SOTAB: The WDC Schema.org Table Annotation Benchmark

Keti Korini[1,*], Ralph Peeters[1] and Christian Bizer[1]

[1]Data and Web Science Group, University of Mannheim, Mannheim, Germany

Abstract

Understanding the semantics of table elements is a prerequisite for many data integration and data discovery tasks. Table annotation is the task of labeling table elements with terms from a given vocabulary. This paper presents the WDC Schema.org Table Annotation Benchmark (SOTAB) for comparing the performance of table annotation systems. SOTAB covers the column type annotation (CTA) and columns property annotation (CPA) tasks. SOTAB provides ~50,000 annotated tables for each of the tasks containing Schema.org data from different websites. The tables cover 17 different types of entities such as movie, event, local business, recipe, job posting, or product. The tables stem from the WDC Schema.org Table Corpus which was created by extracting Schema.org annotations from the Common Crawl. Consequently, the labels used for annotating columns in SOTAB are part of the Schema.org vocabulary. The benchmark covers 91 types for CTA and 176 properties for CPA distributed across textual, numerical and date/time columns. The tables are split into fixed training, validation and test sets. The test sets are further divided into subsets focusing on specific challenges, such as columns with missing values or different value formats, in order to allow a more fine-grained comparison of annotation systems. The evaluation of SOTAB using Doduo and TURL shows that the benchmark is difficult to solve for current state-of-the-art systems.

Keywords

Table Annotation, Column Type Annotation, Column Property Annotation, Schema.org

1. Introduction

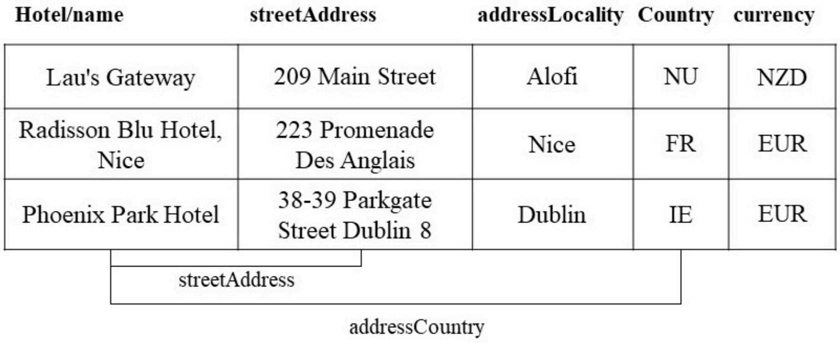
Tables containing structured data are widely used on the Web. Understanding the semantics of tables is useful for a variety of data integration and data discovery tasks such as knowledge base augmentation [1] or dataset search [2]. Table annotation is the task of annotating a table with terms from given vocabulary, knowledge graph, or database schema. Table annotation includes tasks such as *Column Type Annotation* (CTA) and *Column Property Annotation* (CPA). CTA is the annotation of table columns with the type of the entities contained in a column. CPA refers to the annotation of pairs of table columns with labels that indicate the relationship between the main column of the table and another column. Figure 1 shows an example of a table describing hotels with CTA labels shown above the table and CPA labels below.

SemTab@ISWC 2022, October 23–27, 2022, Hangzhou, China (Virtual)

[*]Corresponding author.

kkorini@uni-mannheim.de (K. Korini); ralph.peeters@uni-mannheim.de (R. Peeters); christian.bizer@uni-mannheim.de (C. Bizer)
0000-0002-2158-0070 (K. Korini); 0000-0003-3174-2616 (R. Peeters); 0000-0003-2367-0237 (C. Bizer)

CEUR Workshop Proceedings (CEUR-WS.org)



Hotel/name	streetAddress	addressLocality	Country	currency
Lau's Gateway	209 Main Street	Alofi	NU	NZD
Radisson Blu Hotel, Nice	223 Promenade Des Anglais	Nice	FR	EUR
Phoenix Park Hotel	38-39 Parkgate Street Dublin 8	Dublin	IE	EUR

streetAddress
addressCountry

Figure 1: An example table from SOTAB showcasing CTA and CPA labels

This paper presents the *WDC Schema.org Table Annotation Benchmark* (SOTAB) for comparing the performance of table annotation systems on the CTA and CPA tasks. The CTA dataset consists of 59,548 tables covering 17 Schema.org [3] types of which 162,351 columns have been annotated using 91 Schema.org types and properties. The CPA dataset consists of 48,379 tables where 174,998 column pairs are annotated using 176 Schema.org properties. The tables used in the benchmark originate from the WDC Schema.org Table Corpus which was created by extracting Schema.org annotations from the December 2020 version of the Common Crawl. Each table in the corpus contains all entities of a specific Schema.org type that are provided by a specific host, for example all movies annotated on *imdb.com*. The columns of the table are the attributes that are used by the host for describing the entities. Overall, the SOTAB tables contain data gathered from 74,215 different hosts which makes the benchmark data quite heterogeneous.

2. Related Work

This section provides an overview of the existing benchmarks for evaluating table annotation systems and compares them to SOTAB in Table 1. The ToughTables, HardTables, BioDiv and GitTables datasets are used by the *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching* (SemTab) which is a benchmark competition for table annotation systems that takes place every year as part of the International Semantic Web Conference [4]. The *ToughTables* (2T) [5] dataset's tables are divided into easily solvable tables and tough tables harder to predict and are annotated for CTA. The *Hard Tables* [6] dataset was generated querying DBpedia using SPARQL queries to create tables that resemble tables found in the Web. The *GitTables* [7] corpus is made of 1M tables from GitHub annotated with DBpedia and Schema.org [3] classes and properties. A subset of this corpus was annotated for CTA. The *BioDiv* [8] dataset's tables belong to the biodiversity domain and contain numerical data and abbreviations in their column values. Further datasets for benchmarking table annotation systems include: *T2Dv2* [9] which consists of Web tables annotated with DBpedia properties; *WikiTables-TURL* [10] which consists of Wikipedia tables annotated using terms from Freebase [11]; and *RedTab* [12] which offers tables belonging to the music and literature domain and being manually annotated for CPA.

Table 1
Overview of existing CTA and CPA benchmarks. The Labels column reports the number of unique labels used to annotate table columns. The KG/VOC column names the vocabulary or knowledge graph used for annotation: DBpedia (DBP), WikiData (WD), Schema.org (SCH) or Freebase (FB).

		Median			CTA		CPA	
Benchmark	Tables	Rows	Cols	KG/VOC	Columns	Labels	Columns	Labels
T2Dv2 [9]	779	4	18	DBP	-	-	670	119
Hard Tables [6]	8,957	7	2	WD	9,398	2,235	14,531	472
GitTables [7]	1,101	25	11	SCH/DBP	721/2,533	59/122	-	-
2T [5]	180	89	4	DBP/WD	540	39/276	-	-
BioDiv [8]	50	99	17	WD	614	92	-	-
WikiTable [10]	580,171	8	5	FB	654,670	255	67,201	121
REDTab [12]	9,149	5	18	-	-	-	22,236	23
SOTAB (ours)	107,927	42	8	SCH	162,351	91	174,998	176

3. Creation of the SOTAB Benchmark

This section describes the selection of the SOTAB tables from the WDC Schema.org Table Corpus, the assignment of lables to the tables, as well as well the selection of challenging columns into specific subset of the test set.

The WDC Schema.org Table Corpus was created using Schema.org data that was extracted from the December 2020 version of the Web Data Commons Microdata and JSON-LD corpus. The Schema.org Table Corpus consists of 4.2 million relational tables covering 43 Schema.org types. Each table in the corpus contains the descriptions of all entities of a specific Schema.org type that were extracted for a specific host, e.g. all movie records from imdb.com or product records from ebay.com. All extracted entities of one Schema.org type are collected per host and subsequently passed through a pipeline of processing and cleansing steps. For more details about the Schema.org Table Corpus we refer the reader to the project website[1].

Table Selection. We begin building SOTAB with a language identification phase on the tables from the Schema.org Table Corpus. The language identification phase aims to filter out non-English rows. For this purpose, we use the fastText language identification model [13, 14] and keep rows from tables where the model is at least 50% confident that the language is English. Finally, we remove all tables with less than 10 remaining rows and less than 3 columns.

Label Generation. As the WDC Schema.org Table Corpus uses Schema.org properties as column headers, these properties can be directly used as labels for the CPA task. We derive the CTA label for a column from its CPA label using the Schema.org vocabulary definition which specifies the types that are allowed as property values. In cases where the vocabulary definition allows multiple types, a manual selection of the most appropriate type is done. Lastly,

[1] http://webdatacommons.org/structureddata/schemaorgtables/

we add some CTA annotations that are not included in the Schema.org vocabulary such as *IdentifierAT*, *MusicArtistAT* and *Museum/name*. This is done with the purpose of including more fine-grained labels instead of for example simply *name*, so that an annotation system needs to better understand the semantics to select the correct annotation. After assigning the labels, another filtering step is performed based on label frequency: We only keep the columns that have CPA and CTA labels that are used at least 50 times.

Test Sets for Specific Challenges. In addition to the full test sets for both tasks, we provide subsets of the test sets that measure how good the systems can handle specific annotation challenges. We provide test sets for the following challenges: (i) *Missing Values*: This set contains columns having a value density between 10 and 70 percent and are thus harder to predict, (ii) *Format Heterogeneity*: columns whose values are represented using different value formats such as date or weight columns and (iii) *Corner Cases*: columns that are difficult to annotate as their values are very similar to the values of other columns. Examples of corner cases include startDate versus endDate and currenciesAccepted versus priceCurrency.

Table 2
Statistics of the SOTAB Datasets

		Training Large	Training Small	Validation	Test (Full)	Test (MV)	Test (FH)	Test (CC)	Test (RC)
CTA	Tables	46,790	11,517	5,732	7,026	1,369	475	1,754	4,563
	Columns	130,471	33,004	16,840	15,040	2,808	619	3,015	8,598
CPA	Tables	37,128	9,435	4,771	6,480	1,479	1,101	2,129	5,024
	Columns	134,425	33,643	17,417	23,156	4,032	1,593	3,492	14,039

4. Profiling the SOTAB Benchmark

The selection and labeling phase results in 46,790 tables annotated for the CTA task and 48,379 tables annotated for the CPA task which cover 17 Schema.org classes such as *Person*, *Event* or *JobPosting*. The selected table columns include three data types: textual values, numerical values and DateTime values. All CPA tables include a main (subject) column in the first column position used in pairs with other columns for the CPA task. 91 CTA labels are used to annotate 162,351 columns in the CTA tables such as *Organization*, *Date* and *Offer*, and 176 CPA labels are used to annotate 174,998 main column/column pairs in the CPA tables like *price*, *datePublished* and *productID*. Detailed statistics about the number of tables per Schema.org class as well as number of columns per label are provided on the SOTAB project page[2]. We split the CTA and CPA tables with a ratio of 80:10:10 into fixed training, validation and test sets by using multi-label stratification to include examples of all labels in every set. We further split the training set into a smaller subset using the same method and provide the Small training set with the goal of comparing how methods perform when trained on less examples. Furthermore, we provide subsets of columns in the test set for specific annotation challenges. These subsets are created by grouping the columns in the test set by the challenges mentioned in Section 3:

[2] http://webdatacommons.org/structureddata/sotab/

test columns that have missing values (*Test MV*), test columns that are corner cases (*Test CC*), test columns that include values in different formats (*Test FH)* and test columns that are chosen randomly (*Test RC*). Statistics of all splits are given in Table 2.

5. Benchmark Evaluation

We evaluate the difficulty of SOTAB using three supervised methods. The first is a simple Random Forest classifier using TF-IDF-weighted words as features. The second is TURL [10], which is pre-trained on a large corpus of Wikipedia tables using a combination of the Masked Language Model objective of BERT and a Masked Entity Recovery objective introduced by the authors. The last is Doduo [15] which uses multi-task learning for the simultaneous fine-tuning of BERT for the CTA and CPA tasks. We fine-tune TURL for 50 epochs and Doduo for 30 epochs using a learning rate of 5e-5. We report the micro-F1 score on the test sets for CTA and CPA in Table 3. In both CTA and CPA tasks, there is a significant difference in the F1 scores among the methods, with Transformer methods achieving at least 17 percentage points more than Random Forest. This can be an indicator that incorporating table context helps the model make better decisions for both tasks. The corner cases columns (CC) and columns with missing values (MV) appear to be the subsets where both methods in both tasks make more prediction errors. Finally, training with the smaller set leads to 3-8 percentage points lower scores on both tasks. The F1 scores show that SOTAB is challenging for all methods.

Table 3
SOTAB results for Random Forest (RF), TURL and Doduo (DO) using the large and small training sets

	CTA						CPA					
	Large			Small			Large			Small		
	RF	Turl	DO	RF	Turl	DO	RF	Turl	DO	RF	Turl	DO
Full	58.58	78.96	84.82	55.57	72.16	76.27	44.80	72.93	79.96	41.44	66.30	75.38
MV	60.47	73.14	83.28	58.04	66.98	69.55	44.59	66.24	78.27	42.06	59.97	74.34
CC	55.15	73.59	78.03	51.97	68.19	74.00	38.00	62.54	71.24	35.39	57.87	66.73
FH	64.78	90.14	92.98	62.35	87.88	85.95	45.69	77.15	83.50	43.69	69.86	77.38
RC	58.72	81.93	87.12	55.53	74.11	81.82	46.46	76.96	82.25	42.51	69.81	77.64

6. Conclusion and Availability

This paper introduced the WDC SOTAB benchmark. The aim of SOTAB is to complement the set of publicly available table annotation benchmarks with a CTA and CPA benchmark covering various entity types of general interest, e.g. products, local business, job postings, and to provide training data from many independent data sources for these types in order to reflect the full heterogeneity of the values that are used to describe entities. The WDC SOTAB benchmark is available for public download on the project page. The code that was used for the creation of the benchmark is provided on github[3].

[3] https://github.com/wbsg-uni-mannheim/wdc-sotab

References

[1] D. Ritze, O. Lehmberg, Y. Oulabi, C. Bizer, Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 251–261.

[2] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, et al., Dataset search: A survey, The VLDB Journal 29 (2020) 251–272.

[3] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: Evolution of structured data on the web, Communications of the ACM 59 (2016) 44–51.

[4] V. Cutrona, J. Chen, V. Efthymiou, O. Hassanzadeh, E. Jimenez-Ruiz, et al., Results of SemTab 2021, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, volume 3103, CEUR-WS, 2022, pp. 1–12.

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

[6] 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, Springer International Publishing, 2020, pp. 514–530.

[7] M. Hulsebos, Ç. Demiralp, P. Groth, GitTables: A Large-Scale Corpus of Relational Tables, arXiv:2106.07258 (2022).

[8] 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, volume 3103, CEUR-WS, 2021, pp. 13–18.

[9] D. Ritze, C. Bizer, Matching Web Tables To DBpedia - A Feature Utility Study, in: Proceedings of the 20th International Conference on Extending Database Technology, 2017, pp. 210–221.

[10] X. Deng, H. Sun, A. Lees, Y. Wu, C. Yu, TURL: Table understanding through representation learning, Proceedings of the VLDB Endowment 14 (2020) 307–319.

[11] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2008, pp. 1247–1250.

[12] S. Singh, A. F. Aji, G. Singh, C. Christodoulopoulos, A relation extraction dataset for knowledge extraction from web tables, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 2319–2327.

[13] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, 2017, pp. 427–431.

[14] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, et al., Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).

[15] Y. Suhara, J. Li, Y. Li, D. Zhang, Ç. Demiralp, et al., Annotating columns with pre-trained language models, in: Proceedings of the 2022 International Conference on Management of Data, 2022, pp. 1493–1503.