

Table Interpretation

SIGIR 2019 tutorial - Part II

Shuo Zhang and **Krisztian Balog**

University of Stavanger

Outline for this Part

Definition

Table interpretation encompasses methods that aim to make tabular data processable by machines.

Three specific subtasks:






- 1 Column type identification (a.k.a. column-to-concept matching)
- 2 Entity linking in tables
- 3 Relation extraction

Column Type Identification

Definition

Column type identification is concerned with determining the types of columns, including locating the core column.

All-time (List of Grand Slam men's singles champions)

Rank ↕	Player ↕	Total ↕	Years ↕
1	 Roger Federer	20	2003–2018
2	 Rafael Nadal	17	2005–2018
3	 Pete Sampras	14	1990–2002
	 Novak Djokovic	14	2008–2018
5	 Roy Emerson	12	1961–1967



Person 

<http://dbpedia.org/ontology/Person>

Single-concept vs. Multi-concept Relational Tables

- Most existing work assumes the presence of a single core column (a.k.a. *single-concept relational tables*)
- In some cases, a relational table might have multiple core columns that may be located at any position in the table, called *multi-concept relational table* (Braunschweig et al., 2015)
- **We focus on single-concept relational tables in this tutorial**

Comparison of Column Type Identification Studies

Reference	Knowledge base	Method
Venetis et al. (2011)	Automatically built IS-A DB	Majority vote
Mulwad et al. (2010)	Wikitology	Entity search
Fan et al. (2014)	Freebase	Concept-based + crowdsourcing
Wang et al. (2012)	Probase	Heading-based search
Lehmberg and Bizer (2016)	DBpedia	Feature-based classification
Zhang (2017)	Wikipedia	Unsupervised featured-based
Zhang and Chakrabarti (2013)	-	Semantic graph method

Approaches for Column Type Identification

- **Majority vote** (Venetis et al., 2011)
- Search-based
 - **Entity search** (Mulwad et al., 2010)
 - Heading search
- Feature-based
 - Unsupervised
 - Supervised
- **Crowdsourcing** (Fan et al., 2014)

- They argue that the meaning of web tables is “only described in the text surrounding them. Header rows exist in few cases, and even when they do, the attribute names are typically useless.”
- Key underlying idea: use facts extracted from text on the Web to interpret tables
- An IS-A database is built, consisting of (instance, class) pairs, by examining specific linguistic patterns on the Web
- A column A is labelled with class C if a substantial fraction of the cells in a column A are labeled with class C in the IS-A database
- Using a knowledge base (YAGO) is found to result in higher precision, while annotating against the IS-A database has better coverage (i.e., higher recall)

- Key idea: obtain possible class labels by utilizing entities in a knowledge base (here: Wikitology ([Syed, 2010](#)))
- Each cell's value in a column is mapped to a ranked list of classes, and then a single class which best describes the whole column is selected
 - Retrieve top- k entities from the KB using a complex query, and consider their classes
 - Then, a PageRank-based method is used to compute a score for the entities' classes, from which the one with the highest score is regarded as the class label

- Issue: Because of the inherent semantic heterogeneity in web tables, not all tables can be matched to a knowledge base using pure machine learning methods
- Idea: use machine learning for “easy” cases and defer to crowdsourcing for “difficult” ones
- A *column difficulty estimator* component determines the columns that will be most beneficial for crowdsourcing, based on
 - Difficulty to determine the concept for the column
 - The degree of influence of the column, if verified by the crowd, on inferring the concepts of other columns

- Each microtask contains a table column and its candidate concepts

T_1 *Top Rated Movies*

Title	Directed By	Language
Les Misérables	T. Hooper	EN
Life of PI	A. Lee	EN
Inception	C. Nolan	EN

↓

Which *Concepts* the column is most likely refer to?

☒ *Film/Title*
☐ *Book/Title*
☐ *None of the Above*

Figure: Crowdsourcing microtask interface in (Fan et al., 2014)

Take-away Points for Column Type Identification






- Most relational tables are single-concept
- Methods typically rely on public knowledge bases
- Low coverage of knowledge bases is an open issue

Entity Linking in Tables

Definition

Recognizing and disambiguating specific entities (such as persons, organizations, locations, etc.), a task commonly referred to as *entity linking*, is a key step to uncovering semantics.

All-time (List of Grand Slam men's singles champions)

Rank ↕	Player ↕	Total ↕	Years ↕
1	 Roger Federer	20	2003–2018
2	 Rafael Nadal	17	2005–2018
3	 Pete Sampras	14	1990–2002
	 Novak Djokovic	14	2008–2018
5	 Roy Emerson	12	1961–1967

http://dbpedia.org/page/Rafael_Nadal

B

Overview

Reference	Knowledge base	Method
Limaye et al. (2010)	YAGO catalog, DBpedia, and Wikipedia tables	Inference of five types of features ^a
Bhagavatula et al. (2015)	YAGO	Graphical model
Wu et al. (2016)	Chinese Wikipedia, Baidu Baike, and Hudong Baike	Probabilistic method ^b
Efthymiou et al. (2017)	DBpedia	Vectorial representation and ontology matching
Zhang (2017)	Wikipedia	Optimization
Mulwad et al. (2010)	Wikilogy	SVM classifier
Lehmberg et al. (2016)	Google Knowledge Graph	-
Ibrahim et al. (2016)	YAGO	Probabilistic graphical model
Zhang et al. (2013)	DBpedia	Instance-based schema mapping
Hassanzadeh et al. (2015)	DBpedia, Schema.org, YAGO, Wikidata, Freebase	Ontology overlap ^c
Ritze and Bizer (2017)	DBpedia	Feature-based method
Ritze et al. (2015, 2016)	DBpedia	Feature-based method
Lehmberg and Bizer (2017)	DBpedia	Feature-based method

^a Designed for table search

^b Multiple KBs

^c KB comparison

Approaches for Entity Linking in Tables

- **Probabilistic graphical models** (Bhagavatula et al., 2015)
- **Feature-based methods** (Ritze and Bizer, 2017)
- Optimization
- Look-up based and ontology matching

TabEL (Bhagavatula et al., 2015)

- Traditional entity linking pipeline
 - Mention identification
 - Candidate generation
 - Disambiguation
- Disambiguation technique tailored to tables
 - Collective classification technique, optimizing all entity decisions jointly (iterative inference over the graphical model)
 - Soft constraints encourage disambiguations of mentions in the same row and column to be related to one another

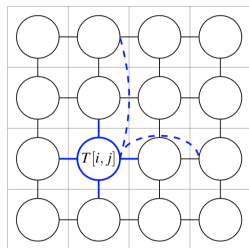


Figure: Graphical model used for disambiguation. Circles represent variables and edges represent their dependencies. For brevity, non-adjacent dependencies are only shown for the cell $T[i, j]$.

TabEL (Bhagavatula et al., 2015)

- Experiments both on Web and Wikipedia tables (based on (Limaye et al., 2010))
- Web tables dataset
 - 9,000 test mentions from 428 tables from the Web
 - Re-labeled erroneous gold annotations
 - Reported accuracy is 92.9% (vs. commonness baseline of 88.6%)
- Wikipedia tables dataset (WIKI_LINKS-RANDOM)
 - 50,000 test mentions from around 3,000 tables randomly drawn from Wikipedia
 - Existing links are removed and treated as gold annotations
 - Reported accuracy is 96.1% (vs. commonness baseline of 87.8%)
 - (Another variant, TABEL_35K, considers unlinked mentions, while retaining existing ones)
- Resources: <http://websail-fe.cs.northwestern.edu/TabEL/>

Web table features for EL (Ritze and Bizer, 2017)

- Features found in the table (T) or outside the table (C)
- Single table features (TS) refer to a value in a single cell while multiple features combine values coming from more than one cell (TM)

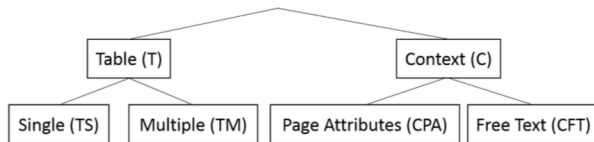


Figure: Categorization of web table features in (Ritze and Bizer, 2017)

Web table features for EL (Ritze and Bizer, 2017)

Feature	Description	Cat.
Entity label	The label of an entity	TS
Attribute label	The header of an attribute	TS
Value	The value that can be found in a cell	TS
Entity	The entity in one row represented as a bag-of-words	TM
Set of attr. labels	The set of all attribute labels in the table	TM
Table	The text of the table content without considering any structure	TM
URL	The URL of the web page from which the table has been extracted	CPA
Page title	The title of the web page	CPA
Surrounding words	The 200 words before and after the table	CFT

Table: Web table features in (Ritze and Bizer, 2017)

KB Features for EL (Ritze and Bizer, 2017)

Feature	Description
Instance label	The name of the instance mentioned in the <code>rdfs:label</code>
Property label	The name of the property mentioned in the <code>rdfs:label</code>
Class label	The name of the class mentioned in the <code>rdfs:label</code>
Value	The literal or object that can be found in the object position of triples
Instance count	The number of times an instance is linked in the Wikipedia corpus
Instance abstract	The DBpedia abstract describing an instance
Instance classes	The DBpedia classes (including the superclasses) to which an instance belongs to
Set of class instances	The set of instances belonging to a class
Set of class abstracts	The set of all abstracts of instances belonging to a class

Table: DBpedia features in (Ritze and Bizer, 2017)

Entity Linking for Web Tables ([Ritze and Bizer, 2017](#))

- Taking as many features into account as possible is beneficial
- Table features are generally considered more important than context features (which may add a lot of noise)
- These methods tend to perform better for large tables. To overcome this, [Lehmberg and Bizer \(2017\)](#) stitch tables, i.e., merge tables from the same website as a single large table, in order to improve the matching quality.

Resources for Table Matching by Ritze et al.

- WDC Web Tables Corpus
(<http://webdatacommons.org/webtables/>)
- T2D gold standard (Ritze et al., 2015)
 - Schema-level gold standard: 1,748 tables of which 762 can be matched with DBpedia classes and 7,983 columns which correspond to DBpedia properties
 - Entity-level gold standard: 233 tables, including 26 124 row-to-entity correspondences to DBpedia resources
- T2D v2 gold standard (Ritze and Bizer, 2017)
 - Includes tables that cannot be matched to the KB (deciding whether a table can be matched or not is part of the problem)

Take-away Points for Entity Linking in Tables






- Entity linking for tables needs special entity linker instead of using a textual entity linker
 - The notion of entity coherence needs to be captured differently
 - To understand the semantics of a table, it is often necessary to partly understand the content of the web page that embeds the table
- Many of the features are tailored specifically to relational tables; it is unclear whether they would work well for other types of tables

Relation Extraction

Definition

Relation extraction refers to the task of associating a pair of columns in a table with the relation that holds between their contents and/or extracting relationship information from tabular data and representing them in a structured format (e.g., as subject-predicate-object triples).

All-time (List of Grand Slam men's singles champions)

Rank ↕	Player ↕	Total ↕	Years ↕
1	 Roger Federer	20	2003–2018
←2	 Rafael Nadal	17	2005–2018
3	 Pete Sampras	14	1990–2002
	 Novak Djokovic	14	2008–2018
5	 Roy Emerson	12	1961–1967

→ <Peter_Sampras, careerYears, 1990-2002>



Overview of Relation Extraction Studies

Reference	KB	Method	Source of extraction
Venetis et al. (2011)	IS-A DB	Frequency-based	Core + attribute columns
Mulwad et al. (2010)	DBpedia	Utilizing CTI and EL	Any pair of columns
Mulwad et al. (2013)	DBpedia	Semantic passing	Any pair of columns
Zhang (2017)	Wikipedia	Optimization	Any pair of columns
Sekhavat et al. (2014)	YAGO, PATTY		Any pair of entities in the same row
Muñoz et al. (2014)	DBpedia	Look-up based	Any pair of entities in the same row
Zwicklbauer et al. (2013)	DBpedia	Frequency-based	
Chen and Cafarella (2013)	-	Classification	All columns

- They leverage a relations database of (*argument1*, *predicate*, *argument2*) triples
 - Triples are extracted from the Web, using an open information extraction system, TextRunner (Banko and Etzioni, 2008)
- For binary relationships, the relationship between columns A and B is labeled with R if a substantial number of pairs of values from A and B occur in the relations database
- Annotation quality is evaluated on a table search task (extrinsic evaluation)
 - Queries seek a property of a set of instances or entities (e.g., “wheat production of African countries”)
 - High precision but low recall (only a small portion of a whole table corpus was possible to annotate)
 - They discover that the vast majority of these tables are either not useful for answering entity-attribute queries, or can be labeled using a handful of domain-specific methods

Summary of this Part

- Table interpretation is the first step for many table-related tasks (knowledge base population, QA, etc.)
- Existing methods are based on a (strong) assumption that the column types and relations expressed in a table can be mapped to pre-defined types and relations in a reference KB. In practice, KBs suffer from limited coverage
- It remains unclear what relations are actually “useful”
- Another open question is whether all three subtasks (column type identification, entity linking, and relation extraction) can be performed jointly in a sound way

Bibliography I

- Michele Banko and Oren Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, June 2008.
- Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Tabel: Entity linking in web tables. In *Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366*, pages 425–441, New York, NY, USA, 2015. Springer-Verlag New York, Inc.
- Katrin Braunschweig, Maik Thiele, and Wolfgang Lehner. From web tables to concepts: A semantic normalization approach. In *Conceptual Modeling*, IC3K '16, pages 247–260, 2015.
- Zhe Chen and Michael Cafarella. Automatic web spreadsheet data extraction. In *Proceedings of the 3rd International Workshop on Semantic Search Over the Web, SS@ '13*, pages 1–8, New York, NY, USA, 2013. ACM.
- Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In *Proceedings of the 16th International Semantic Web Conference, ISWC '17*, pages 260–277. Springer, 2017.
- Ju Fan, Meiyu Lu, Beng Chin Ooi, Wang-Chiew Tan, and Meihui Zhang. A hybrid machine-crowdsourcing system for matching web tables. In *Proceedings of the IEEE 30th International Conference on Data Engineering, ICDE '14*, pages 976–987, 2014.

Bibliography II

- Okkie Hassanzadeh, Michael J. Ward, Mariano Rodriguez-Muro, and Kavitha Srinivas. Understanding a large corpus of web tables through matching with knowledge bases: an empirical study. volume 1545 of *CEUR Workshop Proceedings*, pages 25–34. CEUR-WS.org, 2015.
- Yusra Ibrahim, Mirek Riedewald, and Gerhard Weikum. Making sense of entities and quantities in web tables. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1703–1712, New York, NY, USA, 2016. ACM.
- Oliver Lehmberg and Christian Bizer. Web table column categorisation and profiling. In *Proc. of WebDB '16*, pages 4:1–4:7, 2016.
- Oliver Lehmberg and Christian Bizer. Stitching web tables for improving matching quality. *Proc. VLDB Endow.*, 10(11):1502–1513, August 2017.
- Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*, pages 75–76, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
- Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. *Proc. VLDB Endow.*, 3(1-2):1338–1347, September 2010.

Bibliography III

- Emir Muñoz, Aidan Hogan, and Alessandra Mileo. Using linked data to mine rdf from wikipedia's tables. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 533–542, 2014.
- Varish Mulwad, Tim Finin, Zareen Syed, and Anupam Joshi. Using linked data to interpret tables. In *Proceedings of the First International Conference on Consuming Linked Data - Volume 665*, COLD'10, pages 109–120, Aachen, Germany, Germany, 2010. CEUR-WS.org.
- Varish Mulwad, Tim Finin, and Anupam Joshi. Semantic message passing for generating linked data from tables. In *Proceedings of the 12th International Semantic Web Conference - Part I*, ISWC '13, pages 363–378, New York, NY, USA, 2013. Springer-Verlag New York, Inc.
- Dominique Ritze and Christian 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.*, pages 210–221, 2017.
- Dominique Ritze, Oliver Lehmberg, and Christian Bizer. Matching html tables to dbpedia. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, WIMS '15, pages 10:1–10:6, New York, NY, USA, 2015. ACM.
- Dominique Ritze, Oliver Lehmberg, Yaser Oulabi, and Christian Bizer. Profiling the potential of web tables for augmenting cross-domain knowledge bases. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 251–261, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

Bibliography IV

- Yoones A. Sekhavat, Francesco Di Paolo, Denilson Barbosa, and Paolo Merialdo. Knowledge base augmentation using tabular data. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, 2014.
- Zareen Saba Syed. *Wikilogy: A Novel Hybrid Knowledge Base Derived from Wikipedia*. PhD thesis, Catonsville, MD, USA, 2010. AAI3422868.
- Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. Recovering semantics of tables on the web. *Proc. VLDB Endow.*, 4 (9):528–538, June 2011. ISSN 2150-8097.
- Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Q. Zhu. Understanding tables on the web. In *Proceedings of the 31st International Conference on Conceptual Modeling, ER'12*, pages 141–155, Berlin, Heidelberg, 2012. Springer-Verlag.
- Tianxing Wu, Shengjia Yan, Zhixin Piao, Liang Xu, Ruiming Wang, and Guilin Qi. Entity linking in web tables with multiple linked knowledge bases. In *Semantic Technology*, pages 239–253. Springer International Publishing, 2016.
- Meihui Zhang and Kaushik Chakrabarti. Infogather+: Semantic matching and annotation of numeric and time-varying attributes in web tables. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13*, pages 145–156, New York, NY, USA, 2013. ACM.

- X. Zhang, Y. Chen, X. Du, and L. Zou. Mapping entity-attribute web tables to web-scale knowledge bases. *Database Systems for Advanced Applications*, pages 108–122, 2013.
- Ziqi Zhang. Effective and efficient semantic table interpretation using tableminer+. *Semantic Web*, 8:921–957, 2017.
- Stefan Zwicklbauer, Christoph Einsiedler, Michael Granitzer, and Christin Seifert. Towards disambiguating web tables. In *Proceedings of the 12th International Semantic Web Conference (Posters & Demonstrations Track) - Volume 1035*, ISWC-PD '13, pages 205–208, Aachen, Germany, Germany, 2013. CEUR-WS.org.