TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data

Pengcheng Yin* **Graham Neubig**
Carnegie Mellon University
{pcyin,gneubig}@cs.cmu.edu

Wen-tau Yih **Sebastian Riedel**
Facebook AI Research
{scottyih,sriedel}@fb.com

https://github.com/facebookresearch/TaBERT

⇒之前的预训练模型大多数都是在Natural Language Text上进行预训练，如果任务设计结构化or半结构化数据呢？如Text2SQL等。

Challenges：

⇒ 如果直接用在Text上预训练的模型用于下游含结构化数据的任务，可能不太匹配。
⇒ 通常一个结构化数据（比如表格）会有很多行，直接naively encode效率很低。
⇒ 可能非常domain-specific，如人物bio，体育比赛介绍等有特定的schema。

TaBERT

⇒还是基于BERT（用来初始化模型中的Transformer）
⇒通过平行语料的对其关系学习utterance 和 structure data的contextual representation

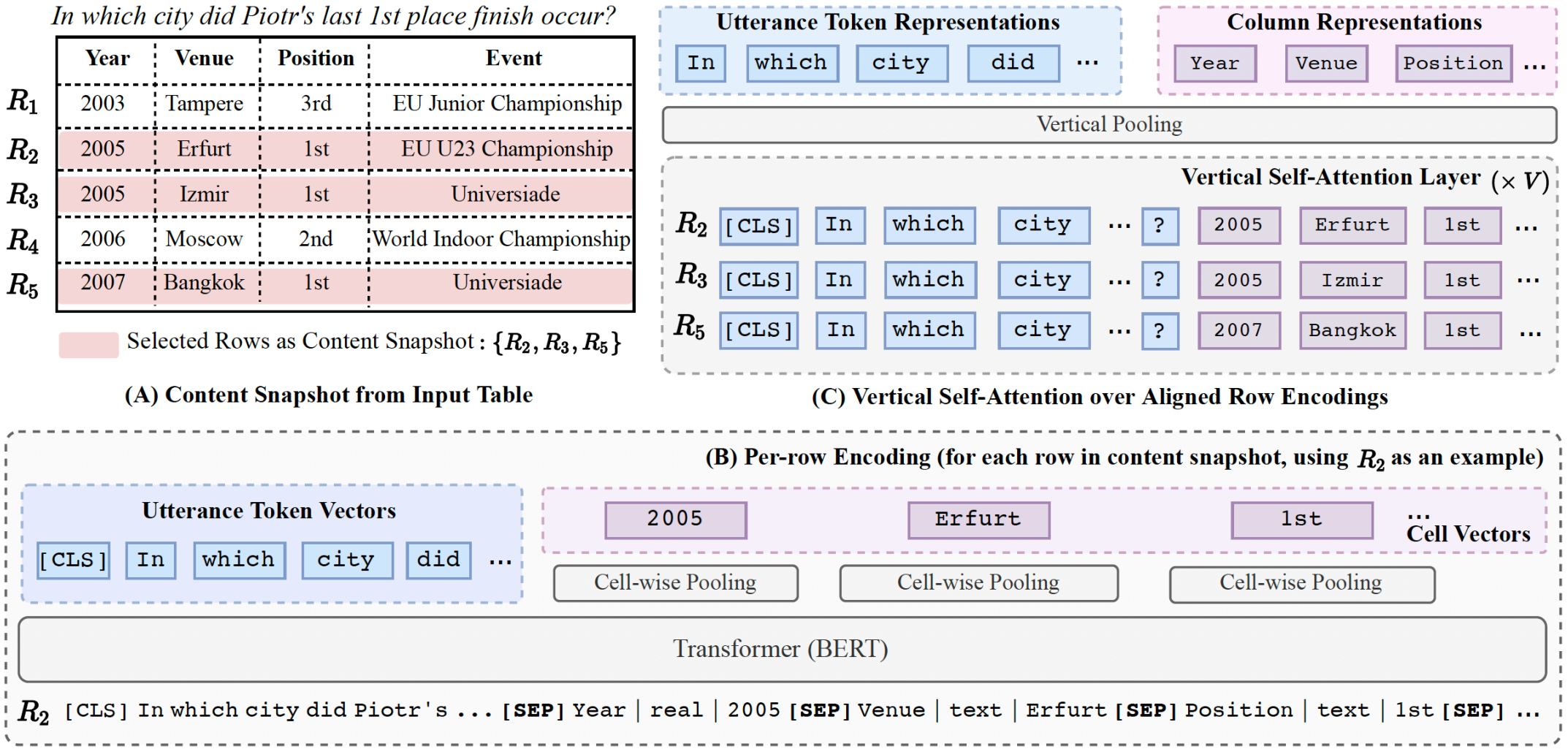


Figure 1: Overview of TABERT for learning representations of utterances and table schemas with an example from WIKITABLEQUESTIONS[3]. **(A)** A content snapshot of the table is created based on the input NL utterance. **(B)** Each row in the snapshot is encoded by a Transformer (only R_2 is shown), producing row-wise encodings for utterance tokens and cells. **(C)** All row-wise encodings are aligned and processed by V vertical self-attention layers, generating utterance and column representations.

Content Snapshot

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 : $\{R_2, R_3, R_5\}$

(A) Content Snapshot from Input Table

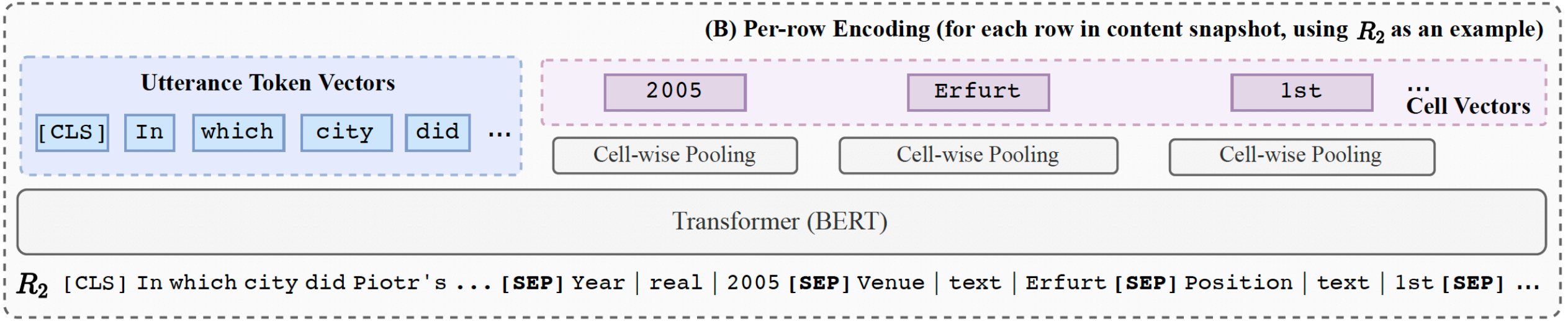
如何采样？两种策略。

策略1: 采样top-K rows跟utterance n-gram overlap最大的（n<=3)

策略2: 对于每一列选择n-gram overlap最大的，构建一个虚拟行

Table一般很大，无法全部信息都利用上。
⇒之前有些工作只用了column name，但是实际上content也会有有用的信息
⇒采样一些信息量大的，与input utterance最相关的row

Row Linearization

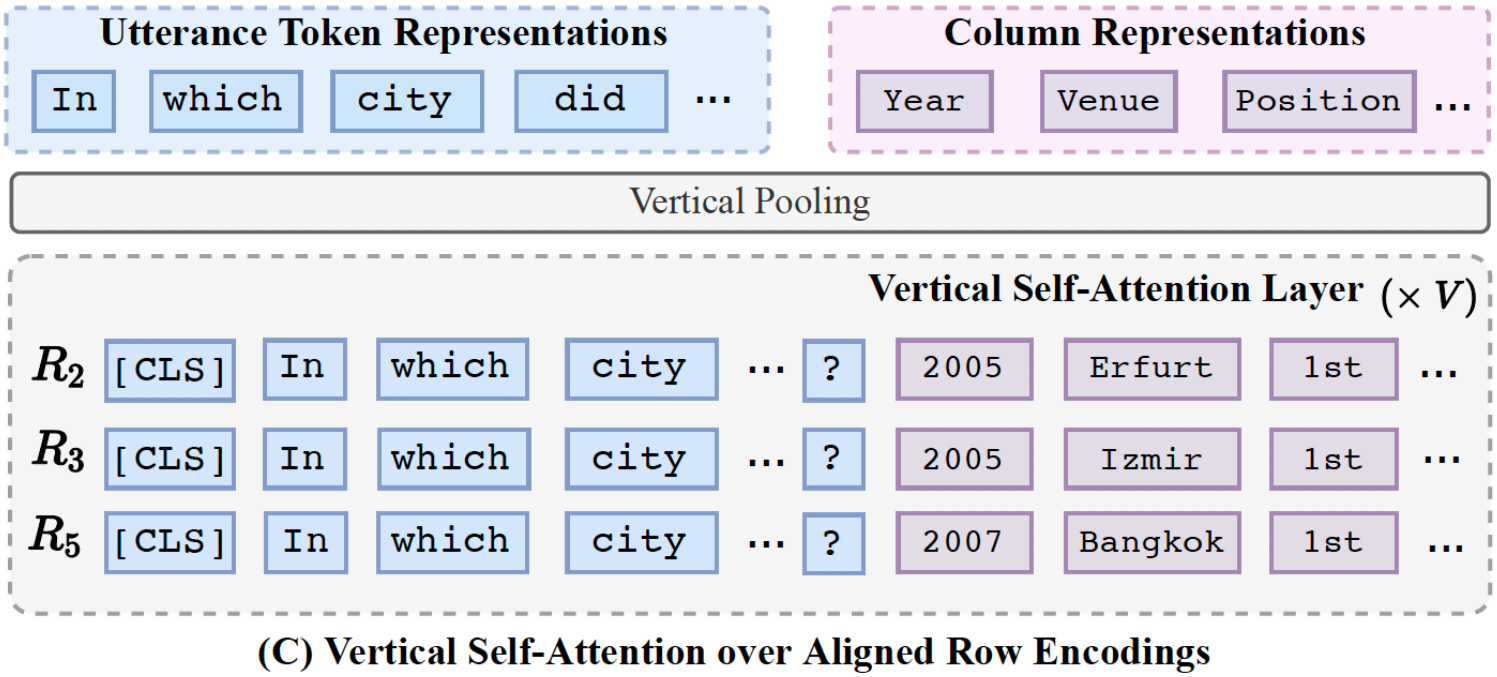


Year | real | 2005

Column Name | Column Type | Cell Value

Row information + utterance 送入Transformer

Vertical Self-attention



(C) Vertical Self-Attention over Aligned Row Encodings

=> 每个cell对应的output平均后作为其表示，并在前面concat上utterance，得到每个row的表示

=> 每个列做self-attention（cross-row information aggregation），share parameters

=> Mean-pooling得到一个最终的表示，包括utterance的表示和table的表示

Training Data (~26 millions)：

1. Wikipedia
2. WDC WebTable Corpus

Training Objective:

1. 对于utterance方面，用标准的MLM
2. 对于Table方面，两个Objective
    1. Masked Column Prediction: 把content snap的20% column name/type MASK，并预测
    2. Cell Value Recovery: 利用最终cell representation position embedding预测cell中对应位置的word。输入中没有把cell value mask掉，因为难度太大而且不太合理。

实验部分：

1. SPIDER => Text2SQL 任务，输入是utterance和Table，
   输出是把utterance对应的SQL
2. WIKITABLEQUESTIONS =>

Our task is as follows: given a table t and a question x about the table, output a list of values y that answers the question according to the table. Example inputs and outputs are shown in Figure 1. The system has access to a training set $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$ of questions, tables, and answers, but the tables in test data do not appear during training.

Accuracy

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**

Table 1: Execution accuracies on WIKITABLEQUESTIONS. †Results from Liang et al. (2018). (TA)BERT models are evaluated with 10 random runs. We report mean, standard deviation and the best results. TEST↦BEST refers to the result from the run with the best performance on DEV. set.

⇒使用TABERT增强比使用BERT效果显著

⇒K=3比K=1的效果要好，但是K=1也毕竟能取得明显优于BERT的效果

⇒-content snapshot表示线性化的时候只有column name/type没有value

⇒可以看到-content snapshot效果下降，说明利用table content有必要

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/ $\text{BERT}_{\texttt{Base}}$ (K = 1)	61.8 ±0.8	62.4
− content snapshot	59.6 ±0.7	60.3
w/ $\text{TABERT}_{\texttt{Base}}$ (K = 1)	63.3 ±0.6	64.2
− content snapshot	60.4 ±1.3	61.8
w/ $\text{TABERT}_{\texttt{Base}}$ (K = 3)	63.3 ±0.7	64.1
w/ $\text{BERT}_{\texttt{Large}}$ (K = 1)	61.3 ±1.2	62.9
w/ $\text{TABERT}_{\texttt{Large}}$ (K = 1)	64.0 ±0.4	64.4
w/ $\text{TABERT}_{\texttt{Large}}$ (K = 3)	**64.5** ±0.6	**65.2**

Table 2: Exact match accuracies on the public development set of SPIDER. Models are evaluated with 5 random runs.

预训练目标的影响

Learning Objective	WIKIQ.	SPIDER
MCP only	51.6 ±0.7	62.6 ±0.7
MCP + CVR	51.6 ±0.5	63.3 ±0.7

Table 5: Performance of pretrained TABERT$_{\texttt{Base}}$ (K = 3) on DEV. sets with different pretraining objectives.

A Relation-Specific Attention Network for Joint Entity and Relation Extraction

Yue Yuan[1,2] , **Xiaofei Zhou**[1,2*] , **Shirui Pan**[3] , **Qiannan Zhu**[1,2] , **Zeliang Song**[1,2] and **Li Guo**[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, School of Cyber Security
[3]Faculty of Information Technology, Monash University
{yuanyue,zhouxiaofei}@iie.ac.cn, shirui.pan@monash.edu

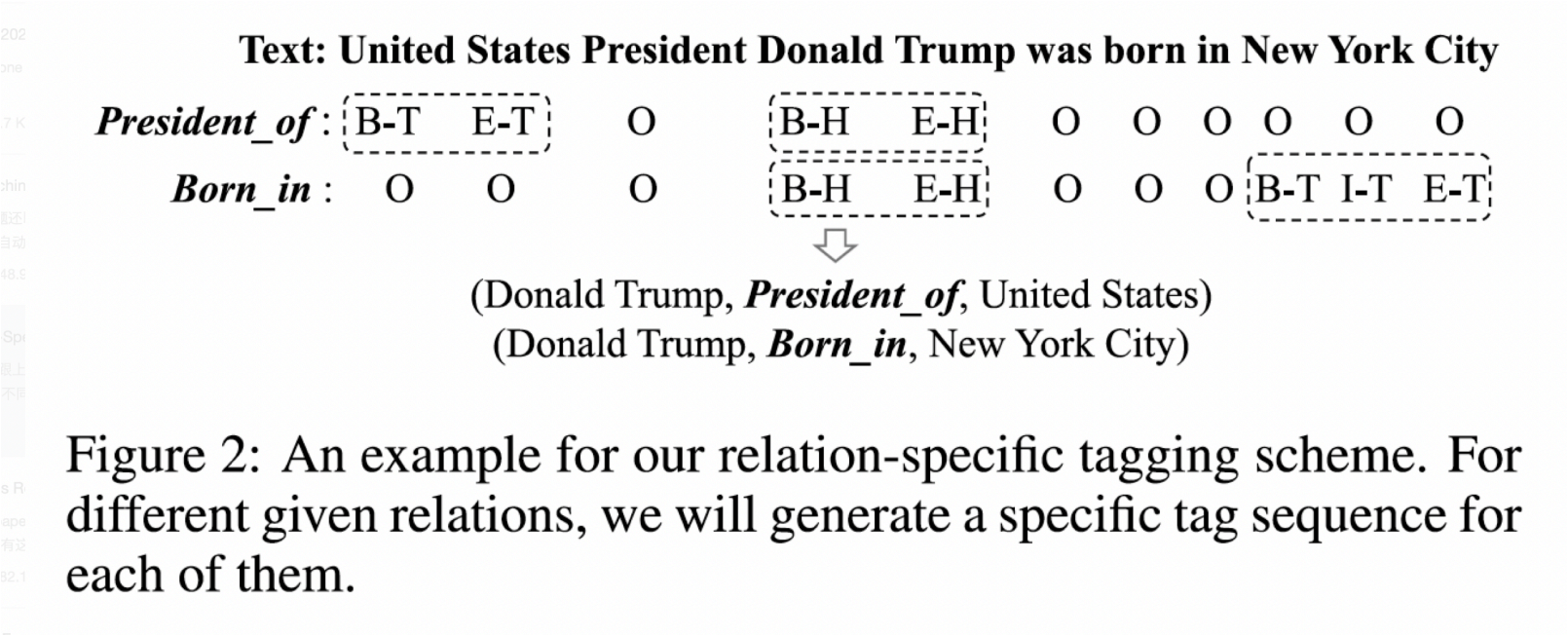


Figure 2: An example for our relation-specific tagging scheme. For different given relations, we will generate a specific tag sequence for each of them.

Attention as Relation: Learning Supervised Multi-head Self-Attention for Relation Extraction

Jie Liu[1*], **Shaowei Chen**[1], **Bingquan Wang**[1], **Jiaxin Zhang**[1], **Na Li**[2] and **Tong Xu**[3]

[1]College of Artificial Intelligence, Nankai University, Tianjin, China

[2]College of Computer Science, Nankai University, Tianjin, China

[3]University of Science and Technology of China, Hefei, China

jliu@nankai.edu.cn, {chenshaowei, wangbq, nkuzjx, 2120180458}@mail.nankai.edu.com, tongxu@ustc.edu.cn

提出现有sentence-level RE三大问题：
=> EPO & SPO
=> Entity在relation可能充当不同语义角色
=>. Entity pair可能有多个不同relations

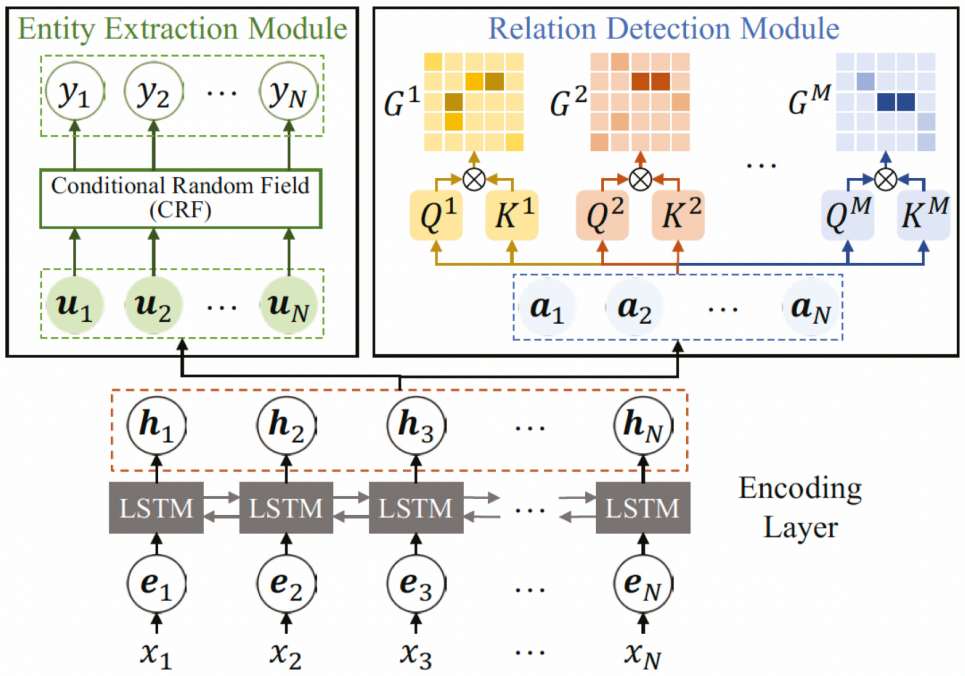


Figure 2: The framework of our model, which consists of an encoding layer, an entity extraction module, and a relation detection module.

什么意思呢，就是把encode后的分别映射到各个relation的subspace，做self-attention。

似乎有一个趋势是：大家偏向于对不同relation在各自subspace上进行tail entity的预测