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

Neural Representation Learning on Tables

Huan Sun

The Ohio State University

USC Viterbi School of Engineering

The Ohio State University College of Engineering

Outline: Neural Representation Learning on Tables

- Background

Outline: Neural Representation Learning on Tables

- Background

- Representative Methods (pre-training objectives, table types, tasks)
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

Outline: Neural Representation Learning on Tables

- Background

- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- Summary

- Resources

Outline: Neural Representation Learning on Tables

- **Background**

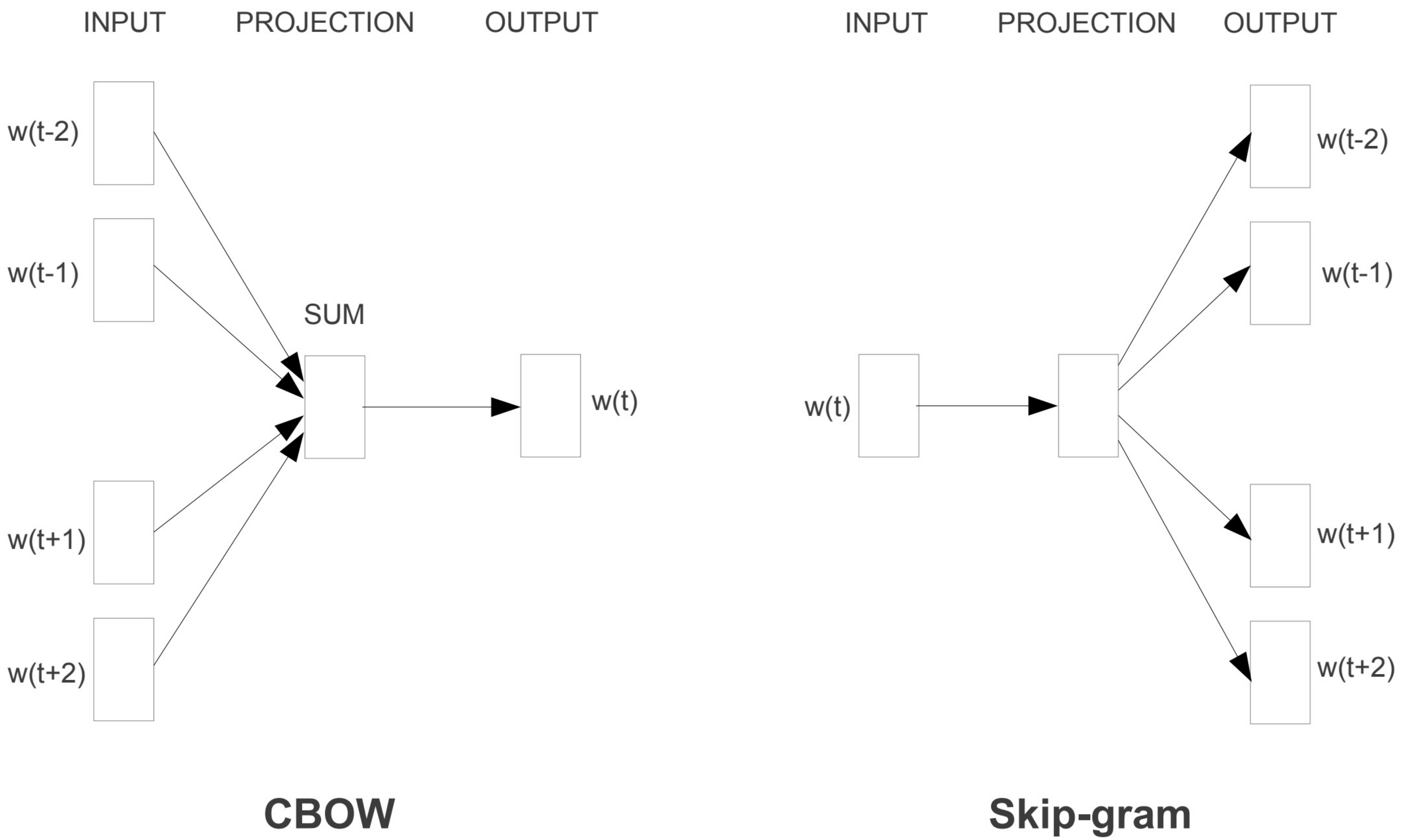
- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- Summary

- Resources

Background: Representation Learning on Text

- Word2Vec [Mikolov et al., 2013]



Background: Contextualized Representation Learning and Pre-training

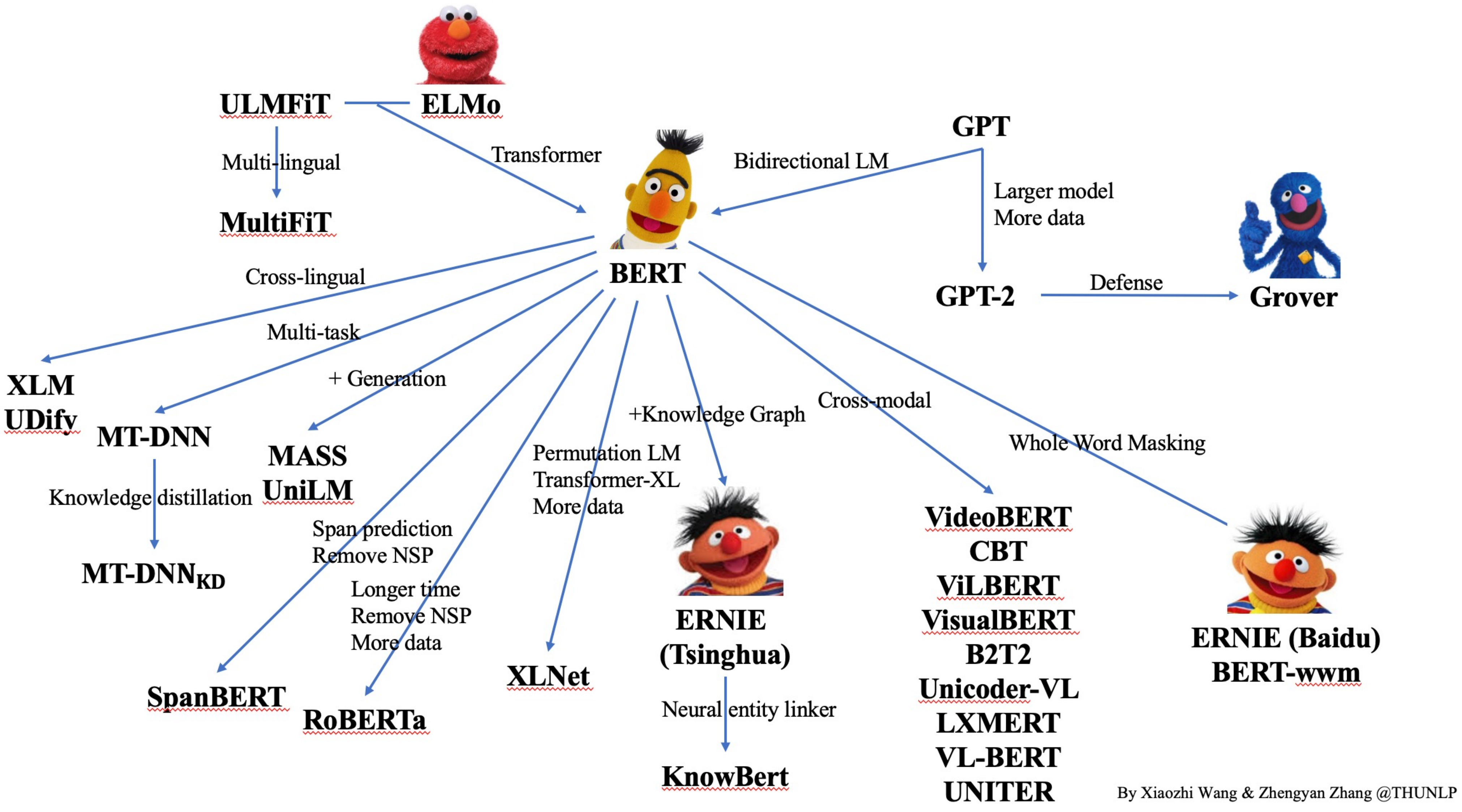
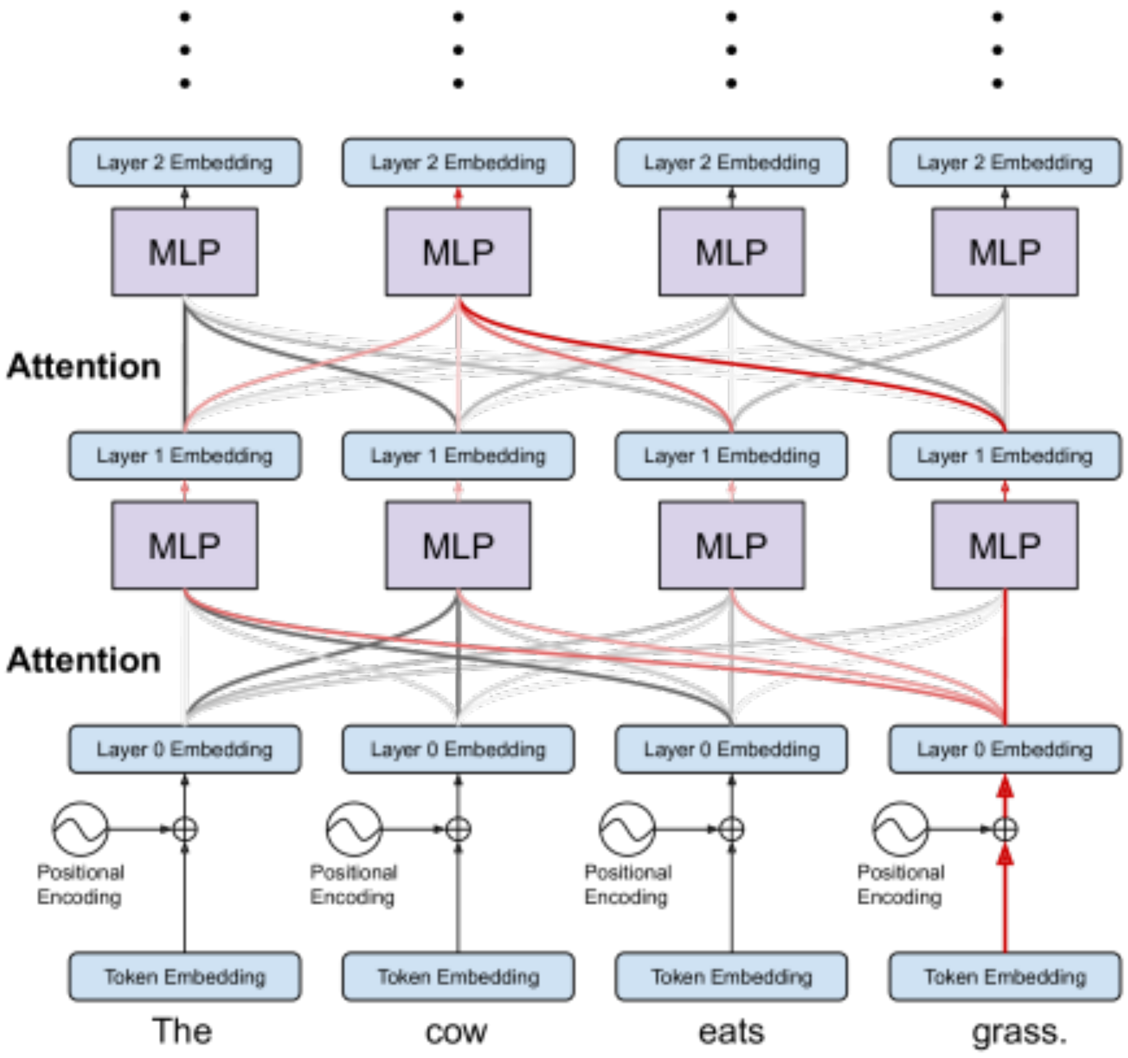


Figure credit: https://github.com/thunlp/PLMpapers

Background: Transformer [Vaswani et al., 2017]



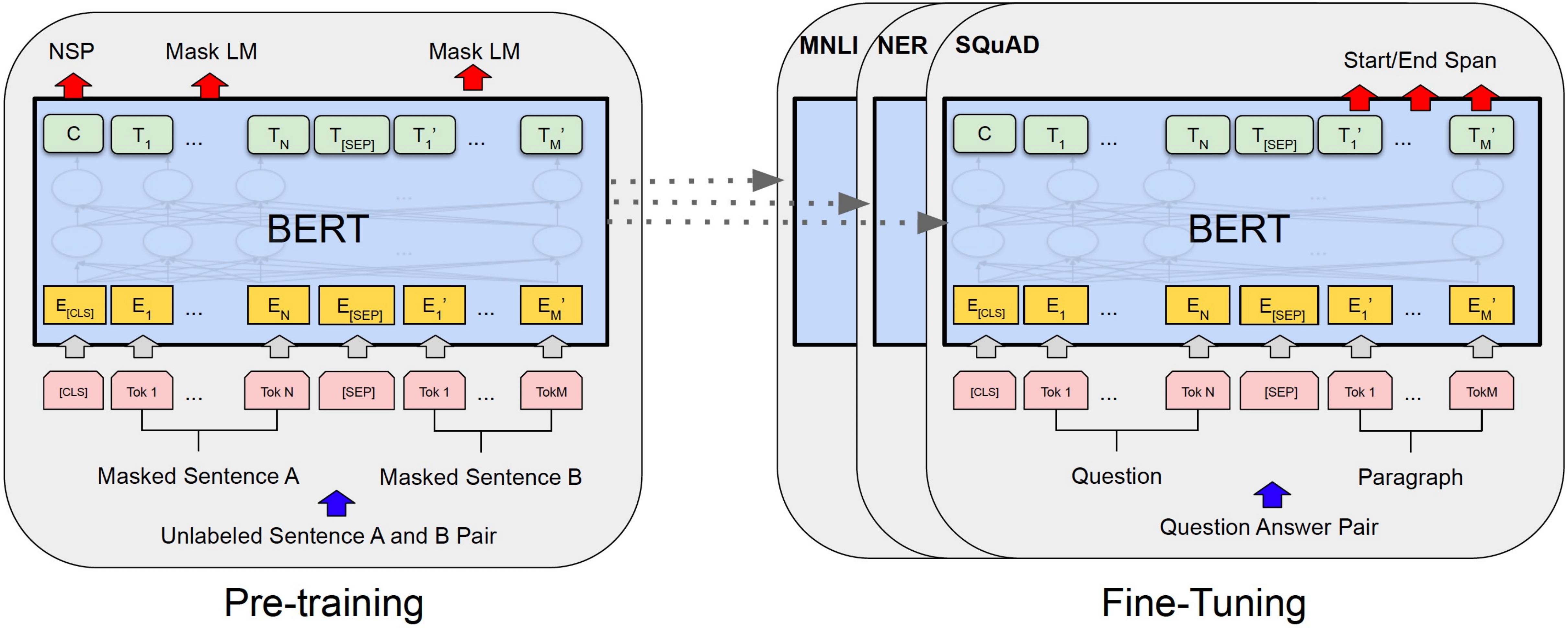
$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V)$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}}$$

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}}$$

Figure credit: https://www.arxiv-vanity.com/papers/1908.04211

Background: BERT [Devlin et al., 2019]



1. Self-supervised objectives: Masked language model (MLM) and Next Sentence Prediction (NSP)
2. Encoder: Transformer encoder [Vaswani et al., 2017]
3. Various downstream tasks such as Question Answering, Named Entity Recognition, etc.

Background: Previous tutorials

- Tutorial on Web Table Extraction, Retrieval and Augmentation
  - Shuo Zhang and Krisztian Balog, SIGIR 2019
  - https://iai-group.github.io/webtables-tutorial/

- Tutorial on Table Extraction and Understanding for Scientific and Enterprise Applications
  - Doug Burdick, Alexandre V Evfimievski, Nancy Wang, Yannis Katsis, Marina Danilevsky, VLDB 2020
  - https://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=10534

For more about definitions of various tasks and other approaches that are not based on table representation learning

Neural Representation Learning on Tables

- Different from "Using Deep Learning for Table-based Tasks"

- General-purpose, not task-specific representations of table elements

- "Unsupervised" representation learning
  - No human annotation
  - Self-supervised data
  - Self-supervised tasks

Outline: Neural Representation Learning on Tables

- Background

- **Representative Methods**
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- Summary

- Resources

Outline: Neural Representation Learning on Tables

- Background

- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

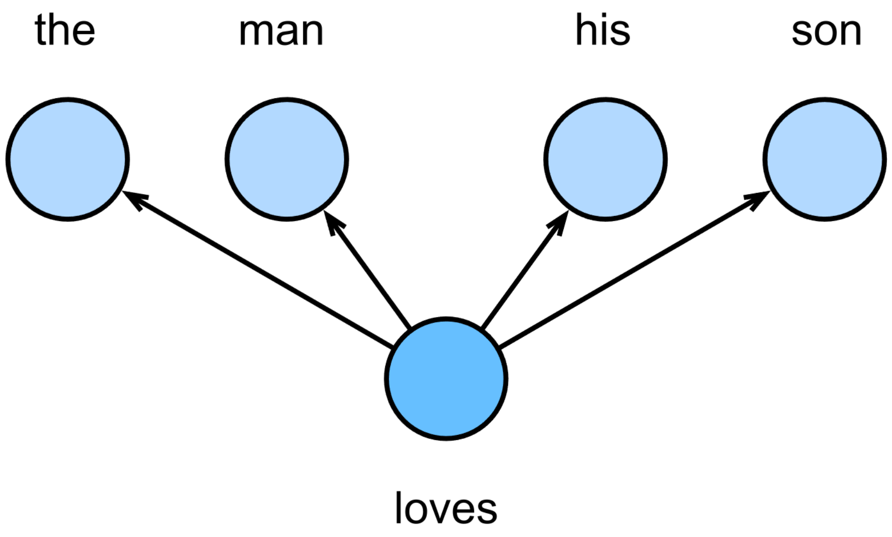
- Summary

- Resources

Table2Vec: Neural Word and Entity Embeddings

- Backbone algorithm: Word2Vec (Skip-Gram) [Mikolov et al., NIPS'13]

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$



[Deng et al., SIGIR'19 (short)]

Figure credit: https://d2l.ai/chapter_natural-language-processing-pretraining/word2vec.html

Table2Vec: Neural Word and Entity Embeddings

- Backbone algorithm: Word2Vec [Mikolov et al., NIPS'13]

- Four variants, depending on what info in a table is modeled

Method	Input	Semantic repr.
Table2VecW	all table data	word embeddings
Table2VecH	table headings	heading embeddings
Table2VecE	all entities	entity embeddings
Table2VecE*	core column entities	entity embeddings

[Deng et al., SIGIR'19 (short)]

Table2Vec: Neural Word and Entity Embeddings

- Backbone algorithm: Word2Vec [Mikolov et al., NIPS'13]

- Four variants

- Tasks
  - Row population
  - Column population
  - Table retrieval

Some results will be discussed later together with more advanced methods'

[Deng et al., SIGIR'19 (short)]

Outline: Neural Representation Learning on Tables

- Background

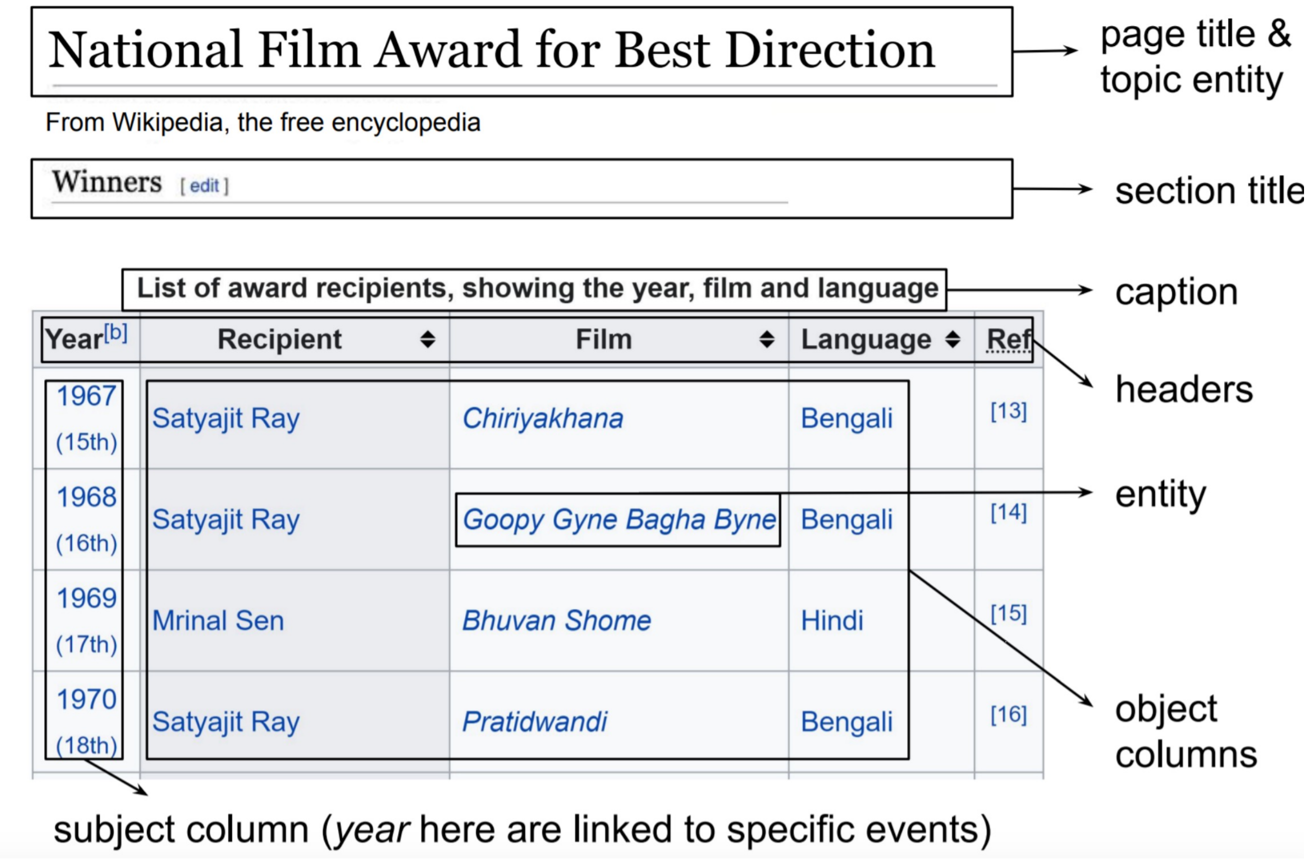
- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- Summary

- Resources

TURL: Table Understanding through Representation Learning

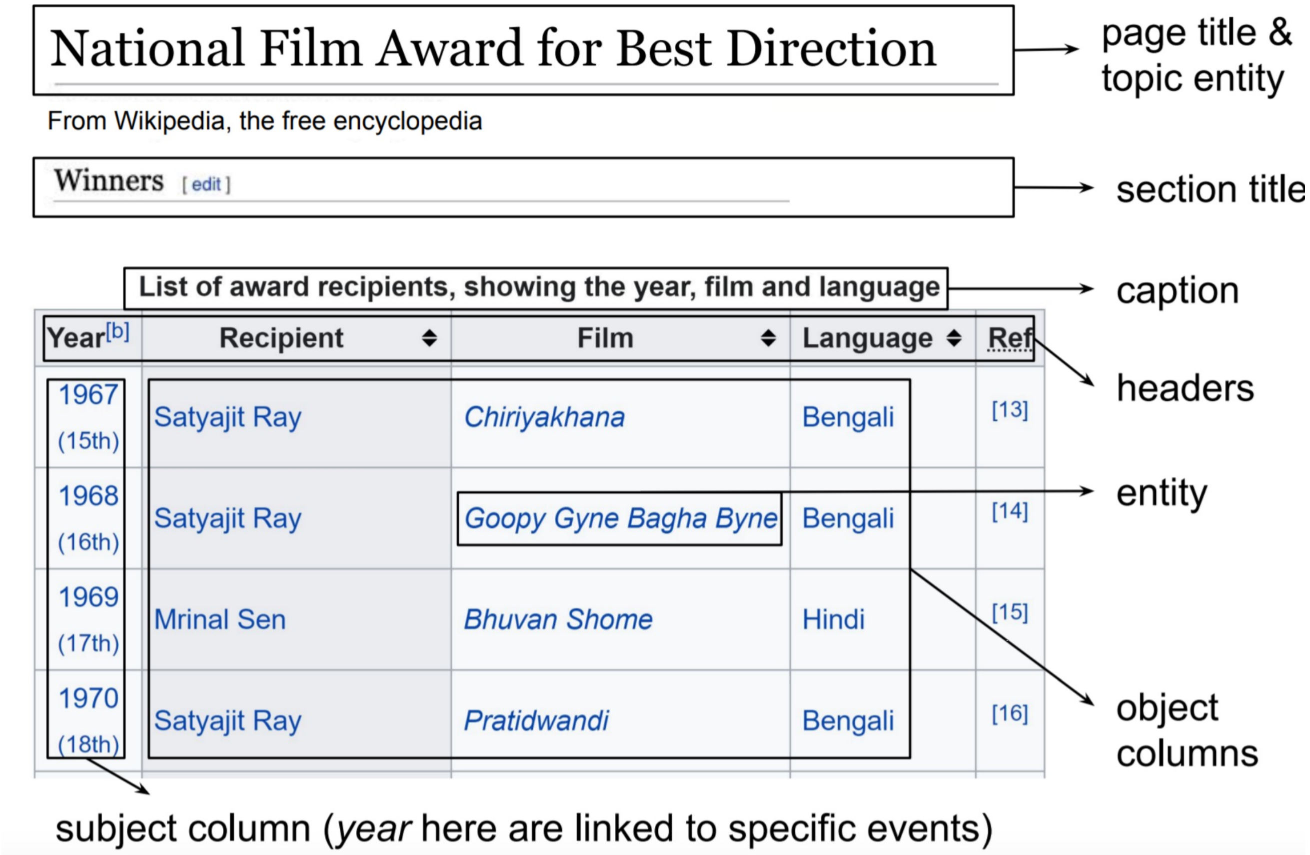
- Goal: (1) To learn deep *contextualized* representations of table elements;



[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

- Goal: (1) To learn deep *contextualized* representations of table elements;
  - To capture row-and-column structure
  - To capture factual knowledge in relational tables



[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

- Goal: (I) To learn deep *contextualized* representations of table elements;

  (II) Pre-training / Fine-tuning paradigm for table-based tasks to reduce feature engineering effort;

[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

- Goal: (I) To learn deep *contextualized* representations of table elements;

(II) Pre-training / Fine-tuning paradigm for table-based tasks to reduce feature engineering effort;

(III) New datasets for a series of table understanding tasks

[Deng et al., VLDB'21]

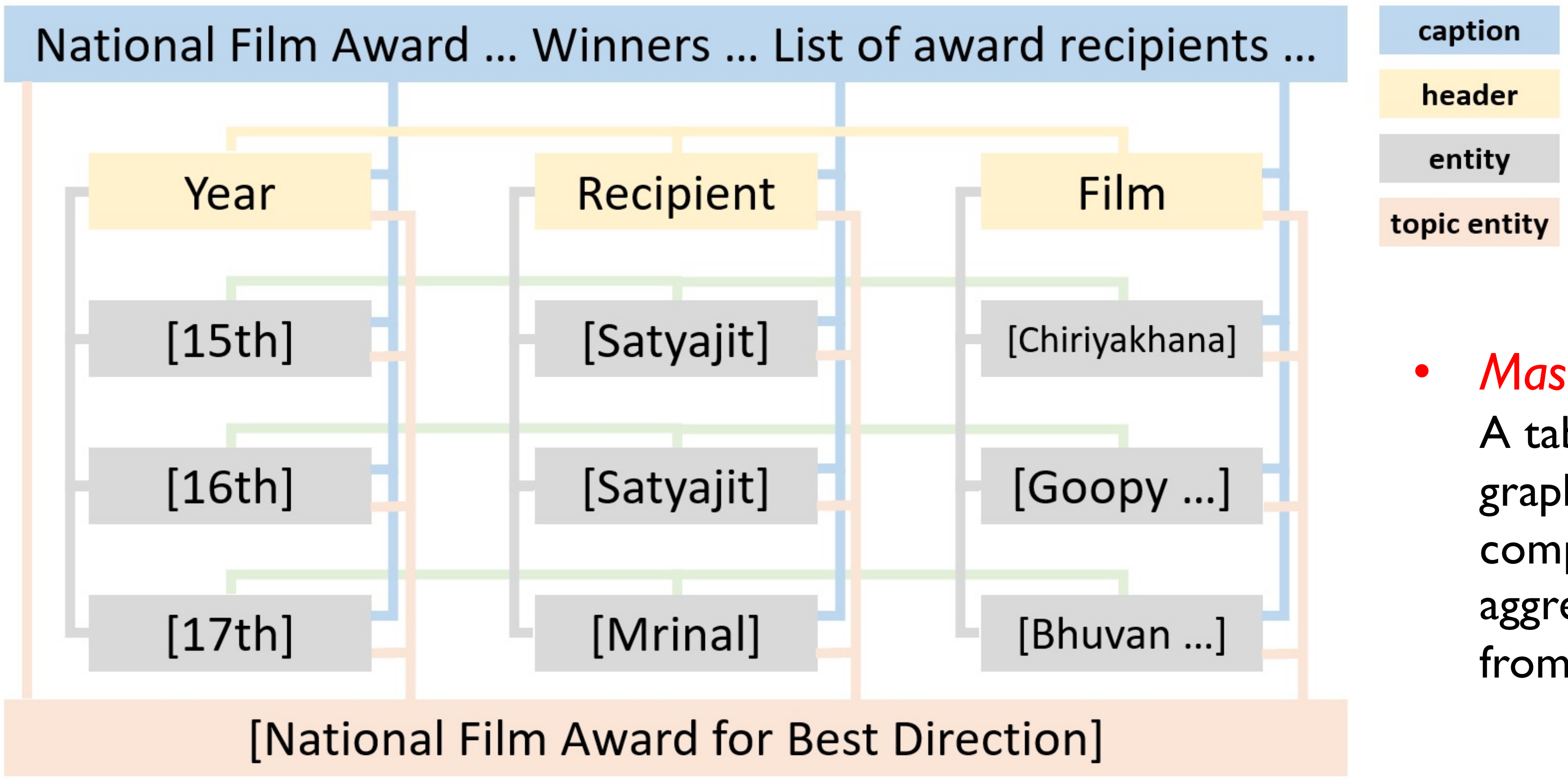
TURL: Table Understanding through Representation Learning

- Model: Transformer [Vaswani et al., NeurIPS 2017] with masked self-attention

[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

- Model: Transformer [Vaswani et al., NeurIPS 2017] with masked self-attention

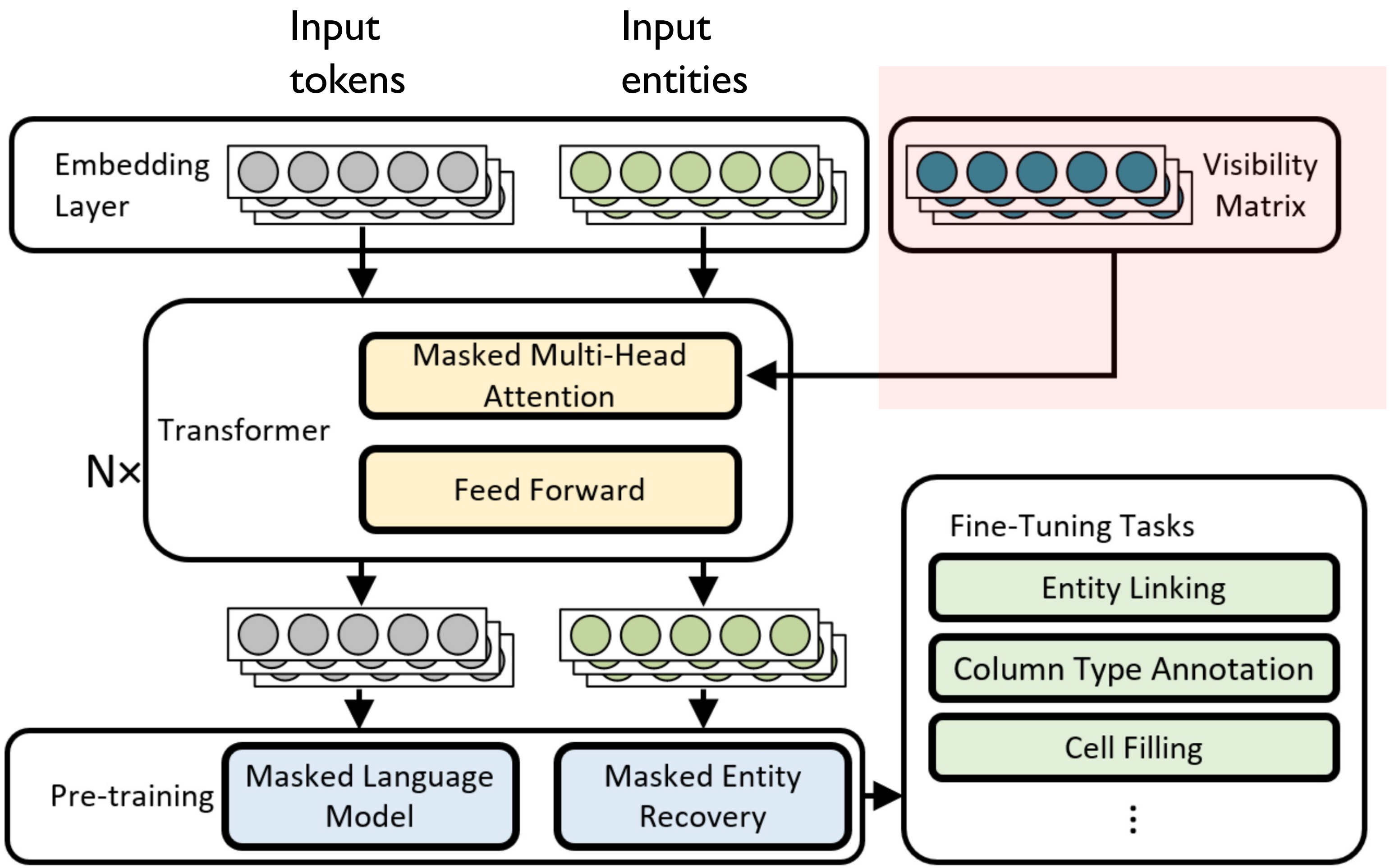


- *Masked self-attention:* A table is treated as a graph and each component can only aggregate information from its neighbors

[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

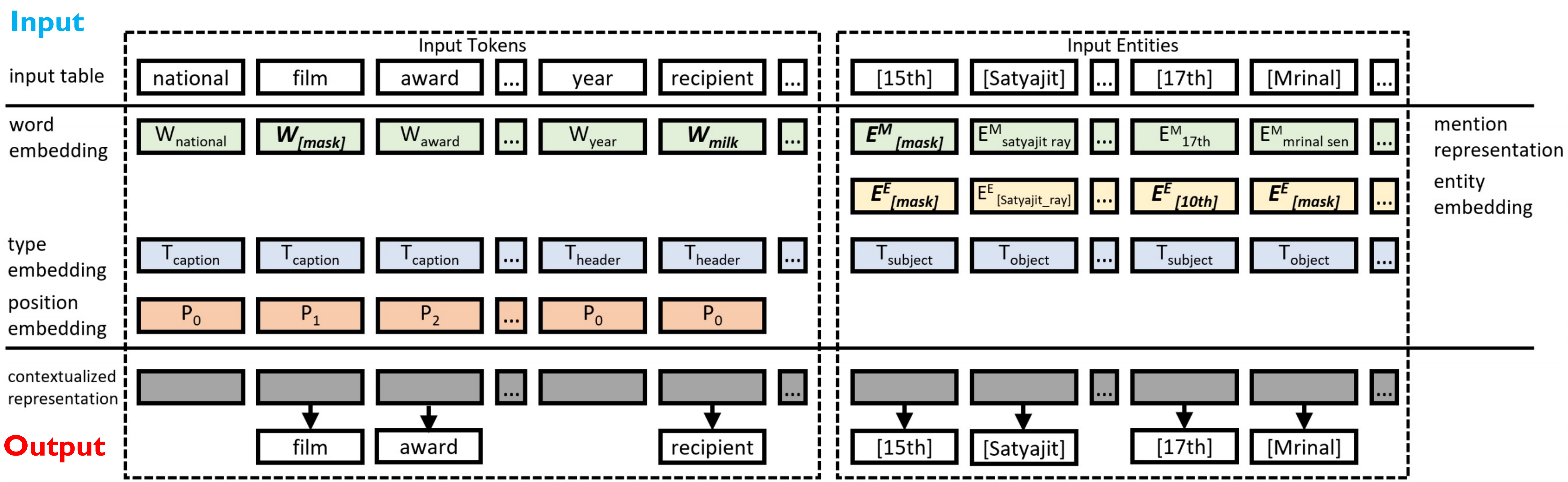
- Model: Transformer [Vaswani et al., NeurIPS 2017] with masked self-attention



[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

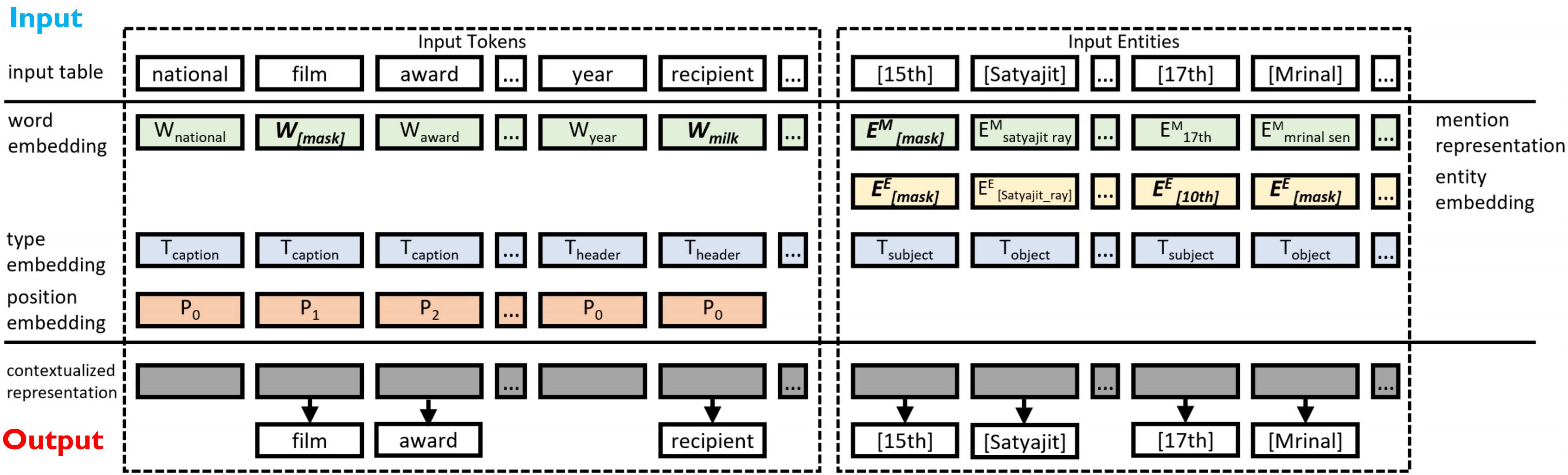
[Deng et al., VLDB'21]



- Type and position embeddings to mark different parts of a table and their relative positions
- Reuse pre-trained embeddings when possible (e.g., initialized with TinyBERT [Jiao et al., 2020])
- Each entity has a unique entity embedding and one mention embedding (obtained from its surface form in the table)

TURL: Table Understanding through Representation Learning

[Deng et al., VLDB'21]



- Pre-training objectives:
  - Masked Language Model (MLM): predict masked tokens in table metadata (i.e., table caption + headers)
  - Masked Entity Recovery (MEM): predict the entity in a masked table cell

TURL: Table Understanding through Representation Learning

[Deng et al., VLDB'21]

- Pre-training data
  - Entity focused relational tables from Wikipedia that contains factual knowledge
  - 570171 / 5036 / 4964 tables for pre-training / validation / testing

	split	min	mean	median	max
# row	train	1	13	8	4670
	dev	5	20	12	667
	test	5	21	12	3143
# ent. columns	train	1	2	2	20
	dev	3	4	3	15
	test	3	4	3	15
# ent.	train	3	19	9	3911
	dev	8	57	34	2132
	test	8	60	34	9215

Dataset statistics (per table) in pre-training

TURL: Table Understanding through Representation Learning

[Deng et al., VLDB'21]

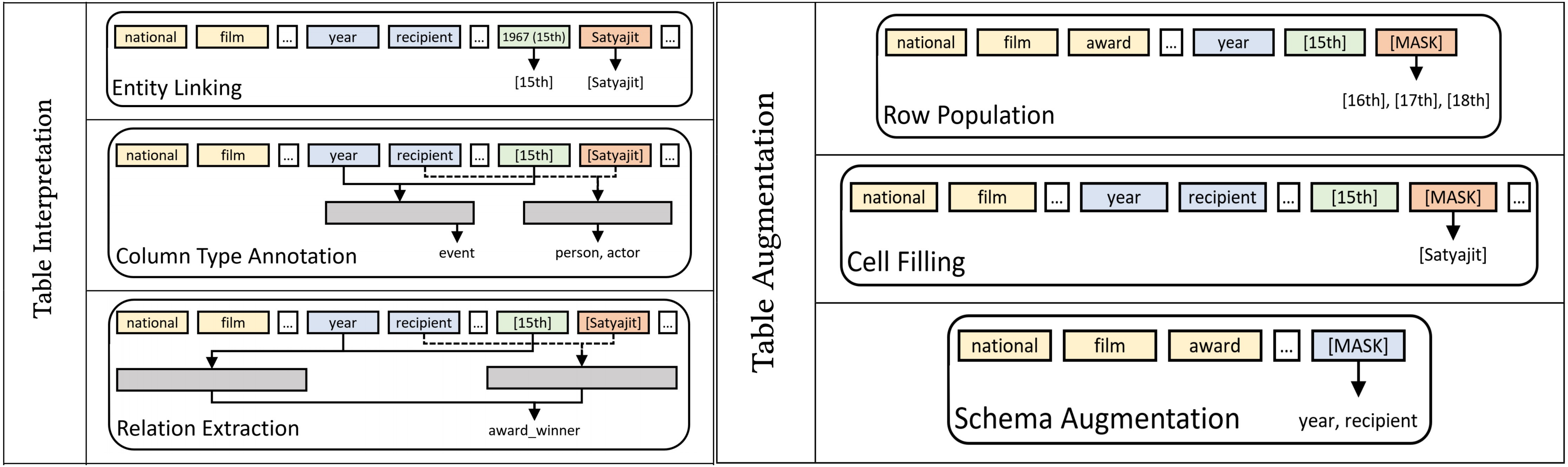
- Pre-training data
  - Entity focused relational tables from Wikipedia that contains factual knowledge
  - **570171** / 5036 / 4964 tables for pre-training / validation / testing

With TinyBERT model architecture, smaller model, smaller pre-training datasets, greener data science

TURL: Table Understanding through Representation Learning

- Fine-tuning strategy for 6 different tasks

[Deng et al., VLDB'21]



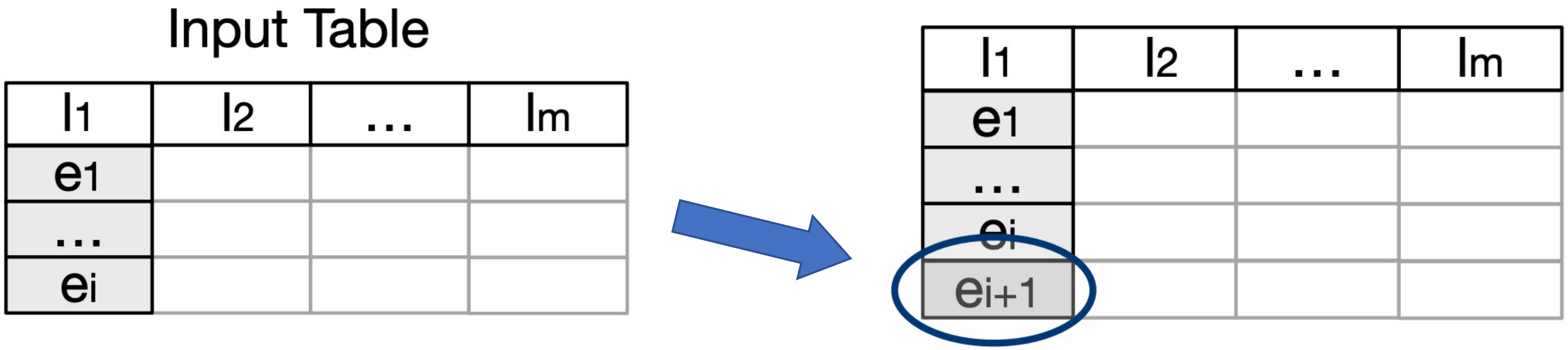
Pre-trained TURL can be applied to all 6 tasks with minimal modification.

TURL: Table Understanding through Representation Learning

[Deng et al., VLDB'21]

- Experiments
  - 6 tasks on publicly available datasets as well as new datasets created in the paper
  - For example, **row population**:

Definition: Given a partial table, and an optional set of seed subject entities, row population aims to retrieve more entities to fill the subject column



See more: https://iai-group.github.io/webtables-tutorial/slides/part-5.pdf

TURL: Table Understanding through Representation Learning

[Deng et al., VLDB'21]

- Experiments
  - 6 tasks on publicly available datasets as well as new datasets created in the paper
  - For example, **row population**:

# seed	0		1	
Method	MAP	Recall	MAP	Recall
EntiTables [Zhang et al., 2017]	17.90	63.30	42.31	78.13
Table2Vec [Deng et al., 2019]	-	63.30	20.86	78.13
TURL + fine-tuning	**40.92**	63.30	**48.31**	78.13

All methods share the same candidate generation module, hence the same recall.

TURL: Table Understanding through Representation Learning

[Deng et al., VLDB'21]

- Experiments
  - 6 tasks on publicly available datasets as well as new datasets created in the paper
  - For example, **entity linking**:

Definition: Given a table T and a knowledge base KB, entity linking aims to link each potential mention in cells of T to its referent entity $e \in$ KB.

List of award recipients, showing the year, film and language

Year[b]	Recipient	Film	Language
$\boxed{\text{1967 (15th)}}$	Satyajit Ray	*Chiriyakhana*	Bengali
1968 (16th)	Satyajit Ray	*Goopy Gyne Bagha Byne*	Bengali
1969 (17th)	Mrinal Sen	*Bhuvan Shome*	Hindi

https://en.wikipedia.org/wiki/15th_National_Film_Awards

TURL: Table Understanding through Representation Learning

- Experiments
  - 6 tasks on publicly available datasets as well as new datasets created in the paper
  - For example, **entity linking**:

Takeaways:

- Outperforms all baselines [Ritze et al., 2015, Efthymiou et al., 2017] on the challenging WikiGS [Efthymiou et al., 2017] and their new Test Set.

Method	WikiGS			Our Test Set			T2D		
	F1	P	R	F1	P	R	F1	P	R
T2K [35]	34	70	22	-	-	-	82	**90**	76
Hybrid II [16]	64	69	**60**	-	-	-	**83**	85	**81**
Wikidata Lookup	57	67	49	62	62	60	80	86	75
TURL + fine-tuning	**67**	**79**	58	**68**	**71**	**66**	78	83	73
w/o entity desc.	60	70	52	60	63	58	-	-	-
w/o entity type	66	78	57	67	70	65	-	-	-
+ reweighting	-	-	-	-	-	-	82	88	77
WikiLookup (Oracle)	74	88	64	79	82	76	90	96	84

[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

- Experiments
  - 6 tasks on publicly available datasets as well as new datasets created in the paper
  - For example, **entity linking**:

Takeaways:

- Outperforms all baselines [Ritze et al., 2015, Efthymiou et al., 2017] on the challenging WikiGS [Efthymiou et al., 2017] and their new Test Set.

- Generalizes well to general Web Tables (T2D [Lehmberg et al., 2016]). Improves Wikidata Lookup and obtain similar performance as state-of-the-art models.

Method	WikiGS			Our Test Set			T2D		
	F1	P	R	F1	P	R	F1	P	R
T2K [35]	34	70	22	-	-	-	82	**90**	76
Hybrid II [16]	64	69	**60**	-	-	-	**83**	85	**81**
Wikidata Lookup	57	67	49	62	62	60	80	86	75
TURL + fine-tuning	**67**	**79**	58	**68**	**71**	**66**	78	83	73
w/o entity desc.	60	70	52	60	63	58	-	-	-
w/o entity type	66	78	57	67	70	65	-	-	-
+ reweighting	-	-	-	-	-	-	82	88	77
WikiLookup (Oracle)	74	88	64	79	82	76	90	96	84

[Deng et al., VLDB'21]

TURL: Table Understanding through Representation Learning

- Summary:

(I) Learning deep *contextualized* representations of table elements

  - With a focus on relational tables (to be extended)
  - Modeling factual knowledge about named entities (Masked Entity Recovery)
  - Masked self-attention mechanism to model table structure

(II) Pre-training / fine-tuning paradigm works well for table-based tasks and greatly reduce feature engineering effort

(III) A benchmark (new datasets for 6 tasks) that is important for model development and evaluation

[Deng et al., VLDB'21]

Outline: Neural Representation Learning on Tables

- Background

- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- Summary

- Resources

TABBIE: Pretrained Representations of Tabular Data

- Idea: [Iida et al., NAACL'21]
  - Using the pretraining objective Corrupt Cell Detection [Clark et al., 2020]

step 1: corrupt 15% of cells

Size	Medals
France	3.6
Italy	5
Spain	4

TABBIE

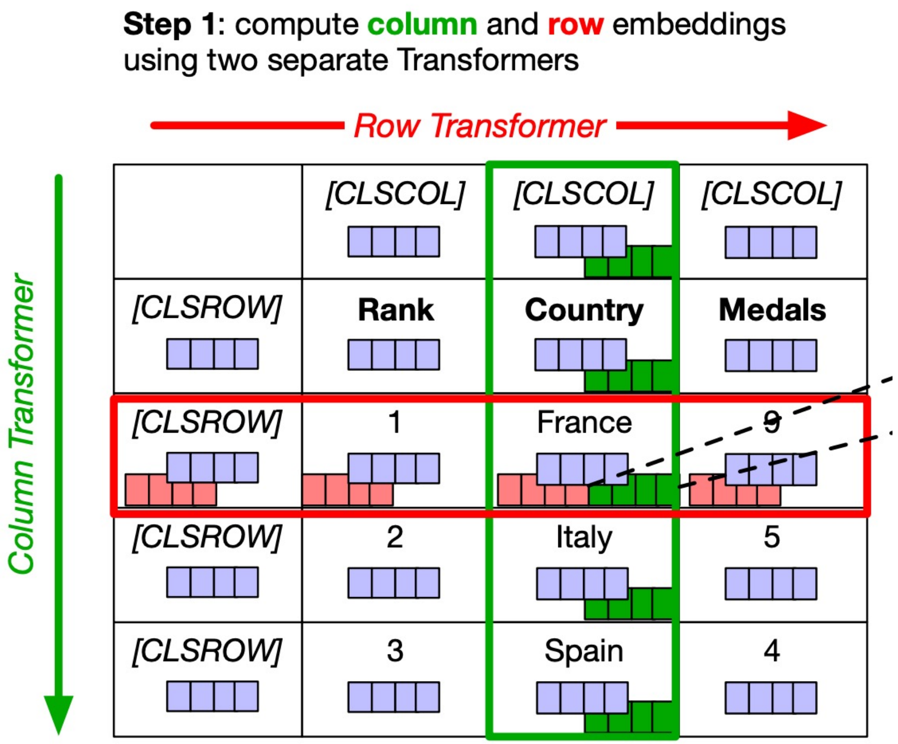
step 2: embed the table with **TABBIE**

step 3: train **TABBIE** to identify the corrupted cells

corrupt!	*real*
real	*corrupt!*
real	*real*
real	*real*

TABBIE: Pretrained Representations of Tabular Data

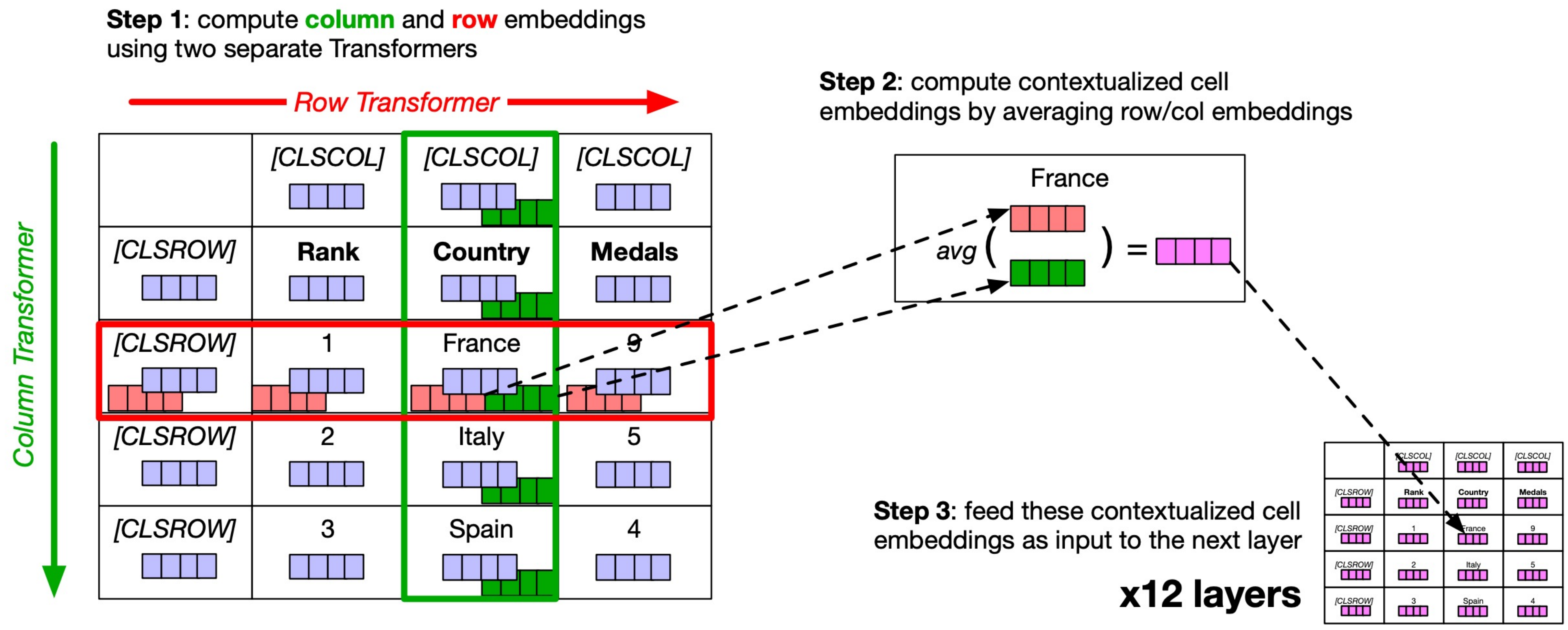
- Model: [Iida et al., NAACL'21]
  - At one layer:
    - Two transformers acting row-wise and column-wise respectively



TABBIE: Pretrained Representations of Tabular Data

[Iida et al., NAACL'21]

- Model:
  - At one layer:
    - Two transformers acting row-wise and column-wise respectively
    - Average row and column result to create cell representation and pass to next layer



TABBIE: Pretrained Representations of Tabular Data

[Iida et al., NAACL'21]

- Tasks
  - **Row population**
  - Column population
  - Cell type annotation

N	Method	MAP	MRR	Ndcg-10	Ndcg-20
1	Entitables	36.8	45.2	-	-
	TaBERT	**43.2**	**55.7**	**45.6**	**47.7**
	TABBIE (FREQ)	42.8	54.2	44.8	46.9
	TABBIE (MIX)	42.6	54.7	45.1	46.8
2	Entitables	37.2	45.1	-	-
	TaBERT	43.8	56.0	46.4	48.8
	TABBIE (FREQ)	**44.4**	**57.2**	**47.1**	**49.5**
	TABBIE (MIX)	43.7	55.7	46.2	48.6
3	Entitables	37.1	44.6	-	-
	TaBERT	42.9	55.1	45.6	48.5
	TABBIE (FREQ)	**43.4**	**56.5**	**46.6**	**49.0**
	TABBIE (MIX)	42.9	55.5	45.9	48.3

Row population. N: number of seed rows

Outline: Neural Representation Learning on Tables

- Background

- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- Summary

- Resources

TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD'21]

- Goal
  - To encode generally structured tables

- Tables of varying structure

Island	Nickname	Area	Population (as of 2010)
Hawaiʻi[29]	The Big Island	4,028.0 sq mi (10,432.5 km^2)	185,079
Maui[30]	The Valley Isle	727.2 sq mi (1,883.4 km^2)	144,444
Oʻahu[31]	The Gathering Place	596.7 sq mi (1,545.4 km^2)	953,207

(a) A vertical relational web table in Wikipedia

Kobe Bryant	
Height	6 ft 6 in
Weight	212 lb
Position	Shooting guard
Number	8, 24
Born	August 23, 1978 Pennsylvania
Died	January 26, 2020 California

(b) A horizontal entity web table

Method	WikiGS			Our Test Set		
	F1	P	R	F1	P	R
T2K [35]	34	70	22	-	-	-
Hybrid II [16]	64	69	**60**	-	-	-
Wikidata Lookup	57	67	49	62	62	60
TURL + fine-tuning	**67**	**79**	58	**68**	**71**	**66**
w/o entity desc.	60	70	52	60	63	58
w/o entity type	66	78	57	67	70	65
+ reweighting	-	-	-	-	-	-
WikiLookup (Oracle)	74	88	64	79	82	76

(c) A matrix PDF table in arXiv

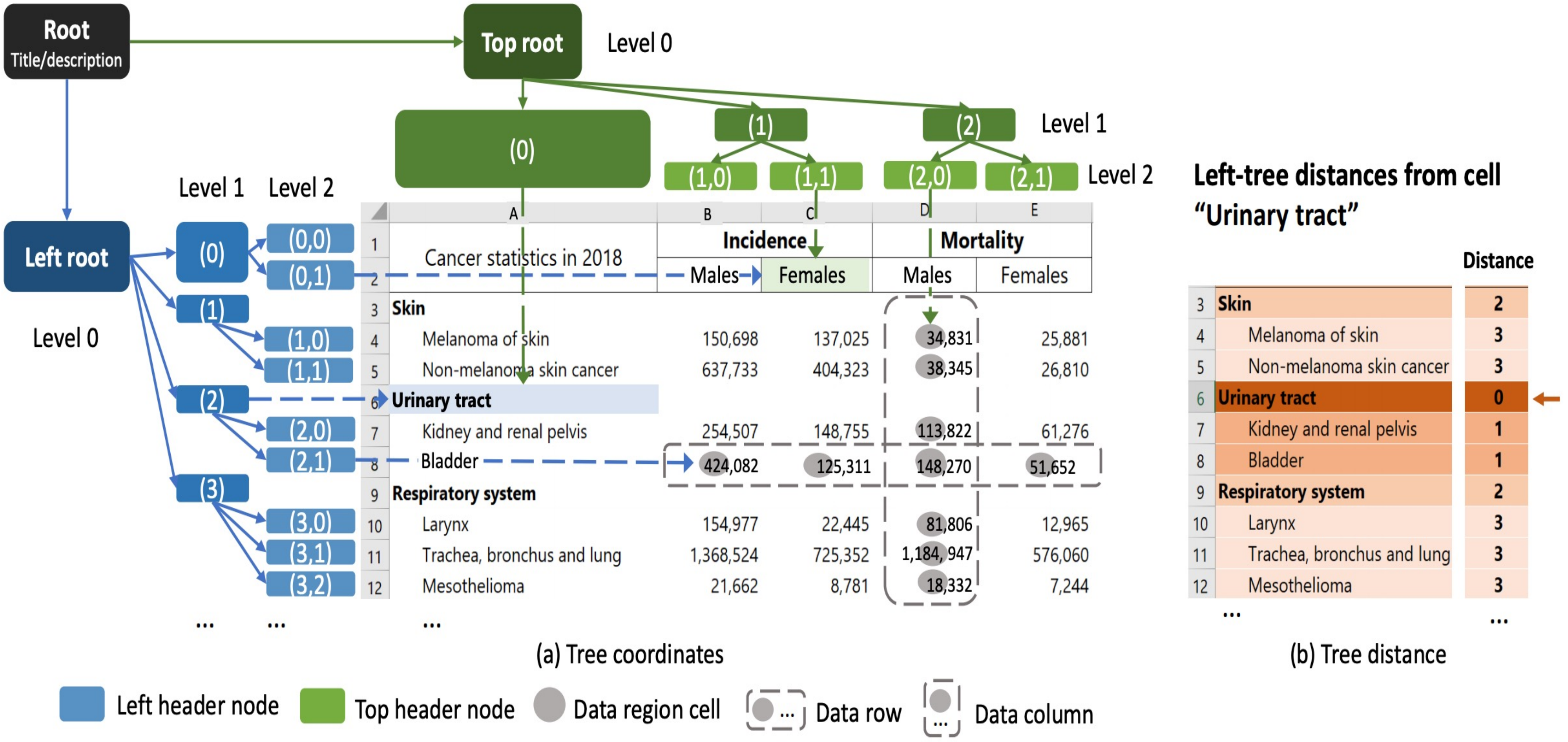
	A	B	C	D	E
1	Cancer statistics in 2018	Incidence		Mortality	
2		Males	Females	Males	Females
3	Skin				
4	Melanoma of skin	150,698	137,025	34,831	25,881
5	Non-melanoma skin cancer	637,733	404,323	38,345	26,810
6	Urinary tract				
7	Kidney and renal pelvis	254,507	148,755	113,822	61,276
8	Bladder	424,082	125,311	148,270	51,652
9	Respiratory system				
10	Larynx	154,977	22,445	81,806	12,965
11	Trachea, bronchus and lung	1,368,524	725,352	1,184,947	576,060
12	Mesothelioma	21,662	8,781	18,332	7,244

(d) A matrix spreadsheet table

TUTA: Tree-based Transformers for Generally Structured Table Pre-training

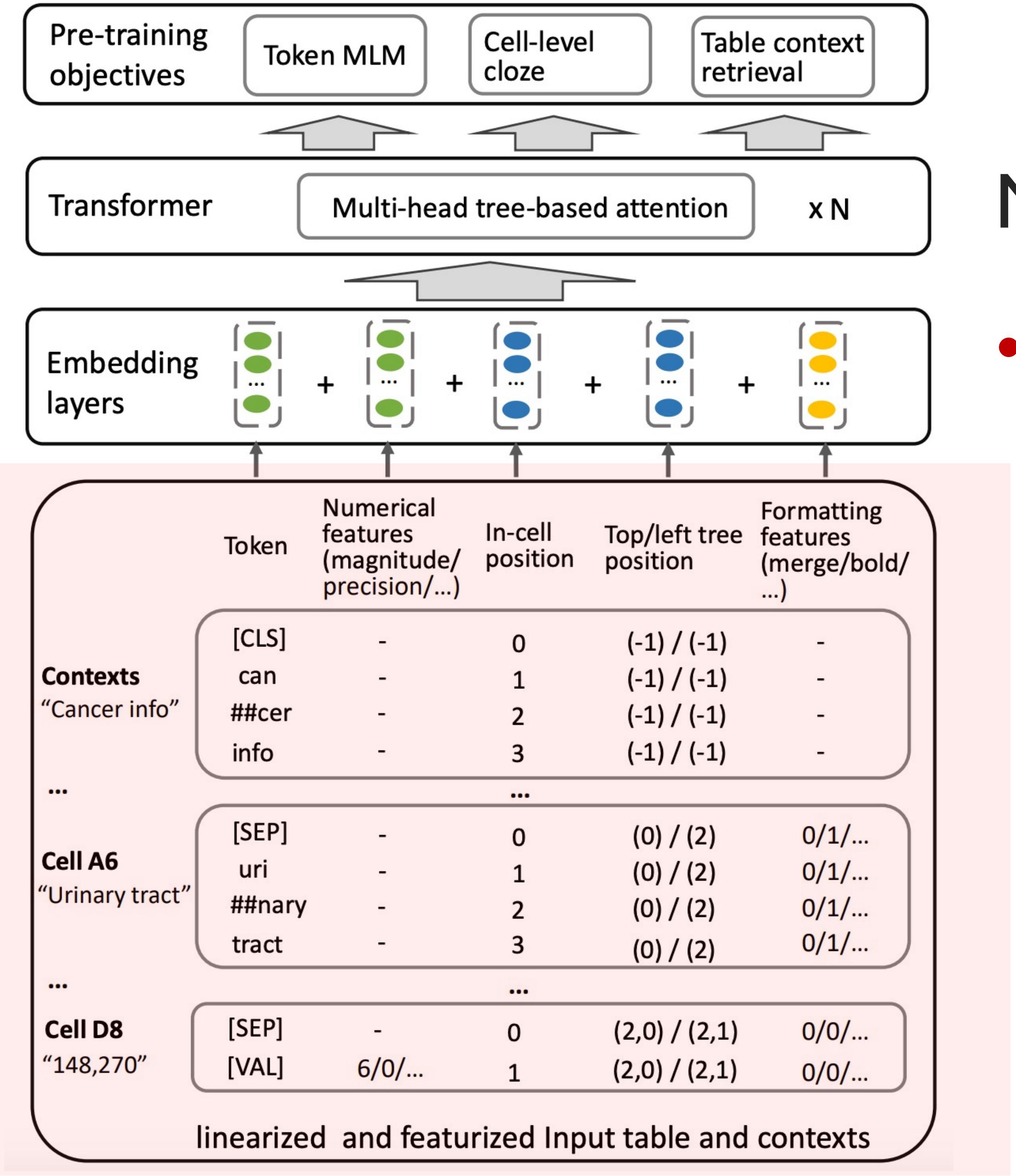
[Wang et al., SIGKDD'21]

- Bi-dimensional coordinate tree of a table



TUTA: Tree-based Transformers for Generally Structured Table Pre-training

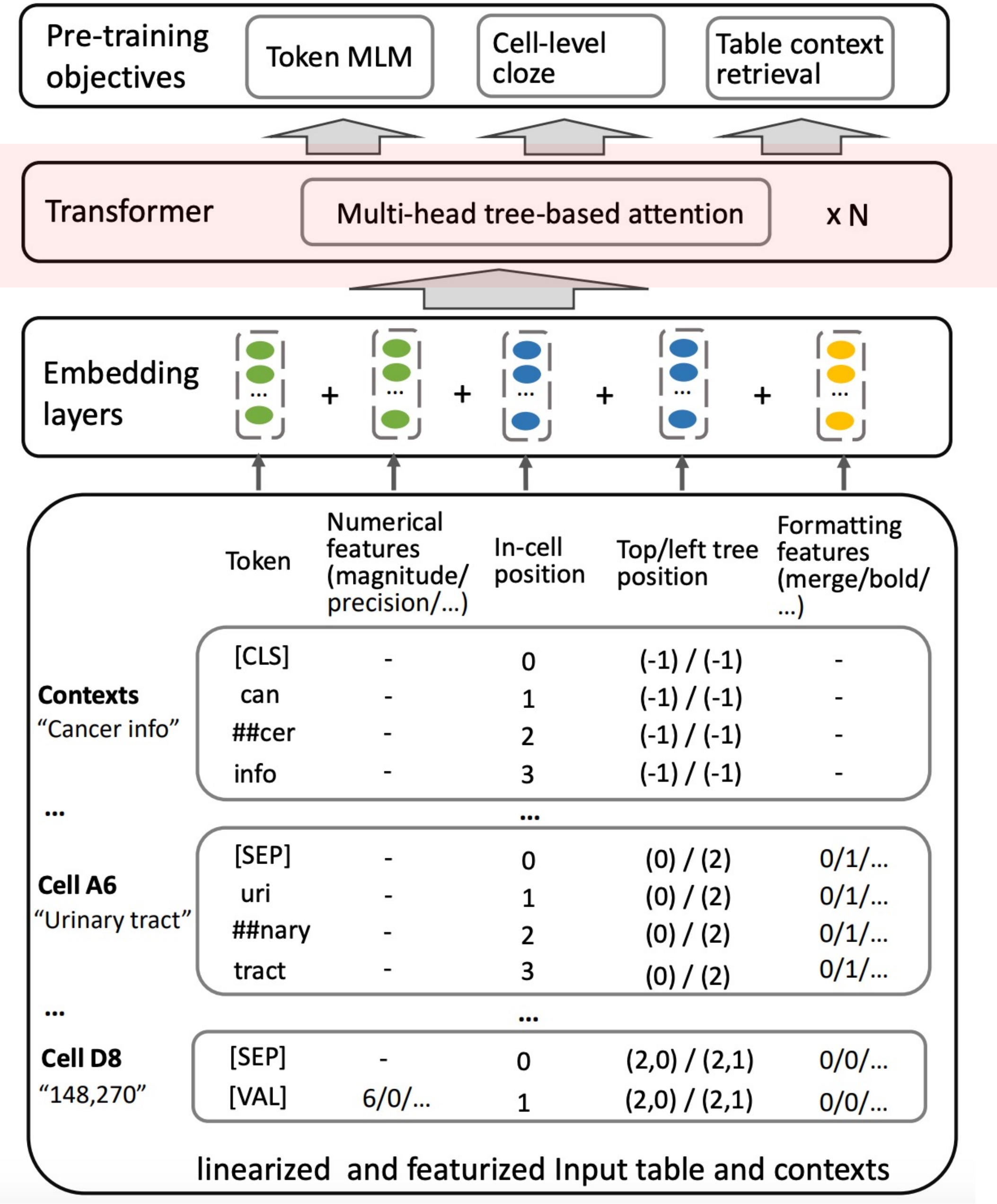
[Wang et al., SIGKDD'21]



Model highlights:

- Model Input
  - Numerical features
  - Top/left tree position
  - Formatting features

TUTA: Tree-based Transformers for Generally Structured Table Pre-training



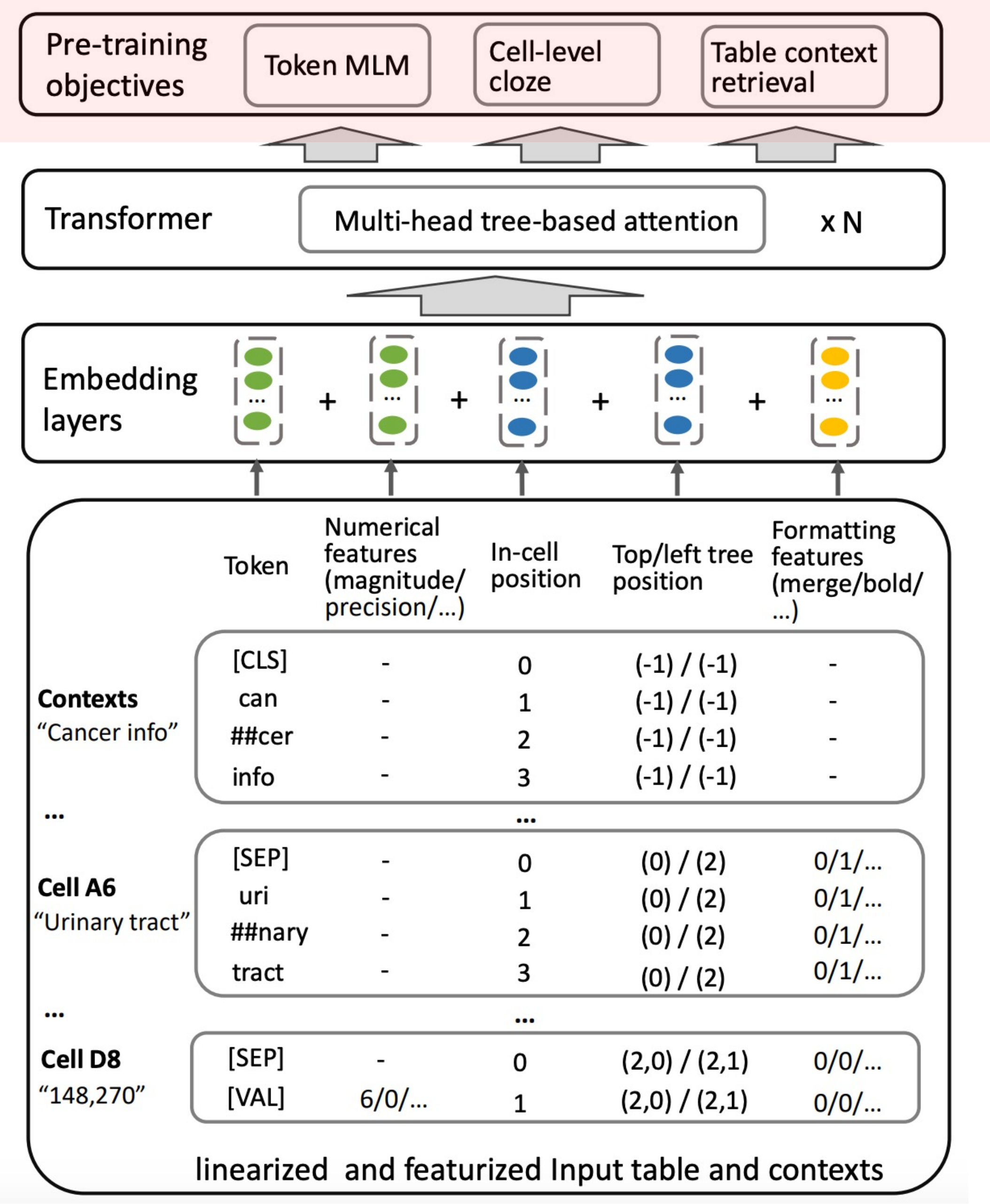
[Wang et al., SIGKDD'21]

Model highlights:

- Model Input
  - Numerical features
  - Top/left tree position
  - Formatting features
- Tree-based attention
  - Only attend to structural neighborhood, defined base on tree distance

TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD'21]



Model highlights:

- Model Input
  - Numerical features
  - Top/left tree position
  - Formatting features
- Tree-based attention
  - Only attend to structural neighborhood, defined base on tree distance

TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD'21]

- Three Pre-training Objectives

Cancer incidence and mortality statistics worldwide in 2018.

	A	B	C	D	E
1	Cancer statistics in 2018	Incidence		(A)	
2		Males	Females	Males	Females
3	**Skin**				
4	Melanoma of skin	150,698	137,025	34,831	25,881
5	Non-melanoma skin cancer	637,733	404,323	38,345	26,810
6	(B)				
7	(D)	254,507	148,755	113,822	61,276
8	Bladder	424,082	125,311	148,270	51,652
9	**Respiratory system**				
10	Larynx	154,977	22,445	81,806	12,965
11	Trachea, bronchus and lung	1,368,524	(F)	1,184,947	576,060
12	(C)	21,662	8,781	18,332	7,244
13	**Digestive organs**				
14	(E)	1,026,215	823,303	484,224	396,568
15	Oesophagus	399,699	172,335	357,190	151,395
22	Pancreas	243,033	215,885	226,910	205,332

Masked language modeling: (I)

Predict randomly masked tokens from the whole vocabulary

Cell-level cloze: (II)

Fill the randomly masked cell locations from cell strings:

A. Mortality
B. Urinary tract
C. Mesothelioma
D. Kidney and renal pelvis
E. Colorectum and anus
F. 725,352

Table context retrieval: (III)

Retrieve table titles and descriptions from randomly selected text snippets:

- ✗ Family Households by Size, Type, and Age of Householder.
- ✓ It includes major cancer types including skin, respiratory, etc...
- ✗ Full-time Law Enforcement Officers.

...

TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD'21]

- Pre-training Objectives
  - Masked Language Model (MLM)
  - Cell-level cloze: Map cell location to cell string
  - Table context retrieval: Match cell representation (masked) to positive and negative candidate segments.

- Tasks (Table Structure Understanding)
  - Cell type classification (metadata, notes, data, top attribute, etc.)
  - Table type classification (relational, list, entity, etc.)

	A	B	C	D	E	F	G	H	I	J
1	Table 15. Family Households by Size, Type, and Age :2011									
2	(Numbers in thousands. Civilian noninstitutionalized population)[1]									
3										
4	Age of householder	Household size								
5		Two people		Three people		Four people		Five or more peop		Total
6		Numb	Percen	Numb	Percen	Numb	Perce	Numb	Perce	Numb
7	15 to 19 years	20	34.5%	28	48.3%	7	12.1%	3	5.2%	58
8	20 to 24 years	436	42.7%	284	15.3%	188	18.4%	112	11.0%	1,020
9	25 to 29 years	1,064	31.8%	944	16.1%	818	24.4%	522	15.6%	3,348
10	30 to 34 years	1,029	19.8%	1,262	14.3%	1,604	30.8%	1,308	25.1%	5,203
11	34 to 39 years	873	15.0%	1,139	11.9%	2,087	35.9%	1,722	29.6%	5,821
12	40 years and ove	22,145	52.0%	7,725	9.9%	7,412	17.4%	5,305	12.5%	42,587
13										
14	[1] Plus armed forces living off post or with their families on post.									

Metadata | Top attribute | Left attribute | Data | Derived | Notes

Index | Index name | Value name

Outline: Neural Representation Learning on Tables

- Background

- Representative Methods
  - Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  - TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB'21; OSU & Google]
  - TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL'21; Sony Co. & Adobe Research & UMass Amherst]
  - TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD'21; MSR & CMU & PKU]

- **Summary**

- **Resources**

Summary

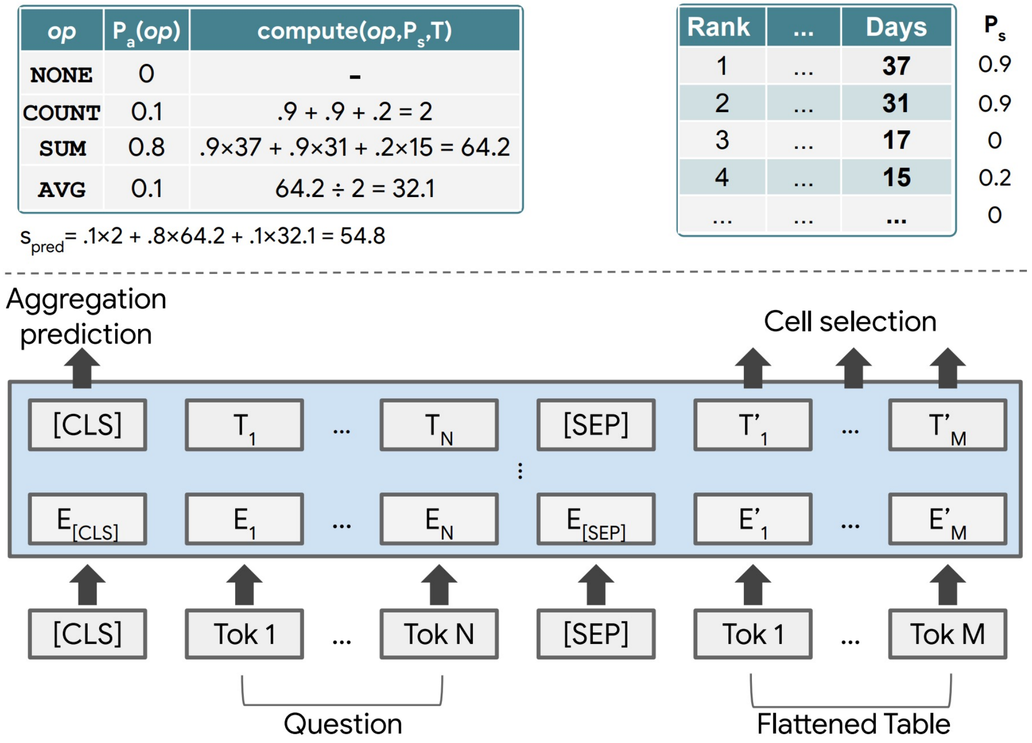
Designs		Table2Vec	TURL	TUTA	TABBIE
Backbone Alg./Model		Word2Vec	Transformer with masked self-attention	Transformer with tree-based attention	Transformer (row-wise, column-wise)
Pretraining Objectives	Masked Language Model (as in BERT)		Yes	Yes	
	Cell-level Cloze			Yes	
	Table Context Retrieval			Yes	
	Masked Entity Recovery		Yes		
	Corrupt Cell Detection				Yes
	Word2Vec	Yes			
Pretraining Corpus		WikiTable: 1.6M tables	WikiTable: **570K** tables	WikiTable + WDC WebTable + web-crawled spreadsheet: **57.9M** tables	WikiTable + Common Crawl : 26.6M tables

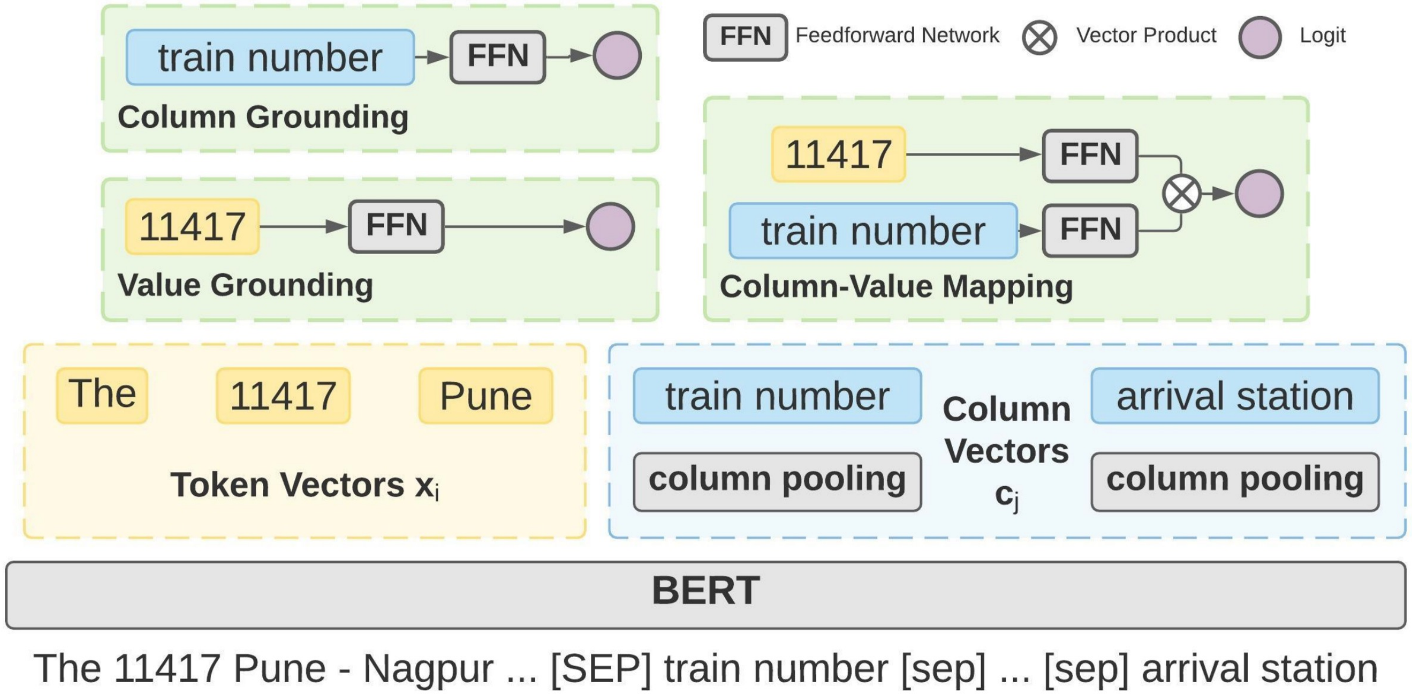
Summary

Designs		Table2Vec	TURL	TUTA	TABBIE
Downstream Tasks	Entity Linking		Yes		
	Column Type Annotation		Yes		Yes
	Relation Extraction		Yes		
	Row Population	Yes	Yes		Yes
	Cell Filling		Yes		
	Schema Augmentation		Yes		
	Column Population	Yes			Yes
	Table Type Classification			Yes	
	Cell Type Classification			Yes	
	Corrupt Cell Detection				Yes
Fine-tuning		No	Yes	Yes	Yes

Summary

- Methods for Joint Representation Learning of Text and Tables





1. TAPAS: Weakly Supervised Table Parsing via Pre-training [Herzig et al., 2020]
2. StruG: Structure-Grounded Pretraining for Text-to-SQL [Deng et al., 2021]
3. TABERT: Pretraining for Joint Understanding of Textual and Tabular Data [Yin et al., 2020]
4. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing [Yu et al., 2021]

Summary

- Future directions
  - Systematic and fair comparison among different methods
    - For different types of tables
    - For different tasks
    - Easy-to-use benchmark
    - Considering pre-training data size, model size, training/inference time, etc.
    - Requires open-source code and datasets
  - Better model architecture & pre-training objectives overall
  - Joint representation of text, tables, and knowledge bases

Resources

- Table2Vec: https://github.com/iai-group/sigir2019-table2vec
- TURL: https://github.com/sunlab-osu/TURL
  (including pre-processed pre-training data, pre-trained model, new datasets for downstream tasks)
- TABBIE: https://github.com/SFIG611/tabbie
- TUTA: https://github.com/microsoft/TUTA_table_understanding/ (to be released after internal review)

References

- Lei Min Deng, Shuo Zhang, and Krisztian Balog. 2019. Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval. In SIGIR'19.
- Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. In VLDB'21.
- Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. Structure-grounded pretraining for text-to-sql. In NAACL'21.
- Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. In International Semantic Web Conference, 2017.
- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In ACL'20.
- Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. TABBIE: Pretrained Representations of Tabular Data. In NAACL'21.
- Dominique Ritze, Oliver Lehmberg, and Christian Bizer. 2015. Matching HTML Tables to DBpedia. In WIMS '15.
- Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In ACL'20.
- Zhiruo, Wang Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. Structure-aware pre-training for table understanding with tree-based transformers. In SIGKDD'21.
- Tao Yu, Wu Chien-Sheng, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. In ICLR'21.
- Shuo Zhang and Krisztian Balog. Entitables: Smart assistance for entityfocused tables. In SIGIR'17.

Thank you!

More Experiments from TURL

[Deng et al., VLDB'21]

- **Column Type and Relation Extraction**
  - Classification based on column representation.
  - Takeaways:
  - Outperforms Sherlock [Hulsebos et al. 2019] and BERT-based RE model.
  - Handle fine-grained types well (e.g., actor, citytown).

Column Type Annotation (Overall)

Method	F1	P	R
Sherlock (only entity mention)	78.47	88.40	70.55
TURL + fine-tuning (only entity mention)	88.86	90.54	87.23
TURL + fine-tuning	**94.75**	**94.95**	**94.56**
w/o table metadata	93.77	94.80	92.76
w/o learned embedding	92.69	92.75	92.63
only table metadata	90.24	89.91	90.58
only learned embedding	93.33	94.72	91.97

Relation Extraction

Method	F1	P	R
BERT-based	90.94	91.18	90.69
TURL + fine-tuning (only table metadata)	92.13	91.17	93.12
TURL + fine-tuning	**94.91**	**94.57**	**95.25**
w/o table metadata	93.85	93.78	93.91
w/o learned embedding	93.35	92.90	93.80

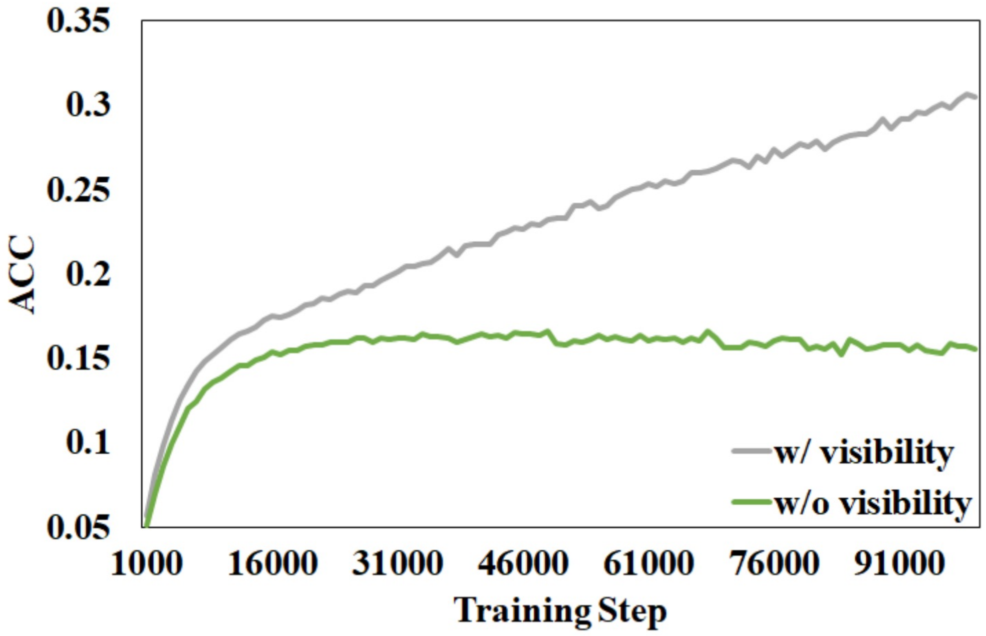
Column Type Annotation (Detailed)

Method	person	pro_athlete	actor	location	citytown
Sherlock	96.85	74.39	29.07	91.22	55.72
TURL + fine-tuning	**99.71**	**91.14**	**74.85**	**99.32**	**79.72**
only entity mention	98.44	87.11	58.86	96.59	60.13
w/o table metadata	99.63	90.38	74.46	99.01	77.37
w/o learned embedding	99.38	90.56	71.39	98.91	75.55
only table metadata	98.26	88.80	70.86	98.11	72.54
only learned embedding	98.72	91.06	73.62	97.78	75.16

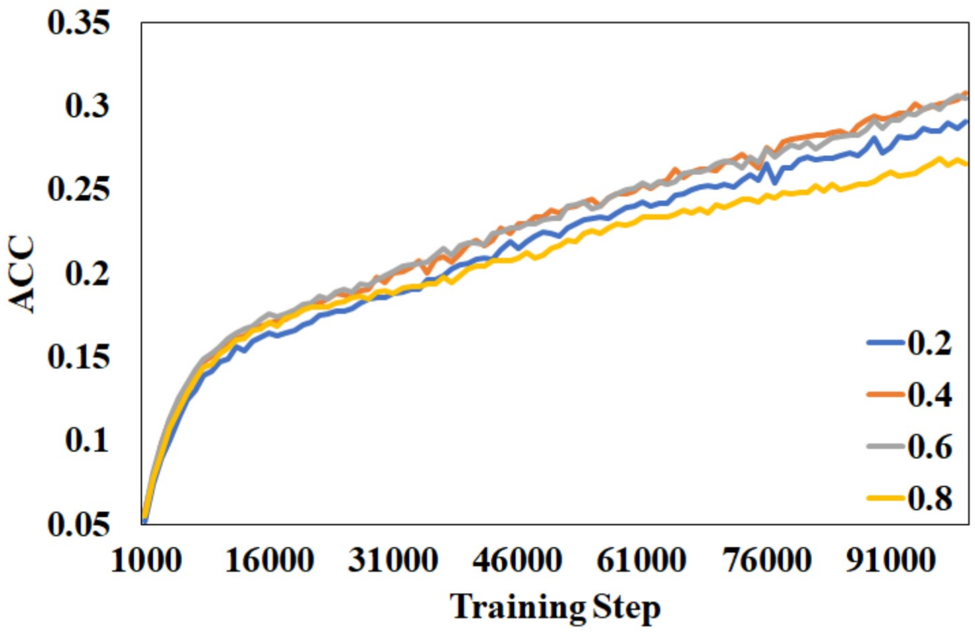
More Experiments from TURL

[Deng et al., VLDB'21]

- **Ablation study**
  - Masked self-attention (visibility matrix) is important for modeling table structure.
  - MER mask ratio can affect the model performance.



(a) Effect of visibility matrix.



(b) Effect of different MER mask ratios.

Top-1 object entity prediction accuracy on validation set