表格数据处理

任务

Caption：2019年世界主要国家经济状况

Header：

Cells：

国家	GDP（万亿美元）	对外贸易额（万亿美元）
美国	21.37	5.1
中国	14.34	3.95
德国	3.85	2.69
日本	5.08	1.35

Query：2019年美国国内生产总值比中国高多少?

Content-Based Table Retrieval for Web Queries

Zhao Yan[†*], Duyu Tang[‡], Nan Duan[‡], Junwei Bao[+*], Yuanhua Lv[§], Ming Zhou[‡], Zhoujun Li[†]

[†]Beihang University [‡]Microsoft Research, Beijing, China

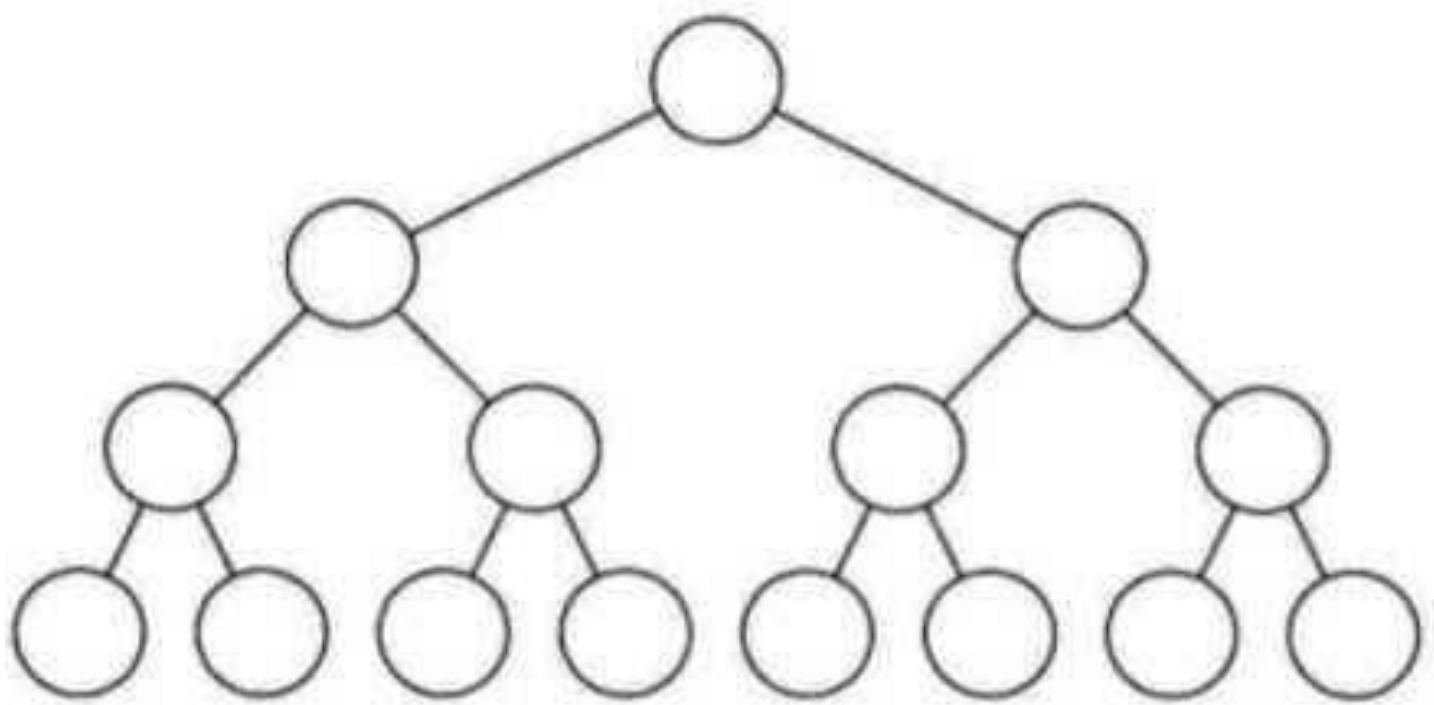
[+]Harbin Institute of Technology [§]Microsoft AI and Research, Sunnyvale CA, USA

[†]{yanzhao, lizj}@buaa.edu.cn [+]baojunwei001@gmail.com

[‡§]{dutang, nanduan, yuanhual, mingzhou}@microsoft.com

用决策树计算表格和问题的相关性得分

- 每一个分支选择由一个特征决定



- 得分:

基于规则的特征

- 1.词级别

$$f_{wmt}(t_a, q) = \frac{\sum_{w \in t_a} \delta(w, q) \cdot idf(w)}{\sum_{w' \in t_a} idf(w')}$$

$$f_{wmq}(t_a, q) = \frac{\sum_{w \in t_a} \delta(w, q) \cdot idf(w)}{\sum_{w' \in q} idf(w')}$$

- Idf：词频
- δ（w,q）：w出现在q中，值为1；未出现，值为0

- 2.短语级别

$$f_{pp}(t_a, q) = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{i,j} score(src_{i,n}^{t_q}, src_{j,n}^{q})}{|t_a| - N + 1}$$

$$score(src_x; src_y) = \sum_{PT} p(tgt_k|src_x) \cdot p(src_y|tgt_k)$$

- 3.句子级别

$$f_{s1}(t_a, q) = cosine(cdssm(t_a), cdssm(q))$$

基于神经网络的特征

$$\alpha_i = \frac{exp(tanh(W[m_i; v_q] + b))}{\sum_{j=1}^{k} exp(tanh(W[m_j; v_q] + b))}$$

- $V_{header} = \sum_{i=1}^{k} \alpha_i m_i$

$$f_{nn}(header, q) = NN_1(M_h, v_q)$$

实验数据

	WQT数据集 (WebQueryTable)	WTQ数据集 (WikiTableQuestions)
表格数	273,816	2,108
平均列数	4.55	6.38
最大列数	52	25
最小列数	1	3
平均行数	9.15	28.50
最大行数	1,517	754
最小行数	2	5
问题的数量	21,113	22,033
问题平均长度	4.61	11.25

实验结果

Setting	MAP	P@1
BM25	58.23	47.12
Feature	61.02	47.79
NeuralNet	61.94	49.02
Feature + NeuralNet	67.18	54.15

Setting	Feature		NeuralNet	
	MAP	P@1	MAP	P@1
Header (H)	22.39	9.76	26.03	13.35
Cell (Cel)	28.85	14.95	27.47	12.92
Caption (Cap)	57.12	56.83	60.16	48.48
H + Cel	31.99	17.08	30.73	16.25
H + Cel + Cap	61.02	47.79	61.94	49.02

ACL2020

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

作者主要贡献（为什么不使用已有的Bert）

- 1.Bert只见过自由格式的文本数据，没见过结构化、半结构化的表格数据 -> 预训练
- 2.表格中可能存在大量无关数据 -> 内容快照
- 3.不同领域内的语义解析难以统一 -> 从问题和表格中获取信息

内容快照（k=3）

- 除了表头，还使用单元格信息

Caption：2019年世界主要国家经济状况

Header：

Cells：

国家 text	GDP（万亿美元）real	对外贸易额（万亿美元）real
美国	21.37	5.1
中国	14.34	3.95
德国	3.85	2.69
日本	5.08	1.35

Query：2019年美国国内生产总值比中国高多少？

N(tri)-gram overlap

S1： november

S2： december

- Trigram：

S1： nov ove vem emb mbe ber

S2： dec ece cem emb mbe ber

S1 ∪ S2 = 3

S1 ∩ S2 = 9

overlap = 3/9

内容快照（k=1）

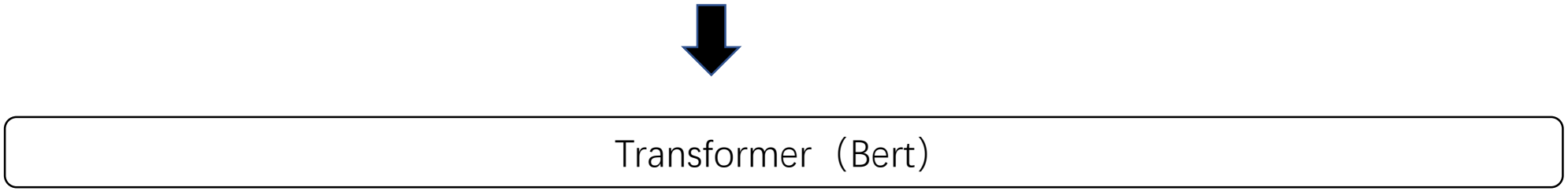
- 寻找最相关的单元格值，拼接成行

国家 text	GDP（万亿美元）real	对外贸易额（万亿美元）real
美国	14.34	5.1

行线性化

Encoding：

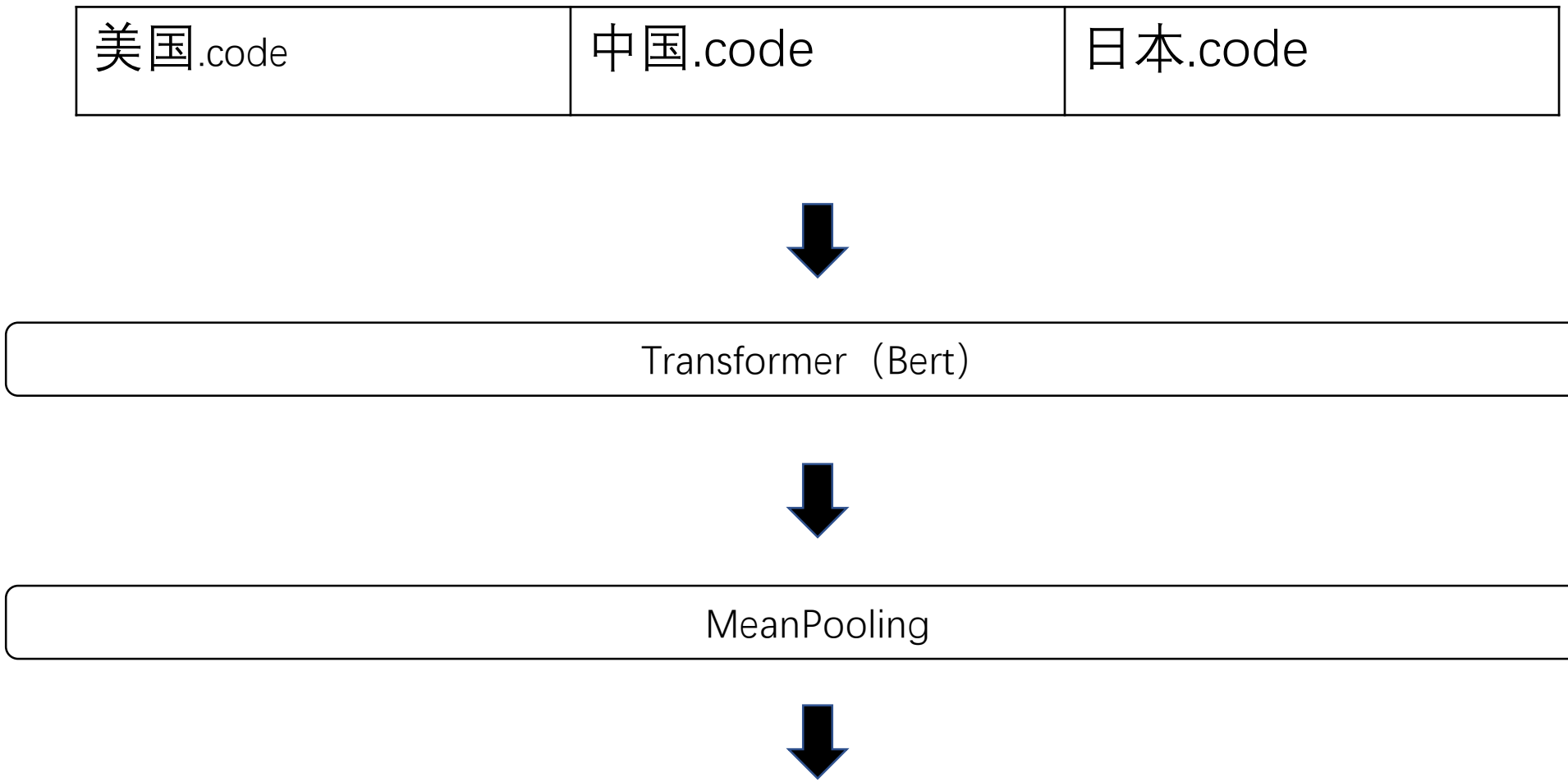
- [CLS] 2019 … [SEP] 国家 | text | 美国 [SEP] GDP | real | 21.37 …



垂直自注意机制

- 聚合不同行之间的信息
- 垂直：同一列

美国 avg(‘美’,‘国’).coding	21.37	5.1
中国	14.34	3.95
日本	5.08	1.35


美国.code
中国.code
日本.code
Transformer（Bert）
MeanPooling
列表示

预训练

- 2600万组 表格+问题

国家 text	【GDP（万亿美元）real】尝试恢复（20%）	对外贸易额（万亿美元）real
美国	21.37（尝试恢复）	5.1
中国	14.34（尝试恢复）	3.95
德国	3.85	2.69
日本	5.08 （尝试恢复）	1.35

实验

- 1. SPIDER 有监督训练 结构化数据 可能涉及多个表

- 2. WIKI TABLE QUESTIONS 弱监督（只知道对错）半结构化数据 可能涉及单个表内跨行推理

实验结果

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**

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**

实验结果

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

Table4(statement_id: **75**)
1644 is in panel 1: 2005/2006–2009/2010 in outsourcing (t0)
statement_type: **refuted**

Predictors of outsourcing.

	Panel 1: 2005/2006–2009/2010			Panel 2: 2009/2010–2013/2014		
	Outsourcing (t0)	Outsourcing (t+1)	Outsourcing (t+2)	Outsourcing (t0)	Outsourcing (t+1)	Outsourcing (t+2)
Financial constraints (t0)	1.224 (0.275)	1.528 (0.422)	1.713 (0.583)	1.554 * (0.389)	2.260 *** (0.652)	3.290 *** (1.156)
Competence constraints (t0)	1.876 ** (0.465)	2.182 *** (0.646)	1.967 * (0.718)	1.806 ** (0.493)	0.998 (0.346)	1.281 (0.517)
Firm-size	1.131 (0.110)	1.035 (0.116)	0.899 (0.120)	1.241 * (0.136)	1.249 * (0.167)	1.150 (0.171)
Firm-type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Model χ 2	85.61	61.99	43.33	60.40	52.27	42.03
p	0.00	0.00	0.00	0.00	0.00	0.00
n	2744	2109	1518	2123	1644	1199