AST Based Sequence to Sequence Natural Language Question to SQL Method

Kai Yang, Tao Yu, Professor Dragomir Radev

Problem Description - Dataset

- WikiSQL: very simple SQL queries, only contains select … from table where …
- Atis/Geoquery/Scholar
  - Same SQL queries show multiple times in dataset
  - Same SQL queries show in both training and test dataset
  - All SQL queries is corresponding to only one database

Problem Description - Evaluation Methods

- SQL queries exactly match
  - order-insensitive in select clause
  - Extra space or bracket does not change effectiveness of the queries
- SQL queries execution result match
  - Different SQL queries could generate same query result (like empty set)

Data

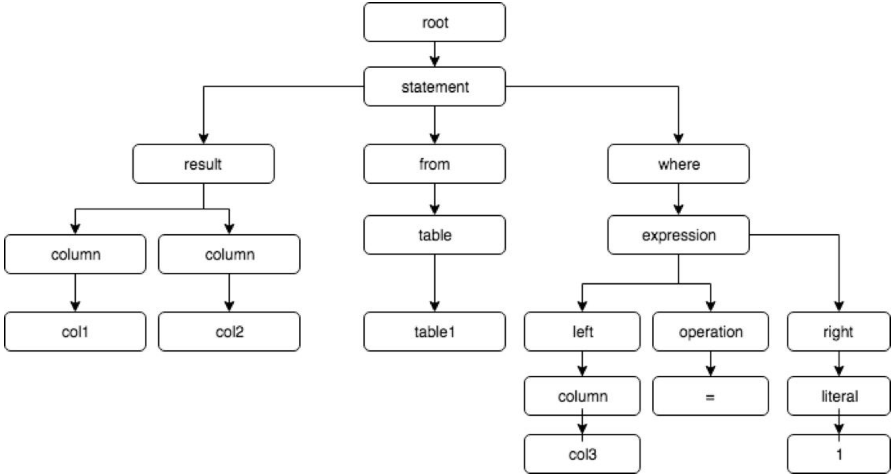
- Prepared by LILY Project members
- About 60 Database
- About 5000 natural language and SQL query pairs

Evaluation

- Parse golden queries and generated queries into syntax tree
- Separate SQL queries into several components
- For each component, compare the different in tree structure of golden queries and generated queries

Approach

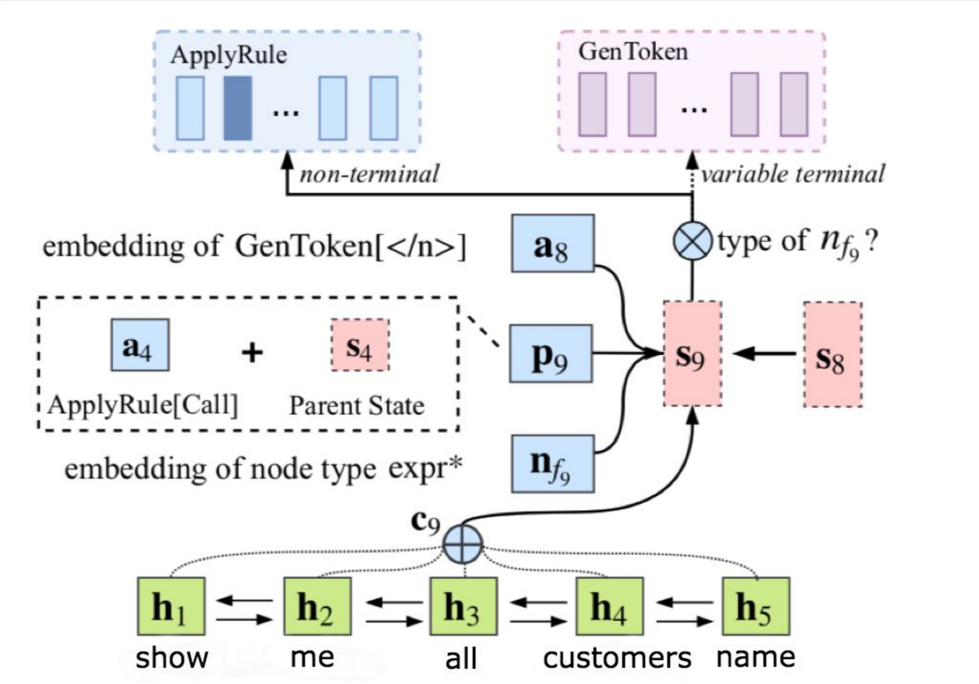
Parse train SQL into AST and generate a sequence of ordered grammar rules (from top to down and from left to right)



```
root->statement
statement->result,from,where
result->column,colum
column->col1
column->col2
from->table
table->table1
where->expression
expression->left,operation,right
left->column
column->col3
operation->=
right->literal
literal->1
```

select col1,col2 from table1 where col3=1;

Approach



$$s_t = f_{LSTM}([a_{t-1} : c_t : p_t : n_t], s_{t-1})$$

Result

- Can only predict column name that shows in training data
- Could not predict complex SQL queries (nested, compound clause)

SQL Components	Accuracy
select	0.216
select_without_agg	0.140
select_agg	0.424
where_expression	0.029
where_operator	0.183
where_nested	0.040
group	0.047
order by	0.250
compound	0.000

Conclusion and Future Work

- We proposed a new dataset and evaluation method
- The model served as a baseline of our NL2SQL task
- The capacity of this model is limited
- We could make progress to improve the result