universität freiburg

TABERT: Pretraining for Joint Understanding of Textual and Tabular Data

Pengcheng Yin
Wen-tau Yih
Graham Neubig
Sebastian Riedel

Giacomo Mossio

18/07/2023

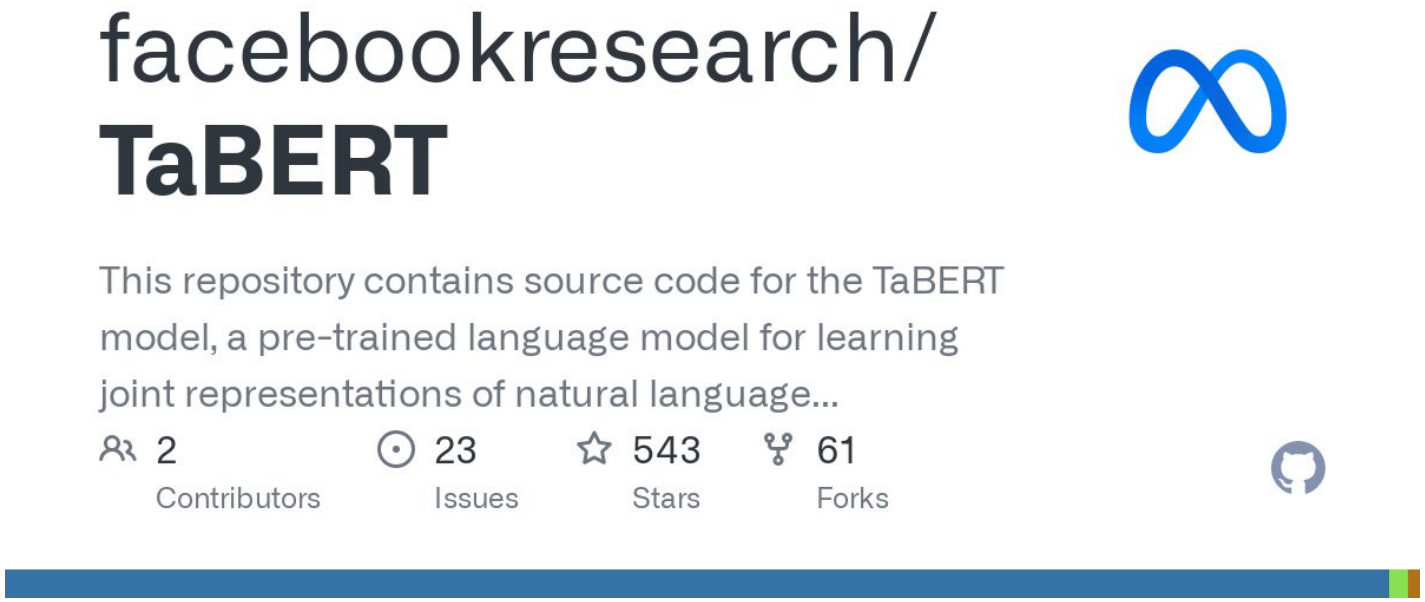
Motivation

- **Pretrained Large Language Models (LLMs) are achieving impressive results**
- **LLMs like BERT (May 2019), GPT-3 (May 2020) and LLaMA (February 2023) are free-text based**

What about Language Models for tabular data?

Introducing TABERT

- **Published in May 2020 by Facebook AI Research in collaboration with Carnegie Mellon University**
- **Pretrained LM that jointly learns representations for NL sentences and (semi-)structured tables**
- **Is used as an encoder in feature representation layers**
- **Is built on top of BERT**
- **Achieved SOTA results in 2020 for tasks involving tabular data**



Agenda

1. Background: BERT

2. Goal: Semantic parsing for databases

3. Architecture: Overview and Architecture

4. Pretraining : How has the model been trained

5. Applications: What it can be used for

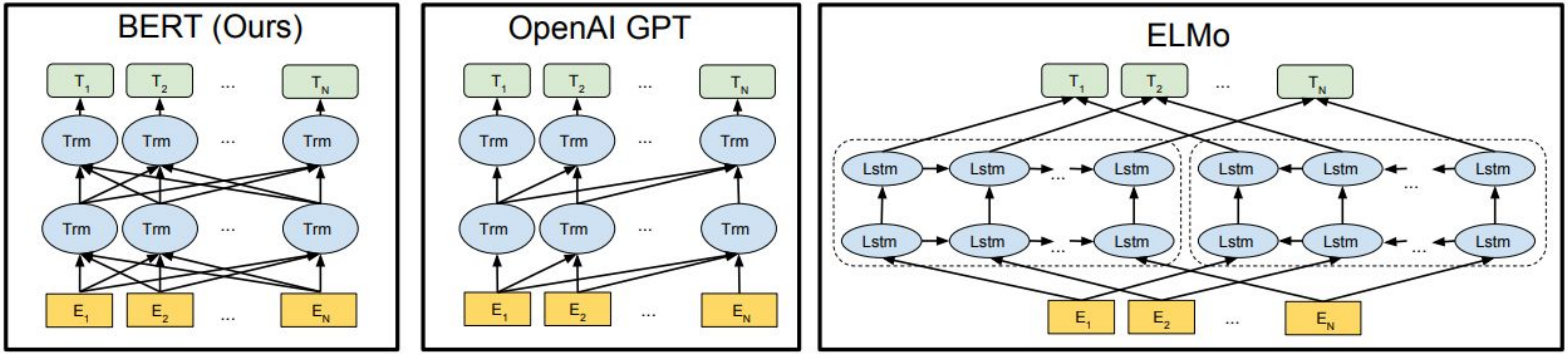
6. Results: Results from the experiments

7. Conclusions: Limitations and future directions

Background

BERT

- Bidirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) is a LLM
- Can be used for Sentence Classification, Named Entity Recognition, Question Answering, Text Generation, Text Summarization, Text Similarity, Semantic Search and other tasks

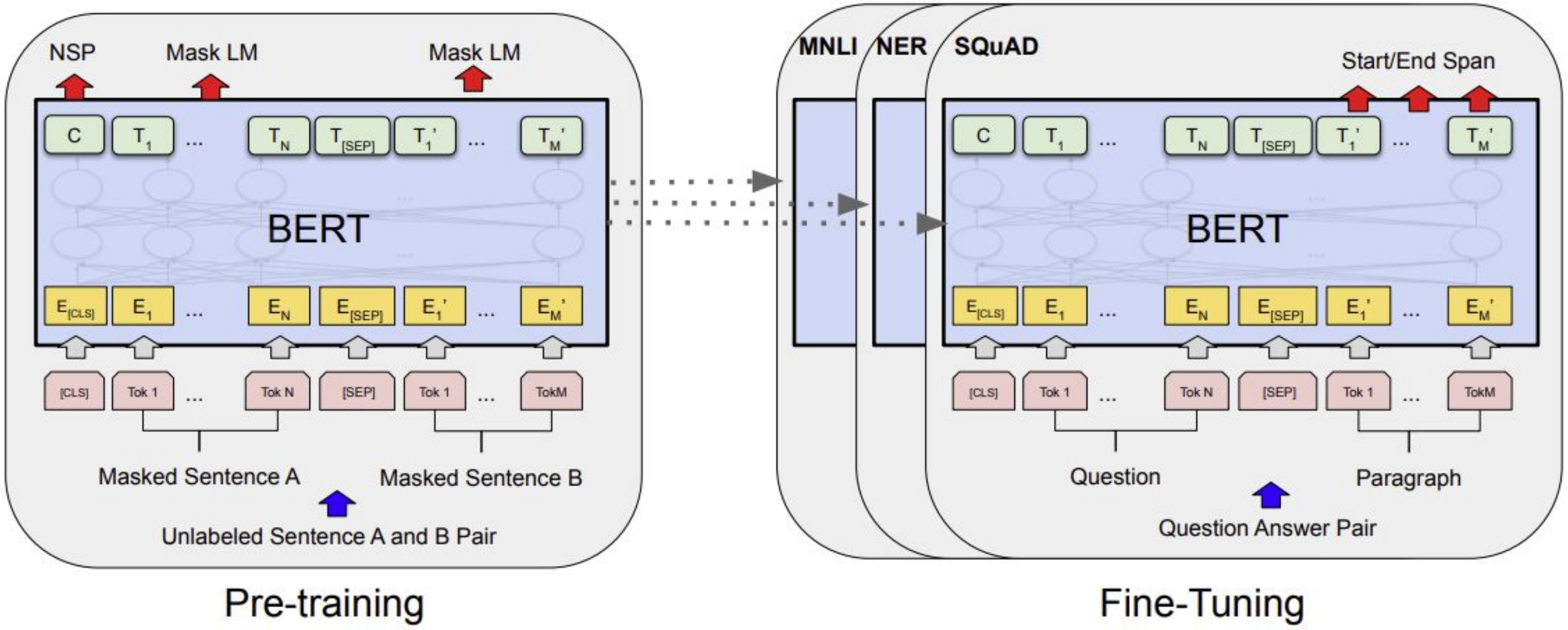


[Source: Devlin et al. 2019]

Background

BERT Training

- Is trained on large-scale text corpora using Masked Language Modeling (MLM) and Next Sentence Prediction objectives
- It learns contextual representations of words and sentences, capturing both syntax and semantics



[Source: Devlin et al. 2019]

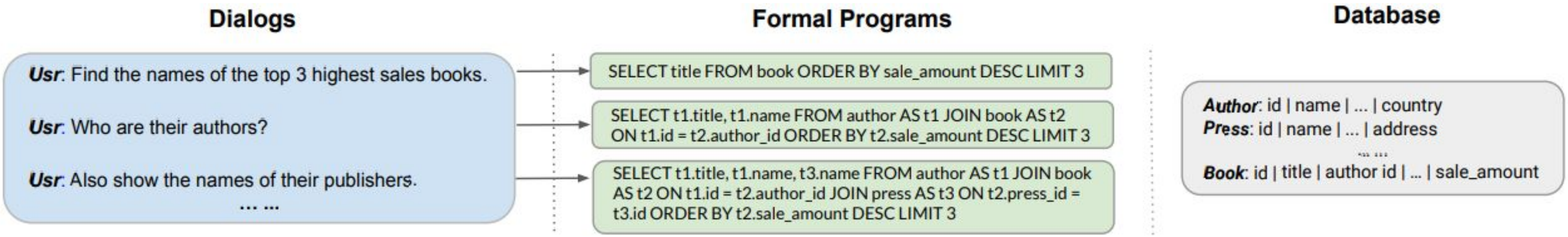
Agenda

1. Background: BERT
2. **Goal: Semantic parsing for databases**
3. Architecture: Overview and Architecture
4. Pretraining : How has the model been trained
5. Applications: What it can be used for
6. Results: Results from the experiments
7. Conclusions: Limitations and future directions

Goal

Semantic Parsing for Tables

- Semantic parsing for tables means convert utterances (= sentences) into structured queries
- Traditional approaches have always struggled
- TABERT does not output the query, it is just used as encoder in the semantic parser



[Source: Yu et al. 2021]

Agenda

1. Background: BERT

2. Goal: Semantic parsing for databases

3. Architecture: Overview and Architecture

4. Pretraining : How has the model been trained

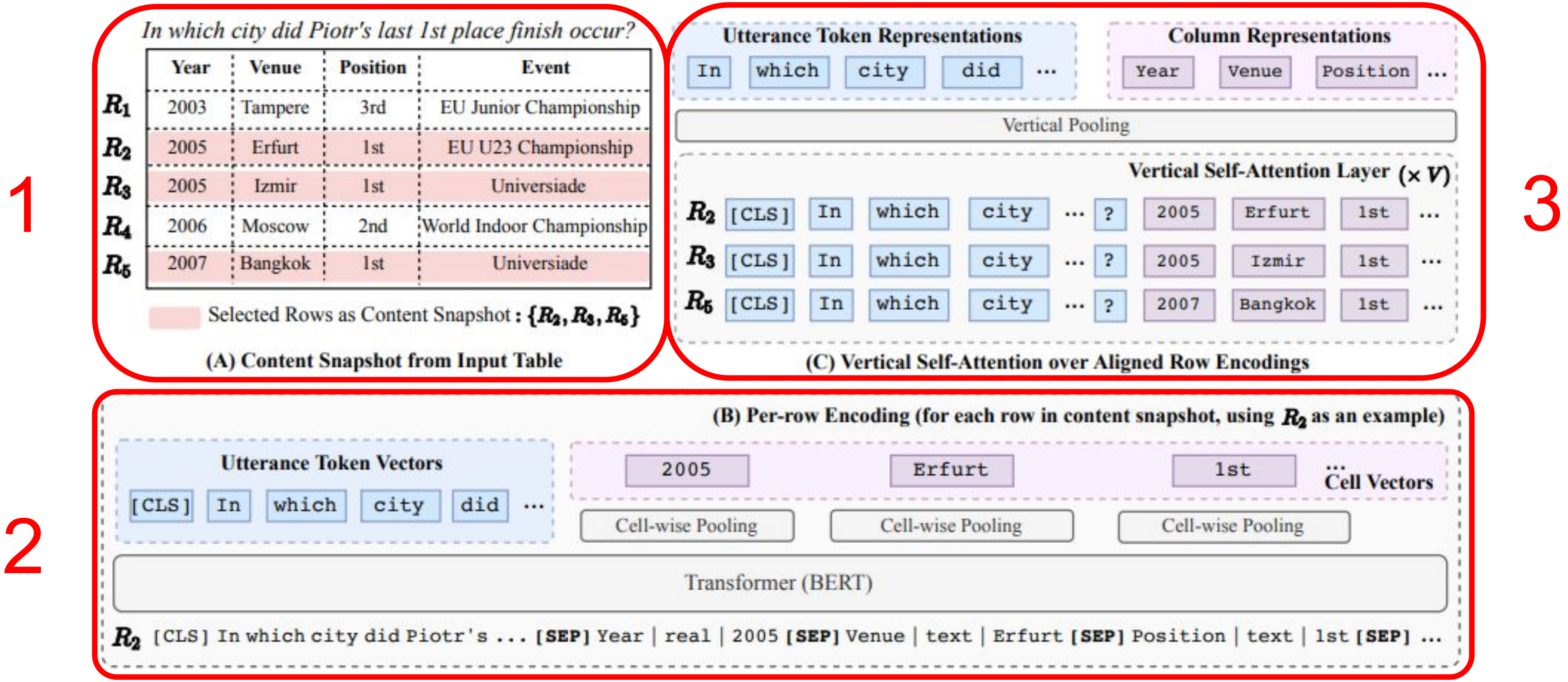
5. Applications: What it can be used for

6. Results: Results from the experiments

7. Conclusions: Limitations and future directions

Architecture

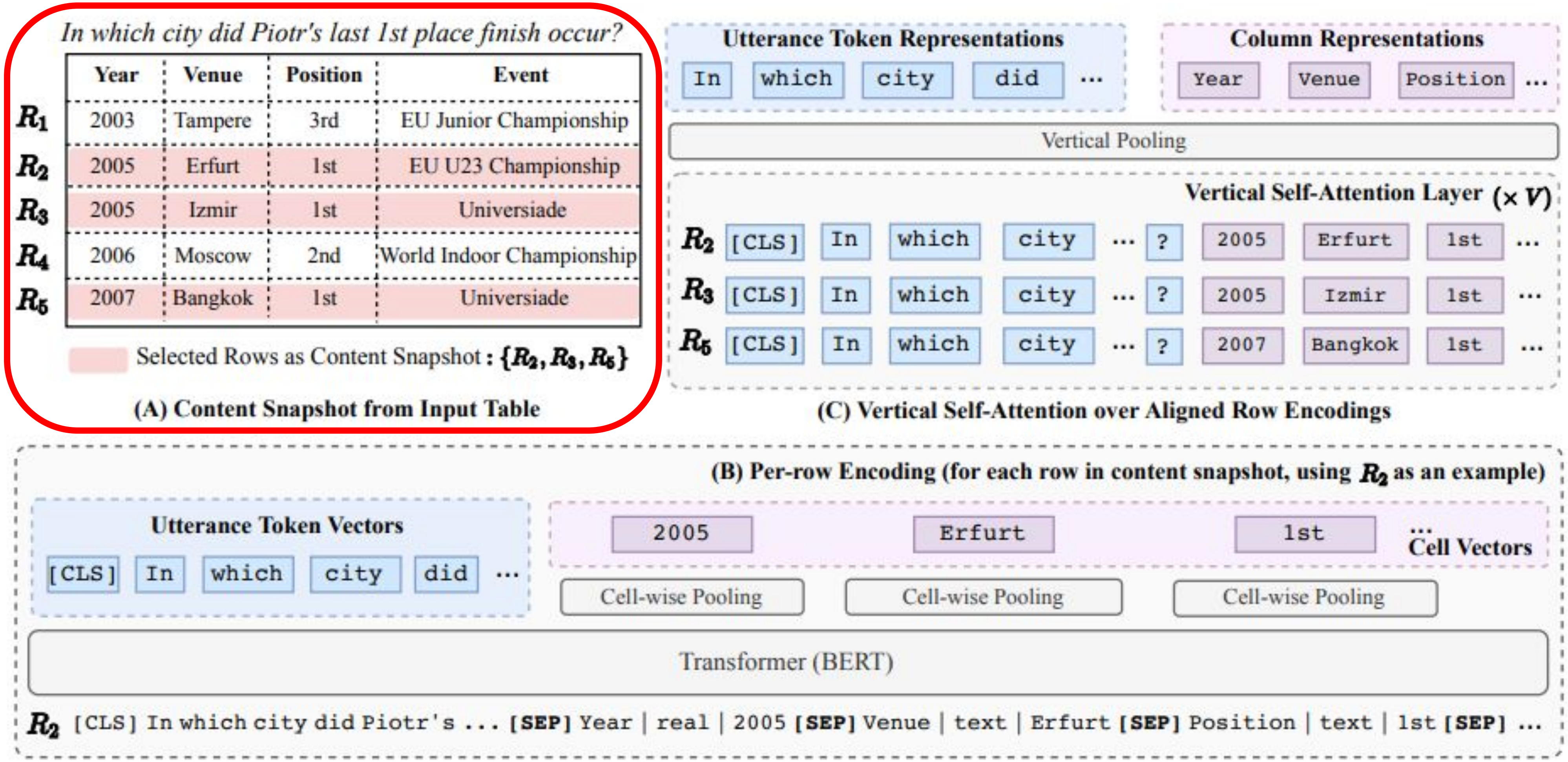
Schematic Overview



Architecture

Schematic Overview

1



Architecture

Content Snapshot

- Content of tables, not just column names
- Input the whole table would be too much
- Selecting rows based on highest n-gram overlap with the utterance
- Selecting top K rows. K=1 we have a synthetic row

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

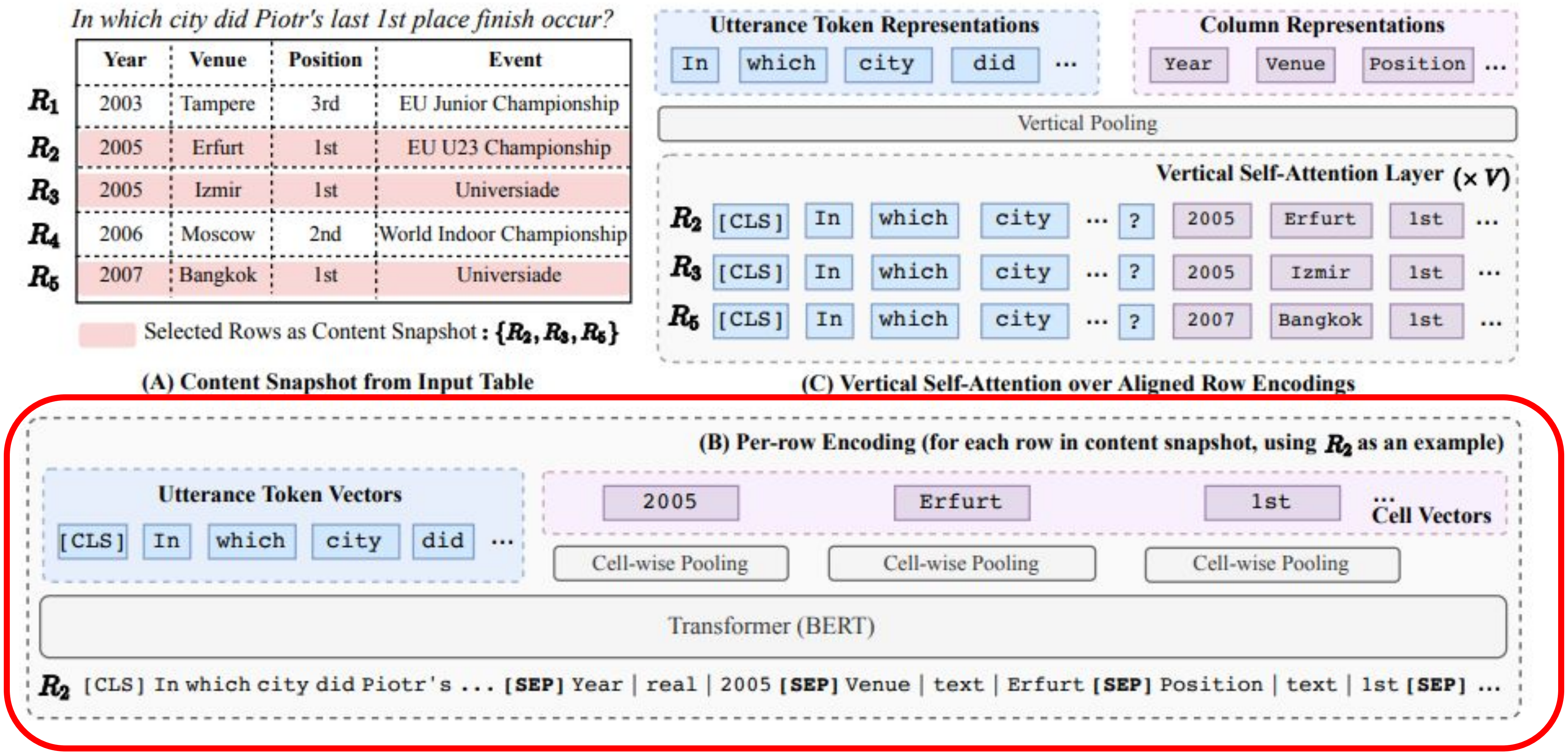
	Year	Venue	Position	Event
R_1	2003	Tampere	3rd	EU Junior Championship
R_2	2005	Erfurt	1st	EU U23 Championship
R_3	2005	Izmir	1st	Universiade
R_4	2006	Moscow	2nd	World Indoor Championship
R_5	2007	Bangkok	1st	Universiade

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

Architecture

Schematic Overview

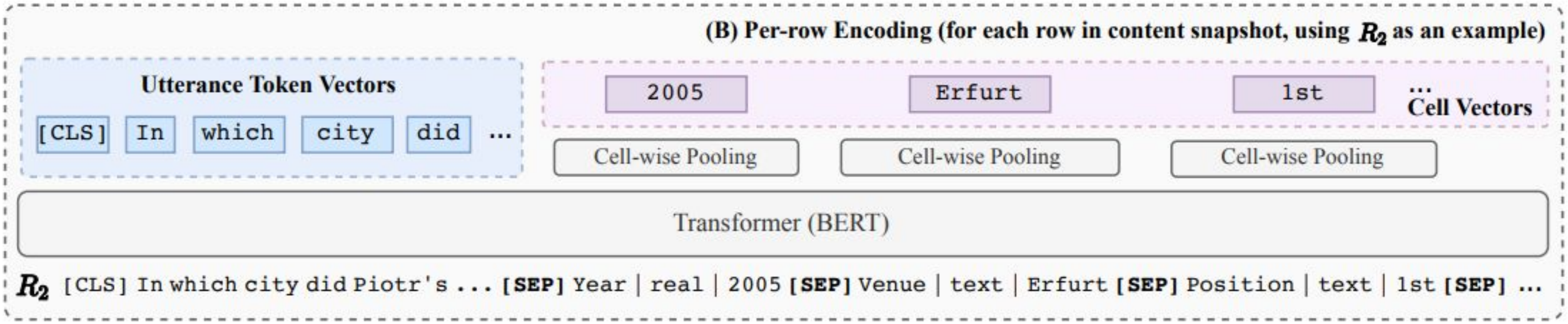
2



Architecture

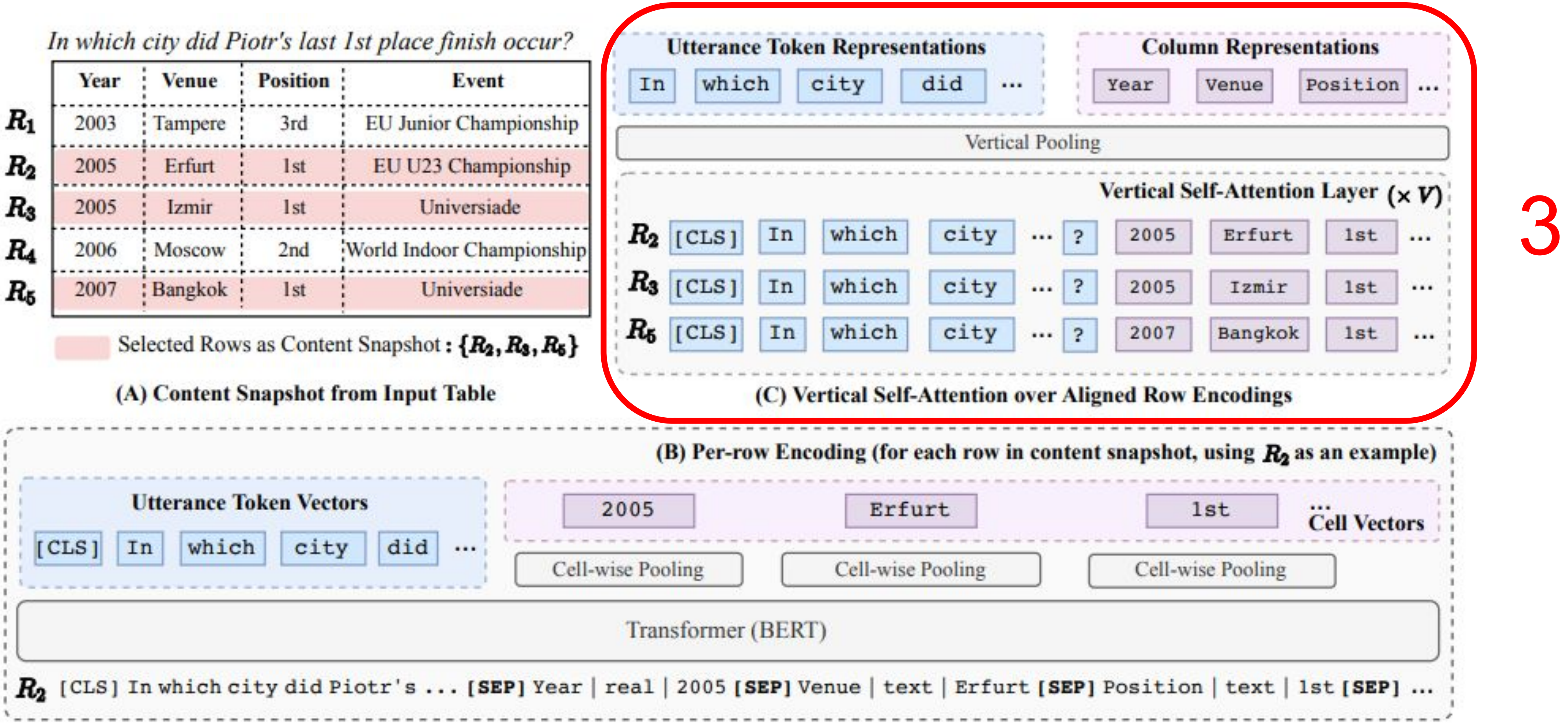
Row Linearization

- One linearized sequence for each row from the snapshot
- Sequence = utterance + (columns + their cell values)
- Specifically each cell is like $\underbrace{\texttt{Year}}_{\text{Column Name}} \mid \underbrace{\texttt{real}}_{\text{Column Type}} \mid \underbrace{\texttt{2005}}_{\text{Cell Value}}$
- Cell-wise pooling to create a single representation for each cell



Architecture

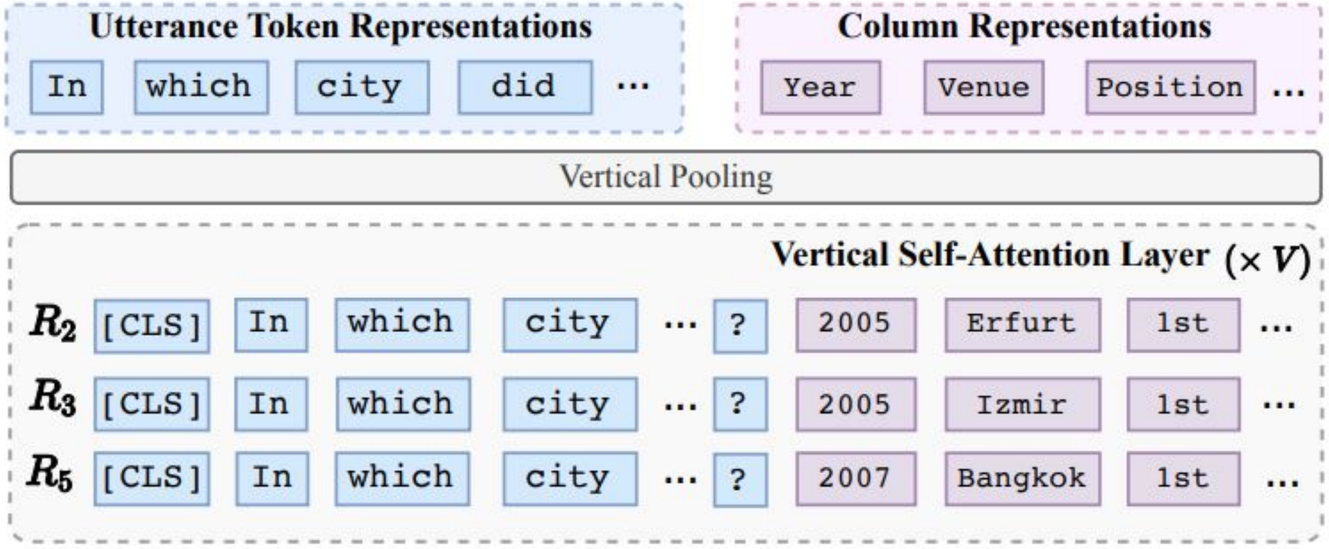
Schematic Overview



Architecture

Vertical Self-Attention

- Words vectors concatenated with cell vectors are the input to V Vertical Self-Attention layers
- OUTPUTS are:
  1. Utterance token representations by mean-pooling
  2. Column representations by mean-pooling
  3. Optionally, fixed length table representation with [CLS]



(C) Vertical Self-Attention over Aligned Row Encodings

Agenda

1. Background: BERT
2. Goal: Semantic parsing for databases
3. Architecture: Overview and Architecture
4. **Pretraining : How has the model been trained**
5. Applications: What it can be used for
6. Results: Results from the experiments
7. Conclusions: Limitations and future directions

Pretraining

Data and Hyper-Parameters

- 26.6 million parallel examples of english only tables and NL sentences
- From Wikipedia and WDC WebTable Corpus

Parameter	TaBERT$_{\text{Base}}$(K = 1)	TaBERT$_{\text{Large}}$(K = 1)	TaBERT$_{\text{Base}}$(K = 3)	TaBERT$_{\text{Large}}$(K = 3)
Batch Size	256	512	512	512
Learning Rate	2×10^{-5}	2×10^{-5}	4×10^{-5}	4×10^{-5}
Max Epoch	10			
Weight Decay	0.01			
Gradient Norm Clipping	1.0			

Pretraining

Learning Objectives

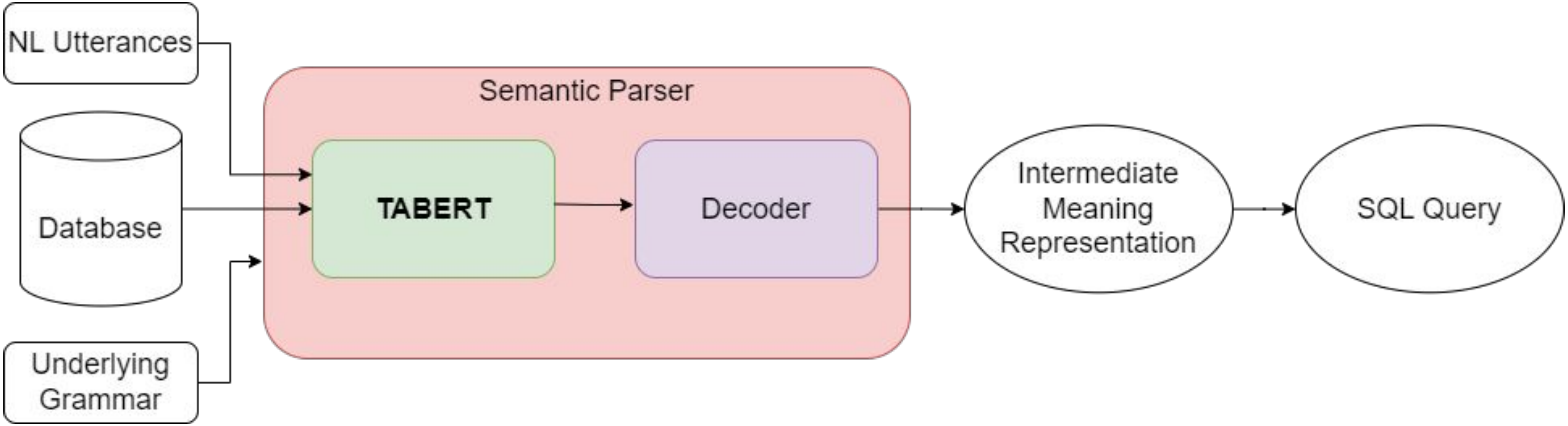
1. Masked Language Modeling (MLM) objective for NL contexts
2. Masked Column Prediction (MCP) objective for recovering names and types of columns
3. Cell Value Recovery (CVR) objective to ensure information of cell values in content snapshots is retained

Agenda

1. Background: BERT
2. Goal: Semantic parsing for databases
3. Architecture: Overview and Architecture
4. Pretraining : How has the model been trained
5. **Applications: What it can be used for**
6. Results: Results from the experiments
7. Conclusions: Limitations and future directions

Applications

Encoder Inside a Semantic Parser



Applications

Supervised Learning on SPIDER Dataset

- TranX is a semantic parser for translation from NL into intermediate representations
- TABERT is used as the encoder for utterances and tables in TranX
- SPIDER is a dataset with 10,181 examples across 200 databases

Complex question What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL

```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

[Source: Yu et al. 2019]

Applications

Weakly-Supervised Learning on WIKITABLEQUESTIONS

- MAPO is a different semantic parser that uses Reinforcement Learning
- TABERT replaces the original LSTM encoder in MAPO
- WikiTableQuestions is a dataset with 22,033 questions and 2108 tables from wikipedia
- The task of weakly supervised semantic parsing is harder

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

[Source: ppasupat github]

Agenda

1. Background: BERT

2. Goal: Semantic parsing for databases

3. Architecture: Overview and Architecture

4. Pretraining : How has the model been trained

5. Applications: What it can be used for

6. Results: Results from the experiments

7. Conclusions: Limitations and future directions

Results

- TABERT_base vs TABERT_large based on BERT_base vs BERT_large
- Different content snapshots dimension K

SPIDER

Top-ranked Systems on Spider Leaderboard		
Model		Dev. Acc.
Global–GNN (Bogin et al., 2019a)		52.7
EditSQL + BERT (Zhang et al., 2019a)		57.6
RatSQL (Wang et al., 2019a)		60.9
IRNet + BERT (Guo et al., 2019)		60.3
+ Memory + Coarse-to-Fine		61.9
IRNet V2 + BERT		63.9
RyanSQL + BERT (Choi et al., 2020)		**66.6**
Our System based on TranX (Yin and Neubig, 2018)		
	Mean	Best
w/ $\text{BERT}_{\text{Base}}$ (K = 1)	61.8 ±0.8	62.4
− content snapshot	59.6 ±0.7	60.3
w/ $\text{TaBERT}_{\text{Base}}$ (K = 1)	63.3 ±0.6	64.2
− content snapshot	60.4 ±1.3	61.8
w/ $\text{TaBERT}_{\text{Base}}$ (K = 3)	63.3 ±0.7	64.1
w/ $\text{BERT}_{\text{Large}}$ (K = 1)	61.3 ±1.2	62.9
w/ $\text{TaBERT}_{\text{Large}}$ (K = 1)	64.0 ±0.4	64.4
w/ $\text{TaBERT}_{\text{Large}}$ (K = 3)	**64.5** ±0.6	**65.2**

WIKITABLEQUESTIONS

Previous Systems on WikiTableQuestions				
Model	Dev		Test	
Pasupat and Liang (2015)	37.0		37.1	
Neelakantan et al. (2016)	34.1		34.2	
Ensemble 15 Models	37.5		37.7	
Zhang et al. (2017)	40.6		43.7	
Dasigi et al. (2019)	43.1		44.3	
Agarwal et al. (2019)	43.2		44.1	
Ensemble 10 Models	–		46.9	
Wang et al. (2019b)	43.7		44.5	
Our System based on MAPO (Liang et al., 2018)				
	Dev	Best	Test	Best
Base Parser†	42.3 ±0.3	42.7	43.1 ±0.5	43.8
w/ $\text{BERT}_{\text{Base}}$ (K = 1)	49.6 ±0.5	50.4	49.4 ±0.5	49.2
− content snapshot	49.1 ±0.6	50.0	48.8 ±0.9	50.2
w/ $\text{TaBERT}_{\text{Base}}$ (K = 1)	51.2 ±0.5	51.6	50.4 ±0.5	51.2
− content snapshot	49.9 ±0.4	50.3	49.4 ±0.4	50.0
w/ $\text{TaBERT}_{\text{Base}}$ (K = 3)	51.6 ±0.5	52.4	51.4 ±0.3	51.3
w/ $\text{BERT}_{\text{Large}}$ (K = 1)	50.3 ±0.4	50.8	49.6 ±0.5	50.1
w/ $\text{TaBERT}_{\text{Large}}$ (K = 1)	51.6 ±1.1	52.7	51.2 ±0.9	51.5
w/ $\text{TaBERT}_{\text{Large}}$ (K = 3)	**52.2** ±0.7	**53.0**	**51.8** ±0.6	**52.3**

Results

Impact of Configurations

CONTENT SNAPSHOT

u: How many years before was the film Bacchae out before the Watermelon?

Input to TaBERT$_{\text{Large}}$ (K = 3) ▷ *Content Snapshot with Three Rows*

Film	Year	Function	Notes
The Bacchae	2002	Producer	Screen adaptation of...
The Trojan Women	2004	Producer/Actress	Documutary film...
The Watermelon	2008	Producer	Oddball romantic comedy...

Input to TaBERT$_{\text{Large}}$ (K = 1) ▷ *Content Snapshot with One Synthetic Row*

Film	Year	Function	Notes
The Watermelon	2013	Producer	Screen adaptation of...

ROW LINEARIZATION

Cell Linearization Template	WikiQ.	Spider
Pretrained TaBERT$_{\text{Base}}$ Models (K = 1)		
`Column Name`	49.6 ±0.4	60.0 ±1.1
`Column Name` | `Type`† (−content snap.)	49.9 ±0.4	60.4 ±1.3
`Column Name` | `Type` | `Cell Value`†	51.2 ±0.5	63.3 ±0.6
BERT$_{\text{Base}}$ Models		
`Column Name` (Hwang et al., 2019)	49.0 ±0.4	58.6 ±0.3
`Column Name is Cell Value` (Chen19)	50.2 ±0.4	63.1 ±0.7

PRETRAINING OBJECTIVE

Learning Objective	WikiQ.	Spider
MCP only	51.6 ±0.7	62.6 ±0.7
MCP + CVR	51.6 ±0.5	63.3 ±0.7

Agenda

1. Background: BERT
2. Goal: Semantic parsing for databases
3. Architecture: Overview and Architecture
4. Pretraining : How has the model been trained
5. Applications: What it can be used for
6. Results: Results from the experiments
7. **Conclusions: Limitations and future directions**

Conclusions

Thoughts and Future Directions

Strengths:

1. Learns joint contextual representations of NL and structured tables
2. Content-aware representations
3. Can be applied to various tasks

Limitations:

1. Dimensions of the model, data for pretraining
2. It relies specifically on data of tables and NL context in pretraining
3. No handling of dynamic updates of tables

Future Works:

1. New tasks such as table retrieval and table-to-text generation
2. Other table linearization strategies
3. Foreign languages
4. Evaluate it on more datasets

Thank you for your attention!

Feedbacks: