TAKCO: A Platform for Extracting Novel Facts from Tables

Benno Kruit
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
kruit@cwi.nl

Peter Boncz
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
boncz@cwi.nl

Jacopo Urbani
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
jacopo@cs.vu.nl

ABSTRACT

Web tables contain a large amount of useful knowledge. Takco is a new large-scale platform designed for extracting facts from tables that can be added to Knowledge Graphs (KGs) like Wikidata. Focusing on achieving high precision, current techniques are biased towards extracting *redundant* facts, i.e., facts already in the KG. Takco aims to find more *novel* facts, still at high precision. Our demonstration has two goals. The first one is to illustrate the main features of Takco's novel interpretation algorithm. The second goal is to show to what extent other state-of-the-art systems are biased towards the extraction of redundant facts using our platform, thus raising awareness on this important problem.

ACM Reference Format:
Benno Kruit, Peter Boncz, and Jacopo Urbani. 2021. TAKCO: A Platform for Extracting Novel Facts from Tables. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3442442.3458611

1 INTRODUCTION

Knowledge Graphs (KGs) are the main vehicle to publish large volumes of factual knowledge on the web. The largest KGs (e.g., Wikidata) contain billions of statements which are widely used to enhance tasks like question answering, data integration, semantic search, or named entity recognition.

The usefulness of KGs depends on how much accurate knowledge they contain. Therefore, it is important to discover missing facts at high precision. While some KGs are created manually (curated), large-scale KG construction should be automated, e.g., by mining data from the web. To this end, we would like to exploit the large amount of knowledge that is available as (HTML) tables on web pages, as spreadsheets, or as publicly available datasets.

Performing KG extraction from web tables can be decomposed into two main operations: *table interpretation* and *slot filling* [3]. The first task consists of linking the structure and content of the table with concepts and relations in the KG. Table interpretation has been the subject of several prior works, e.g. [5, 8], most of which focus primarily on the interpretation of *entity tables*, which are tables where each row describes one entity and columns represent attributes. A newer class of systems focus instead on *n-ary* tables, i.e., tables that express relations are more complex than binary ones [4]. N-ary tables are more difficult to interpret; indeed research on this problem is at its infancy with many challenges still viewed as open research questions.

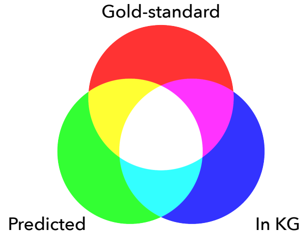


Figure 1: Difference between novel (yellow) and redundant extractions (white) returned by an extraction pipeline

After the table is correctly interpreted, we can extract facts that should be added to the KG. This operation is also known as *slot filling*, as the empty 'slots' in the KG are filled with new facts [6]. Table interpretation strongly affects the quality of slot filling since errors in the former can no longer be corrected. Because of this, state-of-the-art techniques aim for high precision by pruning out many potential assignments.

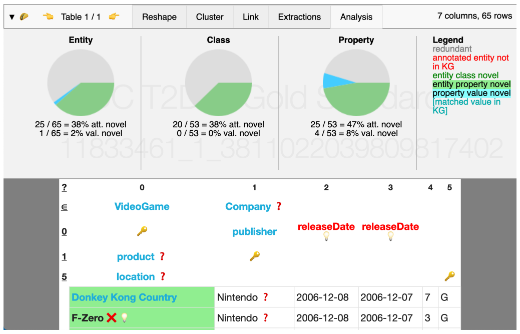
While maximizing precision is desirable, it has been observed that this strategy leads to a high number of redundant extractions [2, 7]. Fig. 1 provides a graphical definition of such extractions: Given an existing KG (blue), the facts returned by an extraction system (green), and a gold standard (red), the redundant facts are the ones which are already in the KG (white). The bias towards redundant extractions is exacerbated when more of the information in the table is novel, which is known as the *knowledge gap* problem. This introduces a trade-off: One the one hand we wish to return extractions with high precision to avoid errors. On the other hand, we would like to avoid returning facts which we already know.

With the goal of achieving high precision without sacrificing the number of novel extractions, we recently introduced a new system called Takco [3, 4]. Takco is a new end-to-end platform to extract knowledge both from entity and n-ary tables. The goal of our demonstration is twofold: First, we would like to show the capabilities of our engine using realistic Web datasets. Since Takco outperformed other methods on standard benchmarks, this demonstration also illustrates the performance of the state-of-the-art for knowledge extraction from tables. The second goal is to use our tool to visually show the amount and quality of *novel* extractions. In this way, we can provide a demonstration of the effect of the redundancy bias mentioned above, hoping to raise awareness on this important issue and to stimulate further research.

We identified three classes of users that should be interested on our demo: *Researchers on knowledge base completion*, *contributors to open KGs*, and *benchmark creators*. We provide below a short description of what we intend to demonstrate for each category of

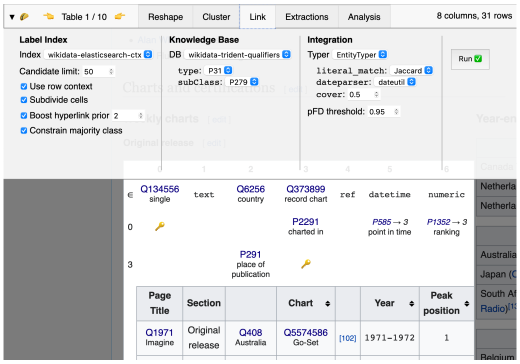
WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3458611



?	0	1	2	3	4	5
∈	VideoGame	Company ?				
0	🔑	publisher	releaseDate	releaseDate		
1	product ?	🔑				
5	location ?					🔑
	Donkey Kong Country	Nintendo ?	2006-12-08	2006-12-07	7	G
	F-Zero ✗	Nintendo ?	2006-12-08	2006-12-07	3	G

Figure 2: Measuring the novelty of extracted facts



Label Index
Index: wikidata-elasticsearch-ctx
Candidate limit: 50
☑ Use row context
☑ Subdivide cells
☑ Boost hyperlink prior 2
☑ Constrain majority class

Knowledge Base
DB: wikidata-trident-qualifiers
type: P31
subClass: P279

Integration
Typer: EntityTyper
literal_match: Jaccard
dateparser: dateutil
cover: 0.5
pFD threshold: 0.95

	0	1	2	3	4	5	6
∈	Q134556 single	text	Q6256 country	Q373899 record chart	ref	datetime	numeric
0	🔑			P2291 charted in		P585 → 3 point in time	P1352 → 3 ranking
3			P291 place of publication	🔑			
	Page Title	Section		Chart		Year	Peak position
	Q1971 Imagine	Original release	Q408 Australia	Q5574586 Go-Set	[102]	1971–1972	1

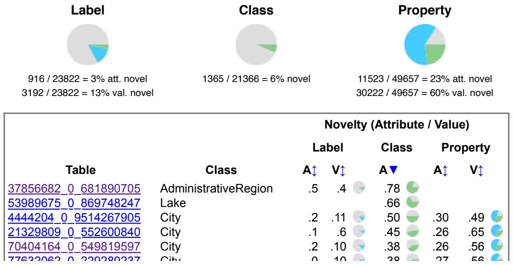
Figure 3: Fine tuning interpretation on Wikipedia tables

users. A short video that gives a high-level overview of what we plan to showcase is available online[1]. In addition, we refer to our public code repository[2] and documentation[3] for more details about the system.

2 TAKCO

TAKCO is a system that we recently introduced with the goal of extracting novel facts from web tables. Like other existing systems, TAKCO is designed to extract knowledge from tables that describe entities (entity tables), but is also capable of processing tables that describe n-ary relations. Details can be found in [3] and [4]. Here, we limit ourselves to provide a brief description of the main algorithms that highlight TAKCO's main features.

For entity tables, the system first identifies a pool of candidate entities from a KG. Then, it calculates a prior probability distribution by matching the attributes of the candidate entity in the KG to other cell values in the same row, and then re-weights these matches by the salience of the relations in the table. Finally, it perform entity linking by constructing a Probabilistic Graphical Model (PGM) and performing a collective disambiguation of all cells with Loopy Belief Propagation (LBP) [1]. This way, TAKCO leverages the assumption that entities that occur together in a column are alike, resulting in more coherent predictions. Since TAKCO relies on both label salience and entity coherence, it was able to outperform the competitors by disambiguating the entities with higher recall, while retaining state-of-the-art precision. Moreover, TAKCO does not rely on any assumptions w.r.t. the structure of the KG, as other systems do.


Label
916 / 23822 = 3% att. novel
3192 / 23822 = 13% val. novel
Class
1365 / 21366 = 6% novel
Property
11523 / 49657 = 23% att. novel
30222 / 49657 = 60% val. novel

		Novelty (Attribute / Value)				
		Label		Class	Property	
Table	Class	A↕	V↕	A▼	A↕	V↕
37856682_0_681890705	AdministrativeRegion	.5	.4	.78		
53989675_0_869748247	Lake			.66		
4444204_0_9514267905	City	.2	.11	.50	.30	.49
21329809_0_552600840	City	.1	.6	.45	.26	.65
70404164_0_549819597	City	.2	.10	.38	.26	.56
77632062_0_229289237	City	0	.10	.38	.27	.56

Figure 4: Analysing the novelty of benchmark datasets

The interpretation of n-ary tables proceeds differently. First, the system applies several heuristics to re-shape the table in a "normal" form. Then, it performs *schema matching* and *functional dependency discovery* to compute a first rudimentary interpretation of the table. Finally, it *clusters* similar tables using multiple schema blocking and matching components to improve the quality of the interpretation. This approach is particularly interesting as it is the first end-to-end method that is able to interpret n-ary tables.

The empirical evaluation reported in [3] and [4] have shown that TAKCO has competitive performance. For instance, it was able to improve the recall of entity linking by 17% over the state-of-the-art, while retaining 92% precision on the T2D-v2 benchmark [5]. TAKCO has also reported good performance in terms of novel extractions, i.e., extracted facts that are not yet in the KG. In particular, it managed improved recall by 25% compared to the state-of-the-art [3].

3 DEMONSTRATION SCENARIOS

Our demonstration is primarily intended for researchers and practitioners who are interested in knowledge extraction from tables, and more generally on the problem of automatic KG construction. We categorized them as system developers, KG contributors, and/or benchmark creators (note that the categories are often overlapping).

System Developers. Our platform supports rapid exploration of configurations, and this is useful for system developers to evaluate the impact of design decisions on extraction quality and novelty. We will demonstrate the performance of TAKCO against other engines on common datasets, and show how the performance varies with different components in the KG completion pipeline. Finally, we will use our platform to identify easy and hard tables with the analytical tool shown in Fig. 2, and show the impact of changes of their configuration in different contexts, e.g., tables of different sizes, varying novelty and a different choice of KG.

KG Contributors. For people interested in contributing novel facts to an existing KG, we would like to show how to tune an tabular KG completion pipeline for their own use-case. For instance, we would demonstrate *domain adaptation*, i.e., how well a pipeline tuned for one dataset would perform on another (See Fig. 3 for a screenshot of the interface). Because practitioners are interested in the generalizability on real data, the fact that TAKCO is KG-agnostic

[1] https://www.dropbox.com/s/iaftfhww1z03wz8/takco-demo-www2021.mp4?dl=0
[2] https://github.com/karmaresearch/takco
[3] https://takco.readthedocs.io/

and that it can load datasets from different domains would allow us to discuss real-world trade-offs.

Benchmark Creators. With the growing popularity of table interpretation, it is fundamental to design better benchmark tools that can stress the systems in key areas. To this end, we argue that benchmark creators need fine-grained control over the knowledge gap in order to tune for desirable properties of competing systems, such as extraction novelty. To support our position, we will use our tool, especially the interface shown in Fig. 4, to illustrate the characteristics of existing benchmarks, and offer some insights into what makes a novel facts from tables easy or difficult to extract.

REFERENCES

[1] Daphne Koller, Nir Friedman, and Francis Bach. 2009. *Probabilistic graphical models: principles and techniques.* MIT press.

[2] Benno Kruit, Peter Boncz, and Jacopo Urbani. 2018. Extracting New Knowledge from Web Tables: Novelty or Confidence? *KBCOM* (2018).

[3] Benno Kruit, Peter Boncz, and Jacopo Urbani. 2019. Extracting Novel Facts from Tables for Knowledge Graph Completion. In *ISWC*. 364–381.

[4] Benno Kruit, Peter A. Boncz, and Jacopo Urbani. 2020. Extracting N-ary Facts from Wikipedia Table Clusters. In *CIKM 2020*. 655–664.

[5] Dominique Ritze, Oliver Lehmberg, and Christian Bizer. 2015. Matching HTML Tables to DBpedia. In *WIMS*.

[6] Dominique Ritze, Oliver Lehmberg, Yaser Oulabi, and Christian Bizer. 2016. Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases. In *WWW*. 251–261.

[7] Shuo Zhang, Edgar Meij, Krisztian Balog, and Ridho Reinanda. 2020. Novel Entity Discovery from Web Tables. In *WWW*. 1298–1308.

[8] Ziqi Zhang. 2017. Effective and efficient semantic table interpretation using tableminer+. *Semantic Web* 8, 6 (2017), 921–957.