1. Overview

This topic is about the automatic conversion of user stories to application, which is essentially the study of natural language processing. The whole process is divided into two stages: first, the conversion of human natural language into function-specific programmatic language. Second, the collection of programming language of each function module obtained from the conversion is used to generate the overall application. Currently, our group's research focuses on the first stage, which is the generation of programming language from natural language.

2. Text2SQL

Currently, our group has set the main research direction as text-to-SQL conversion, i.e. Text2SQL. Text2SQL is a technique for converting natural language descriptions into SQL query statements, and its main purpose is to reduce the threshold for business people to process, use, and analyze data, thus optimizing the utilization of data resources as well as the actual business output. In today's era, a large amount of data is stored in relational databases, from financial and e-commerce to medical fields. Therefore, there are many application scenarios for querying databases using natural language. For example, self-service dashboards and dynamic analytics can be used to obtain the most relevant information to the business through natural language. In the financial sector, business people are empowered by Text2SQL technology, which allows them to directly perform data lookup, data processing, and data analysis through natural language. Salesmen can ask questions directly to the database and have Text2SQL convert the natural language into executable SQL language. For example, when a financial salesman enters, "What are the recovery times for over $100 million in loans?" , the Text2SQL model translates this natural language into "SELECT date FROM loan recovery WHERE amount > $1x10^8$ ". In the traditional program, the salesman needs to propose data requirements to the technical staff first, and then the technical staff will transform the natural language into SQL language and then call the specific data before delivery to the salesman for use.

3. Dataset

The Text to SQL dataset is manually tagged using natural language questions and the corresponding SQL queries. Natural language questions are questions that are restricted to the domain in which the data in the database is located and whose answers come from its database. That is, the questions describe SQL queries. Executing a SQL query will fetch the answer to the question from its database.

A dataset includes the following concepts:

a) Database: A database contains multiple data tables, and a data table contains multiple fields.

b) Fields: Database data source scenarios, which can be divided into single and multiple fields according to the number of scenarios involved, such as two fields for restaurant data and tourist attractions.

c) Single table/multiple tables: According to the number of tables involved in the natural language problem, it is divided into single and multiple tables. In multiple

tables, SQL generation involves table selection.
d) SQL complexity: The data set is classified into simple and complex problems according to the SQL complexity corresponding to the natural language problem, where the problem complexity is determined by the number of keywords, the nesting level, and the number of clauses involved in the SQL to determine the query statement.
e) Conversation rounds: The dataset is classified into single and multiple rounds according to the number of conversation rounds required for complete SQL generation.
f) Combined Conversations: If SQL generation is combined with progressive conversations, the dataset is marked as "Combined Conversations".

4. Frameworks
(1) Rule-based approach: early research was mainly based on a rule-based approach. Because the SQL query statement itself is a programming language with a strong paradigm, and because it is a language, it has a certain syntactic structure.

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

Here, SELECT means which columns are to be selected and how each of these columns is to be processed. WHERE means the selection method or condition that determines which cells of these columns can be selected. COLUMN indicates the target column to be queried. WOP indicates the association rule "and/or" between multiple conditions. COLUMN,OP and VALUE constitute the query conditions, representing the conditional column, conditional operators, respectively.
On the basis of this template, some regular expressions can be designed to match the user input questions to extract the content of each paragraph of the SQL template. It has the disadvantage of requiring enumeration of various expression paradigms and high workload. Therefore, much work has now shifted to end-to-end neural network models.
(2) Seq2Seq-based framework: this technique is proposed because Text2SQL is similar to a translation task. The current mainstream translation models are implemented based on the Seq2Seq framework: the input is first encoded to get the semantic information, and then the semantic encoding is recursively decoded word by word to get the translated sentence. Considering that the same information in the database will have different descriptions in the problem, while the corresponding SQL forms are the same, the form information involved in the question is first recognized, and then the recognized token is anonymized and input to the model.