M-SQL: Multi-Task Representation Learning for Single-Table Text2sql Generation

亮点：

（1）论文是国内第一届Text2SQL比赛中的冠军方案

（2）TableQA数据集比传统的WikiSQL数据集更复杂，现有的模型无法很好地解决TableQA，而M-SQL可以。

（3）论文把任务分成8个子任务，使用了基于预训练的BERT来建模，是一个多任务学习模型。

处理Text2SQL的两种方案

- Seq2seq，比如
  - Seq2SQL：把任务视为一个文本转SQL的翻译任务。encoder用来编码文本信息获取语义表示，decoder用来解码文本的语义表示来生成SQL。没有考虑SQL的语法结构，准确率低
- Sketch-based
  - SQL有固定的结构，只需要预测关键部分填入模板即可。
  - SQLNet：把任务分成6个子任务，每个子任务预测模板的一部分。条件值的预测是seq2seq模型，其它部分是分类模型
  - SQLova and X-SQL也用了类似的任务分解，并引入了预训练BERT。基本上能够解决WikiSQL数据集。

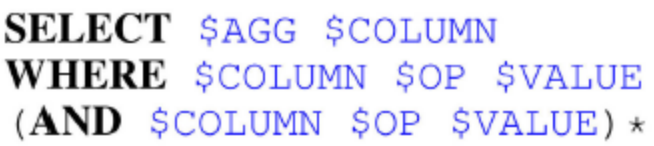


FIGURE 1. Sketch-based approach.

WikiSQL相比现实应用场景有很多简化的地方

- 假设select的列只能是1个
- 多条件的样本很少
- 假设where条件之间的关系只能是AND，不考虑OR
- 假设数据库的内容一定会出现在query中。

Table

Player	No.	Nationality	Position	Years in Toronto	School/Club Team
Antonio Lang	21	United States	Guard-Forward	1999-2000	Duke
Voshon Lenard	2	United States	Guard	2002-03	Minnesota
Martin Lewis	32, 44	United States	Guard-Forward	1996-97	Butler CC(KS)
Brad Lohaus	33	United States	Guard-Center	1996	Iowa
Art Long	42	United States	Guard-Center	2002-03	Cincinnati

Sample data

Situation 1: WikiSQL: Who is the player that wears number 42? ComplexSQL: Who is the player that wears number 42, and which country?
Situation 2: WikiSQL: Who was born in 1996? ComplexSQL: Who was born in 1996 and played for the Duke?
Situation 3: WikiSQL: Who was born in 1996 and played for the Duke? ComplexSQL: Who was born in 1996 or played for the Duke?
Situation 4: WikiSQL: Which player's nationality is United States? ComplexSQL: Who is the American player?

FIGURE 2. WikiSQL data description.

追一科技的Text2SQL比赛数据集TableQA中有更复杂的情况，包括了上述四种。

Data 1:

Query_en: The average daily volume of ChangSha in 2011 was 3.17, so what is the volume in the past week?

Query_zh: 长沙 2011 年平均每天成交量是 3.17，那么近一周的成交量是多少

SQL_en: SELECT 'Seven days trading' WHERE 'City' == ChangSha AND 'Daily trading' == 3.17

SQL_zh: SELECT '七日成交' WHERE '城市' == 长沙 AND '每日成交' == 3.17

Data 2:

Query_en: Please check the situation of the weekly fluctuations of SouFang and RenRen

Query_zh: 请查一查搜房网和人人网的周涨跌幅的情况

SQL_en: SELECT 'Weekly Fluctuation' WHERE 'Name == SouFang Or 'Name' == RenRen

SQL_zh: SELECT '周涨跌幅' WHERE '名称' == 搜房网 AND '名称' == 人人网

FIGURE 3. **TableQA data samples.**

SQLova and X-SQL不能很好地处理TableQA数据集

- TableQA需要两个额外的子任务：预测select的列数量，预测where条件之间的关系
- 现有的模型是基于column representation抽取值的，如果query中有多个值，且这些值属于不同的列，模型就不能准确地抽取值了。M-SQL把这个任务分成两个部分：value extraction和value-column matching.
- TableQA的query形式更加随意，且数据库的内容不一定出现在query中。

提出M-SQL，有8个子模型：S-num,S-col, S-col-agg,W-num-op,W-col,W-col-op,W-col-val and W-cal-match

问题定义

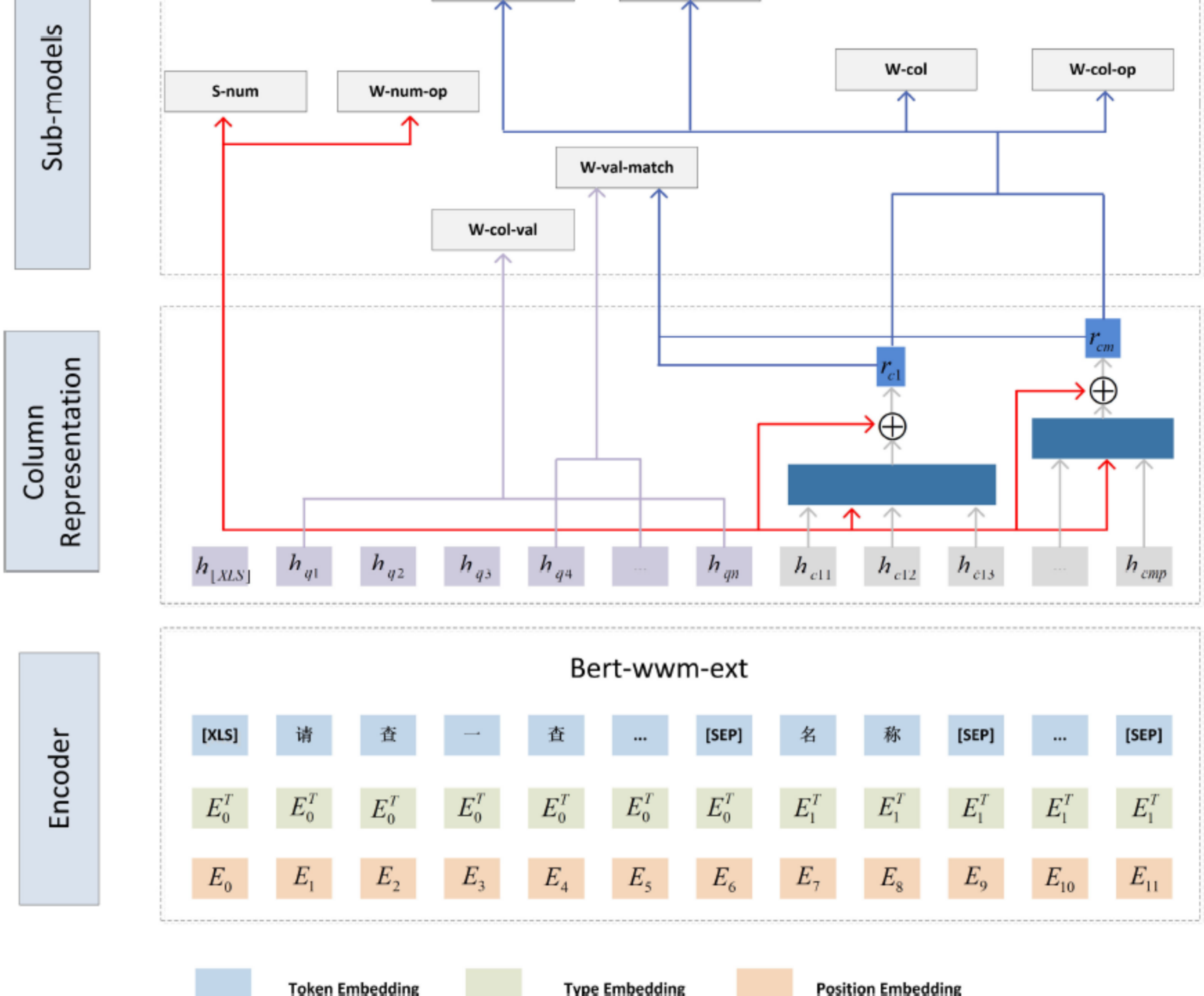
- SQL模板如图所示
- 假设每个SQL都有SELECT和WHERE条件
- $WOP是条件列之间的关系，["", "AND", "OR"]，空值表示没有关系
- $COLUMN是数据库的列名，这里分为selected column和conditional column
- $AGG表示selected column的操作，["", "AVG", "MAX", "MIN", "COUNT", "SUM"]
- $OP是条件列的操作符，[">", "<", "==", "! ="]
- $VALUE是条件列对应的值，如果是字符串类型的话，必须存在于数据库中
- *代表数量，这里假设select的列的数量可以是[1, 2]，条件列的数量可以是[1, 2, 3]

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

FIGURE 4. TableQA sketch.

M-SQL包括三个部分：encoder，column representation和几个sub-models。

- encoder用了BERT-wwm-ext treats the Chinese word as a masking unit（而不是character）
- use the ``CONTENT REINFORCING LAYER'' in X-SQL as the column semantic representation
- 8个子模型：S-num（二类分类，[1, 2]）,S-col（单分类，是否选择）, S-col-agg（预测select列的聚合函数）, W-num-op（预测条件列之间的关系和条件列的数量，7类分类）,W-col（预测条件列，单分类，是否选择）,W-col-op（预测条件列的操作符）,W-col-val（从query中抽取条件列的值） and W-cal-match（match条件列和抽取的值）


Sub-models
S-num
W-num-op
W-col
W-col-op
W-val-match
W-col-val
Column Representation
r_{cm}
r_{c1}
$h_{[XLS]}$
h_{q1}
h_{q2}
h_{q3}
h_{q4}
...
h_{qn}
h_{c11}
h_{c12}
h_{c13}
...
h_{cmp}
Encoder
Bert-wwm-ext
[XLS]
请
查
一
查
...
[SEP]
名
称
[SEP]
...
[SEP]
E_0^T
E_1^T
E_0
E_1
E_2
E_3
E_4
E_5
E_6
E_7
E_8
E_9
E_{10}
E_{11}
Token Embedding
Type Embedding
Position Embedding

Encoder

- 输入[XLS], T1, T2, … , TL, [SEP], H11, H12, … , [SEP], … , [SEP], Hn1, Hn2, … , [SEP]

COLUMN REPRESENTATION

- 用全局信息xls来加强每列的语义表示（attention）

The attention weights about the global information $h_{[XLS]}$ and the t-th column are as follows:

$$s_{ti} = dot(Uh_{[XLS]}, Vh_{cti}) \quad (2)$$

$$a_{ti} = \frac{s_{ti}}{\sum_{j=1}^{n_t} s_{tj}} \quad (3)$$

Both $U, V \in R^{d \times d}$, dot is the dot product. n_t is the number of tokens in the t-th column. s_{ti} is the similarity between $h_{[XLS]}$ and the i-th token in the t-th column. a_{ti} is the attention weight of the i-th token in the t-th column.

The representation of the t-th column is:

$$\overline{r}_{ct} = \sum_{i=1}^{n_t} a_{ti} h_{cti} \quad (4)$$

$$r_{ct} = \overline{r}_{ct} + h_{[XLS]} \quad (5)$$

where n_t is the length of the t-th column. The final column representation r_{ct} is obtained by adding $\overline{r}_{ct}$ and $h_{[XLS]}$.

因为h[XLS]是S-num and W-num-op这两个子任务的输入，这里加上h[XLS]可以在子任务之间建立关联，提高多任务学习的能力

SUB-TASK OUTPUT

S-num（预测Select的列，二类分类，`[1，2]`）

W-num-op（预测条件列之间的关系和条件列的数量，7类分类，`["null-1"，"OR-1"，"AND-1"，"OR-2"，"AND-2","OR-3"，"AND-3"].`）

$$p_1 = sigmoid(W_1 h_{[XLS]}) \quad (6)$$

$$p_2 = softmax(W_2 h_{[XLS]}) \quad (7)$$

S-col（单分类，是否选择） and W-col（预测条件列，单分类，是否选择）。The probability that the i-th column belong to the target column is as follows.

$$p_3 = sigmoid(W_3 r_{ci}) \tag{8}$$

$$p_4 = sigmoid(W_4 r_{ci}) \tag{9}$$

S-col-agg（预测select列的聚合函数，`["", "AVG", "MAX", "MIN", "COUNT", "SUM"]`）

W-col-op（预测条件列的操作符，`[">", "<", "==", "! ="]`）

$$p_5 = softmax(W_5 r_{ci}) \tag{10}$$

$$p_6 = softmax(W_6 r_{ci}) \tag{11}$$

W-col-val（从query中抽取条件列的值，是否抽取query中的第i个token，此时不考虑column representation）

$$p_7 = sigmoid(W_7 h_{qi}) \tag{12}$$

W-val-match（match条件列和抽取的值，如果是匹配的value和column，则标签是1，否则是0，s和e代表抽取的值v的起始位置和结束位置）

$$h_v = \frac{\sum_{i=s}^{e} h_{qi}}{l} \tag{13}$$

$$match_i = sigmoid(u \cdot tanh(W_8 h_v + W_9 rc_i)) \tag{14}$$

where h_v is the value representation. $match_i$ is the match score about the extracted value and the i-th conditional column. W_8, W_9 and u are learnable parameters. W_8 and $W_9 \in R^{d \times d}$. $u \in R^{1 \times d}$. l represents the length of the extracted value span.

用execution-guided decoding strategy删掉那些不合理的SQL。

the highest probability as the output. There are some restrictions on the construction of SQL statements, such as string-type column cannot have numeric operations(<, >). So we use the execution-guided decoding strategy [17] to remove unreasonable SQL statements from the candidate SQLs in the SQL generation stage. In **select** clause, we assume that when the selected column is string-type, the aggregation operator cannot be the numeric operator, such as SUM, MIN, MAX. Similarly, in **where** clause, we assume that, when the conditional column is string-type, the aggregation operator cannot be the numeric operator(>, <). Through data analysis, we find that the selected columns and conditional columns are not coincident. We view this discovery as a filtering rule. We filter the SQL candidates which do not meet the above rules, and select the SQL statement with the highest join probability as the final output.

实验

训练，开发，测试：41,522, 2,198 and 2,198 respectively.

batch size is 32 and the lr is 2e-5

评估指标：Logical-form accuracy(LX)，Execution accuracy(X)，Mean accuracy(MX)（前面两个的均值）.

TABLE 1. The performance of various models on TableQA.

Model	Dev LX(%)	Dev X(%)	Dev MX(%)	Test LX(%)	Test X(%)	Test MX(%)
SQLNet [6]	61.28	66.20	63.74	61.42	67.24	64.33
Coarse2Fine [8]	72.98	76.89	74.94	72.61	76.71	74.66
MQAN [9]	75.66	79.21	77.44	74.84	78.75	76.80
SQLova [10]	81.39	85.26	83.33	81.71	85.76	83.74
X-SQL [11]	82.85	86.99	84.92	83.30	87.58	85.44
M-SQL(ours)	**89.13**	**91.86**	**90.50**	**89.31**	**92.13**	**90.72**
M-SQL-Ens(ours)	**90.54**	**93.40**	**91.97**	**90.49**	**93.31**	**91.90**

TABLE 2. The performance of sub-tasks on TableQA test data.

	S-num(%)	S-col(%)	S-col-agg(%)	W-num-op(%)	W-col(%)	W-col-op(%)	W-col-value(%)	Test LX(%)
M-SQL(ours)	99.50	97.82	98.91	97.45	98.50	99.10	96.95	89.31
M-SQL-Ens(ours)	99.55	98.36	98.91	97.68	99.09	99.27	97.00	90.49

ABLATION STUDY消融实验

TABLE 3. The results of ablation study.

Model	Test LX(%)	Test X(%)	Test MX(%)
M-SQL	89.31	92.13	90.72
– BERT-wwm-ext + BERT-base	88.90	91.45	90.18
– [XLS] + [CLS]	88.90	91.63	90.27
– 2-type	89.13	91.81	90.47
– 2-type + 3-type	\	\	\
– enhance	88.81	91.63	90.22
– BERT-0/1 + BERT-CRF	87.85	90.63	89.24
– BERT-0/1 + BERT-BILSTM-CRF	87.99	90.63	89.31
– BERT-0/1 + BERT-pointer	88.90	91.81	90.36
– lr + rouge-L	85.90	88.49	87.20
– lr + svr	74.75	77.02	75.89
– lr + bayes	86.35	88.67	87.51

2-type是指区分query和column的，3-type是区分query, string-type column and real-type column，没有收敛

enhance：加入数据库的内容来加强列的representation，使得区分度更强。

- 对于每一列，选择和query最相似的cell，把column和cell content concat起来作为一个新的representation
- rouge-L作为相似度计算函数，阈值为0.6
- the "region" column can be enhanced to "region, Nanning"

Chinese version

商户类型	地区	区域	商户名称	地址
百货	广西	防城港	防城港港口区家惠超市	兴港大道 95-1 号
百货	广西	南宁	青秀南城百货	民族大道 64 号
百货	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 地址 WHERE 商户名称=青秀南城百货 AND 区域=Nanning

Answer：民族大道 64 号

English version

Type	Area	Region	Name	Address
Merchandise	Guangxi	Fangchenggang	Jiahui Supermarket in Fangchenggang Port Area	No. 95-1 Xinggang Avenue
Merchandise	Guangxi	Nanning	Qingxiu Nancheng Department Store	No. 64 Minzu Avenue
Merchandise	Guangxi	Nanning	Baisha Nancheng Department Store	No. 20 Baisha Avenue, Nanning
……	……	……	……	……

QUERY：Where is Qingxiu Nancheng Department Store in Nanning?

SQL：SELECT Address WHERE Name= Qingxiu Nancheng Department Store AND Region=Nanning

Answer：No. 64 Minzu Avenue

FIGURE 6. A typical data sample in TableQA.

抽取value的方法

- 看做序列标注问题，用CRF，BILSTM
- 预测start和end位置，pointer
- 0/1标注

抽取的值可能和数据库中的不一致，需要矫正

- 规则匹配+rouge-L计算相似度选择最高相似度的
- 通过机器学习方法，利用统计特征从数据库里选择匹配的value
  - 统计特征：抽取值和数据库值之间的rouge-L recall和precision。query和数据库值之间的rouge-L recall和precision，抽取值和数据库值之间的共现字符数
  - lr，svr（支持向量），bayes

rouge-L：分子是X和Y的最长公共子序列的长度，分母是m和n分别代表是Recall和Precision。m,n分别表示参考摘要和自动摘要的长度（一般就是所含词的个数）

Q&A