Introduction to NL2SQL

Zecheng Zhan

2019/09/08

CONTENT

What

Why

How

END

What is NL2SQL

Semantic Parsing : NL2SQL

- 输入：[自然语言，要查询的schema]
- 输出：SQL语句

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	Forward-Center	1996	Iowa
Art Long	42	United States	Forward-Center	2002-03	Cincinnati

Question:

Who is the player that wears number 42?

SQL:

```
SELECT player
WHERE no. = 42
```

Result:

Art Long

Figure 1: An example of the WikiSQL task.

"SQLNet: GENERATING STRUCTURED QUERIES FROM NATURAL LANGUAGE WITHOUT REINFORCEMENT LEARNING" 2017

How to NL2SQL

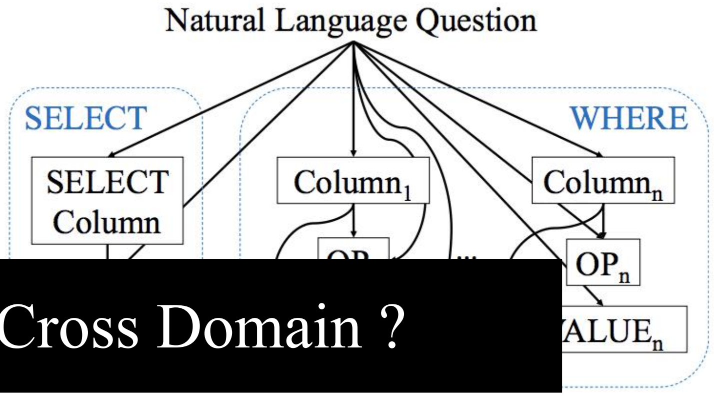
Implication: NL2SQL

- Sketch-Based
    - SQLNet
    - TypeSQL
    - IncSQL
    - SQLova
- Generated-Based
    - Seq2SQL
    - SyntaxSQLNet
    - Coares2Fine
    - IRNet

How to NL2SQL: Sketch-Based

Implication: NL2SQL:Sketch-Based

- Sketch-Based
  - SQLNet
  - TypeSQL
  - IncSQL
  - SQLova



What about Complex SQL with Cross Domain ?

(a) SQL Sketch (b) Graphical illustration of the dependency in a sketch

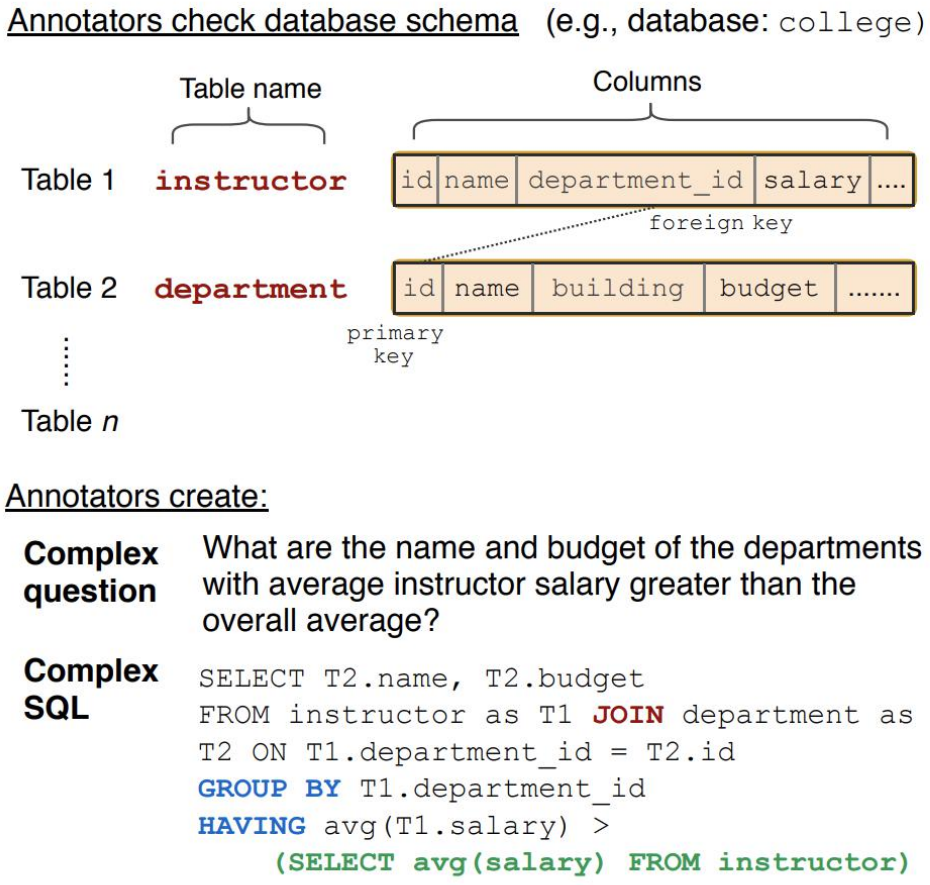
Figure 2: Sketch syntax and the dependency in a sketch

$$R\left(q\left(y\right), q_g\right) = \begin{cases} -2, & \text{if } q\left(y\right) \text{ is not a valid SQL query} \\ -1, & \text{if } q\left(y\right) \text{ is a valid SQL query and executes to an incorrect result} \\ +1, & \text{if } q\left(y\right) \text{ is a valid SQL query and executes to the correct result} \end{cases}$$

How to NL2SQL: Spider DateSet

Yale Semantic Parsing and Text-to-SQL Challenge

- Complex SQL queries
  - Different SQL Components
    - GroupBy;OrderBy;Limit;Having
  - Nested Query(Sub-query)
- Cross Domain Databases
  - Multi-Table

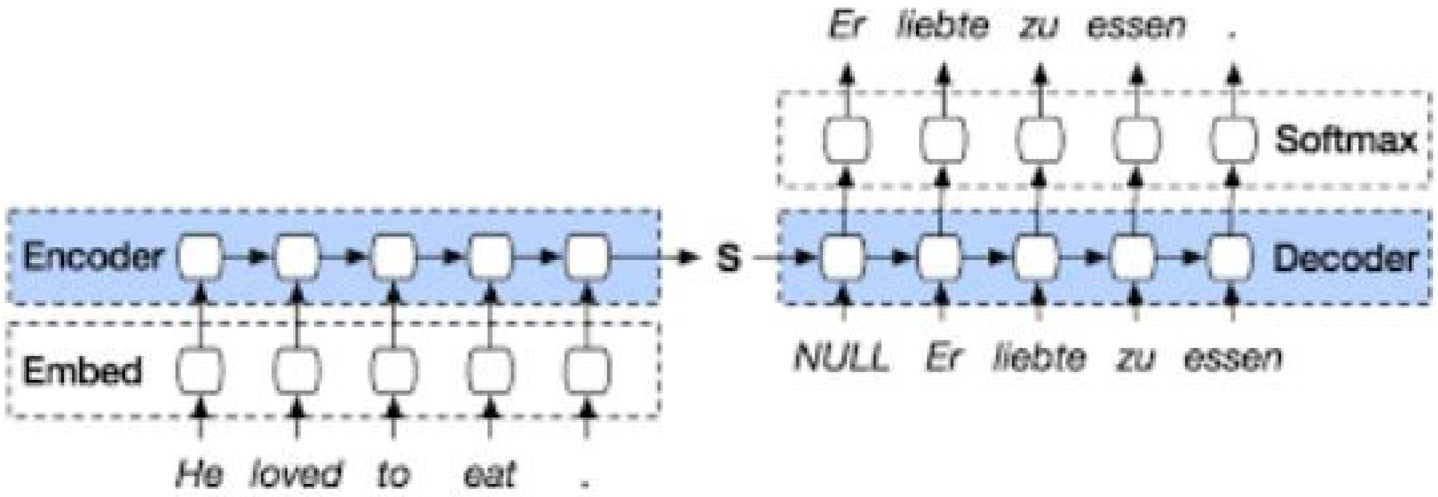


"Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task" EMNLP'18

How to NL2SQL: Generated-Based

Implication: NL2SQL:Generated-Based

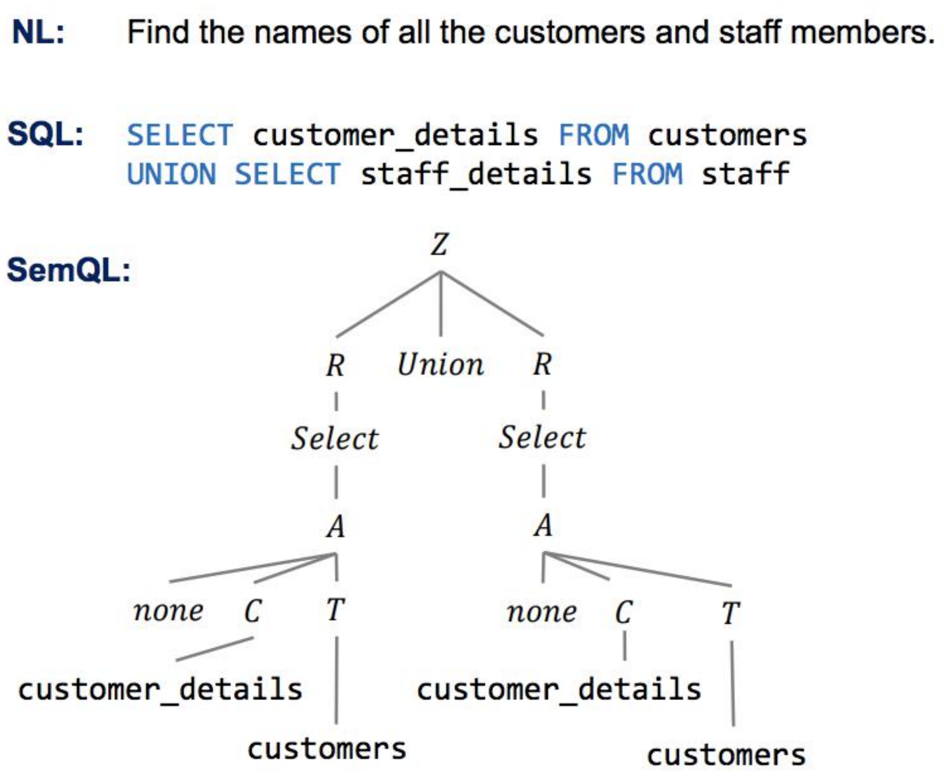
- 符合SQL语法规范
  - Vanilla Seq2Seq [X]
  - Seq2Tree (Search Constrained Seq2Seq)
- 如何处理 Cross-domain
  - Schema Linking

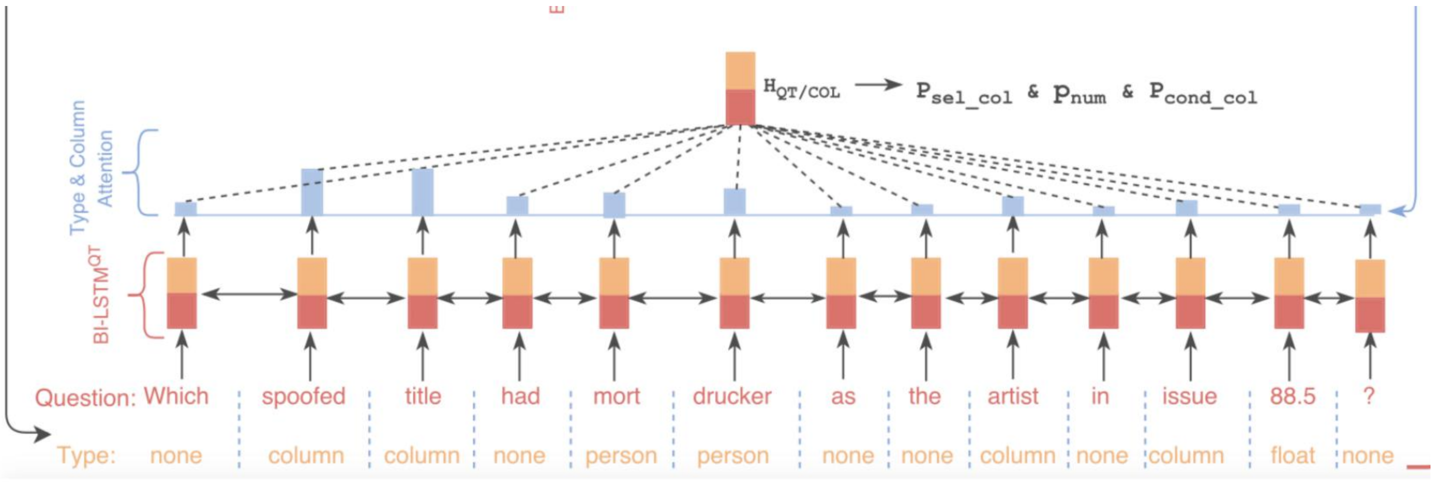


How to NL2SQL: Generated-Based

Implication: NL2SQL:Generated-Based

- 符合SQL语法规范
  - Vanilla Seq2Seq [X]
  - Seq2Tree (Search Constrained Seq2Seq)
- 如何处理 Cross-domain
  - Schema Linking





How to NL2SQL: Generated-Based

AMC
AI ML Club

Implication: NL2SQL:Generated-Based

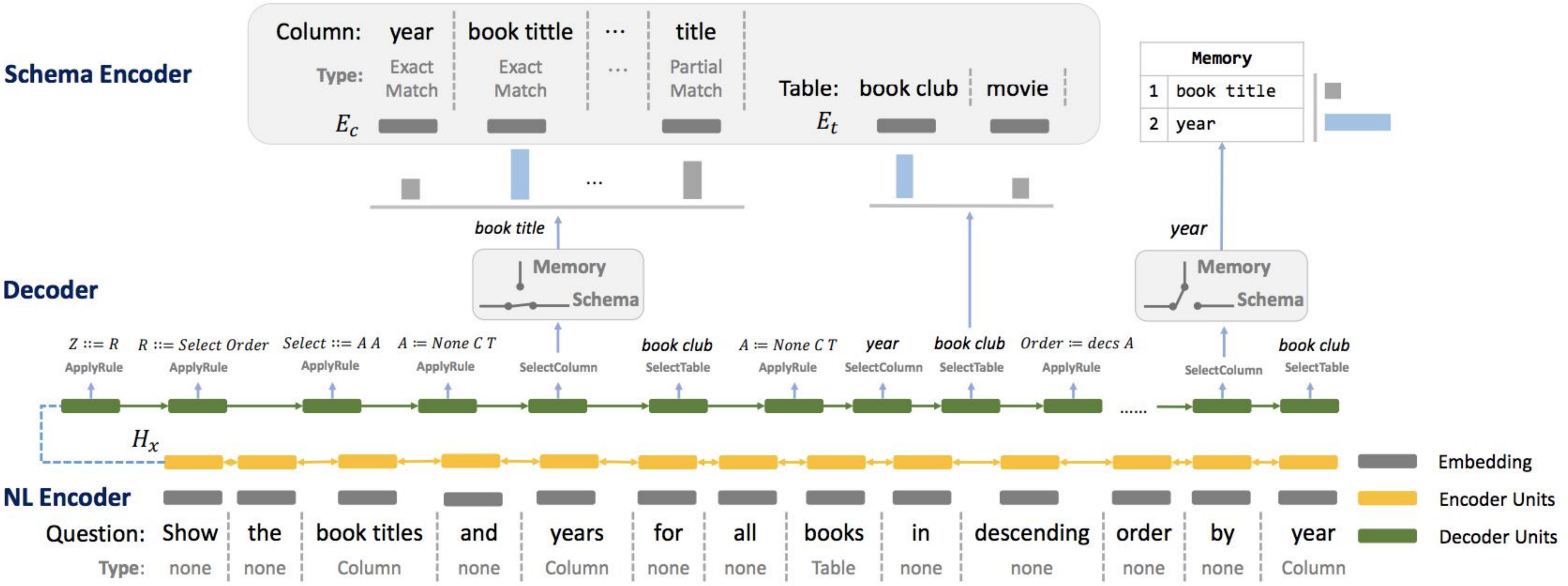


Figure 4: An overview of the neural model to synthesize SemQL queries. Basically, IRNet is constituted by an NL encoder, a schema encoder and a decoder. As shown in the figure, the column 'book title' is selected from the schema, while the second column 'year' is selected from the memory.

References

1. A Syntactic Neural Model for General-Purpose Code Generation
2. Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation
3. SEQ2SQL: GENERATING STRUCTURED QUERIES FROM NATURAL LANGUAGE USING REINFORCEMENT LEARNING
4. SQLNet: GENERATING STRUCTURED QUERIES FROM NATURAL LANGUAGE WITHOUT REINFORCEMENT LEARNING
5. TypeSQL: Knowledge-based Type-Aware Neural Text-to-SQL Generation
6. TRANX: A Transition-based Neural Abstract Syntax Parser for Semantic Parsing and Code Generation
7. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task
8. Representing Schema Structure with Graph Neural Networks for Text-to-SQL Parsing
9. SyntaxSQLNet: Syntax Tree Networks for Complex and Cross-DomainText-to-SQL Task
10. Global Reasoning over Database Structures for Text-to-SQL Parsing

Thanks