Wikary

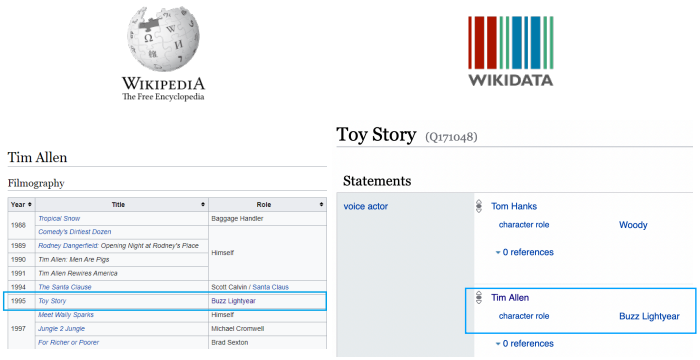
A Dataset of N-ary Wikipedia Tables Matched to Qualified Wikidata Statements

Igor Mazurek, Berend Wiewel and **Benno Kruit**

My name is Benno Kruit and this talk is about Wikary, A Dataset of N-ary Wikipedia Tables Matched to Qualified Wikidata Statements which was created by Igor Mazurek, Berend Wiewel and me at the Vrije Universiteit in Amsterdam.

Wikary

N-ary **Wikipedia** tables matched to **Wikidata**



Many tables on the web express statements about more than two values. For example, this table on Wikipedia expresses the statement that the actor Tim Allen was the voice of Buzz Lightyear in the 1995 movie Toy Story. In the Wikidata Knowledge Base, the binary statement that that Allen was a voice actor in Toy Story is extended with his specific role to make it ternary with three values.

N-ary statements

"The album *Thriller* by Michael Jackson reached the top-1 position in the US Billboard 200 chart on February 26th, 1983."

N-ary statements like this come in many forms. Consider this statement here: "The album Thriller by Michael Jackson reached the top-1 position in the US Billboard 200 chart on February 26th, 1983."

N-ary statements

with Wikidata qualifiers

"The album *Thriller* by Michael Jackson reached the top-1 position in the US Billboard 200 chart on February 26th, 1983."

On the left, you can see what looks like this in Wikidata. The Thriller album is connected to the US Billboard 200 by the "charted in" property. Two qualifiers are added to this statement: "point in time" and "ranking". To facilitate SPARQL querying, this is represented in RDF as a blank node using reification, as seen on the right.

N-ary statements

reified as RDF blank nodes

```
# Thriller (album)
wd:Q44320 p:P2291 [

  # charted in: US Billboard 200
  ps:P2291 wd:Q188819;

  # point in time: 26th February 1983
  pq:P585 "1983-02-26"^^xsd:date;

  # ranking: 1
  pq:P1352 "1"
].
```

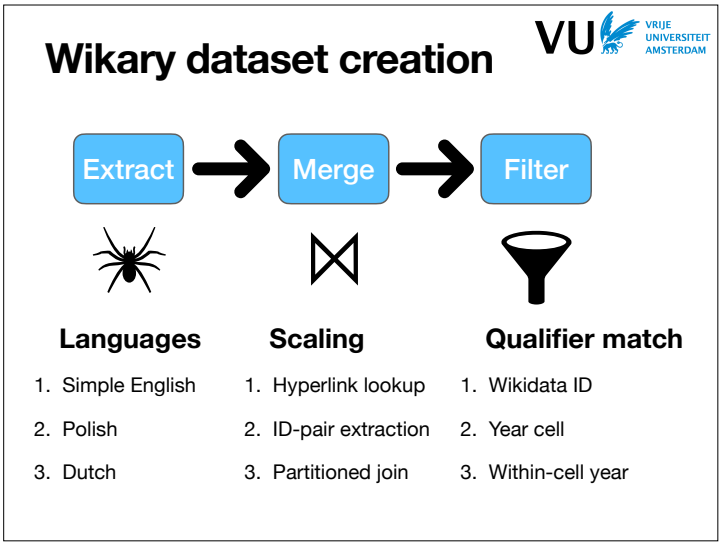
This is what it looks like using RDF Turtle syntax. You can see that values of different datatypes are connected in a coherent way.

N-ary statements

in Wikipedia tables

On the Wikipedia article of *Thriller*, this statement is also mentioned in a table about chart positions. The table has a pretty complicated structure, but it contains many statements of the same form that are not in Wikidata yet. Our goal is to automatically annotate these types of tables with the right properties and qualifiers, in order to extract those statements to add them to Wikidata. But there's currently no large-scale datasets of such tables for training and evaluating annotation systems, so that's why we created our

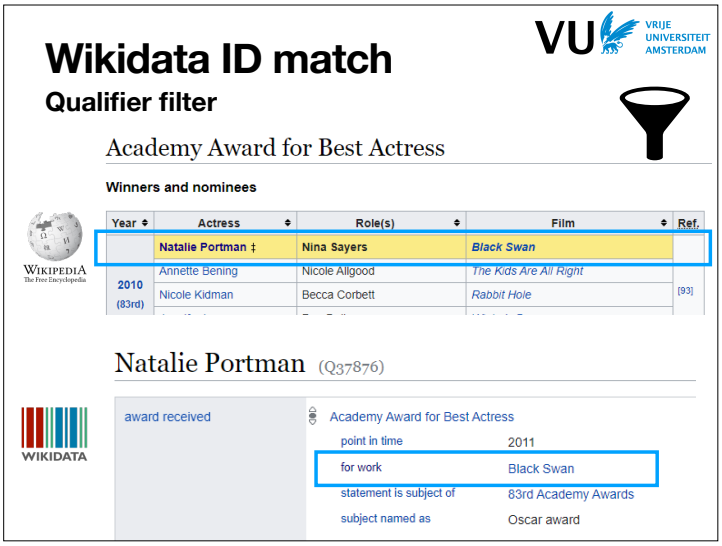
dataset called Wikary.



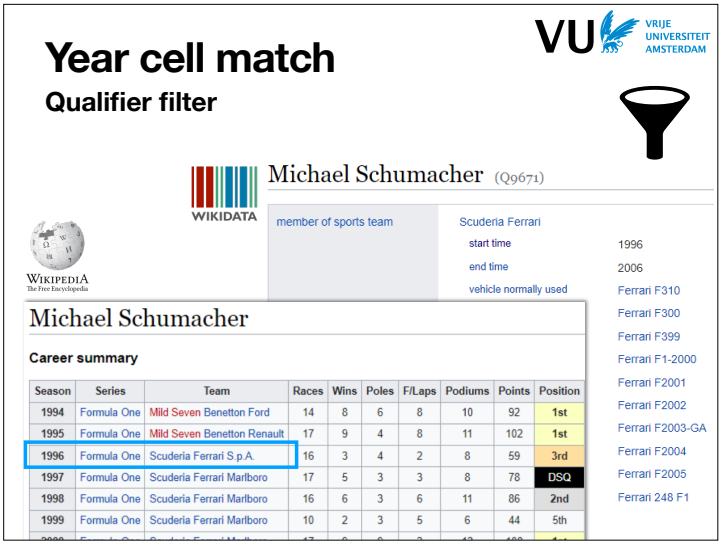
The method we used to create this dataset has three steps:

- First, we **extract** tables from a local HTML copy of Wikipedia to retain all their original formatting. We used three language versions: Simple English, Polish and Dutch, because this is what our annotators spoke. We opted for the Simple English version instead of the Full English because its tables contain less complicated formatting, and the dataset is smaller and thus easier to process. We aim to create a dataset from the Full Engish version soon.
- Second, we **merge** the tables with Wikidata. We look up the Wikidata IDs for all hyperlinks of the tables, and then extract all pairs of IDs from every row which appear in two different columns or the page ID. This is followed by a database-style join between those pairs and the subject-object pairs of Wikidata statements with qualifiers. We partitioned this over several compute nodes for scalability.
- Finally, we **filter** the tables based on three heuristics. We keep only

tables for which any of these heuristics match the joined row to a qualifier value of the joined Wikidata statement.
I will next explain the Wikidata ID, year cell, and within-cell year match heuristics.



The Wikidata ID match-heuristic retains tables in which a Wikidata ID in the joined row matches a qualifier value of the joined statement. In this example about the Academy Awards, you can see that the table was joined on the pair of entities “Natalie Portman” and “Academy Award for Best Actress”. The heuristic then keeps this table because “Black Swan” matches the qualifier “for work”.



The Year cell match-heuristic retains tables for which a four-digit cell in the joined row matches a qualifier value of the joined statement. In this example, the table was joined on “Michael Schumacher” and “Scuderia Ferrarri”, and the heuristic keeps this table based on the value “1996”, which is the value of the “start time” qualifier.

Within-cell year match

Qualifier filter

Grand Theft Auto

List of titles

Title	Developer	Platforms	First Released
Grand Theft Auto: Chinatown Wars	Rockstar North	Nintendo DS, PSP	March 17, 2009 (Nintendo DS) October 20, 2009 (PSP)
Grand Theft Auto: The Ballad of Gay Tony	Rockstar North	PlayStation 3, Xbox 360	October 29, 2009 (Xbox 360) April 13, 2010 (PlayStation 3)
Grand Theft Auto V	Rockstar North	PlayStation 3, Xbox 360	September 17, 2013

Grand Theft Auto V (Q17452)

platform	Xbox 360	
	publication date	17 September 2013
	1 reference	
	PlayStation 3	
	publication date	17 September 2013

The Within-cell year match-heuristic is similar, and retains tables for which a four-digit string anywhere in the joined row matches part of a qualifier value of the joined statement. In this example, we can see that “Grand Theft Auto V” and “Playstation 3” joined this table to Wikidata, and then it was kept based on the value “2013”. This allowed for higher recall, but we wanted to keep it optional for if the precision became too low.

Quality Evaluation

with our annotation interface

- Runs in Jupyter notebook
- Shows original table with highlighted match
- Used to annotate 750 matches total
- Filtering:
  - 98.4% precision
  - 23.3% recall

N-ary | Not N-ary | Incorrect formatting | Skip

prev | next

Page title	Division_of_Bendigo	tableIndex: 0 rowIndex: 6		Visit wikipedia page
Subject	Q152666	Billy Hughes	Billy Hughes	Visit wikidata page
Property	P102	member of political party		Visit wikidata page
Object	Q1516976	Nationalist	Nationalist Party of Australia	Visit wikidata page

Member	Party	Results
	Protectionist	1901–1906
Sir John Quick	Independent Protectionist	1906–1909
	Commonwealth Liberal	1909–1913
John Arthur	Labor	1913–1914
Alfred Hampson	Labor	1915–1917
Billy Hughes	Nationalist	1917–1922
Geoffry Hurry	Nationalist	1922–1929
Richard Keane	Labor	1929–1931

We measured precision and recall by annotating the tables with a simple user interface. This was implemented to run in a Jupyter notebook, and shows the table with original HTML formatting, with the maching row highlighted. It was used to annotate 750 tables in total. The filtering step had a precision of 98% and a recall of 23%. The heuristics were designed for precision, because we wanted to use this data for training and evaluating classifiers.

Wikary: data statistics

Extracted Wikipedia tables

	Full Dataset		At least one hyperlink matched to Wikidata			Merged with Wikidata statements with qualifiers		
Lang.	**Tables**	**Pages**	**Tables**	**Pages**	**T. % of dataset**	**Tables**	**Pages**	**T. % of dataset**
Simple EN	35.047	18.194	23.809	14.935	68%	10.023	8.155	**28.6%**
NL	**582.319**	243.893	360.778	153.628	62%	87.584	51.867	15%
PL	537.366	199.791	366.795	149.858	**68.3%**	114.694	62.196	21.3%

Here’s some statistics about the dataset. From left to right, you can see that the Dutch part contains the most tables, the Polish part has the highest percentage of tables that have a hyperlink that can be matched to Wikidata, and the the Simple English part has the highest percentage of tables that can be merged with Wikidata statements that have qualifiers.

Wikary: data statistics

Filtering of merged Wikipedia tables

	Full Dataset		Wikidata ID match			Year cell match			Within-cell year match		
Lang.	**Tables**	**Pages**	**Tables**	**Pages**	**T. % of data**	**Tables**	**Pages**	**T. % of data**	**Tables**	**Pages**	**T. % of data**
Simple EN	10.023	8.155	706	605	**7%**	58	53	0.6%	301	218	3%
NL	87.584	51.867	4.790	4.094	5.5%	1.055	904	1.2%	14.490	7.508	**16%**
PL	114.694	62.196	2.420	2.015	2.1%	1.438	1.329	1.3%	10.532	4.804	**9.2%**

In this table you can see the result of the filtering heuristics. The Wikidata ID matching results in the highest percentage of tables in the Simple English part, and the Within-cell year match results in the highest percentage of tables in the Dutch and Polish parts.

Preliminary use-case

Automatically distinguish binary vs. N-ary table

- Trained various linear classifiers
  - N-ary tables **vs.** binary tables from Wikipedia list pages
- Good cross-validation performance
  - 80-90% accuracy
  - Pre-processing for CPA
- **Caveat**: probable data bias
  - samples about different topics

Features
Column names
% of Entity-type columns
% of Numeric-type columns
% of String-type columns
% of Time/date-type columns
Min column-uniqueness
Max column-uniqueness
Mean column-uniqueness

We used this dataset to train various simple linear classifiers on the task of automatically distinguishing binary and n-ary Wikipedia tables. The training data consisted of the n-ary tables from Wikary, and tables collected from Wikipedia List pages which we assume to be binary. We used a features set of column names, column types, and statistics of the number of unique values per column. The best-performing models performed well, and could be useful as pre-processing for the Column-Property Annotation task. However, because the training data is from two separate sources, we suspect the cross-validation scores are not reliable, and the models will prove to be biased on real data, so that's something we still have to fix.

Conclusion

Wikary: N-ary Wikipedia tables matched to Wikidata

- Almost 32.000 tables
- From 3 Wikipedia editions
- Precision-focused qualifier match: 98.4%

Thank you!

Future Work

- More languages (full English)
- Extend to non-matching tables

zenodo.org/record/7025005

Igor Mazurek
Berend Wiewel
Benno Kruit (b.b.kruit@vu.nl)

To conclude, our Wikary dataset consists of almost 32 thousand n-ary tables matched to Qualified Wikidata statements, from 3 Wikipedia editions with high precision. Our dataset is publicly available on Zenodo.
In the future, we'd like to expand this dataset to more languages including the full English Wikipedia, and find n-ary tables that don't overlap with Wikidata, but contain all new statements that could be added to Wikidata to extend its coverage.

No. overall		Title	Original airdate
17	1	"A Slight Case of Reincarnation"	31 December 1966
Adam is determined to free an African leader from the spell of the Face. *			
18	2	"Black Echo"	7 January 1967
Adam visits an exiled Russian Grand Duchess, who is requesting him because he is familiar with a pearl necklace that she once owned, and is now trying to reclaim it. However, upon arrival at the Duchess's home, Adam finds two horrifying secrets from his past are about to return...			
19	3	"Conspiracy of Death"	14 January 1967
Adam investigates the murder of an old wartime friend.			

(a) Cells containing long text, wrapped to next row

Year	Population	Year	Population
1749	200	1885	453
1763	273	1907	344
1780	327	1925	330
1800	273	1945	351
1818	329	1955	295
1869	402		

(b) Repeated header

Catholic, Orthodox, Protestant, and most Oriental Orthodox	Original language (Koine Greek)
Canonical Gospels	
Matthew	Greek (majority view; see note)[1] [Σ]ΟΥΑΓΓΕΛΙΟΝ
Mark	Greek
Luke	Greek
John	Greek
Apostolic History	
Acts	Greek

(c) Sub-headers

Sled	Athletes	Event	Run 1		Run 2		Run 3	
			Time	Rank	Time	Rank	Time	Rank
LAT-1	Sandis Prūsis Mārcis Rullis	Two-man	48.10	13	48.06	12	47.92	8
LAT-2	Intars Dīcmanis Gatis Gūts	Two-man	48.07	12	48.22	18	48.14	13

(d) Hierarchical header structure

Pos.	Rider	Points	POL	ITA	GER	SWE	GBR	DEN
1	(5) Billy Hamill	113	16	20	9	25	18	25
2	(1) Hans Nielsen	111	18	25	25	9	20	14
3	(4) Greg Hancock	88	12	13	13	16	16	18
4	(2) Tony Rickardsson	86	20	18	16	14	7	11
5	(8) Henrik Gustafsson	80	14	4	18	20	11	13

(e) Hybrid relational-matrix table

Extra: Harder tables

Chart (1971)	Rank
Canada Top Singles (*RPM*)[134]	15
Netherlands (Dutch Top 40)[135]	67
Netherlands (Single Top 100)[136]	82