NLP with tabular data

彭凯龙

目录

- 前言
- 数据集（30%）
- 论文介绍（60%）
- 总结（10%）

前言

- 表的形式

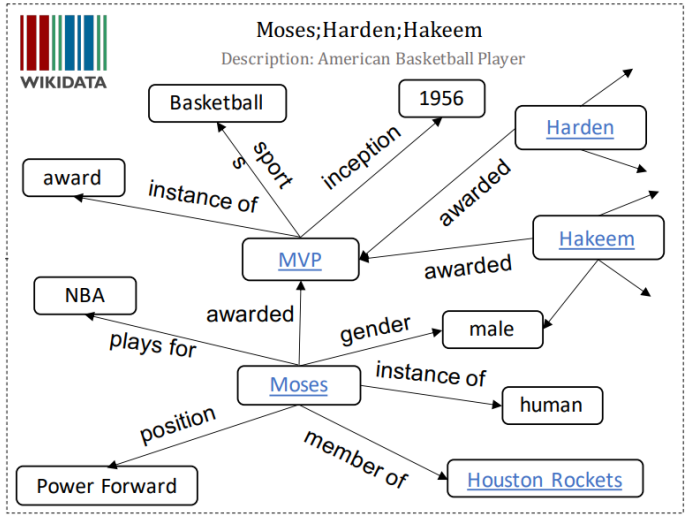
  键值对

  表格

  图

Bernard Keen	
Born	September 5, 1890
Died	August 5, 1981 (aged 90)
Nationality	British
Awards	Fellow of the Royal Society[1]
	Scientific career
Fields	Soil scientist
Institutions	University College London

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



数据集

- **分类**

  TabFact, InfoTabs

- **生成**

  WikiBio, RotoWire, LogicNLG, HybridQA, ToTTo

  WikiSQL, Spider

WikiTableQuestions (2015)

- 变形：SequentialQA (2017)

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: {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}

Compositional Semantic Parsing on Semi-Structured Tables. ACL (1) 2015: 1470-1480

WikiSQL(2017)

- 依据问题生成sql语句
  - Execution Acc
  - Logical form Acc
- 只含select和where，表无外键
- 变形：Spider (2018)

Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...	...	...	...	...

Question:

How many CFL teams are from York College?

SQL:

```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = “York”
```

Result:

2

Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. CoRR abs/1709.00103 (2017)

LogicNLG (2020)

- 生成逻辑推理/符号运算得到的语句

  linked columns

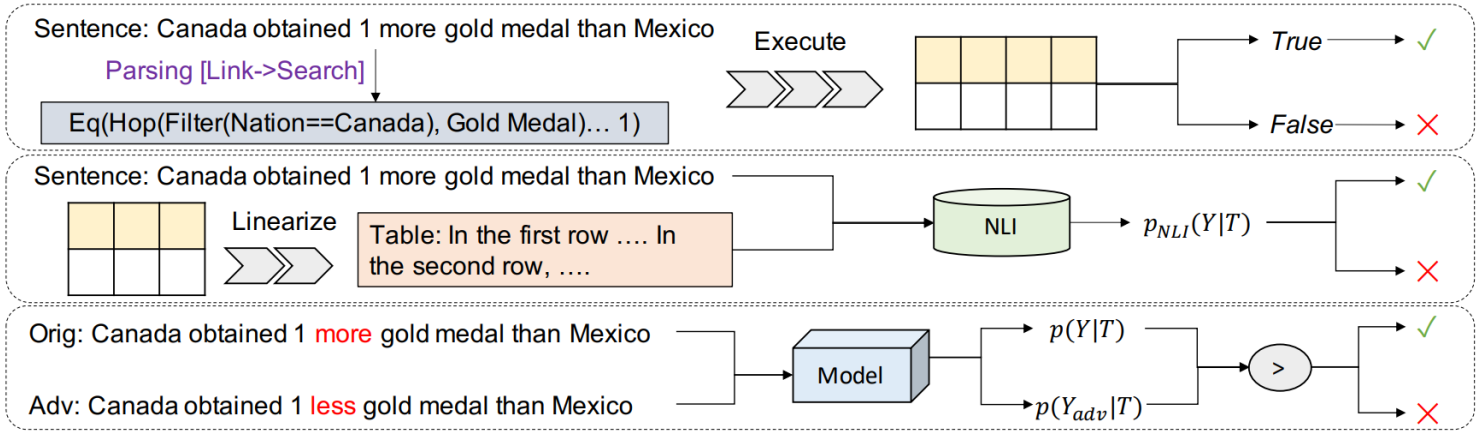
Medal Table from Tournament

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.

- Parsing-based Evaluation
- NLI-based Evaluation
- Adversarial Evaluation



Logical Natural Language Generation from Open-Domain Tables. ACL 2020: 7929-7942

ToTTo (2020)

- 条件生成
- Hallucination

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description:None

		RUSHING					RECEIVING				
YEAR	**TEAM**	**ATT**	**YDS**	**AVG**	**LNG**	**TD**	**NO.**	**YDS**	**AVG**	**LNG**	**TD**
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1050	4.9	62	9	92	1016	11	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	-	**1991**	**8189**	**4.1**	**71**	**56**	**566**	**4911**	**8.7**	**73**	**17**

Target Text: Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

ToTTo: A Controlled Table-To-Text Generation Dataset. EMNLP (1) 2020: 1173-1186

InfoTabs (2020)

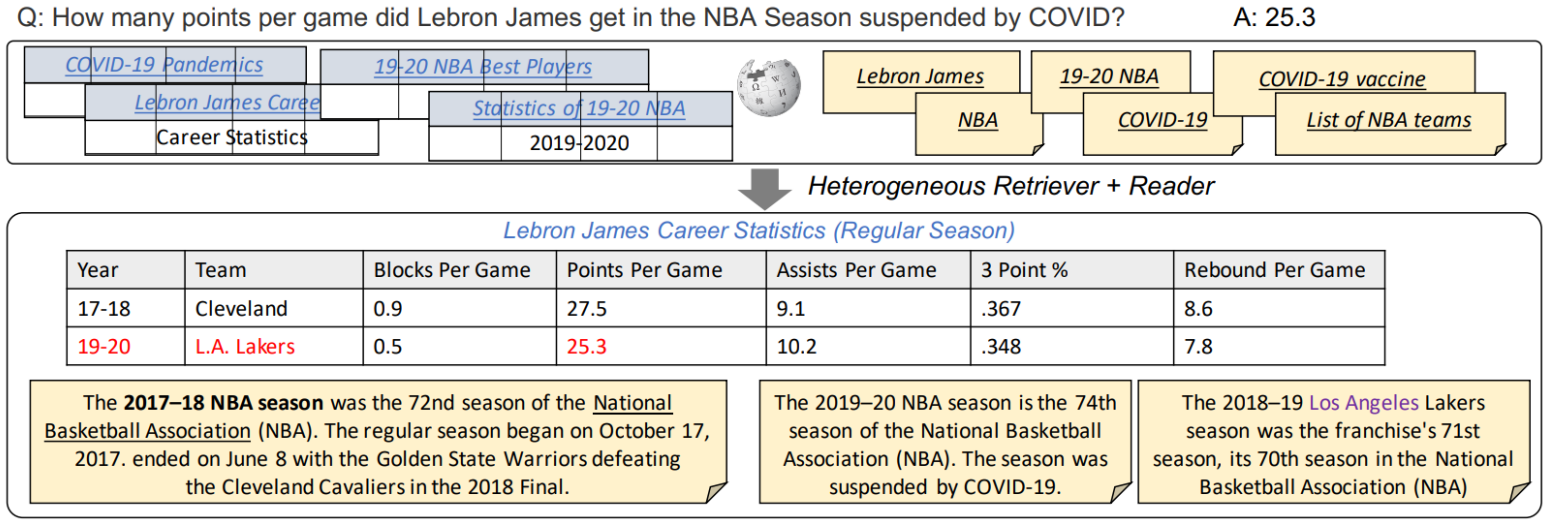
- 考察模型推理能力
- entailment, contradiction and neutral
- Multi-Faceted Evaluation
  - α_1 set：陈述词汇构成、表格涉及领域与训练集分布相近
  - α_2 set：最小限度改变陈述的某些词汇
  - α_3 set：改变表格涉及领域

Premise	Dev	α_1	α_2	α_3
Train with SVM				
Para	59.11	59.17	46.44	41.28
Train with BERT_B				
Para	63.00	63.54	52.57	48.17
Train with RoBERTa_B				
Para	67.2	66.98	56.87	55.36
Train with RoBERTa_L				
WMD-1	65.44	65.27	57.11	52.55
WMD-3	72.55	70.38	62.55	61.33
Para	75.55	74.88	65.55	64.94

INFOTABS: Inference on Tables as Semi-structured Data. ACL 2020: 2309-2324

OTT-QA (2021)

- 异构数据QA



Open Question Answering over Tables and Text. CoRR abs/2010.10439 (2020)

论文介绍

- Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data. ACL 2020: 3864-3870
- Table Fact Verification with Structure-Aware Transformer. EMNLP (1) 2020: 1624-1629
- TCN: Table Convolutional Network for Web Table Interpretation. CoRR abs/2102.09460 (2021)
- CFGNN: Cross Flow Graph Neural Networks for Question Answering on Complex Tables. AAAI 2020: 9596-9603
- GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. CoRR abs/2009.13845 (2020)

Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data (ACL 2020)

- 动机：

  同时适用于结构化数据（WikiBio）与自然文本（SQuAD, NQG）

  简单模型

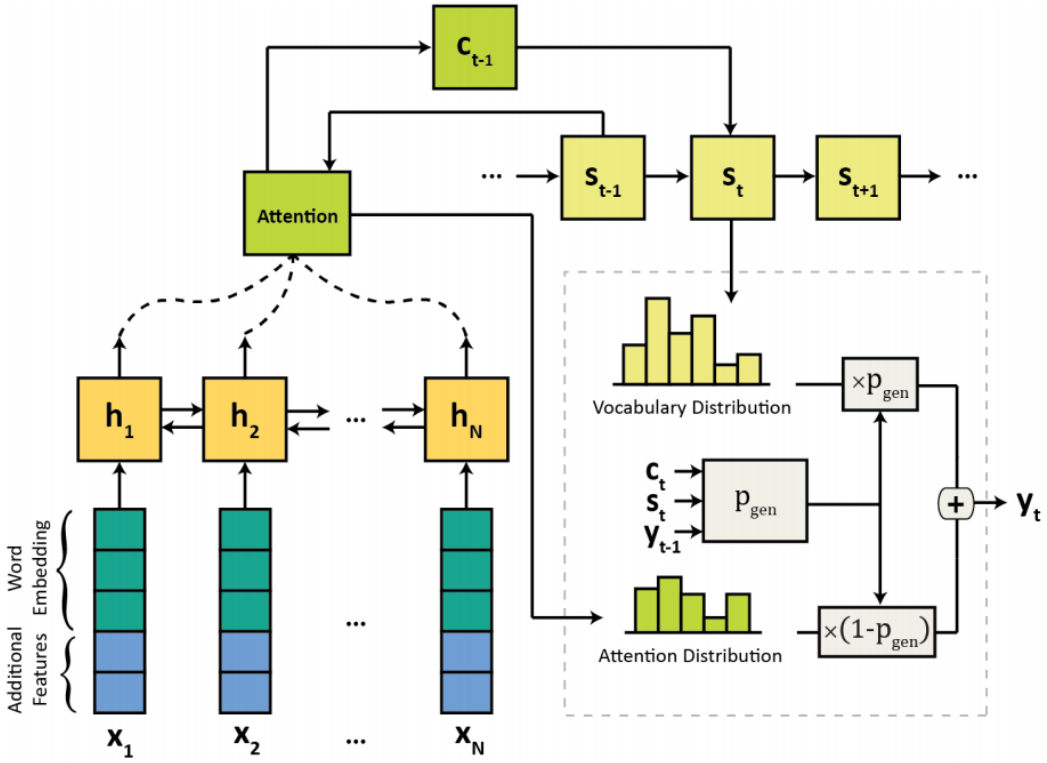
- 模型结构

- Cat additional feature

  Table: field name, p_t^+, p_t^-

  Unstructured: a single bit

Bernard Keen	
Born	September 5, 1890
Died	August 5, 1981 (aged 90)
Nationality	British
Awards	Fellow of the Royal Society[1]
Scientific career	
Fields	Soil scientist
Institutions	University College London



- Copy mechanism

- Exponential Moving Average

  $\bar{\theta} \leftarrow \beta \times \bar{\theta} + (1 - \beta) \times \theta$

Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data (ACL 2020)

- 结果

Models	BLEU-4	ROUGE-4
KN*	2.21	0.38
Template KN**	19.80	10.70
Lebret et al. (2016)	34.70 ± 0.36	25.80 ± 0.36
Bao et al. (2018)	40.26	-
Sha et al. (2018)	43.91	37.15
Liu et al. (2018) Orig.	44.89 ± 0.33	41.21 ± 0.25
Liu et al. (2018) Repl.	44.45 ± 0.11	39.65 ± 0.10
Liu et al. (2019b)	45.14 ± 0.34	41.26 ± 0.37
Our Model	46.07 ± 0.17	41.53 ± 0.30
+ EMA	**46.76 ± 0.03**	**43.54 ± 0.07**

Models	Split-1			Split-2		
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L
Heilman (2011)	-	-	-	9.47	18.97	31.68
Du et al. (2017)	12.28	16.62	39.75	-	-	-
Zhou et al. (2017)	-	-	-	13.29	-	-
Zhou et al. (2018)	-	-	-	13.02	-	44.0
Yao et al. (2018)	-	-	-	13.36	17.70	40.42
Song et al. (2018)	13.98	18.77	42.72	13.91	-	-
Zhao et al. (2018)	15.32	19.29	43.91	15.82	19.67	44.24
Sun et al. (2018)	-	-	-	15.64	-	-
Kumar et al. (2018)	16.17	19.85	43.90	-	-	-
Kim et al. (2019)	16.20 ± 0.32	19.92 ± 0.20	43.96 ± 0.25	16.17 ± 0.35	-	-
Liu et al. (2019a)	-	-	-	**17.55**	21.24	44.53
Our Model	14.81 ± 0.47	19.69 ± 0.24	43.01 ± 0.28	16.14 ± 0.25	20.44 ± 0.20	43.95 ± 0.26
+ EMA	**16.29 ± 0.04**	**20.70 ± 0.08**	**44.18 ± 0.15**	17.47 ± 0.10	**21.37 ± 0.06**	**45.18 ± 0.22**

Table Fact Verification with Structure-Aware Transformer (EMNLP 2020)

- 动机：

  单纯对表格做linearization会丢失结构信息

- 方法：

  将表格结构信息注入self-attention layer的mask中

$$M_{i,j} = \begin{cases} 0 & w_i \sim w_j \\ -\infty & w_i \not\sim w_j \end{cases}$$

$$\boldsymbol{Q}^l, \boldsymbol{K}^l, \boldsymbol{V}^l = \boldsymbol{H}^l \boldsymbol{W}_q, \boldsymbol{H}^l \boldsymbol{W}_k, \boldsymbol{H}^l \boldsymbol{W}_v$$

$$\boldsymbol{A}^l = \text{softmax}(\frac{\boldsymbol{Q}^l {\boldsymbol{K}^l}^T + \boldsymbol{M}}{\sqrt{d_k}})$$

$$\boldsymbol{H}^{l+1} = \boldsymbol{A}^l \boldsymbol{V}^l$$

将符号推理问题转化为匹配问题（summary row）

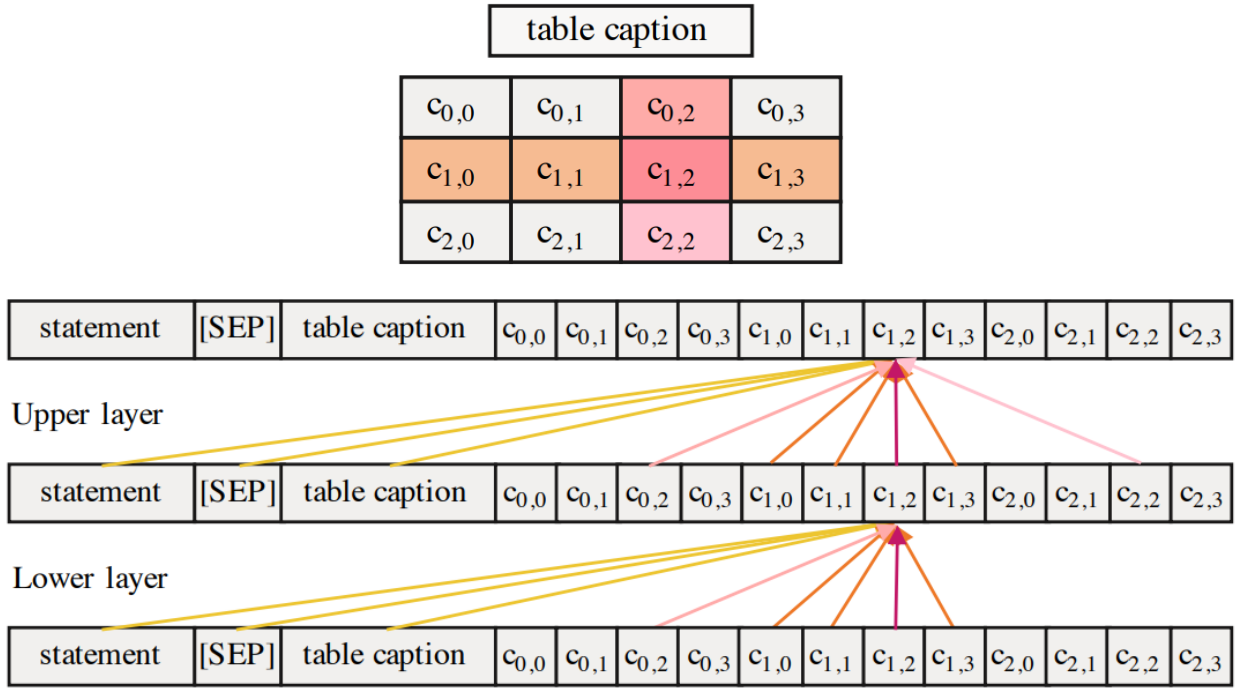


Table Fact Verification with Structure-Aware Transformer (EMNLP 2020)

- 结果

Model	Val	Test	Test(simple)	Test(complex)
LPA(Wenhu et al., 2020)†	65.1	65.3	78.7	58.5
Table-BERT(Wenhu et al., 2020)†	66.1	65.1	79.1	58.2
Table-BERT tuned*	68.38	68.30	82.35	61.48
BERT with cell position encoding	59.31	59.44	63.24	57.58
SAT with Horizontal scan	72.96	72.82	85.44	66.62
- w/o visible matrix	68.41	67.67	75.93	63.61
- w/o summary row	72.00	72.09	85.53	65.49
- w/o visible matrix w/o summary row	66.84	66.01	74.37	61.90
SAT with Vertical scan	**73.31**	**73.23**	**85.46**	**67.23**
- w/o visible matrix	64.21	64.27	68.77	62.06
- w/o summary row	71.71	71.59	84.70	65.15
- w/o summary row and w/o visible matrix	63.03	62.34	66.71	60.19
- all layers w/o cross row attention	72.83	72.26	84.61	66.11
- all layers w cross row attention	72.02	71.82	83.45	66.10

Base	MASK-LM	69.6 $\pm$4.4	69.9 $\pm$3.8	82.0 $\pm$5.9	63.9 $\pm$2.8
Base	SQA	74.9 $\pm$0.2	74.6 $\pm$0.2	87.2 $\pm$0.2	68.4 $\pm$0.4
Base	Counterfactual	75.5 $\pm$0.5	75.2 $\pm$0.4	87.8 $\pm$0.4	68.9 $\pm$0.5
Base	Synthetic	77.6 $\pm$0.2	77.9 $\pm$0.3	89.7 $\pm$0.4	72.0 $\pm$0.2
Base	Counterfactual + Synthetic	**78.6** $\pm$0.3	**78.5** $\pm$0.3	**90.5** $\pm$0.4	**72.5** $\pm$0.3

- 训练耗时

  SAT：50min/epoch, 15-18 epochs (1 V100)

  Tapas：pretraining 3 days (32 Cloud TPU V3)

TCN: Table Convolutional Network for Web Table Interpretation (WWW 2021)

- 动机：

  Intra- and inter-table context

  Schema heterogeneity, context limitation

- 方法：

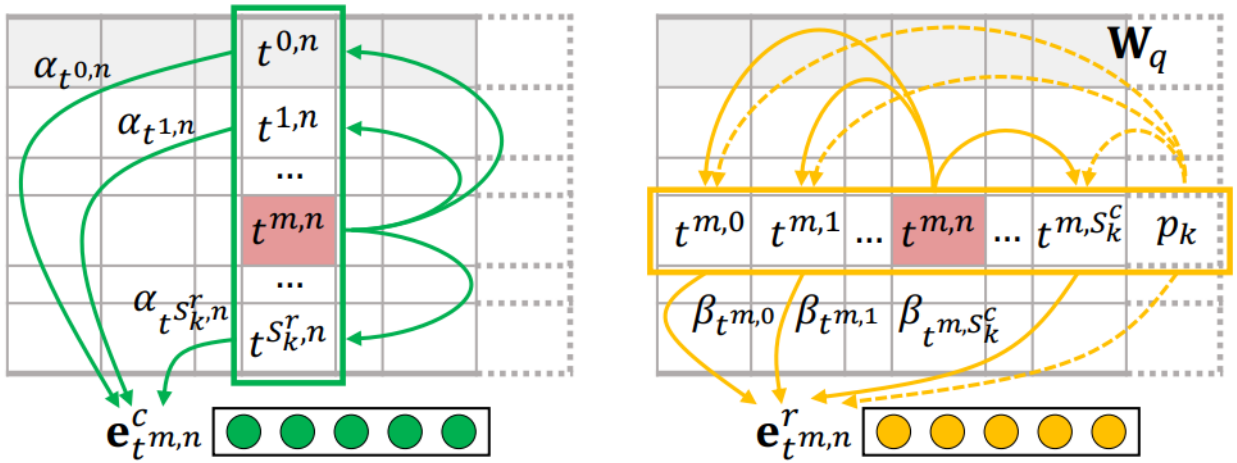
  Intra-table aggregation

  Column: $$\alpha_{t^{m',n}} = \frac{\exp\left(\mathrm{e}_{t^{m',n}}{}^{\top} \cdot \mathrm{e}_{t^{m,n}}\right)}{\sum_{\tilde{m}=0,\tilde{m}\neq m}^{S_k^r} \exp\left(\mathrm{e}_{t^{\tilde{m},n}}{}^{\top} \cdot \mathrm{e}_{t^{m,n}}\right)}$$

  $$\mathrm{e}^c_{t^{m,n}} = \sigma\left(\mathrm{W}_c \cdot \sum_{m'=0,m'\neq m}^{S_k^r} \alpha_{t^{m',n}} \mathrm{e}_{t^{m',n}}\right)$$

  Row: $$\beta_{t^{m,n'}} = \frac{\exp\left(\mathrm{e}_{t^{m,n'}}{}^{\top} \cdot \mathrm{W}_q \cdot (\mathrm{e}_{t^{m,n}} \| \mathrm{e}_{p_k})\right)}{\sum_{\tilde{n}=0,\tilde{n}\neq n}^{S_k^c} \exp\left(\mathrm{e}_{t^{m,\tilde{n}}}{}^{\top} \cdot \mathrm{W}_q \cdot (\mathrm{e}_{t^{m,n}} \| \mathrm{e}_{p_k})\right)}$$

  $$\mathrm{e}^r_{t^{m,n}} = \sigma\left(\mathrm{W}_r \cdot \sum_{n'=0,n'\neq n}^{S_k^c} \left(\beta_{t^{m,n'}} \mathrm{e}_{t^{m,n'}}\right) \| \mathrm{e}_{p_k}\right)$$



(a) Column aggregation (b) Row aggregation

$$\mathrm{e}^a_{t^{m,n}} = \sigma\left(\mathrm{W}_a \cdot (\mathrm{e}^c_{t^{m,n}} \| \mathrm{e}^r_{t^{m,n}})\right) = \mathrm{AGG}_a(t^{m,n})$$

TCN: Table Convolutional Network for Web Table Interpretation (WWW 2021)

- 方法:

Inter-table aggregation

Value:

$$\mathcal{N}_v(t_k^{m,n}) := \{t_{k'}^{\tilde{m},\tilde{n}} \mid t_{k'}^{\tilde{m},\tilde{n}} = t_k^{m,n} \wedge 0 \le k' \le K \wedge k' \neq k\}$$

$$\mathrm{E}^v_{t^{m,n}} \in \mathbb{R}^{|\mathcal{N}_v(t_k^{m,n})| \times D_a}$$

$$\Omega^v_{t^{m,n}} = \mathrm{softmax}\left(\mathrm{W}_s \cdot \left(\mathrm{E}^v_{t^{m,n}}\right)^\top\right)$$

$$\mathrm{e}^v_{t^{m,n}} = \mathrm{mean}(\Omega^v_{t^{m,n}} \cdot \mathrm{E}^v_{t^{m,n}} \cdot \mathrm{W}_b)$$

Position:

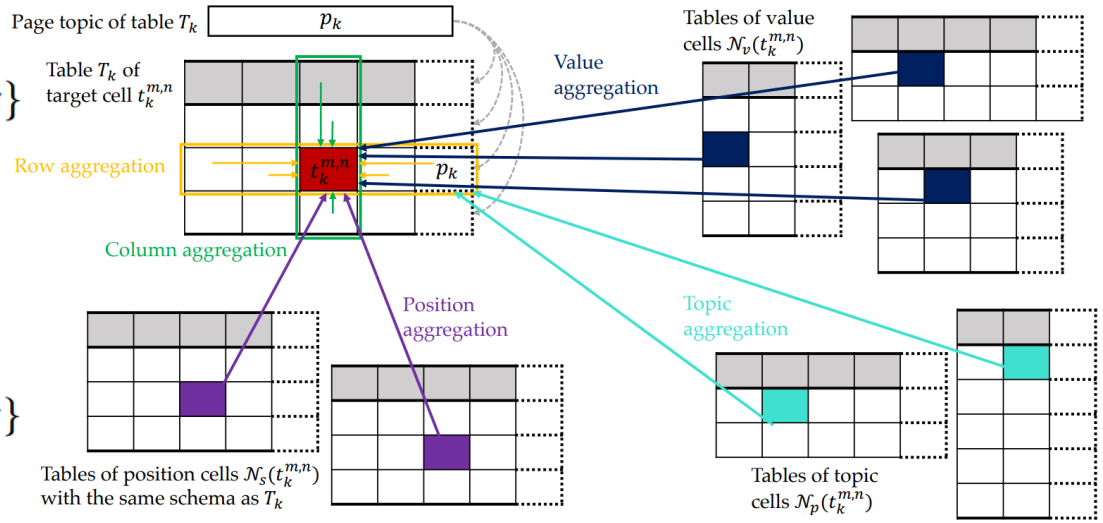
$$\mathcal{N}_s(t_k^{m,n}) := \{t_{k'}^{m,n} \mid \phi(k) = \phi(k') \wedge 0 \le k' \le K \wedge k' \neq k\}$$

$$\mathrm{E}^s_{t^{m,n}} \in \mathbb{R}^{|\mathcal{N}_s(t_k^{m,n})| \times D_a}$$

Topic:

$$\mathcal{N}_p(t_k^{m,n}) := \{t_{k'}^{\tilde{m},\tilde{n}} \mid t_{k'}^{\tilde{m},\tilde{n}} = p_k \wedge 0 \le k' \le K \wedge k' \neq k\}$$

$$\mathrm{e}^p_{p_k} = \mathrm{e}_{p_k} \| \mathrm{e}^p_{t^{m,n}}$$



$$\mathrm{h}_{t^{m,n}} = \sigma\left(\mathrm{W}_h \cdot \left(\mathrm{e}_{t^{m,n}} \| \mathrm{e}^a_{t^{m,n}} \| \mathrm{e}^v_{t^{m,n}} \| \mathrm{e}^s_{t^{m,n}}\right)\right)$$

TCN: Table Convolutional Network for Web Table Interpretation (WWW 2021)

- 训练：

$$\mathrm{h}_{t_k^{*,n}} = \mathrm{AVG}\left(\{\mathrm{h}_{t_k^{m,n}}\}_{m=0}^{S_k^r}\right)$$

Supervised：

Column type detection：

$$\mathcal{J}_k^{\mathcal{C}} = -\sum_{n=0}^{S_k^c}\sum_{c\in C}\mathbb{I}_{c_{t_k^{*,n}}=c}\cdot\log\frac{\exp\left(\mathrm{M}_c\cdot\mathrm{h}_{t_k^{*,n}}\right)}{\sum_{c'\in C}\exp\left(\mathrm{M}_{c'}\cdot\mathrm{h}_{t_k^{*,n}}\right)}$$

Pairwise column relation prediction:

$$\mathcal{J}_k^{\mathcal{R}} = -\sum_{n=1}^{S_k^c}\sum_{r\in\mathcal{R}}\mathbb{I}_{r_{t_k^{*,n}}=r}\cdot\log\frac{\exp\left(\mathrm{M}_r\cdot\left(\mathrm{h}_{t_k^{*,0}}\parallel\mathrm{h}_{t_k^{*,n}}\right)\right)}{\sum_{r'\in\mathcal{R}}\exp\left(\mathrm{M}_{r'}\cdot\left(\mathrm{h}_{t_k^{*,0}}\parallel\mathrm{h}_{t_k^{*,n}}\right)\right)}$$

$$\mathcal{J} = \sum_{k\in\mathcal{B}}\gamma\mathcal{J}_k^{\mathcal{C}} + (1-\gamma)\mathcal{J}_k^{\mathcal{R}}$$

Unsupervised:

$$\mathcal{J} = -\sum_{k\in\mathcal{B}}\sum_{\hat{t}_k^{m,n}\in\hat{T}_k}\sum_{v\in\mathcal{V}}\mathbb{I}_{t_k^{m,n}=v}\cdot\log\frac{\exp\left(\mathrm{M}_v\cdot\mathrm{h}_{\hat{t}_k^{m,n}}\right)}{\sum_{v'\in\mathcal{V}}\exp\left(\mathrm{M}_{v'}\cdot\mathrm{h}_{\hat{t}_k^{m,n}}\right)}$$

TCN: Table Convolutional Network for Web Table Interpretation (WWW 2021)

- 结果

Method	Column type C			Pairwise column relation $\mathcal{R}$		
	Acc.	F1-weighted	Cohen's kappa κ	Acc.	F1-weighted	Cohen's kappa κ
TABLE2VEC	.832	.820	.763	.822	.810	.772
TABERT	.908	.861	.834	.877	.870	.846
TURL	.914	.877	.876	.890	.889	.838
HNN	.916	.883	.869	.848	.843	.794
SHERLOCK	.922	.895	.863	.831	.818	.802
TCN-intra	.911	.881	.873	.893	.894	.869
TCN-$\mathcal{N}_v$	.939 (+3.1%)	.916 (+4.0%)	.897 (+2.8%)	.920 (+3.0%)	.920 (+2.9%)	.898 (+3.3%)
TCN-$\mathcal{N}_s$	.934 (+2.5%)	.908 (+3.1%)	.894 (+2.4%)	.908 (+1.7%)	.912 (+2.0%)	.881 (+1.4%)
TCN-$\mathcal{N}_p$	.923 (+1.3%)	.890 (+1.0%)	.880 (+0.8%)	.906 (+1.4%)	.904 (+1.1%)	.875 (+0.7%)
TCN	**.958 (+5.2%)**	**.938 (+6.5%)**	**.913 (+4.6%)**	**.934 (+4.6%)**	**.925 (+3.5%)**	**.905 (+4.1%)**

	#tables K	Avg. # rows S_k^r	Avg. # cols. S_k^c	#column types $\lvert C\rvert$	#pairwise relations $\lvert\mathcal{R}\rvert$
$\mathcal{D}^m$	128,443	7.0	3.6	8	14
$\mathcal{D}^w$	55,318	16.1	2.4	201	121
	#schemas U	**Avg. #value cells $\lvert\mathcal{N}_v\rvert$**	**Avg. #position cells $\lvert\mathcal{N}_s\rvert$**		**Avg. #topic cells $\lvert\mathcal{N}_p\rvert$**
$\mathcal{D}^m$	11	10.7	5048.1		2.4
$\mathcal{D}^w$	6,538	7.2	0.9		1.9

Method	$\mathcal{D}^m$		$\mathcal{D}^w$	
	Type C	Relation $\mathcal{R}$	Type C	Relation $\mathcal{R}$
TCN (supervised)	.938	.925	.933	.941
TCN + TABERT for pre-training	.948	.933	.942	.949
TCN + TURL for pre-training	.945	.934	.946	.953
TCN full w/ pre-training & fine-tuning	.957	.946	.951	.960

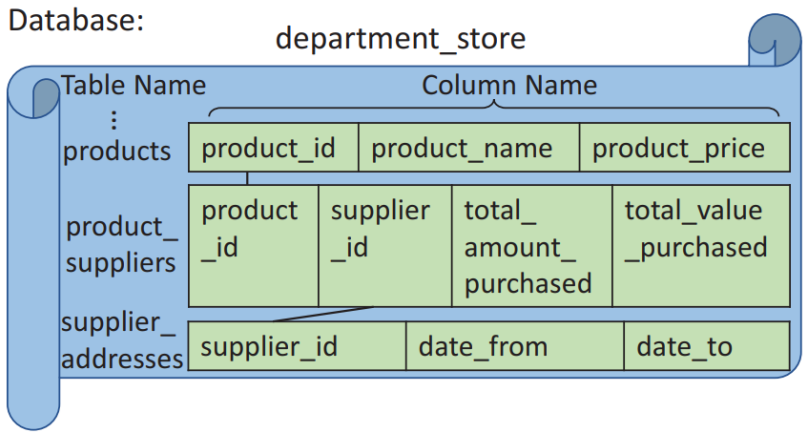
CFGNN: Cross Flow Graph Neural Networks for Question Answering on Complex Tables (AAAI 2020)

- 动机：

  用图对多表格间的关系进行建模

  利用同级结点间的关系

  反复思考结点间关系



Question: What are the average amount purchased and value purchased for the supplier who supplies the most products.

SQL:

```
SELECT avg(total_amount_purchased) ,
       avg(total_value_purchased)
FROM product_suppliers
WHERE supplier_id  =
    (SELECT supplier_id
    FROM Product_Suppliers
    GROUP BY supplier_id
    ORDER BY count(*) DESC LIMIT 1)
```

CFGNN: Cross Flow Graph Neural Networks for Question Answering on Complex Tables (AAAI 2020)

- 模型结构：

Cross Flow Graph Neural Network

blue node: table term　green: column

purple curve: Bi-directional RNN among neighbor

$$\overrightarrow{\mathbf{h}}^{k}_{u,i} = \overrightarrow{\mathrm{RNN}}(\overrightarrow{\mathbf{h}}^{k}_{u,i-1}, \mathbf{W}^{k}\mathbf{h}^{k-1}_{u,i})$$

$$\overleftarrow{\mathbf{h}}^{k}_{u,|N(v)|-i-1} = \overleftarrow{\mathrm{RNN}}(\overleftarrow{\mathbf{h}}^{k}_{u,|N(v)|-i}, \mathbf{W}^{k}\mathbf{h}^{k-1}_{u,|N(v)|-i-1})$$

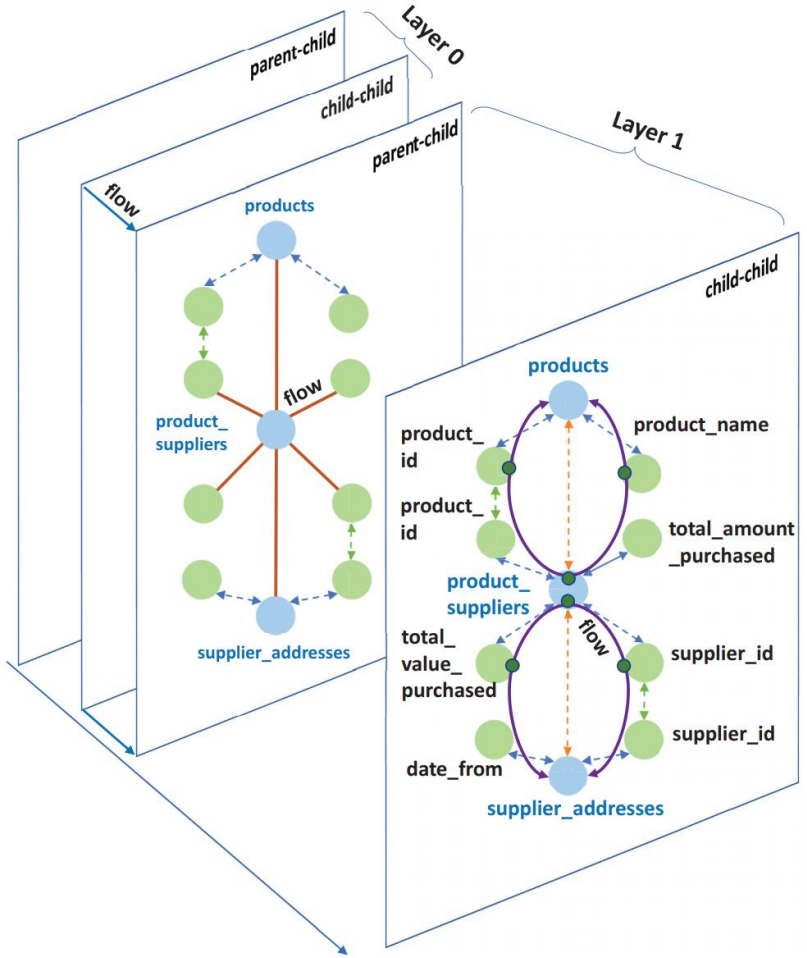
$$\bar{\mathbf{h}}^{k}_{v} = [\overrightarrow{\mathbf{h}}^{k}_{u,|N(v)|-1}, \overleftarrow{\mathbf{h}}^{k}_{u,0}]$$

orange lines: attention of the center node

$$\alpha_{v,u} = \frac{e^{\beta(\mathbf{g}^{T}[\mathbf{W}^{k}\mathbf{h}_{v}, \mathbf{W}^{k}\mathbf{h}_{u}]}}{\sum_{u' \in N(v)} e^{\beta(\mathbf{g}^{T}[\mathbf{W}^{k}\mathbf{h}_{v}, \mathbf{W}^{k}\mathbf{h}_{u'}])}}$$

$$\hat{\mathbf{h}}^{k}_{v} = \gamma(\sum_{u \in N(v)} \alpha_{v,u}\mathbf{W}^{k}\mathbf{h}^{k-1}_{u})$$

$$\mathbf{h}^{k}_{v} = \overrightarrow{\mathrm{RNN}}(\mathbf{h}^{k-1}_{v}, [\bar{\mathbf{h}}^{k}_{v}; \hat{\mathbf{h}}^{k}_{v}])$$



CFGNN: Cross Flow Graph Neural Networks for Question Answering on Complex Tables (AAAI 2020)

- 模型结构：

Hierarchical Encoding Layer

$$emb_i^{\text{BERT}} = \gamma \sum_{d=0}^{D} \alpha_d f(h_i^d)$$

两层RNN

$$c_j^k = \frac{i}{I+1} \sum_{i=0}^{I} e_{j,i}^k$$

$$s_j^i = e_i{}^{\mathrm{T}} c_j$$

$$a_j^i = exp(s_j^i) / \sum_{k=1}^{n} exp(s_k^i)$$

$$h_i^{attn} = \sum_{j=1}^{n} a_j^i c_j$$

$$\hat{h}_i = \text{BiRNN}(\hat{h}_{i-1}, [e_i, h_i^{attn}])$$

Reasoning Layer

column name: 与column type拼接得到$\hat{c}_j$ $\quad \bar{c}_j = \text{CFGNN}(\hat{c}_j)$

question: database-aware representation $\quad \bar{h}_i = \text{BiRNN}(\bar{h}_{i-1}, \hat{h}_i)$

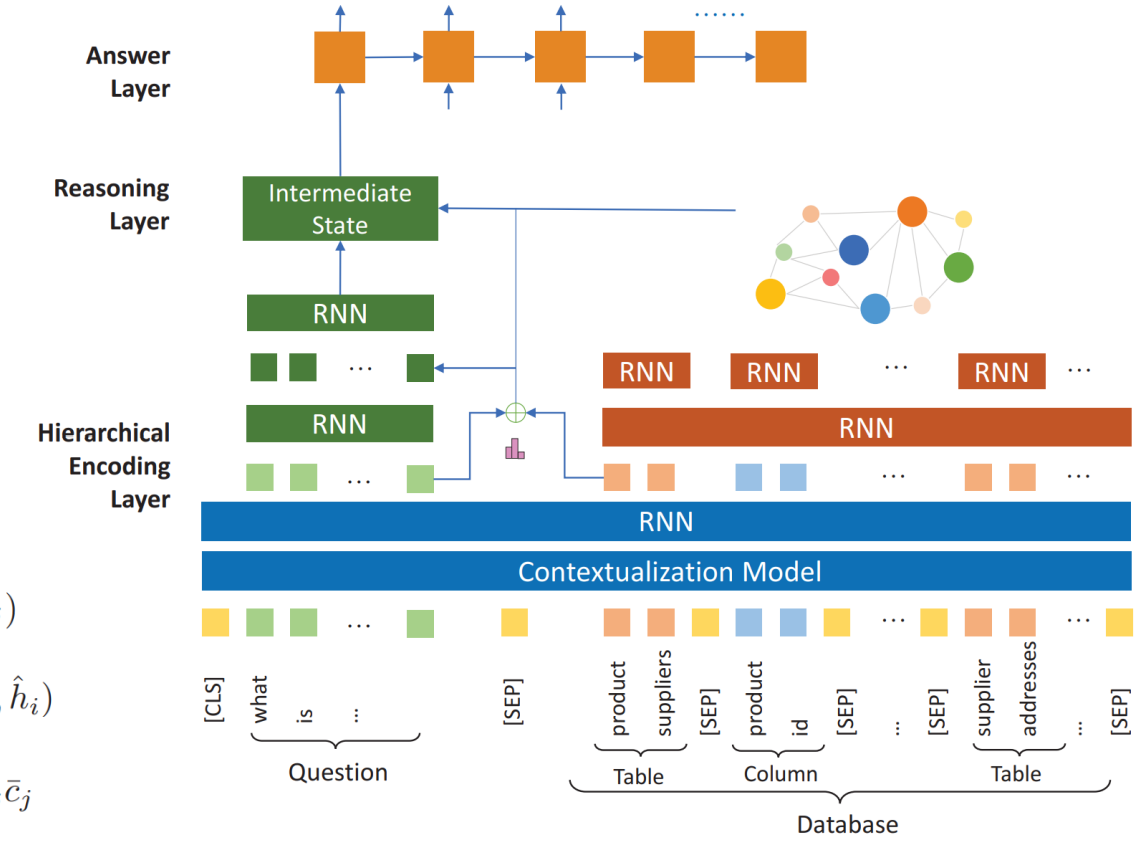
reasoning: CFGNN

$$h_i^{gnn} = \sum_{j=1}^{n} a_j^i \bar{c}_j$$

$$\tilde{h}_i = [\bar{h}_i; h_i^{gnn}]$$

Answer Layer

grammar-based LSTM



CFGNN: Cross Flow Graph Neural Networks for Question Answering on Complex Tables (AAAI 2020)

- 结果：

Method	Easy	Medium	Hard	Extra Hard	All
RCSQL	53.2%	27.0%	20.1%	6.5%	28.8%
Schema GNN	64.8%	41.6%	29.3%	15.9%	40.9%
CFGNN (ours)	**69.2%**	**50.9%**	**34.5%**	**27.6%**	**48.7%**
w/o parent-child flow	68.8%	46.8%	26.4%	21.2%	44.5%
w/o child-child flow	62.8%	44.5%	27.0%	22.9%	42.5%
w/o reasoning flow	66.8%	48.4%	27.0%	21.2%	44.8%
w/o graph structure	68.0%	47.0%	21.8%	15.9%	42.7%
replace CFGNN with GNN	58.8%	39.5%	27.0%	14.1%	37.9%
replace CFGNN with GAT	67.2%	46.8%	28.7%	20.6%	44.4%
replace CFGNN with GGNN	62.8%	39.8%	29.9%	19.4%	40.3%

Table 3: Ablation studies on accuracy of Exact Matching (dev set).

Method	SELECT	WHERE	GROUP BY	ORDER BY	KEYWORDS
RCSQL	68.7%	39.0%	63.1%	63.5%	76.5%
Schema GNN	75.0%	42.8%	71.6%	67.9%	83.1%
CFGNN (ours)	81.9%	**52.5%**	71.5%	**76.6%**	**82.9%**
w/o parent-child flow	81.1%	44.2%	70.9%	73.1%	82.6%
w/o child-child flow	78.3%	46.5%	72.8%	69.1%	82.5%
w/o reasoning flow	**82.3%**	48.6%	73.2%	65.8%	80.9%
w/o graph structure	79.9%	45.8%	72.1%	65.0%	81.0%
replace CFGNN with GNN	73.9%	42.6%	66.9%	65.8%	82.8%
replace CFGNN with GAT	80.1%	43.5%	**73.2%**	70.0%	81.9%
replace CFGNN with GGNN	71.9%	45.1%	68.8%	68.3%	82.7%

Table 4: Ablation studies on F1 scores of Component Matching (dev set).

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing (ICLR 2021)

- 动机：

  使用SCFG处理的数据、MLM、SSP对模型做预训练

- 方法：

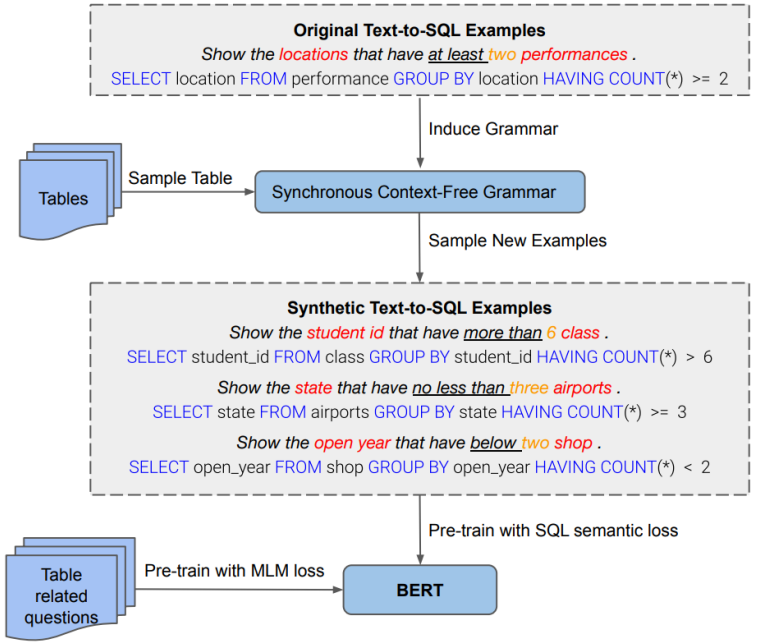
  SCFG：利用Spider数据集中的问题生成推导规则与问题模板

  手动寻找实体/短语

  应用于其他表格数据集

  MLM：对问题与表格标题随机做遮盖

  SQL semantic prediction：由Text预测需要的列与操作



Non-terminals	Production rules
TABLE → t_i	1. ROOT → ⟨"For each COLUMN0 , return how many times TABLE0 with COLUMN1 OP0 VALUE0 ?", SELECT COLUMN0 , COUNT (*) WHERE COLUMN1 OP0 VALUE0 GROUP BY COLUMN0 ⟩
COLUMN → c_i	
VALUE → v_i	
AGG → ⟨ MAX, MIN, COUNT, AVG, SUM⟩	
OP → ⟨ =, ≤, ≠, ... , LIKE, BETWEEN ⟩	
SC → ⟨ ASC, DESC ⟩	2. ROOT → ⟨"What are the COLUMN0 and COLUMN1 of the TABLE0 whose COLUMN2 is OP0 AGG0 COLUMN2 ?", SELECT COLUMN0 , COLUMN1 WHERE COLUMN2 OP0 (SELECT AGG0 (COLUMN2)) ⟩
MAX → ⟨"maximum", "the largest"...⟩	
≤ → ⟨"no more than", "no above"...⟩	
...	

总结

- **数据集**
  - 逻辑、数值推理
- **编码**
  - 不使用预训练模型：融合关键词、位置信息
  - 使用预训练模型：文本线性化、修改模型、编码后处理
- **推理**
  - 视下游任务而定
- **剪枝**
  - 编码前过滤
  - 编码后过滤
- **训练数据、方法**
- **未来方向**
  - 异构数据编码
  - 跨表/多表编码
  - 单元格为单位过滤