追一科技

打通人与结构化数据间壁垒

首届中文NL2SQL挑战赛

汇报人：张啸宇

团队：不上90不改名字
成员：张啸宇，赛斌、王苏宏

2 任务描述

代码	名称	上市地点	收盘价	周涨跌幅	月涨跌幅
SINA.O	新浪	纳斯达克	58.93	-4.52	8.791
BITA.N	易车	纽约证券交易所	18.11	-4.78	-11.742
JRJC.O	金融界	纳斯达克	1.03	-7.21	30.383
TAOP.O	淘屏	纳斯达克	1.09	-9.17	2.834
SFUN.N	搜房网	纽约证券交易所	1.71	-9.52	28.575
RENN.N	人人网	纽约证券交易所	1.61	-9.55	14.18

Query：搜房网和人人网的周涨跌幅分别是多少？

SQL：SELECT 周涨跌幅 WHERE 名称='搜房网' OR 名称='人人网'

3 数据集

- 单表转换：WikiSQL、TableQA
- 多表转换：Spider

4

WikiSQL VS TableQA

- 语种，WikiSQL：英文，TableQA：中文
- 数量，WikiSQL：8w+，TableQA：5w
- 难易，WikiSQL：简单，TableQA：较难
  - Select字段数量：1 vs [1, 2]
  - Where条件数量：以单个为主 vs 以多个为主
  - Where条件操作：AND vs [AND，OR]
  - Value标准度：存在于数据库表中 vs 形式多样

5 WikiSQL

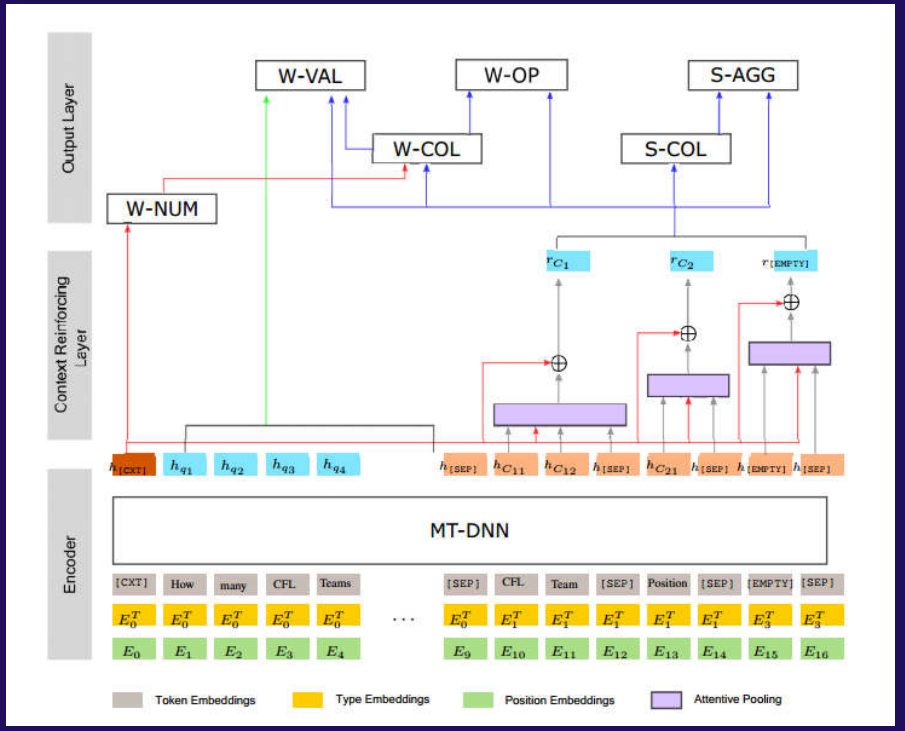
Supervised via logical forms

Model	Dev logical form accuracy	Dev execution accuracy	Test logical form accuracy	Test execution accuracy	Uses execution
X-SQL +Execution-Guided Decoding (He 2019)	86.2	92.3	86.0	91.8	Inference
SQLova +Execution-Guided Decoding (Hwang 2019)	84.2	90.2	83.6	89.6	Inference
IncSQL +Execution-Guided Decoding (Shi 2018)	51.3	87.2	51.1	87.1	Inference
Execution-Guided Decoding (Wang 2018)	76.0	84.0	75.4	83.8	Inference
X-SQL (He 2019)	83.8	89.5	83.3	88.7	
SQLova (Hwang 2019)	81.6	87.2	80.7	86.2	
IncSQL (Shi 2018)	49.9	84.0	49.9	83.7	
MQAN (unordered) (McCann 2018)	76.1	82.0	75.4	81.4	
MQAN (ordered) (McCann 2018)	73.5	82.0	73.2	81.4	

6 SOTA: X-SQL

- 模型框架：6个子任务

```
SELECT $AGG $COLUMN
WHERE $COLUMN $OP $VALUE
(AND $COLUMN $OP $VALUE)*
```



7

SOTA: X-SQL

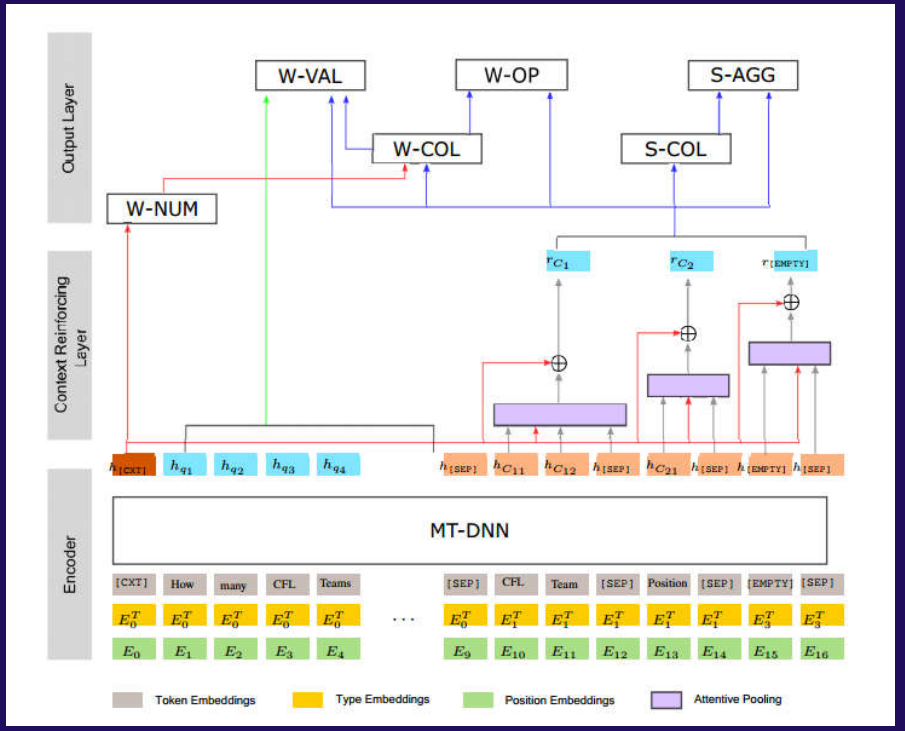
- 模型框架：6个子任务
- 缺陷：
  - 模型框架不完全适配
  - value抽取，column特征不显著，抽取混乱

$$p_{\text{start}}^{\text{W-VAL}}(q_j|C_i) = \text{SOFTMAX}\ g\left(U^{\text{start}} h_{q_j} + V^{\text{start}} r_{C_i}\right)$$

$$p_{\text{end}}^{\text{W-VAL}}(q_j|C_i) = \text{SOFTMAX}\ g\left(U^{\text{end}} h_{q_j} + V^{\text{end}} r_{C_i}\right)$$

  - query：长沙9月1号天气如何？

```
SELECT $AGG $COLUMN
WHERE $COLUMN $OP $VALUE
(AND $COLUMN $OP $VALUE)*
```

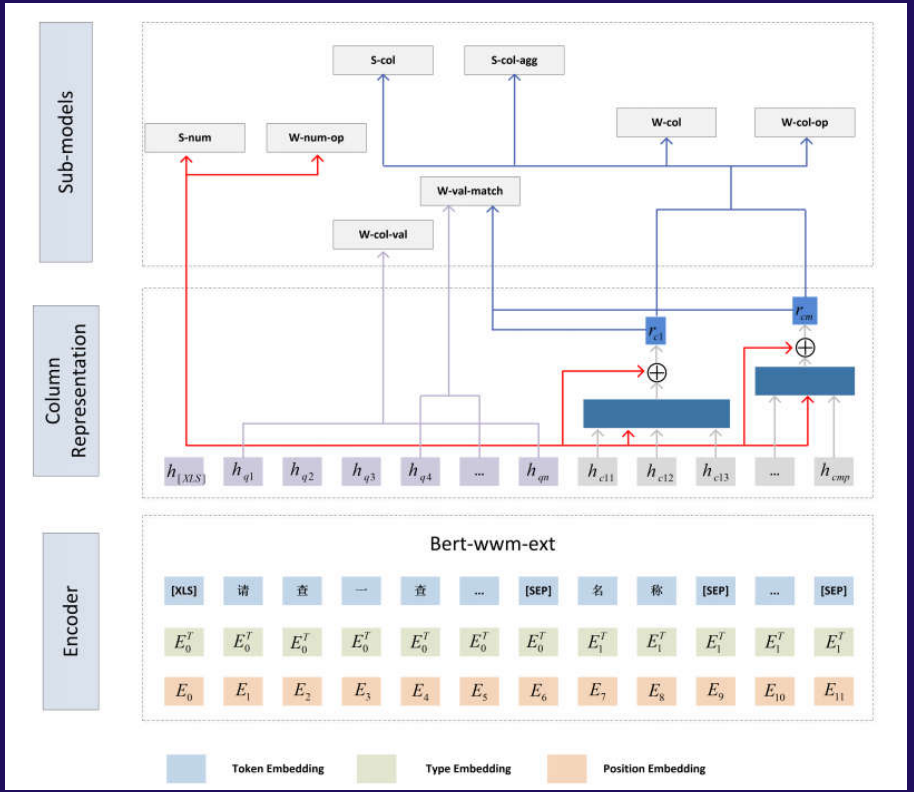


8

M-SQL

- 模型框架：8子任务
- 增加三个子模型
  - S-num
  - Value抽取
  - Value匹配
- W-num -> W-num-op
- X-SQL vs M-SQL： 72 vs 83

SELECT ($AGG $COLUMN)*
WHERE $WOP ($COLUMN $OP $VALUE)*



M-SQL: Multi-Schema Representation Learning for Single-Table Text2SQL Generation

处理细节-1

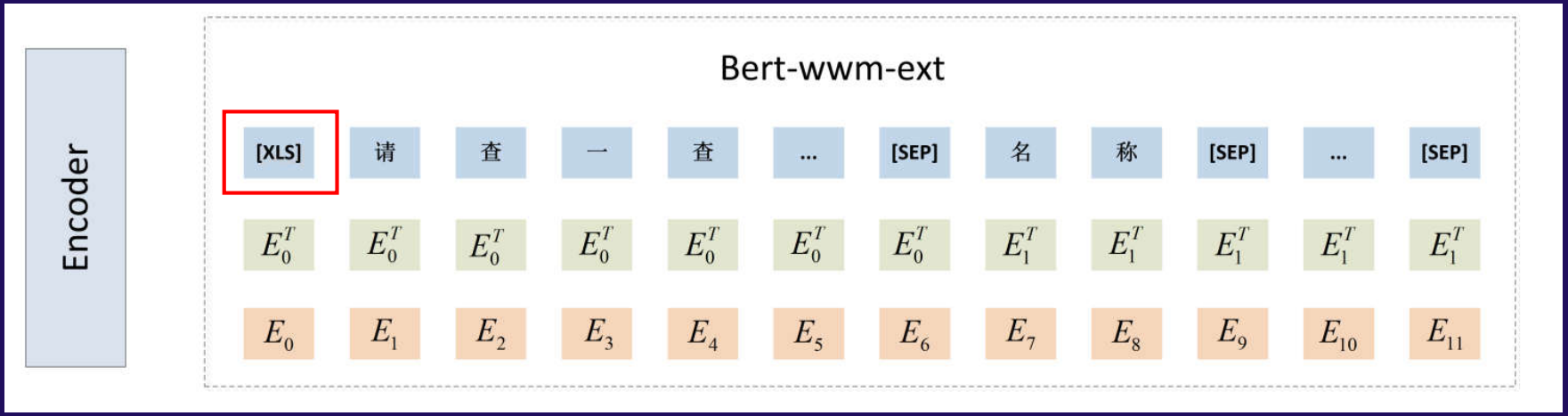
数据预处理（query）

- 数字：哪些城市上一周成交一手房超十五万平？（十五，15）
- 年份：你知道10年的土地成交面积吗？（10年，2010）
- 单位：哪些城市最近一周新盘库存超过5万套？（5万，50000）
- 日期：哪个公司于18年12月28号成立？（18年12月28号，2018/12/28）
- 同义：你能帮我算算芒果这些剧的播放量之和是多少吗？（芒果，芒果TV）
- 应对
  - 规则修正
  - 编辑距离匹配（阈值：0.6）
  - 统一query范式，修query VS 修value：0.3

10

处理细节-2

[CLS] -> [XLS]，+0.3

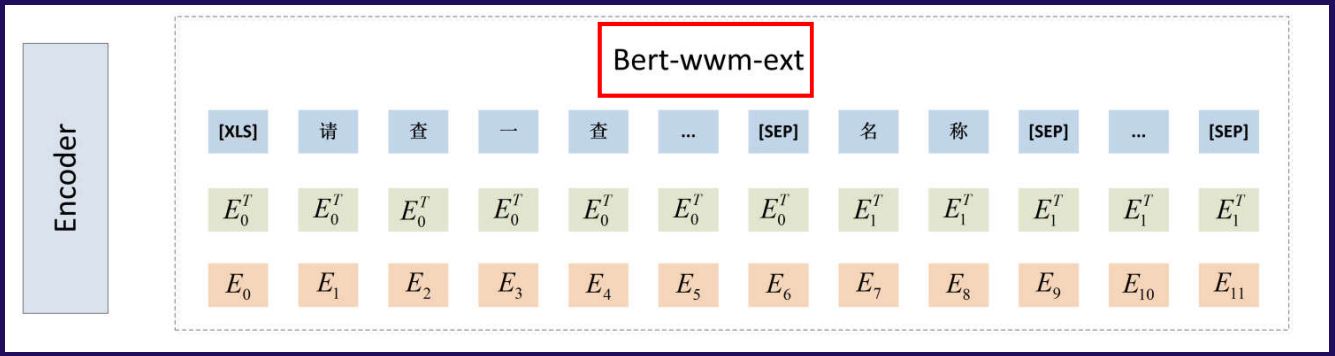


11

处理细节-3

预训练模型

- BERT-base，Google，val_acc: 0.8891
- BERT-wwm，哈工大，val_acc: 0.8892
- BERT-wwm-ext，哈工大，val_acc: 0.8922
- RoBERTa-wwm-ext，哈工大，val_acc: 0.8953

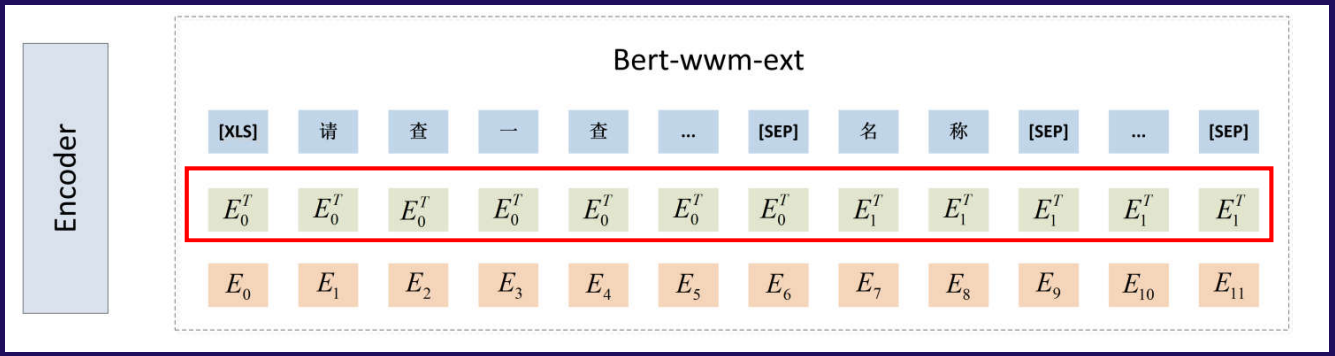


12

处理细节-4

type编码

- 不使用，val_acc: 0.8913
- 2标记，val_acc: 0.8922
- 3标记（X-SQL），不收敛

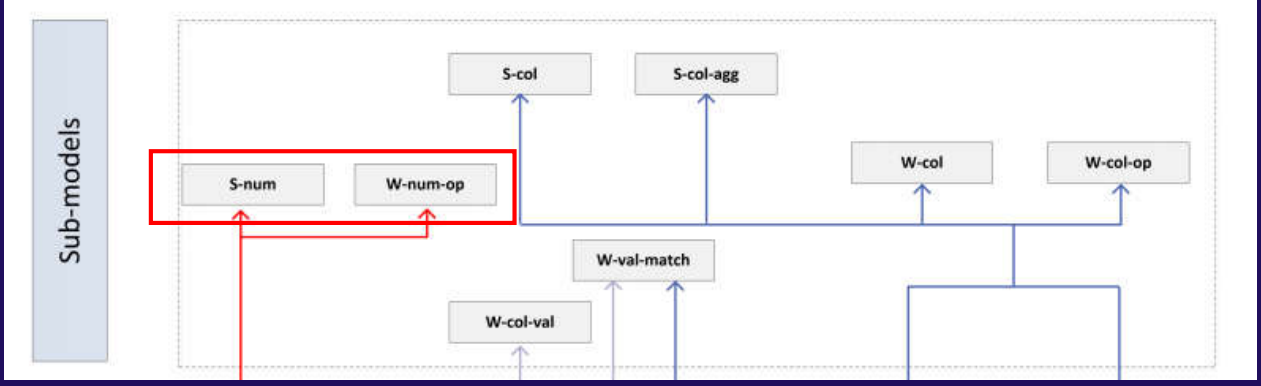


13

处理细节-5

子模型融合（S-num，W-num，W-op）

- S-num: 二分类；W-num: 四分类；W-op：三分类
- 不融合，val_acc: 0.8893
- W-num，W-op融合，七分类，val_acc: 0.8922
- S-num，W-num，W-op融合，十四分类，val_acc: 0.8812

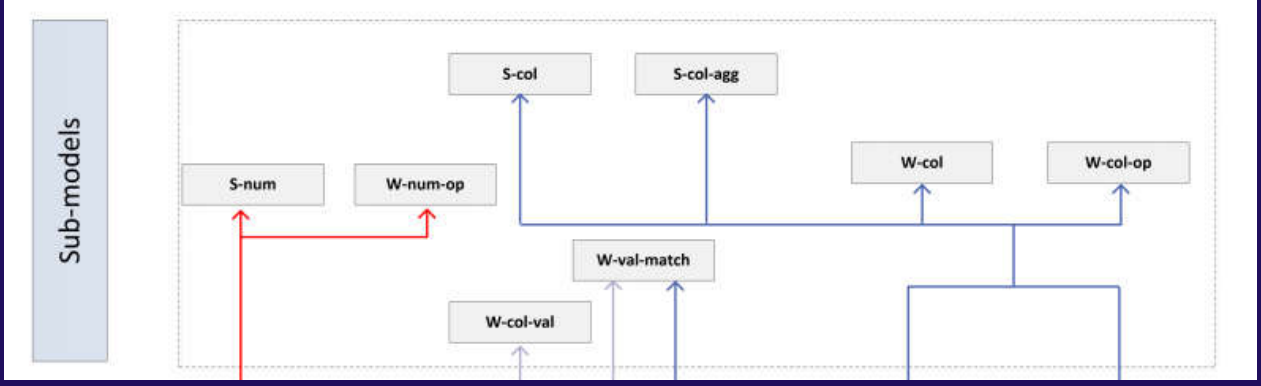


14

处理细节-6

子模型互促

- 单独优化S-num：不收敛
- 同时优化S-num、W-num-op：收敛
- 同时优化S-num、W-num-op、S-col、S-col-agg、W-col、W-col-op，val_acc: 0.9067
- 同时优化所有任务，前六项子任务，val_acc: 0.9145



处理细节-7

table-column 信息增强: column+sim(content)，+0.4

商户类型	地区	区域	商户名称	商户地址
百货类	广西	防城港	防城港港口区家惠超市	港口区兴港大道95-1号
百货类	广西	南宁	青秀南城百货	民族大道64号综合商场一楼
百货类	广西	南宁	白沙南城百货公司	南宁市白沙大道20号
…	…	…	…	…

Query：青秀南城百货有限公司在南宁的哪个位置

SQL：SELECT 商户地址 WHERE 区域='南宁' AND 商户名称='青秀南城百货'

信息增强：

地区->"地区，广西"，

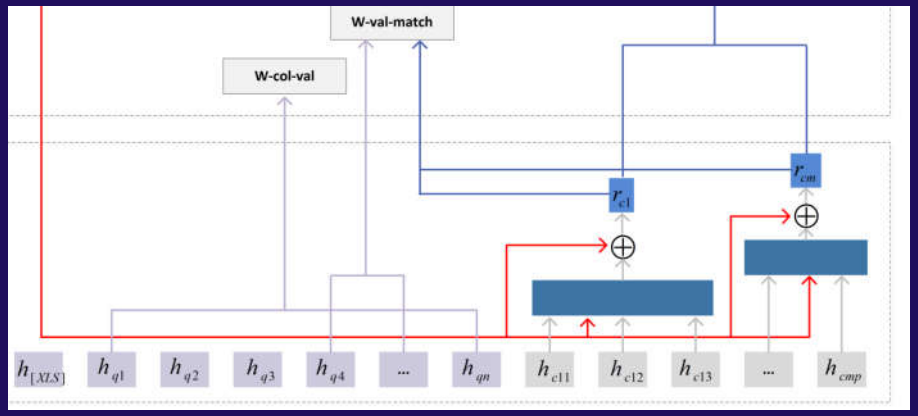
区域->"区域，南宁"

16

处理细节-8

value抽取

- bert + crf， val_acc: 0.8785
- bert + bilstm + crf， val_acc: 0.8801
- bert + 半指针， val_acc: 0.8891
- bert + 0/1标记， val_acc: 0.8922



query：青秀南城百货有限公司在哪？

bert_tokenizer：['[XLS]', '青', '秀', '南', '城', '百', '货', '有', '限', '公', '司', '在', '哪', '？', '[SEP]']

value:：青秀南城百货有限公司

tag：[0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

处理细节-9

value检索

- rouge-L 匹配：0.9213
- 机器学习(5特征)
  - Lr: 0.9772
  - SVR: 0.7755
  - Bayes: 0.9334
- 神经网络
  - 句对模型:0.9852

代码	名称	上市地点	周涨跌幅
SINA.O	新浪	纳斯达克	-4.52
BITA.N	易车	纽约证券交易所	-4.78
SFUN.N	搜房网	纽约证券交易所	-9.52
RENN.N	人人网	纽约证券交易所	-9.55
TENG.N	腾讯	香港联交所	+1.23

Query1：人人周涨跌幅是多少？
Value：人人

Query2: 企鹅在哪上市？
Value：腾讯

处理细节-10

Select-col，Where-col联合修正，+0.21

- 假定：Select-col与Where-col 不重合
- 规则1：如果S-num为2，W-num为2，依据概率进行判断
- 规则2：如果S-num小于W-num，则该字段属于Select部分

19

结果展示

	S_num	S_col	S_col_agg	W_num_op	W_col	W_col_op	W_col_value	join_acc
单模型	0.9952	0.9783	0.9893	0.9745	0.9851	0.9910	0.9697	0.8922
集成模型	0.9955	0.9838	0.9893	0.9768	0.9907	0.9925	0.9701	0.9051

第二赛季 第一赛季

排名	参与者	组织	score	最优成绩提交日
1	不上90不改名字	国防科大	0.8964	2019-08-05
2	BugCreater	国双科技	0.8902	2019-08-02
3	大佬带我飞	华南理工大学	0.8820	2019-08-03
4	guoxz	Jia Li Dun University	0.8704	2019-07-29
5	老哥们不放假吗	浙江大学	0.8659	2019-08-06
6	Model S	观安信息/MioTech	0.8609	2019-07-31

第二赛季 第一赛季

排名	参与者	组织	score	最优成绩提交日
1	不上90不改名字	国防科大	0.9219	2019-09-09
2	BugCreater	国双科技	0.9143	2019-09-08
3	老哥们不放假吗	浙江大学	0.9106	2019-09-07
4	大佬带我飞	华南理工大学	0.9076	2019-09-04
5 ↑1	Model S	观安信息/MioTech	0.9065	2019-09-09
6 ↑1	爆写规则一万行	华中科技大学	0.9018	2019-09-09

Discussion(应用)

- 有无答案
- 内容编码
- 语义匹配
- 多表转换

21

Discussion(应用)

- 有无答案
- 内容编码
- 语义匹配
- 多表转换

代码	名称	上市地点	收盘价	周涨跌幅
SINA.O	新浪	纳斯达克	58.93	-4.52
BITA.N	易车	纽约证券交易所	18.11	-4.78
JRJC.O	金融界	纳斯达克	1.03	-7.21
TAOP.O	淘屏	纳斯达克	1.09	-9.17
SFUN.N	搜房网	纽约证券交易所	1.71	-9.52
RENN.N	人人网	纽约证券交易所	1.61	-9.55

Query：阿里巴巴的周涨跌幅是多少？

22

Discussion(应用)

- 有无答案
- 内容编码
- 语义匹配
- 多表转换

商户类型	地区	区域	商户名称
百货类	广西	防城港	防城港港口区家惠超市
百货类	广西	南宁	青秀南城百货
百货类	广西	南宁	白沙南城百货公司
…	…	…	…

Query：青秀南城百货有限公司在南宁的哪个位置

信息增强：

地区->“地区，广西”，

区域->“区域，南宁”

23

Discussion(应用)

- 有无答案
- 内容编码
- 语义匹配
- 多表转换

代码	名称	上市地点	周涨跌幅
SINA.O	新浪	纳斯达克	-4.52
BITA.N	易车	纽约证券交易所	-4.78
SFUN.N	搜房网	纽约证券交易所	-9.52
RENN.N	人人网	纽约证券交易所	-9.55
TENG.N	腾讯	香港联交所	+1.23

Query2: 企鹅在哪上市？

24

Discussion(应用)

- 有无答案
- 内容编码
- 语义匹配
- 多表转换

Spider 1.0

Yale Semantic Parsing and Text-to-SQL Challenge

What is Spider?

Spider is a large-scale *complex and cross-domain* semantic parsing and text-to-SQL dataset annotated by 11 Yale students. The goal of the Spider challenge is to develop natural language interfaces to cross-domain databases. It consists of 10,181 questions and 5,693 unique complex SQL queries on 200 databases with multiple tables covering 138 different domains. In Spider 1.0, different complex SQL queries and databases appear in train and test sets. To do well on it, systems must *generalize well to not only new SQL queries but also new database schemas.*

Why we call it "Spider"? It is because our dataset is complex and cross-domain like a spider crawling across mutiple complex(with many foreign keys) nests(databases).

Spider Paper (EMNLP'18)

Spider Post

SParC, the context-dependent version of the Spider task, introduces a new Semantic Parsing in Context challenge.

SParC Challenge (ACL'19)

Leaderboard - Exact Set Match without Values

For exact matching evaluation, instead of simply conducting string comparison between the predicted and gold SQL queries, we decompose each SQL into several clauses, and conduct set comparison in each SQL clause. Please refer to the paper and the Github page for more details.

Rank	Model	Dev	Test
1 June 24, 2019	TPNet + BERT *Anonymous*	**63.9**	**55.0**
2 Sep 20, 2019	GIRN + BERT *Anonymous*	60.2	54.8
3 May 19, 2019	IRNet + BERT *Microsoft Research Asia* (Guo and Zhan et al., ACL '19) code	61.9	54.7
4 Sep 19, 2019	RATSQL *Anonymous*	60.6	53.7
5 Sep 1, 2019	EditSQL + BERT *Yale University & Salesforce Research* (Zhang et al., EMNLP '19) code	57.6	53.4
6 June 24, 2019	TPNet *Anonymous*	55.4	48.5

追一科技

Thank you

汇报人：张啸宇

2019.10.12