Bridging Between Tables and Human Languages

From Tables to Knowledge: Recent Advances in Table Understanding (Part IV)

Muhao Chen

Department of Computer Science / Information Sciences Institute

University of Southern California

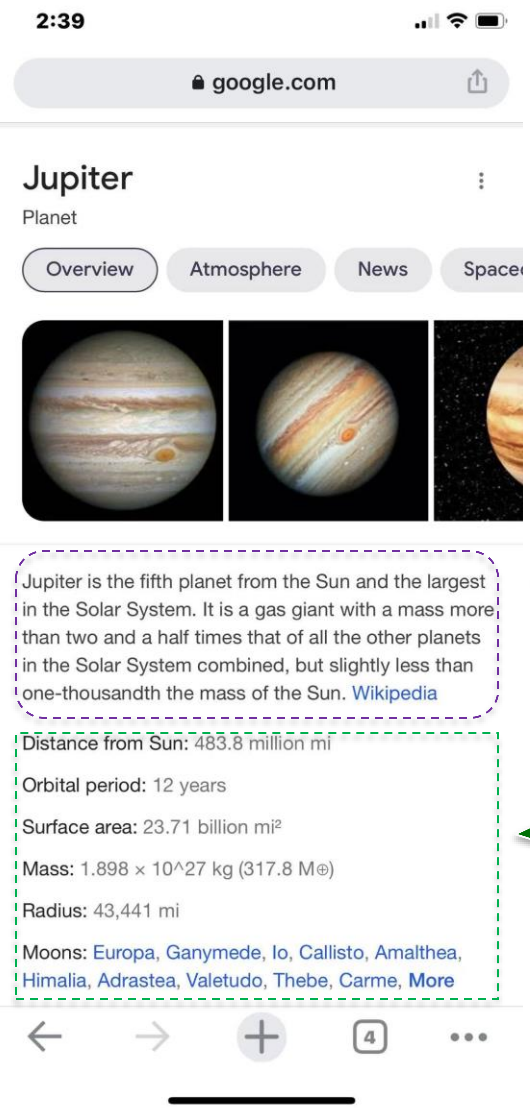
Aug 2021

KDD Tutorials

Recent Advances in Table Understanding

How Do *Table Understanding* Interface with *Natural Language Understanding*?

Table Understanding and NLU Are Related

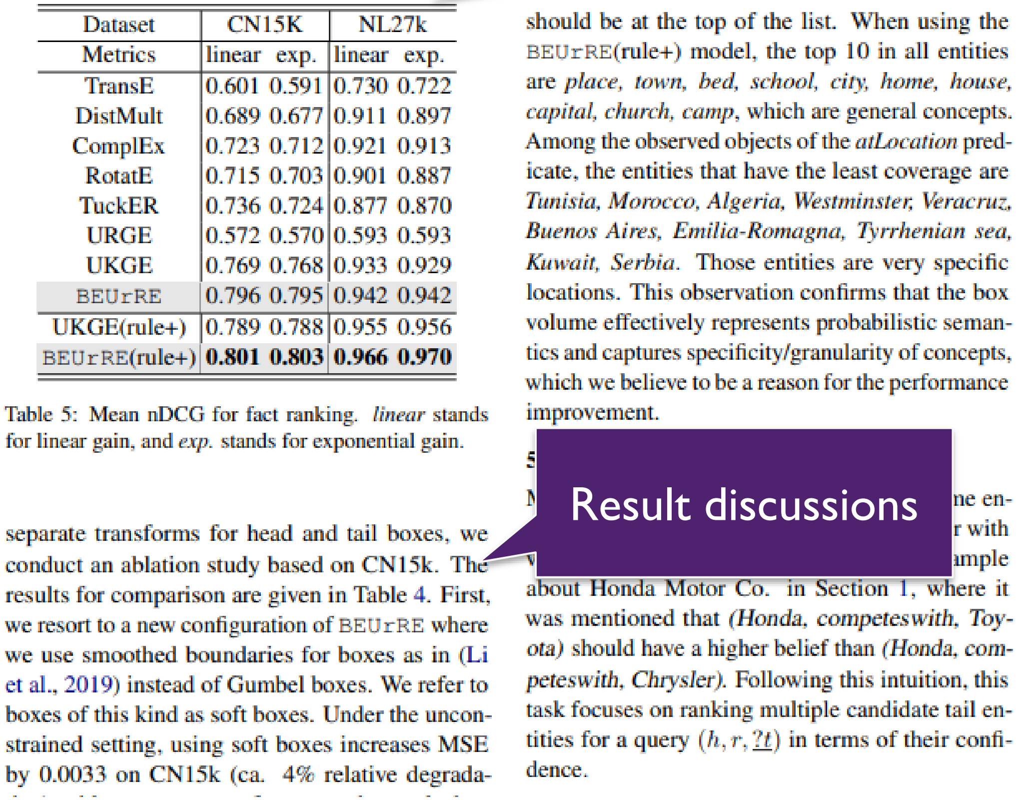


Searching for an entity at Google.

Text description

Attributes in a compact table

Experimental result table(s)



Dataset	CN15K		NL27k	
Metrics	linear	exp.	linear	exp.
TransE	0.601	0.591	0.730	0.722
DistMult	0.689	0.677	0.911	0.897
ComplEx	0.723	0.712	0.921	0.913
RotatE	0.715	0.703	0.901	0.887
TuckER	0.736	0.724	0.877	0.870
URGE	0.572	0.570	0.593	0.593
UKGE	0.769	0.768	0.933	0.929
BEUrRE	0.796	0.795	0.942	0.942
UKGE(rule+)	0.789	0.788	0.955	0.956
BEUrRE(rule+)	**0.801**	**0.803**	**0.966**	**0.970**

Table 5: Mean nDCG for fact ranking. *linear* stands for linear gain, and *exp.* stands for exponential gain.

separate transforms for head and tail boxes, we conduct an ablation study based on CN15k. The results for comparison are given in Table 4. First, we resort to a new configuration of BEUrRE where we use smoothed boundaries for boxes as in (Li et al., 2019) instead of Gumbel boxes. We refer to boxes of this kind as soft boxes. Under the unconstrained setting, using soft boxes increases MSE by 0.0033 on CN15k (ca. 4% relative degrada-

should be at the top of the list. When using the BEUrRE(rule+) model, the top 10 in all entities are *place, town, bed, school, city, home, house, capital, church, camp*, which are general concepts. Among the observed objects of the *atLocation* predicate, the entities that have the least coverage are *Tunisia, Morocco, Algeria, Westminster, Veracruz, Buenos Aires, Emilia-Romagna, Tyrrhenian sea, Kuwait, Serbia*. Those entities are very specific locations. This observation confirms that the box volume effectively represents probabilistic semantics and captures specificity/granularity of concepts, which we believe to be a reason for the performance improvement.

5 ...

N ... ne en- ... r with ... ample about Honda Motor Co. in Section 1, where it was mentioned that (*Honda, competeswith, Toyota*) should have a higher belief than (*Honda, competeswith, Chrysler*). Following this intuition, this task focuses on ranking multiple candidate tail entities for a query $(h, r, ?t)$ in terms of their confidence.

Result discussions

Reading about experiments in a scientific paper.

Tables and text: two views of information, complementary sources of knowledge

Natural Language Interfaces to Tabular Content

Connecting tables and NL lead to a flexible and user-friendly way of accessing tabular content.

Wikitables

The best-selling video game?

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

Semantic retrieval of tables

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

A wii game by Nintendo.

Retrieving cell content

CONSOLIDATED STATEMENTS OF OPERATIONS - USD ($) $ in Thousands	12 Months Ended		
	Jan. 31, 2020	Jan. 31, 2019	Jan. 31, 2018
Income Statement [Abstract]			
Revenue	$ 622,658	$ 330,517	$ 151,478
Cost of revenue	115,396	61,001	30,780
Gross profit	507,262	269,516	120,698
Operating expenses:			
Research and development	67,079	33,014	15,733
Sales and marketing	340,646	185,821	82,707
General and administrative	86,841	44,514	27,091
Total operating expenses	494,566	263,349	125,531
Income (loss) from operations	12,696	6,167	(4,833)
Interest income and other, net	13,666	2,182	1,315
Total	26,362	8,349	(3,518)
Provision for income taxes	1,057	765	304
Net income (loss)	25,305	7,584	(3,822)
Distributed earnings attributable to participating securities	0	0	(4,405)
Undistributed earnings attributable to participating securities	(3,555)	(7,584)	0
Net income (loss) attributable to common stockholders	$ 21,750	$ 0	$ (8,227)
Net income (loss) per share attributable to common stockholders:			
Basic (in dollars per share)	$ 0.09	$ 0.00	$ (0.11)
Diluted (in dollars per share)	$ 0.09	$ 0.00	$ (0.11)
Weighted-average shares used in computing net income (loss) per share attributable to common stockholders:			
Basic (in shares)	233,641,336	84,483,094	78,119,865
Diluted (in shares)	254,298,014	116,005,681	78,119,865

Table showing the growing revenue of Zoom.

Generating summarizations for tables

Tabular Knowledge Assists NLU

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

- The best-selling video game of all time is **Minecraft**. ✓

- The best-selling video game of all time is **Tetris**. ✗

Tables as evidence for natural language claim verification

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x_1: *"Greece held its last Summer Olympics in which year?"*
y_1: $\{2004\}$

x_2: *"In which city's the first time with at least 20 nations?"*
y_2: $\{\text{Paris}\}$

x_3: *"Which years have the most participating countries?"*
y_3: $\{2008, 2012\}$

x_4: *"How many events were in Athens, Greece?"*
y_4: $\{2\}$

x_5: *"How many more participants were there in 1900 than in the first year?"*
y_5: $\{10\}$

Tables as reference for answering questions

Common Challenges for Connecting Tables and Natural Language

Handling heterogeneous structures

Gameloft
Video game publisher

Gameloft SE is a French video game publisher based in Paris, founded in December 1999 by Ubisoft co-founder Michel Guillemot. The company operates 19 development studios worldwide, and publishes games with a special focus on the mobile games market.

Lake	Area
Windermere	5.69 sq mi
Ullswater	3.86 sq mi
Derwent Water	2.06 sq mi

(a) Relational table

Country	United States
State	California
County	Los Angeles
Region	South California

(b) Entity table

	Right-handed	Left-handed
Males	43	9
Females	44	4
Totals	87	12

(c) Matrix table

		To		
		Solid	Liquid	Gas
From	Solid	Solid trans	Melting	Sublimation
	Liquid	Freezing	-	Boiling
	Gas	Deposition	Condensation	-

(d) Nested table

Linear text vs. diverse table layout structures

Weak connections between tables and text

Gameloft

From Wikipedia, the free encyclopedia

Gameloft SE is a French video game publisher based in Paris, founded in December 1999 by Ubisoft co-founder Michel Guillemot. The company operates 19 development studios worldwide, and publishes games with a special focus on the mobile games market. Formerly a public company traded at the Paris Bourse, Gameloft was acquired by media conglomerate Vivendi in 2016.

Contents [hide]

History [edit]

Game development strategy [edit]

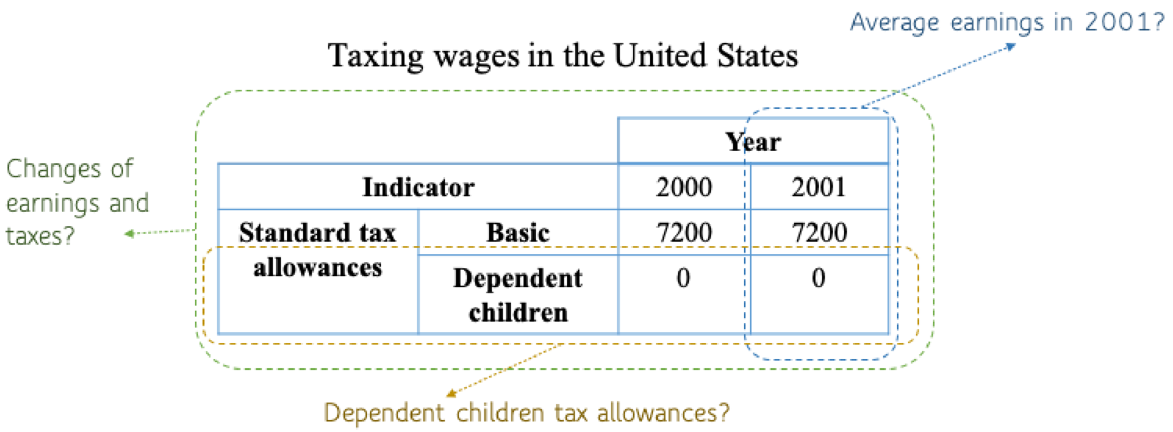
Gameloft was founded by Michel Guillemot, one of the five founders of Ubisoft, on 14 December 1999.[2][3] By February 2009, Gameloft had

Precise alignment rarely exists

Gameloft SE

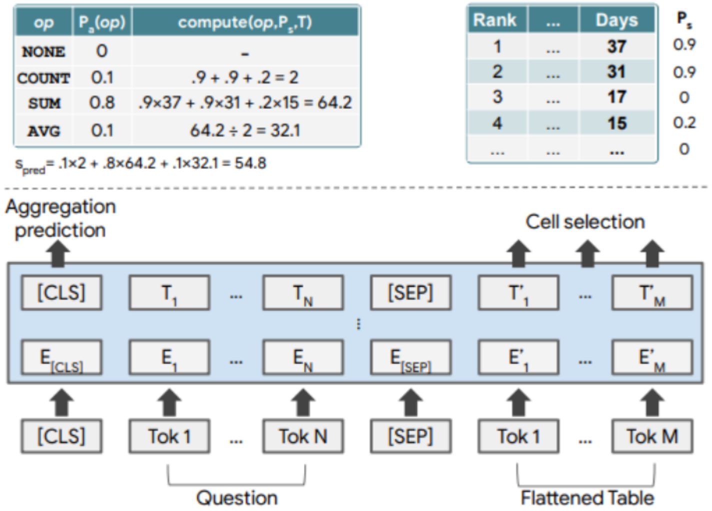
Type	Subsidiary
Industry	Video games
Founded	14 December 1999; 21 years ago
Founder	Michel Guillemot
Headquarters	Paris, France
Area served	Worldwide
Key people	Stéphane Roussel (chairman, CEO) Alexandre de Rochefort (CFO)
Revenue	258,000,000 euro (2017)
Number of employees	4,600[1] (2019)
Parent	Vivendi (2016–present)
Website	gameloft.com

Capturing multi-granular content



Agenda

1. Representation Learning for Tables + Language



2. Natural Language Interface for Tabular Content

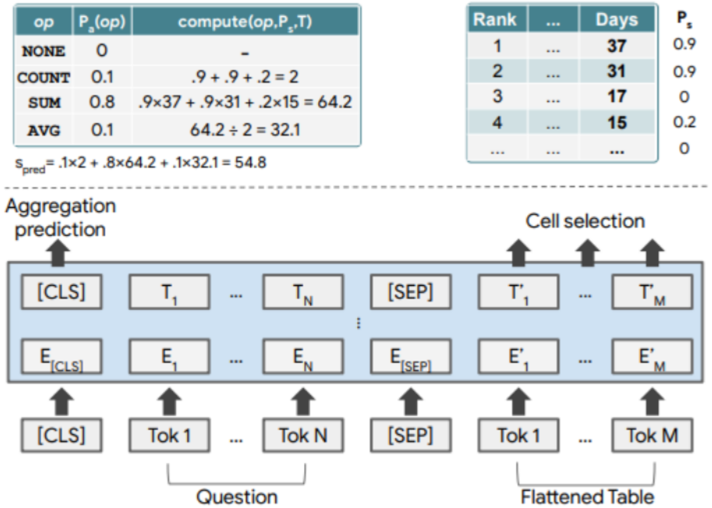
3. Table-assisted Natural Language Understanding

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

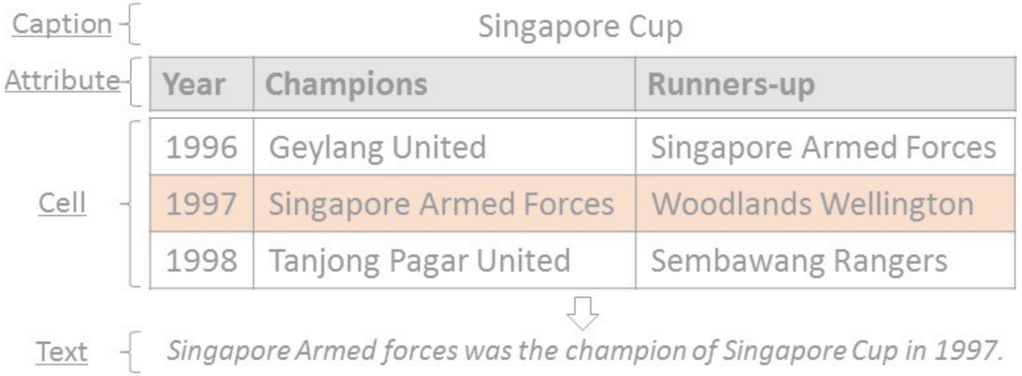
4. Open Research Directions

Agenda

1. Representation Learning for Tables + Language



2. Natural Language Interface for Tabular Content



3. Table-assisted Natural Language Understanding

Rank	Title	Sales	Platform(s)
1	Minecraft	200,000,000	Multi-platform
2	Grand Theft Auto V	135,000,000	Multi-platform
3	Tetris (EA)	100,000,000	Mobile
4	Wii Sports	82,900,000	Wii
5	PlayerUnknown's Battlegrounds	70,000,000	Multi-platform
6	Super Mario Bros.	48,240,000	Multi-platform
7	Pokémon Red / Green / Blue / Yellow	47,520,000	Multi-platform

4. Open Research Directions

Representation Learning for Tables and Text

The backbone of NL interfaces to tables and table-assisted NLU

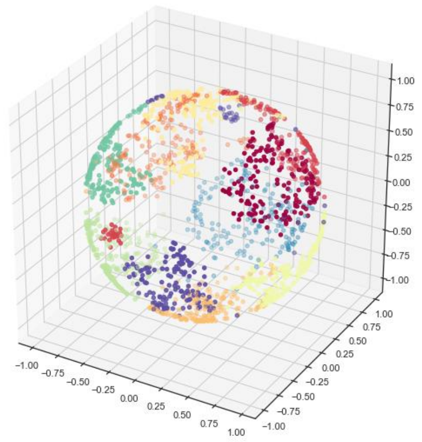
Goal

Tables

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

Natural Language

should be at the top of the list. When using the BEUrRE(rule+) model, the top 10 in all entities are *place, town, bed, school, city, home, house, capital, church, camp*, which are general concepts. Among the observed objects of the *atLocation* predicate, the entities that have the least coverage are *Tunisia, Morocco, Algeria, Westminster, Veracruz, Buenos Aires, Emilia-Romagna, Tyrrhenian sea, Kuwait, Serbia*. Those entities are very specific locations. This observation confirms that the box volume effectively represents probabilistic semantics and captures specificity/granularity of concepts, which we believe to be a reason for the performance improvement.



Joint (latent) representation

Relevance between NL and tabular content

Challenges

- Precise table-text alignment rarely exists.
- Tabular content is presented in different granularities (cells, rows, cols, etc.)
- Linear text vs. structured tables

TaBERT: Joint Language Modeling for Tables and Text

1. Coarse-grained table-text association

×2.6M from **Wikipedia** and **WDC Web Tables**

surrounding text

Coarse-grained association

In which city did Piotr's last 1st place finish occur?

	Year	Venue	Position	Event
R_1	2003	Tampere	3rd	EU Junior Championship
R_2	2005	Erfurt	1st	EU U23 Championship
R_3	2005	Izmir	1st	Universiade
R_4	2006	Moscow	2nd	World Indoor Championship
R_5	2007	Bangkok	1st	Universiade

Selected Rows as Content Snapshot : $\{R_2, R_3, R_5\}$

Top K rows based on ***n*-gram** overlapping with the text utterance (*n ≤ 3)*

2. BERT-based encoding with three pre-training tasks

pre-training objectives

- Masked Language Modeling (MLM) objective
- Masked Column Prediction: recovering column names and data types
- Cell Value Recovery

Transformer (BERT)

R_2 [CLS] In which city did Piotr's ... **[SEP]** Year | real | 2005 **[SEP]** Venue | text | Erfurt **[SEP]** Position | text | 1st **[SEP]** ...

Text utterance

Row linearization: a sequence of (column name, data type, value) tuples

Yin, et al. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. ACL-20

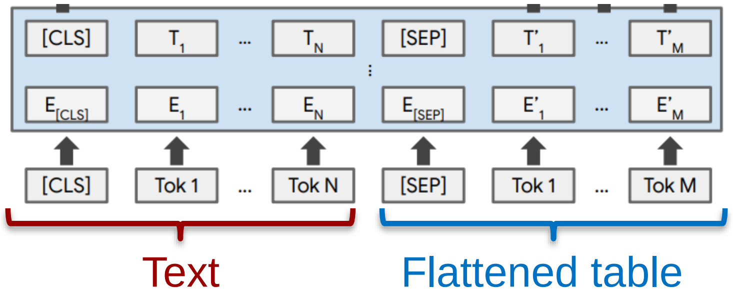
https://github.com/facebookresearch/TaBERT

TaPas: Weakly-supervised Table Question Answering

1. Pretraining

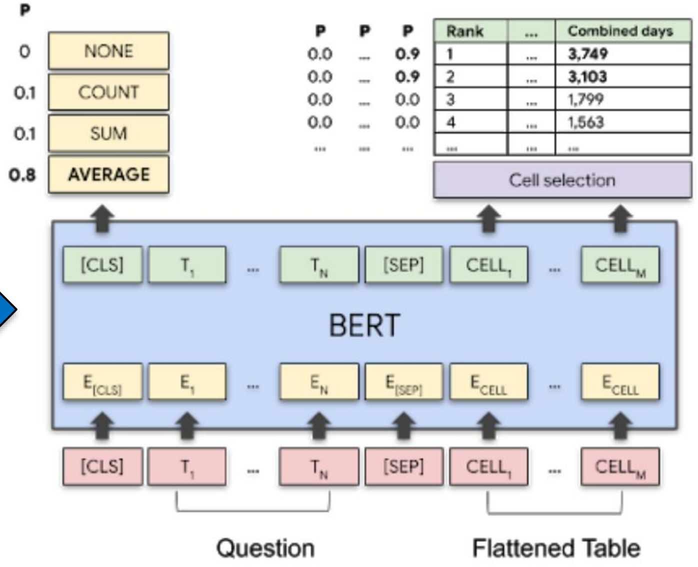
- **6.2M Tables:** 3.3M infoboxs and 2.9M WikiTables
- Table captions, article titles, article descriptions, segment titles and surround segment text

MLM Pretraining on BERT



2. Fine-tuning

WIKITQ (Pasupat+ ACL-15)
SQA (Iyyer+ ACL-17)
WikiSQL (Zhong+ 2017)



Which wrestler had the most number of reigns?	Ric Flair	Cell selection
Average time as champion for top 2 wrestlers?	AVG(3749,3103)=3426	Scalar answer

- Cell selection: selecting subsets of cells
- Scalar answer: estimating a soft scalar outcome over all aggregates with Huber loss

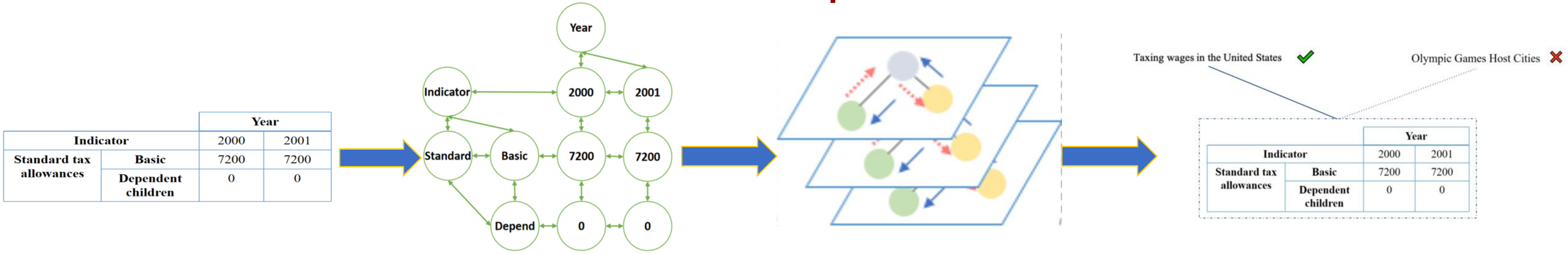
TaPas offers SOTA performance as the backbone model of table-based NLI tasks.

Herzig, et al. TaPas: Weakly Supervised Table Parsing via Pre-training. ACL-20
Eisenschlos, et al. Understanding tables with intermediate pre-training. Findings of EMNLP-20

https://github.com/google-research/tapas

Graph Representation Learning for Complex Tables

What if tables have complex structures?



Complex tables

Cell adjacency graph

Graph representation learning (e.g. Graph Transformer)

Graph-text matching

Comparing to language models

Pros:

- Can handle arbitrary table layout structures
- Can easily summarize multi-granular contents (with global nodes)

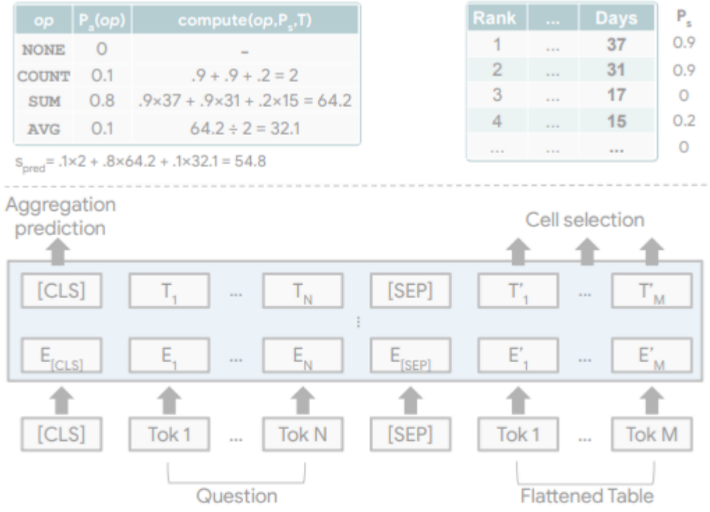
Con:

- Weaker table-text association (semantic shifts between feature spaces of the LM and the graph encoder)

Zhang, et al. A Graph Representation of Semi-structured Data for Web Question Answering. COLING-20
Wang, et al. Retrieving Complex Tables with Multi-Granular Graph Representation Learning. SIGIR-21

Agenda

1. Representation Learning for Tables + Language



2. Natural Language Interface for Tabular Content

3. Table-assisted Natural Language Understanding

4. Open Research Directions

Natural Language Interfaces for Tabular Content

1. Using natural language to retrieve the tabular content

The best-selling video game?

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

2.Describing tabular content with natural language

Caption: Singapore Cup

Attribute / Cell:

Year	Champions	Runners-up
1996	Geylang United	Singapore Armed Forces
1997	Singapore Armed Forces	Woodlands Wellington
1998	Tanjong Pagar United	Sembawang Rangers

Text: *Singapore Armed forces was the champion of Singapore Cup in 1997.*

Semantic Table Retrieval

Changes of taxes in U.S.?

Taxing wages in the United States

		Year	
Indicator		2000	2001
Standard tax allowances	Basic	7200	7200
	Dependent children	0	0

Olympic Games Host Cities

City	Country	Year	Continent
Los Angeles	U.S.	2028	North America
Milan–Cortina d'Ampezzo	Italy	2026	Europe
Paris	France	2024	
Beijing	China	2022	Asia

Input:

- A natural language query
- A set of **tables**, where each table consists of:
  - table body (headers, data cells, etc.)
  - context (captions, footnotes, etc.)

Output:

- A ranked list of **semantically relevant** tables

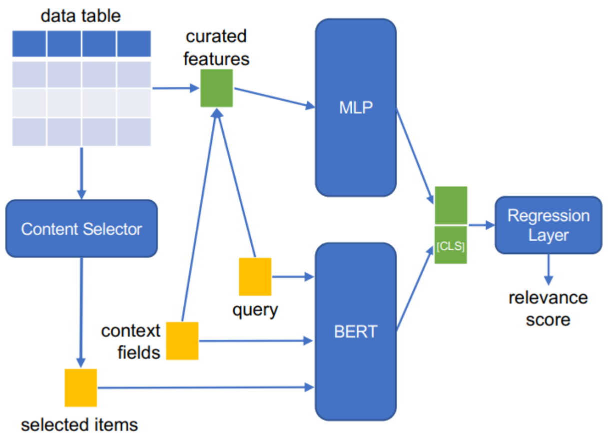
Semantic Table Retrieval

Earlier methods

Lexical matching
- **BM25**: Robertson, et al. Okapi at TREC-3. NIST special publication 500225 (1995)
- **Multi-field doc ranking**: Pimplikar and Sarawagi. 2012. Answering table queries on the web using column keywords. PVLDB-12
- **Lexical Table Retrieval**: Zhang and Balog: Ad hoc table retrieval using semantic similarity. WWW-18

Feature engineering / statistical machine learning
- **Linear regression**: Cafarella et al. Data integration for the relational web. PVLDB-09
- **Tab-Lasso**: Bhagavatula, et al. Methods for exploring and mining tables on wikipedia. KDD-13
- **MDF & GRU-matching**: Sun, et al. Content-based table retrieval for web queries. Neurocomputing 349 (2019), 183–189

Recent language models offer more precise and generalizable retrieval



BERT4TR
- Using BERT to match between queries and linearized tables
- Chen, et al. Table Search Using a Deep Contextualized Language Model. SIGIR-20

TaBERT offers even better performance

Semantic Table Retrieval

More challenges: Complex tables and diverse query intents

Various layout structures

Lake	Area
Windermere	5.69 sq mi
Ullswater	3.86 sq mi
Derwent Water	2.06 sq mi

(a) Relational table

Country	United States
State	California
County	Los Angeles
Region	South California

(b) Entity table

	Right-handed	Left-handed
Males	43	9
Females	44	4
Totals	87	12

(c) Matrix table

		To		
		Solid	Liquid	Gas
From	Solid	Solid trans	Melting	Sublimation
	Liquid	Freezing	-	Boiling
	Gas	Deposition	Condensation	-

(d) Nested table

Diverse query intents

Average earnings in 2001?

Taxing wages in the United States

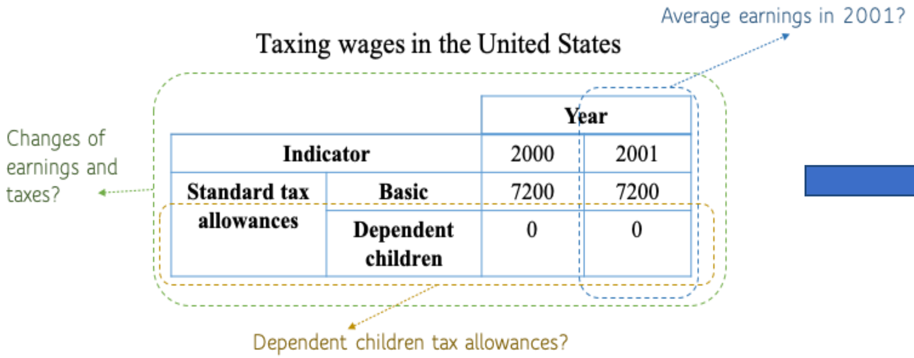
Indicator		Year	
		2000	2001
Standard tax allowances	Basic	7200	7200
	Dependent children	0	0

Changes of earnings and taxes?

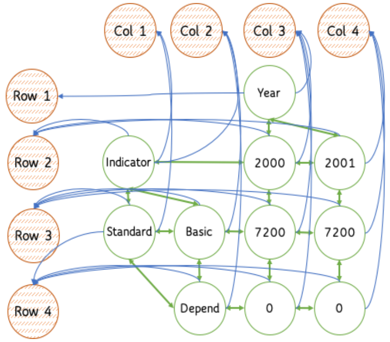
Dependent children tax allowances?

Wang, et al. Retrieving Complex Tables with Multi-Granular Graph Representation Learning. SIGIR, 2021

Semantic Table Retrieval

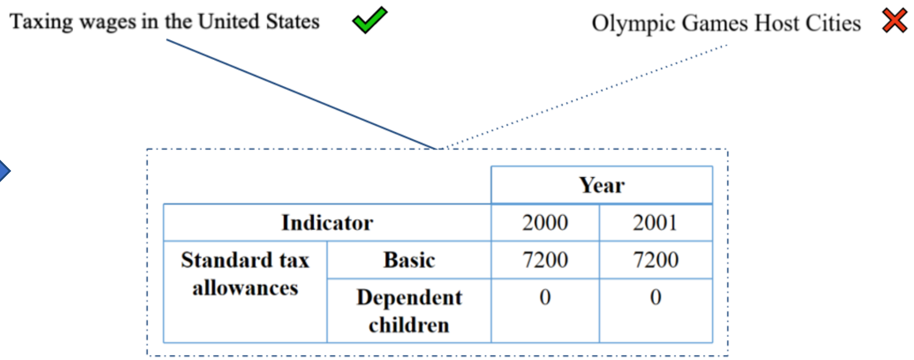


Arbitrary table layouts



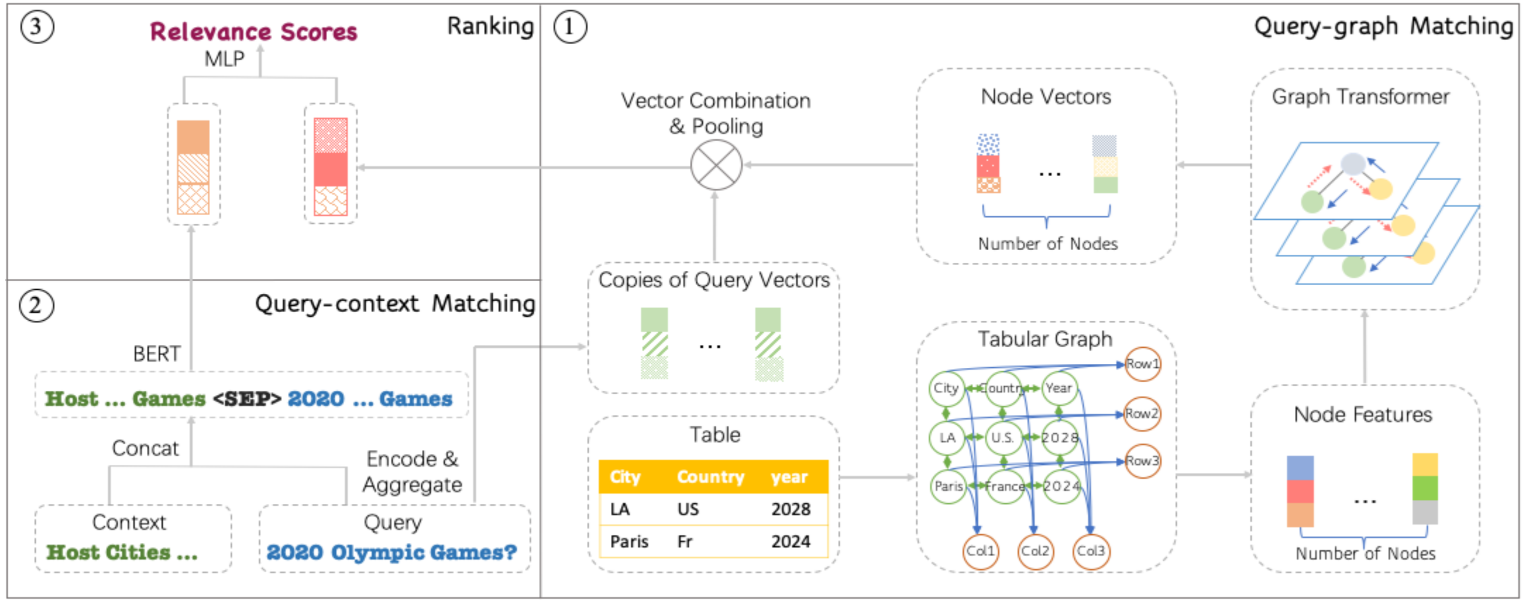
Multi-granular tabular graph

- Cell node adjacency
- Row-/Col- node summarization



Pre-trained graph transformer

- Table-caption matching



Model Architecture

Wang, et al. Retrieving Complex Tables with Multi-Granular Graph Representation Learning. SIGIR, 2021

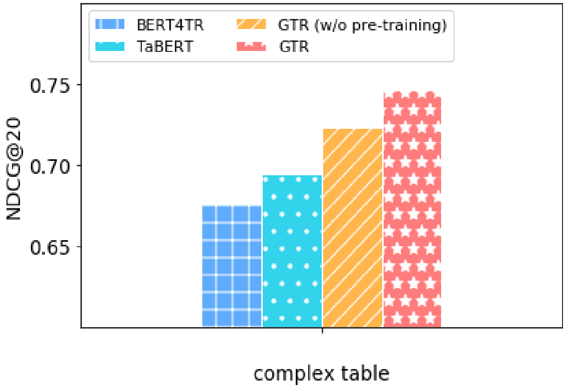
Semantic Table Retrieval

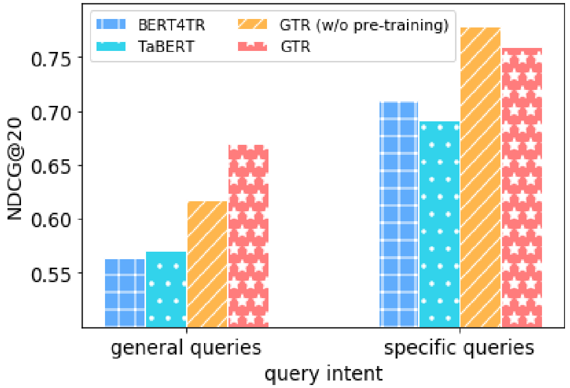
Pre-trained Graph Transformer (GTR)

Results on WikiTables

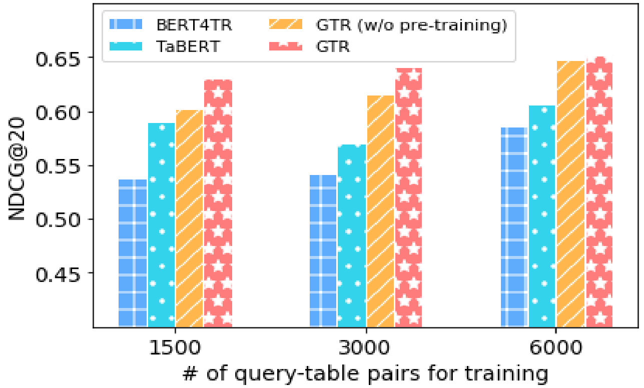
Method	NDCG@5	NDCG@10	NDCG@15	NDCG@20	MAP
BM25	0.3196	0.3377	0.3732	0.4045	0.4260
WebTable	0.2980	0.3150	0.3486	0.3922	-
SDR	0.4573	0.4841	0.5195	0.5534	-
MDR	0.5021	0.5116	0.5451	0.5761	-
Tab-Lasso	0.5161	0.5018	0.5330	0.5481	-
LTR	0.5910	0.5712	0.5858	0.6041	0.5615
TaBERT	0.5926	0.6108	0.6451	0.6668	0.6326
BERT4TR	0.6052	0.6171	0.6386	0.6689	0.6191
GTR (w/o pre-training)	0.6554	0.6747	0.6978	0.7211	0.6665
GTR	**0.6671**	**0.6856**	**0.7065**	**0.7272**	**0.6859**

Better generalization to **complex tables** and **diverse query intents**





Better **cross-dataset generalization**



Graph Transformer vs. Linear Language Models

- >8% relative improvement on all metrics
- better than BERT-based methods even w/o pre-training

Wang, et al. Retrieving Complex Tables with Multi-Granular Graph Representation Learning. SIGIR, 2021

Table-to-text Generation

Generating NL descriptions to summarize tabular content

- WIKIBIO dataset [Lebret+ EMNLP-16]: **surface-level NLG**.
- **Logical NLG** dataset [Chen+ ACL-20]

The emerging challenge: describing logical comparison

Medal Table from Tournament

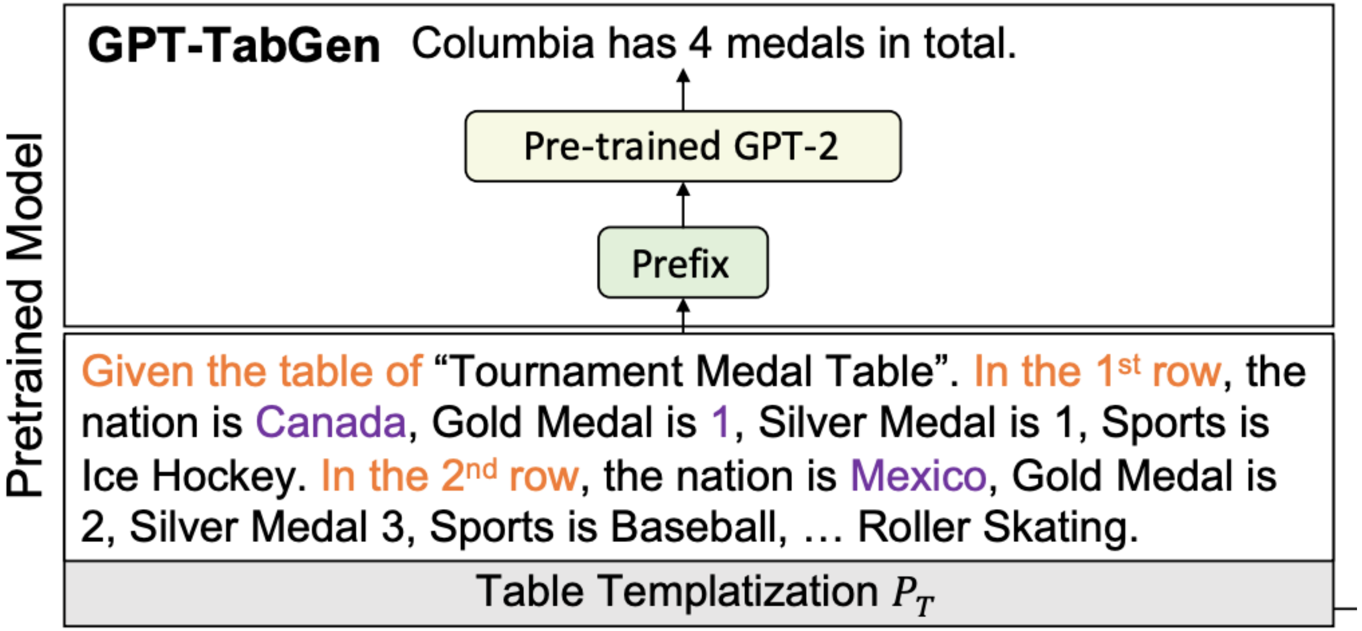
Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation
Sentence: Canada has got 3 gold medals in the tournament. **Sentence:** Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation
Sentence: Canada obtained 1 more gold medal than Mexico. **Sentence**: Canada obtained the most gold medals in the game.

Lebret, et al. Neural Text Generation from Structured Data with Application to the Biography Domain. EMNLP-16
Chen et al. Logical Natural Language Generation from Open-Domain Tables. ACL-20

Table-to-text Generation



GPT-TabGen [Chen+ ACL-20]

1. Generating a per-row (intermediate) description based on a <col name, value> template.
2. Summarize the intermediate description: fulfilling a summary template with GPT-2

Existing models can only achieve 20% logical correctness (according to Chen+ ACL-20)!

Lebret, et al. Neural Text Generation from Structured Data with Application to the Biography Domain. EMNLP-16
Chen et al. Logical Natural Language Generation from Open-Domain Tables. ACL-20

Controlled Table-to-text Generation

Summarizing facts only based on several highlighted cells

- The ToTTo dataset: 121,000 training examples; 7,500 examples each for development and test

Bill Dooley
Head coaching record

Year	Team	Overall	Conference	Standing	Bowl/playoffs	Coaches#	AP°
North Carolina Tar Heels (Atlantic Coast Conference) (1967 – 1977)							
1967	North Carolina	2 – 8	2 – 5	7th			
1968	North Carolina	3 – 7	1 – 6	8th			
1969	North Carolina	5 – 5	3 – 3	T – 3rd			
1970	North Carolina	8 – 4	5 – 2	T – 2nd	L Peach		
1971	North Carolina	9 – 3	6 – 0	1st	L Gator	18	
1972	North Carolina	11 – 1	6 – 0	1st	W Sun	14	12
1973	North Carolina	4 – 7	1 – 5	6th			
1974	North Carolina	7 – 5	4 – 2	T – 2nd	L Sun		
1975	North Carolina	3 – 7 – 1	1 – 4 – 1	6th			
1976	North Carolina	9 – 3	4 – 1	2nd	L Peach		
1977	North Carolina	8 – 3 – 1	5 – 0 – 1	1st	L Liberty	14	17
North Carolina:	69 – 53 – 2	38 – 28 – 2					
Virginia Tech Gobblers / Hokies (NCAA Division I-A Independent) (1978 – 1986)							
1978	Virginia Tech	4 – 7					
1979	Virginia Tech	5 – 6					
1980	Virginia Tech	8 – 4			L Peach		
1981	Virginia Tech	7 – 4					
1982	Virginia Tech	7 – 4					
1983	Virginia Tech	9 – 2					
1984	Virginia Tech	8 – 4			L Independence		
1985	Virginia Tech	6 – 5					
1986	Virginia Tech	10 – 2 – 1			W Peach		20
Virginia Tech:	64 – 38 – 1						
Wake Forest Demon Deacons (Atlantic Coast Conference) (1987 – 1992)							
1987	Wake Forest	7 – 4	4 – 3	T – 3rd			
1988	Wake Forest	6 – 4 – 1	4 – 3	T – 4th			
1989	Wake Forest	2 – 8 – 1	1 – 6	7th			
1990	Wake Forest	3 – 8	0 – 7	8th			
1991	Wake Forest	3 – 8	1 – 6	T – 7th			
1992	Wake Forest	8 – 4	4 – 4	T – 4th	W Independence	25	25
Wake Forest:	29 – 36 – 2	14 – 29					
Total:	163 – 126 – 5						
National championship Conference title Conference division title or championship game berth							
#Rankings from final Coaches Poll. ° Rankings from final AP Poll.							

The challenge: overgeneration (missing descriptions) and under generation (unexpected descriptions).

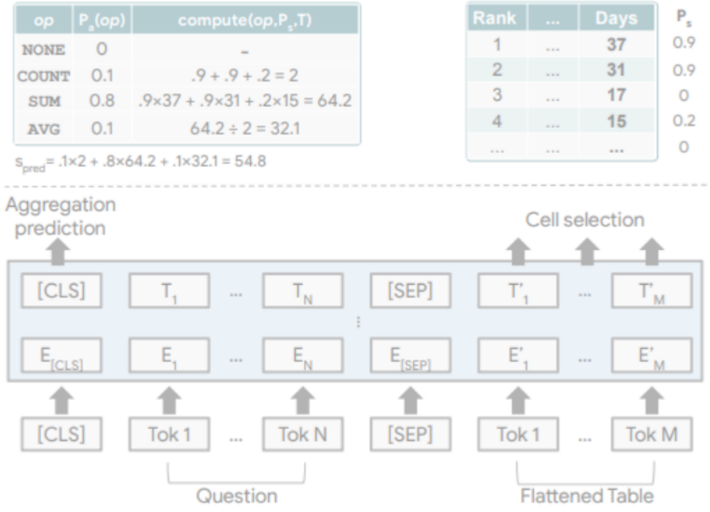
- **GOLD:** Bill Dooley served as the head coach at the North Carolina (1967–1977), Virginia tech (1978–1986) and Wake Forest (1987–1992).
- **BART(sub-table):** Bill Dooley served as the head coach at North Carolina from 1967 to 1974 and at Virginia Tech from 1974 to 1992.
- **BART(full-table):** Bill Dooley served as the head coach at North Carolina from 1967 to 1989 and at Virginia Tech from 1990 to 2005, compiling a career coaching record of 201–151–10.

An open question: graph representation learning as prior?

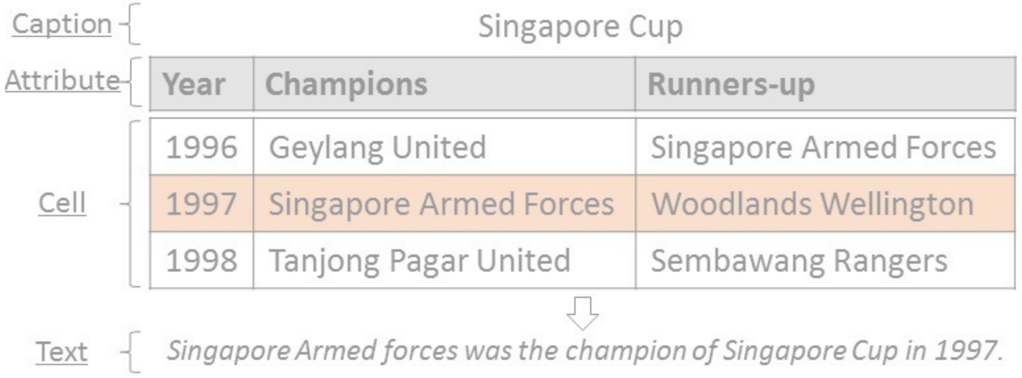
Parikh,, et al. ToTTo: A Controlled Table-To-Text Generation Dataset. EMNLP-20

Agenda

1. Representation Learning for Tables + Language



2. Natural Language Interface for Tabular Content



3. Table-assisted Natural Language Understanding

Minecraft is the best-selling game. (✓/✗)

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

4. Open Research Directions

Table-assisted Natural Language Understanding

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

- The best-selling video game of all time is **Minecraft**. ✓
- The best-selling video game of all time is **Tetris**. ✗

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

1. Web tables as trustworthy evidence for verifying claims

2. Web tables as clean references for answering questions

Table-based Fact Verification

The TabFact dataset: 16k Wikipedia tables as evidence for verifying 118k human annotated statements

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

1. **Table retrieval:** finding evidence table(s)
2. **NLI:** textual entailment using the table as premise and the statement as hypothesis

Chen et al. TabFact: A Large-scale Dataset for Table-based Fact Verification. ICLR-20

Table-based Fact Verification

Logical program based approach: learn to parse NL statements into logical programs, and execute the program on tables

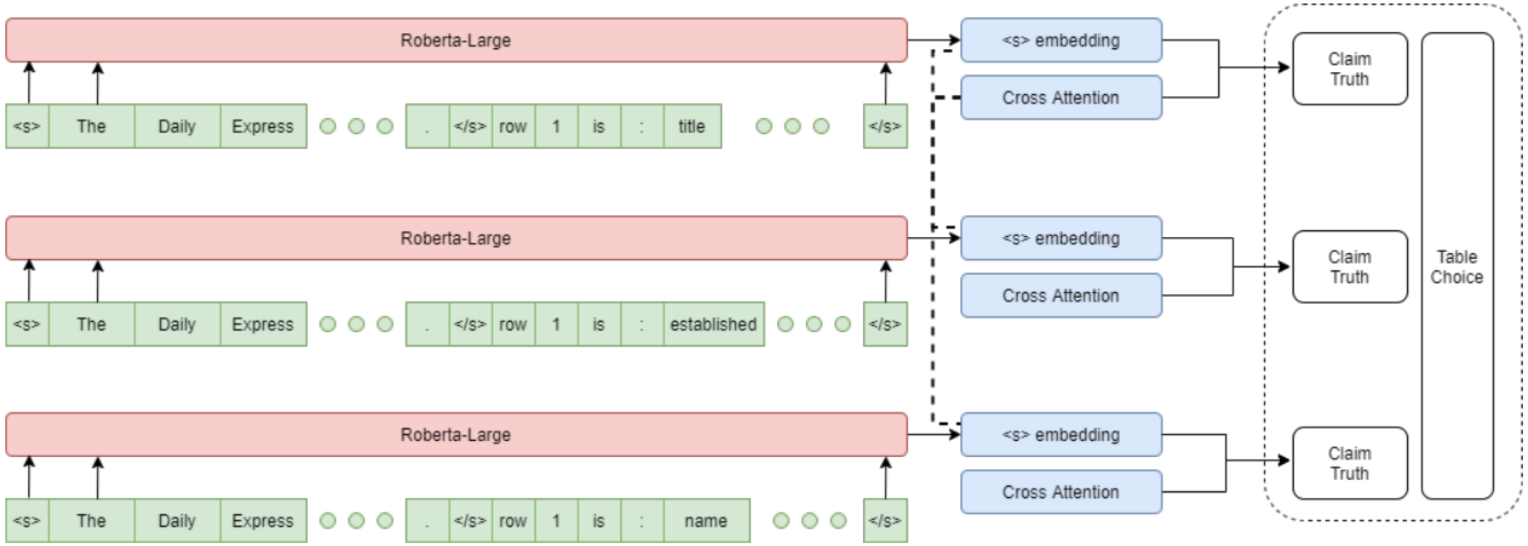
Year	Tournaments Played	Avg. Score	Scoring Rank
2007	22	72.46	81
2008	29	71.65	22
2009	25	71.90	34
2010	18	73.42	92
2011	11	74.42	125

Statement Ji-young Oh played more tournament in 2008 than any other year.

Logical form parser

Program *eq { max { all_rows ; tournaments played } ; hop { filter_eq { all_rows ; year ; 2008 } ; tournaments played } }* = True

Zhong et al. LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network. ACL-20
Yang et al. Program Enhanced Fact Verification with Verbalization and Graph Attention Network. EMNLP-20

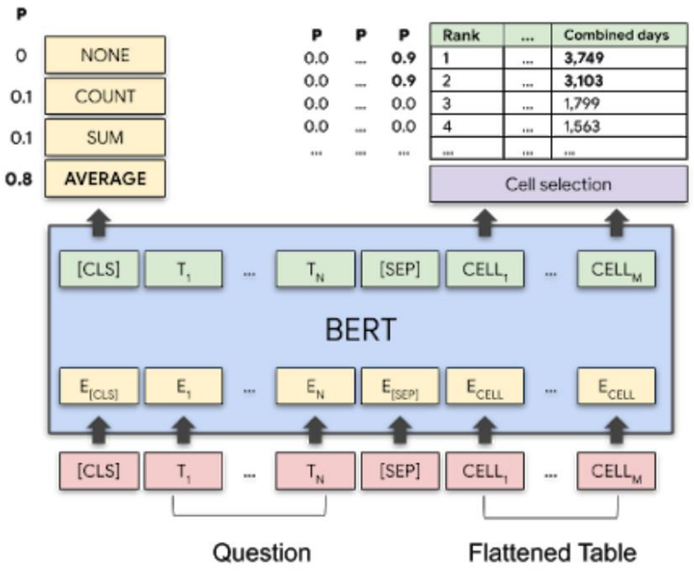


Jointly learning for table retrieval and textual entailment.

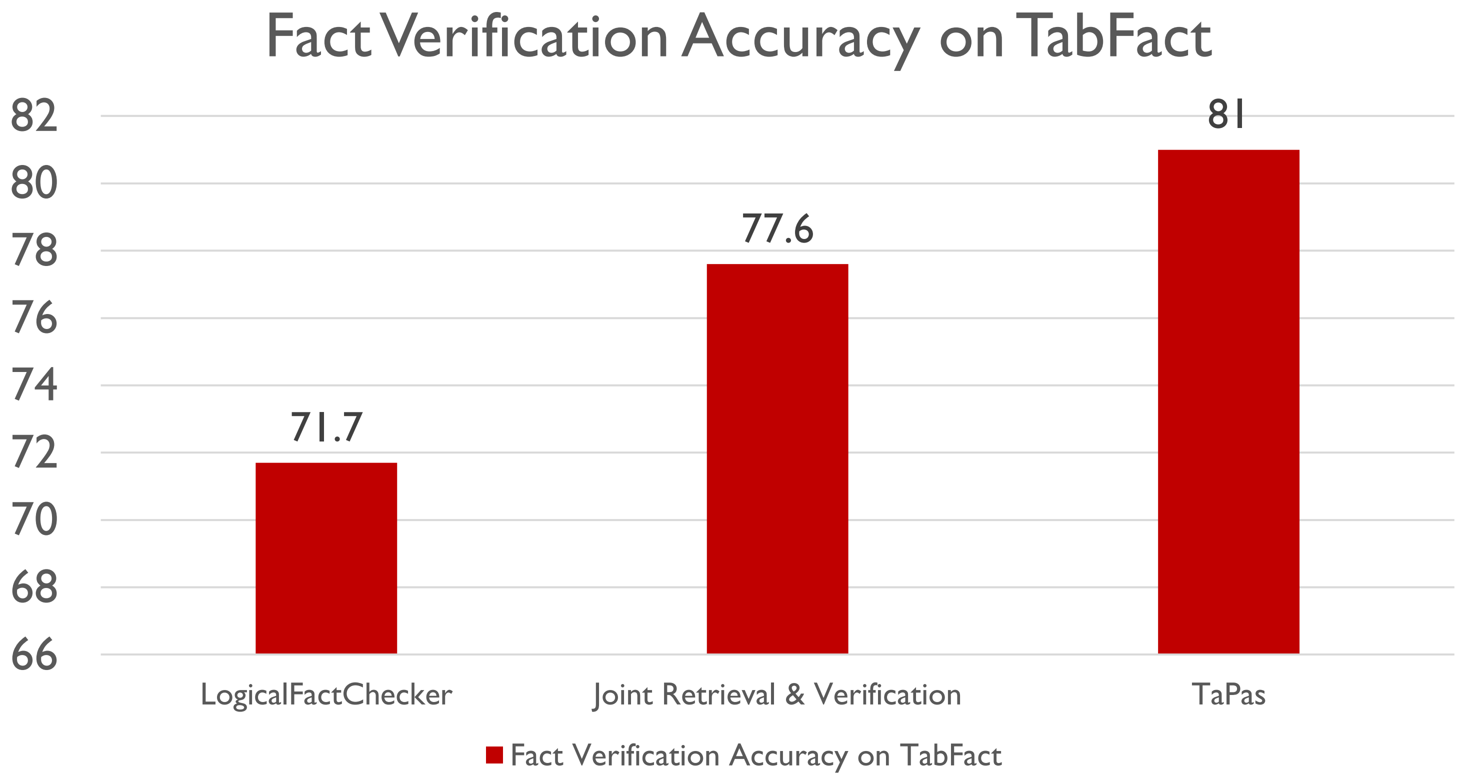
Schlichtkrull, et al. Joint Verification and Reranking for Open Fact Checking Over Tables. 2020

Table-based Fact Verification

Textual entailment seems to be the right direction.
Table-assisted language modeling (TaPas) provides a strong solution.



TaPas



Herzig, et al. TaPas: Weakly Supervised Table Parsing via Pre-training. ACL-20

Table QA

Searching for table cells that answer natural language questions

- TabMCQ [Jauhar+, ACL-16] and WikiTableQuestions [Pasupat and Liang, EMNLP-15]

Given:

Question — What languages do people in France speak?

Table Database from the Web

Country	Capital	Location	Main Language	Currency
Algeria	Algiers	Africa	Arabic, French	Dinar
France	Paris	Europe	French	Euro
Hungary	Budapest	Europe	Hungarian	Forint
Singapore	Singapore	Asia	Malay, Chinese, Tamil	Singapore Dollar

Goal: to find a table cell containing answers.

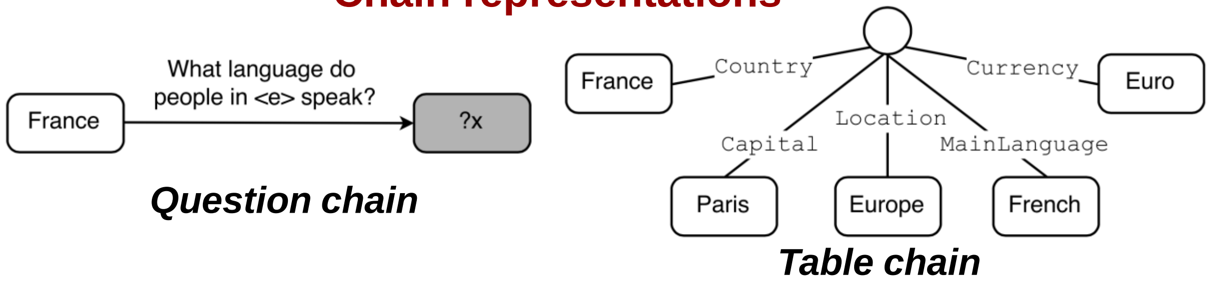
Answer — French

Evidence

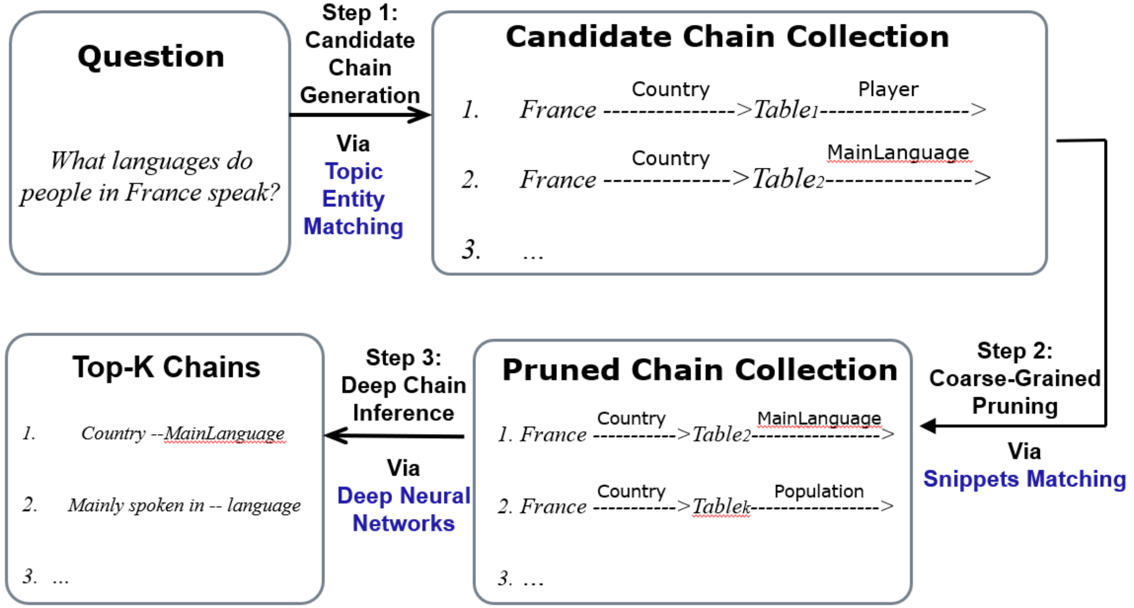
Country	Main Language
France	French

Source: http://hasibul.info/gk/countries.php

Chain representations



Chain matching



Sun, et al. Table Cell Search for Question Answering. WWW-16

Table QA

TaBERT [ACL-20] +Weakly-supervised Semantic Parser (MAPO [Liang+ NIPS-18])

1. Coarse-grained table-text association

×2.6M from **Wikipedia** and **WDC Web Tables**

surrounding text

Coarse-grained association

In which city did Piotr's last 1st place finish occur?

	Year	Venue	Position	Event
R_1	2003	Tampere	3rd	EU Junior Championship
R_2	2005	Erfurt	1st	EU U23 Championship
R_3	2005	Izmir	1st	Universiade
R_4	2006	Moscow	2nd	World Indoor Championship
R_5	2007	Bangkok	1st	Universiade

Selected Rows as Content Snapshot : $\{R_2, R_3, R_5\}$

Top K rows based on ***n*-gram** overlapping with the text utterance (*n* ≤ 3)

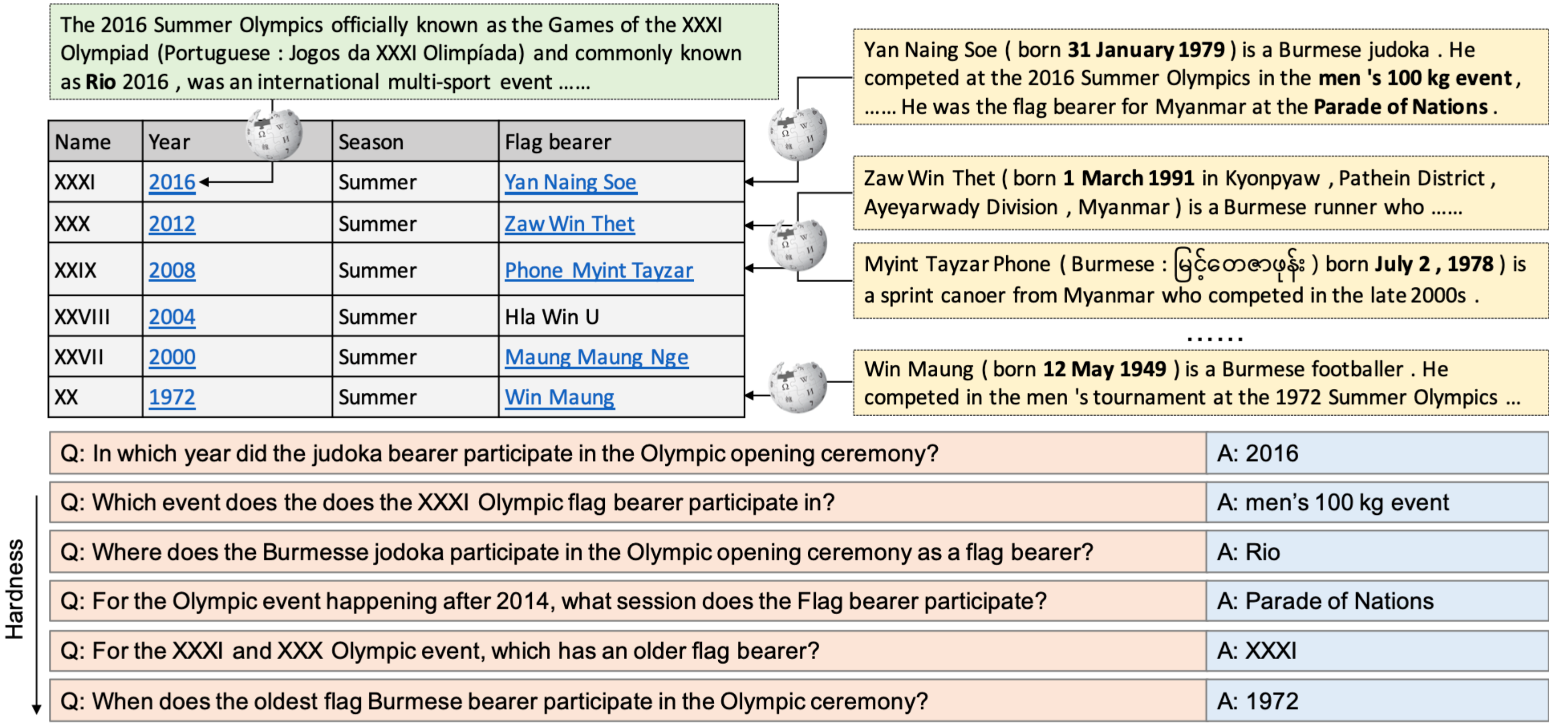
2. TaBERT as encoder for parsing questions into symbolic forms

In which city did Piotr's last 1st place finish occur?

```
Table.contains(column=Position, value=1st)     # Get rows whose 'Position' field contains '1st'
     .argmax(order_by=Year)                    # Get the row which has the largest 'Year' field
     .hop(column=Venue)                        # Select the value of 'Venue' in the result row
```

51.8 testing accuracy on WIKITQ, one of the SOTA's

HybridQA



Answering questions based on complementary information in tables and documents:

- 13K Wiki Tables
- Hyperlinked paragraphs

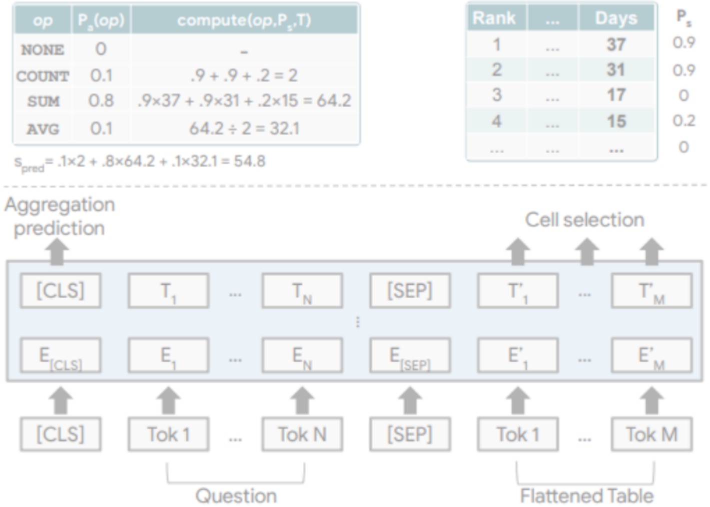
Split	Train	Dev	Test	Total
In-Passage	35,215	2,025	20,45	39,285 (56.4%)
In-Table	26,803	1,349	1,346	29,498 (42.3%)
Computed	664	92	72	828 (1.1%)
Total	62,682	3,466	3,463	69,611

Need to combine both TableQA and Doc QA

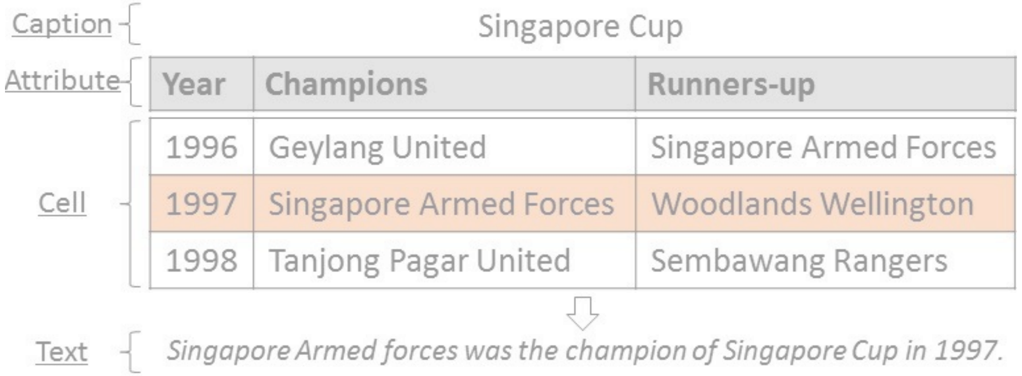
Chen, et al. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. Findings of EMNLP-20

Agenda

1. Representation Learning for Tables + Language



2. Natural Language Interface for Tabular Content



3. Table-assisted Natural Language Understanding

Rank	Title	Sales	Platform(s)
1	Minecraft	200,000,000	Multi-platform
2	Grand Theft Auto V	135,000,000	Multi-platform
3	Tetris (EA)	100,000,000	Mobile
4	Wii Sports	82,900,000	Wii
5	PlayerUnknown's Battlegrounds	70,000,000	Multi-platform
6	Super Mario Bros.	48,240,000	Multi-platform
7	Pokémon Red / Green / Blue / Yellow	47,520,000	Multi-platform

4. Open Research Directions

Language Grounding to Tables

Grounding text spans (in scientific literature) to corresponding tabular content

Table 4: Ablation study of **EVA** based on DBP15k (FR→EN).

model	H@1	H@10	MRR
w/o structure	.391 ±.004	.514 ±.003	.423 ±.004
w/o image	.749 ±.002	.929 ±.002	.817 ±.001
w/o attribute	.750 ±.003	.927 ±.001	.813 ±.003
w/o relation	.763 ±.006	.928 ±.003	.823 ±.004
w/o IL	.715 ±.003	.936 ±.002	.795 ±.004
w/o CSLS	.786 ±.005	.928 ±.001	.838 ±.003
full model	.793 ±.003	.942 ±.002	.847 ±.004

4.3 Ablation Study

We report an ablation study of **EVA** in Tab. 4 using DBP15k (FR→EN). As shown, IL brings ca. 8% absolute improvement. This gap is smaller than what has been reported previously (Sun et al. 2018). This is because the extra visual supervision in our method already allows the model to capture fairly good alignment in the first 500 epochs, leaving smaller room for further improvement from IL. CSLS gives minor but consistent improvement to all metrics during infer-

Scientific Leaderboard Construction

Scientific Publication

A Joint Model for Entity Analysis: Coreference, Typing, and Linking

Abstract: We present a joint model of three core tasks in the entity analysis stack: coreference resolution (within-document clustering), named entity recognition (coarse semantic typing), and entity linking (matching to Wikipedia entities). Our model is formally a structured conditional random field. Unary factors encode local features from strong baselines for each task. We then add binary and ternary factors to capture cross-task interactions, such as the constraint that coreferent mentions have the same semantic type. On the ACE 2005 and OntoNotes datasets, we achieve state-of-the- art results for all three tasks. Moreover, joint modeling improves performance on each task over strong independent baselines.

...

	Dev						Test					
	MUC	B^3	$CEAF_e$	Avg.	NER	Link	MUC	B^3	$CEAF_e$	Avg.	NER	Link
INDEP.	77.95	74.81	71.84	74.87	83.04	73.07	81.03	74.89	72.56	76.16	82.35	74.71
JOINT	**79.41**	**75.56**	**73.34**	**76.10**	**85.94**	**75.69**	81.41	74.70	72.93	76.35	**85.60**	**76.78**
Δ	+1.46	+0.75	+1.50	+1.23	+2.90	+2.62	+0.42	−0.19	+0.37	+0.19	+3.25	+2.07

Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (task-specific factors only)

Leaderboard Annotations

Task	Dataset	Evaluation Metric	Best Result
Named Entity Recognition	ACE 2005 (Test)	Accuracy	85.60
Entity Linking	ACE 2005 (Test)	Accuracy	76.78
Coreference Resolution	ACE 2005 (Test)	Avg. F1	76.35
...	...	...	...

Hou, et al. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. ACL-19

Automated Table Cleaning and Expansion

How to automatically query Web corpora, verify what are in the table and add what are not there?

Rank	Title	Sales	Platform(s)
1	*Minecraft*	200,000,000	Multi-platform
2	*Grand Theft Auto V*	135,000,000	Multi-platform
3	*Tetris* (EA)	100,000,000	Mobile
4	*Wii Sports*	82,900,000	Wii
5	*PlayerUnknown's Battlegrounds*	70,000,000	Multi-platform
6	*Super Mario Bros.*	48,240,000	Multi-platform
7	*Pokémon Red / Green / Blue / Yellow*	47,520,000	Multi-platform

1. Answer-agnostic question generation

- *How many sales does Minecraft have?*
- *What are popular Nintendo Switch games?*

2. Cleaning: Open-domain QA + Claim verification

Web corpora

3. Expansion: Open-domain QA + Answer consolidation

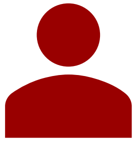
Tables and Dialogue Agents

Table-assisted Dialogue Agent

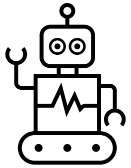
Chain restaurant steakhouses [edit]

North America [edit]

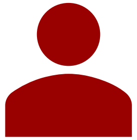
- Black Angus
- Bonanza Steakhouse
- The Capital Grille
- Charlie Brown's Steakhouse
- Claim Jumper
- Del Frisco's Double Eagle Steak House
- Doe's Eat Place
- Fleming's Prime Steakhouse & Wine Bar
- Fogo de Chão
- Harry Caray's Italian Steakhouse
- Hoss's Steak and Sea House
- Houston's Restaurant
- K-Bob's Steakhouse
- The Keg
- Lawry's
- Logan's Roadhouse
- Lone Star Steakhouse & Saloon
- LongHorn Steakhouse
- Montana Mike's
- Morton's The Steakhouse
- Mr. Steak
- Outback Steakhouse
- The Palm
- Ponderosa Steakhouse
- Quaker Steak & Lube
- Rodizio Grill
- Rustler Steak House
- Ruth's Chris Steak House
- Saltgrass Steak House
- Sirloin Stockade
- Sizzler
- Smith & Wollensky
- Steak and Ale Restaurant
- Stoney River Legendary Steaks
- Strip House
- Tahoe Joe's
- Texas de Brazil
- Texas Land and Cattle
- Texas Roadhouse
- Timber Lodge Steakhouse
- Valle's Steak House
- Victoria Station (defunct)
- Western Sizzlin'
- Wolfgang's Steakhouse
- York Steak House



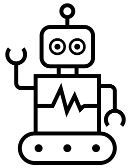
I want to reserve a in Beverly Hills.



What type of food?



A popular steakhouse. But not too expensive.

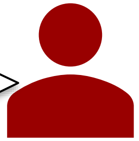


How about *Lawry's the Prime Rib*?

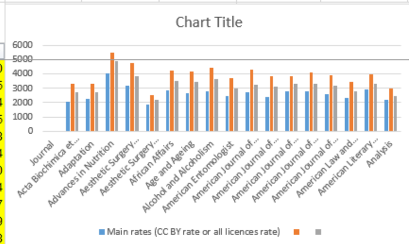
Conversational Spreadsheet Editing

	Main rates (CC BY rate or all licences rate)			Main member rates (CC BY rate or all licences rate)			Licences offered		
Journal	GBP	USD	EUR	GBP	USD	EUR	CC BY	CC BY-NC	CC BY-NC-ND
Acta Biochimica et Biophysica Sinica	2090	3350	2720				x	x	x
Adaptation	2265	3312	2725				x		
Advances in Nutrition	4044	5500	4894	3309	4500	4004	x	x	
Aesthetic Surgery Journal	3177	4765	3875	2530	3800	3100	Available for funded articles only		
Aesthetic Surgery Journal Open Forum	1838	2500	2188	1471	2000	1750	x		
African Affairs	2841	4261	3484				x		
Age and Ageing	2658	4202	3460				Available for fun	x	x
Alcohol and Alcoholism	2791	4413	3634				x		
American Entomologist	2478	3717	3017	1983	2974	2413	x		
American Journal of Clinical Pathology	2690	4290	3219	1759	2812	2286	x		
American Journal of Epidemiology	2363	3833	3098				x	x	
American Journal of Health-System Pharma	2809	3820	3343	2478	3371	2949	x		
American Journal of Hypertension	2757	4136	3343				Available for funded articles only		
American Journal of Legal History	2596	3894	3180				x		
American Law and Economics Review	2307	3460	2812				x		
American Literary History	2892	3955	3341				x		x
Analysis	2200	2992	2442				x	x	x
Animal Frontiers	0	0	0				x		
Annals of Behavioral Medicine	2286	3809	2742	1829	3047	2195	x	x	

delete 6 rows from the beginning
delete the left most two rows
merge the cells from C1 to C3
create line charts using data from B2 through D20



	Main rates (CC BY rate or all licences rate)							
Journal	GBP	USD	EUR					
Acta Biochimica et Biophysica Sinica	2090	3350	2720					
Adaptation	2265	3312	2725					
Advances in Nutrition	4044	5500	4894					
Aesthetic Surgery Journal	3177	4765	3875					
Aesthetic Surgery Journal Open Forum	1838	2500	2188					
African Affairs	2841	4261	3484					
Age and Ageing	2658	4202	3460					
Alcohol and Alcoholism	2791	4413	3634					
American Entomologist	2478	3717	3017					
American Journal of Clinical Pathology	2690	4290	3219					
American Journal of Epidemiology	2363	3833	3098					
American Journal of Health-System Pharma	2809	3820	3343	2478	3371	2949	x	
American Journal of Hypertension	2757	4136	3343				Available for funded articles only	
American Journal of Legal History	2596	3894	3180				x	
American Law and Economics Review	2307	3460	2812				x	
American Literary History	2892	3955	3341				x	
Analysis	2200	2992	2442				x	x



References

Yin, et al. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. ACL-20

Herzig, et al. TaPas: Weakly Supervised Table Parsing via Pre-training. ACL-20

Eisenschlos, et al. Understanding tables with intermediate pre-training. Findings of EMNLP-20

Zhang, et al. A Graph Representation of Semi-structured Data for Web Question Answering. COLING-20

Wang, et al. Retrieving Complex Tables with Multi-Granular Graph Representation Learning. SIGIR-21

Chen, et al. Table Search Using a Deep Contextualized Language Model. SIGIR-20

Lebret, et al. Neural Text Generation from Structured Data with Application to the Biography Domain. EMNLP-16

Chen et al. Logical Natural Language Generation from Open-Domain Tables. ACL-20

Parikh,, et al. ToTTo: A Controlled Table-To-Text Generation Dataset. EMNLP-20

Chen et al. TabFact : A Large-scale Dataset for Table-based Fact Verification. ICLR-20

Schlichtkrull, et al. Joint Verification and Reranking for Open Fact Checking Over Tables. 2020

Zhong et al. LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network. ACL-20

Yang et al. Program Enhanced Fact Verification with Verbalization and Graph Attention Network. EMNLP-20

Sun, et al. Table Cell Search for Question Answering. WWW-16

Chen, et al. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. Findings of EMNLP-20

Iyyer, et al. Search-based neural structured learning for sequential question answering. ACL-17

Zhong, et al. Seq2sql: Generating structured queries from natural language using reinforcement learning. 2017

USC Viterbi
School of Engineering

THE OHIO STATE UNIVERSITY
COLLEGE OF ENGINEERING

Thank You