretation · Semantic Annotations · Tabular data

1 Introduction & Motivation

Much information is conveyed within tables. Just think of the relational databases or tables present on the Web pages. In order to size the spread of tabular data, 2.5M tables have been identified within the Common Crawl repository[1] [3]. The current snapshot of Wikipedia contains more than 3.23M tables from more than 520k Wikipedia articles [1]. The tables may contain high-value data, but due to the lack of contextual information or meta-data, they can be challenging to understand, both for humans and for machines. In order to solve this problem, several techniques have been proposed in the state-of-the-art whose aim is the semantic annotation of tabular data using information extracted from a Knowledge Graph (KG) (*e.g.*, DBpedia[2]). Inside a Semantic Table Interpretation (STI) process, it is possible to identify three main tasks:

1. assigning a semantic type (*e.g.*, a KG class) to a column (Column Type Annotation (CTA));
2. matching a cell to a KG entity (Cell Entity Annotation (CEA));
3. assigning a KG property to the relationship between two columns (Column Predicate Annotation (CPA)).

[1] commoncrawl.org
[2] wiki.dbpedia.org

Although several approaches deal with semantic annotations on tabular data, there are limited Gold Standards (GSs) for the assessment of the quality of these annotations. The main ones are T2Dv2, Limaye, Musicbrainz, IMBD, Taheryan 2015 and SemTab 2019. Table 1 shows statistics for these GSs.

T2Dv2[3] Gold Standard (GS) consists of a manually annotated dataset of 779 Web tables extracted from Web Table Corpora[4]. Inside this dataset, only 234 tables share at least one instance with DBpedia.

Limaye [4] consists of over 6,000 tables extracted from Wikipedia and the general Web. Entities in the tables are annotated with links to Wikipedia articles; columns and relations between columns are annotated by concepts and properties from the YAGO KG[5]. Limaye 200 [6] is a subset of the Limaye dataset; it is composed of 200 tables annotated using a manual and an automatic process. LimayeAll [6] is another version of Limaye, re-annotated through an automatic process. It contains 6,310 tables and the annotations are extracted from Freebase.

MusicBrainz [6] is composed of a set of annotated tables extracted from MusicBrainz record label webpages[6]. Each MusicBrainz record label webpage contains a table listing the music released by a production company. The reference KG is Freebase.

The IMDB [6] is composed of annotations related to a dataset of 7,416 tables about film extracted from a set of web pages of the IMDB[7].

Taheriyan 2016 [5] is composed of two datasets manually annotated. The first dataset contains 29 tables related to museum works annotated through two different ontologies (*i.e.*, CIDOC-CRM and the European Data Model, EDM). In the second, there are 15 tables about weapons interpreted using the Schema.org ontology.

The SemTab[8] challenge [2] presents a common framework to conduct a systematic evaluation of tabular data to KG matching systems. SemTab is composed of several evaluation rounds and relies on an automated method to generate benchmark datasets. The target KG in 2019 was DBpedia, but other KGs will be used in future editions of SemTab (*e.g.*, Wikidata[9] will be introduced in the 2020 edition).

The discrepancy between the KG used for annotations, the structure of the tables, the various storage formats (*e.g.*, CSV, JSON, XML, HTML) and the absence of some types of annotations makes it challenging to use these datasets for the evaluation of STI approaches. Besides, in the state-of-the-art, there are only two scripts[10] to automate the evaluation. One provides only a command-line

[3] webdatacommons.org/webtables/goldstandardV2.html
[4] webdatacommons.org/webtables/
[5] github.com/yago-naga/yago3
[6] musicbrainz.org/label/13a464dc-b9fd-4d16-a4f4-d4316f6a46c7
[7] www.imdb.com
[8] www.cs.ox.ac.uk/isg/challenges/sem-tab/
[9] www.wikidata.org/
[10] *(i)* Web Data INTEgRation Framework: github.com/olehmberg/winter; and *(ii)* SemTab evaluator: github.com/sem-tab-challenge/aicrowd-evaluator

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

GS		Tables	Columns	Rows	Classes	Entities	Predicates	KG
T2Dv2		234	1,157	27,996	39	-	154	DBpedia
Limaye		6,522	-	-	747	142,737	90	Wikipedia and Yago
LimayeAll		6,310	28,547	135,978	-	227,046	-	Freebase
Limaye200		200	903	4,144	615	-	361	Freebase
MusicBrainz		1,406	9,842	-	9,842	93,266	7,030	Freebase
IMDB		7,416	7,416	-	7,416	92,321	-	Freebase
Taheriyan		29	2,467	16,006	-	-	-	CIDOC-CRM EDM Model Schema.org
SemTab 2019	**Round 1**	64	320	9,088	120	8,418	116	DBpedia
	Round 2	11,924	59,620	298,100	14,780	463,796	6,762	
	Round 3	2,161	10,805	153,431	5,752	406,827	7,575	
	Round 4	817	3,268	51,471	1,732	107,352	2,747	

interface, the other, instead, has been integrated into a multi-purpose platform[11] which aims to propose real-world problems as challenges to find collaborative solutions; in this case, the evaluation is provided only in the form of scores.

For this reason, we have implemented STILTool[12], a web application to automate the quality assessment of the annotations produced by STI approaches.

2 Overview of STILTool

The purpose of STILTool is to provide a reliable tool for the evaluation of annotations. The evaluation is carried out by comparing the semantic annotations with one or more GSs.

It is developed as a web application with the Python-based Django framework[13] and MongoDB[14] as a database. The code is freely available through a Git repository[15]. In order to achieve the scalability of the application, and therefore improve efficiency and to facilitate the deployment on servers, STILTool has been installed in a Docker container.

An authentication system has been integrated to allow users to have their own set of annotations and GS stored privately.

STILTool is composed of three main parts: *(i)* loading data, *(ii)* evaluate annotations, and *(iii)* compare results.

Loading data. STILTool allows users to upload both a set of annotations and GSs. A GS is composed at least by the annotations for one of the three main STI tasks (*i.e.*, CTA, CPA, CEA) in CSV format; a score criterion is automatically defined based on the task and the type of annotation.

Furthermore, a GS has to define its availability to other users: it can be *(a) private* - it is accessible only for the owner, or *(b) public* - it is accessible for all

[11] www.aicrowd.com
[12] zoo.disco.unimib.it/stiltool/
[13] www.djangoproject.com
[14] www.mongodb.com
[15] bitbucket.org/disco_unimib/stiltool/

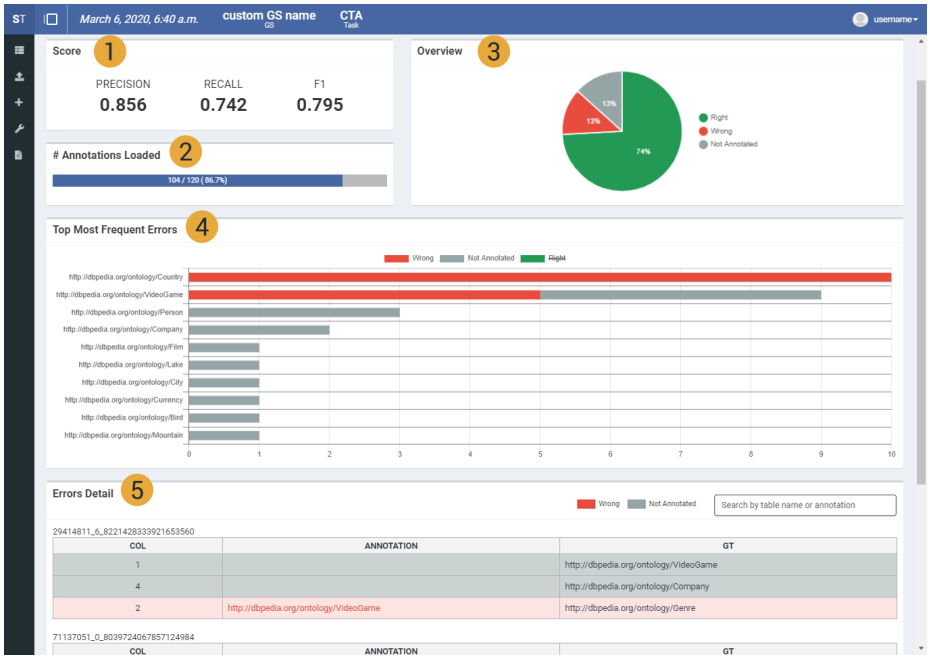

Fig. 1. STILTool "Evaluate Annotations" section: Global Info (1. score, 2. number of annotations loaded, 3. right/wrong/missing annotations chart), 4. Most recurrent errors, 5. Side by side comparison

users. A *public GS* has an additional configuration parameter to define how users can access it: *(a) score mode* - user gets only the score from the evaluation (*e.g.*, during a challenge where GS annotations should not be provided to participants); or *(b) info mode*: user gets the score and some detailed statistics and info about the evaluation.

Uploading annotations (in CSV format) require the user to select the STI task to evaluate and the GS for the comparison.

Evaluate Annotations. This section provides some detailed information and statistics about the annotations provided; it is only available for annotations compared against a *private GS* or *public-info mode GS*. Data displayed are grouped by three main categories: *(i)* global info, *(ii)* most recurrent errors, and *(iii)* side by side comparison of the user and GS annotations.

Global Info. To give the user an overall overview of the annotation evaluations results, some general info such as the obtained score (Figure 1(1)), the total loaded annotations (Figure 1(2)) or the percentage of right/wrong/missing annotations (Figure 1(3)) is displayed.

Most recurrent errors. A list of the ten most wrong and missing annotations is displayed using a bar chart, as showed in Figure 1(4); this visualisation allows the user to visually identify the common error patterns that occur in the annotations.

Side by side comparison. To allow a detailed data analysis, the wrong or missing annotations are displayed side by side with the GS ones, grouped by table (Figure 1(5)).

Compare results. All loaded annotations sets are displayed grouped by GS and task. For each, some global info is displayed (*i.e.*, the obtained score and the completeness of the annotations against the target GS). A line chart is used to display the score across the different uploads to allow a comparison of the results in time. The data displayed in the chart can be filtered.

3 Conclusion

STILTool is a web application which aims to automate the quality assessment of semantic annotations produced by STI approaches. It offers a graphical interface to analyse in detail the results of the evaluation and to track how a STI approach improves in time. Using the different settings provided by the tool, it can be used as a generic evaluation tool or as the underlying platform for a STI challenge. Regarding this, STILTool will be tested during the SemTab 2020 challenge.

As a future development, we could consider extending the functionality of the tool to cover different fields that use similar Gold Standards and metrics or integrate the system with other evaluation tools.

4 Acknowledgments

Special thanks to Andrea Barazzetti and David Chieregato for their support during the development of the project. EJR was supported by the SIRIUS Centre for Scalable Data Access (Research Council of Norway).

References

1. Fetahu, B., Anand, A., Koutraki, M.: TableNet: An Approach for Determining Fine-Grained Relations for Wikipedia Tables. In: The World Wide Web Conference (WWW). p. 2736–2742 (2019)
2. Jimenez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K.: SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In: The Semantic Web: ESWC 2020. Springer International Publishing (2020)
3. Lehmberg, O., Ritze, D., Meusel, R., Bizer, C.: A Large Public Corpus of Web Tables Containing Time and Context Metadata. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 75–76 (2016)
4. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and Searching Web Tables Using Entities, Types and Relationships. Proc. VLDB Endow. **3**(1-2), 1338–1347 (2010)
5. Taheriyan, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: Leveraging linked data to discover semantic relations within data sources. In: The Semantic Web – ISWC. pp. 549–565 (2016)
6. Zhang, Z.: Effective and Efficient Semantic Table Interpretation using TableMiner+. Semantic Web **8**(6), 921–957 (2017)