NL to SQL translator

1. Introduction

Natural Language Processing(NLP) aims to help machines to understand the human text and respond it by themselves-in much the same way human do.

NLP combines with several computational linguistics--the model of human language, machine learning and deep learning models. All these technologies enable computer full ‘understand’ the meaning of text written by human language.

Structured query language(SQL) is a database query and programming language used to access data and query, update and manage relational database systems. Different from other programming languages (such as C language, Java, etc.), SQL is composed of few keywords, and each SQL statement is composed of one or more keywords which can help organization to understand the text’s meaning easier.

1. Background

In the era of big data, fast extraction and data anlysis is becoming a focus of attention. However, the anlysis nd the manipulation of database usually require a basic knowledge of coding, which set a high threshold for these people who know nothing about coding.A number of conversational data analysis applications based on SQL generation techniques have been proposed by industry to access and analyze large-scale data by simply typing natural language.

1. Usage Scenario

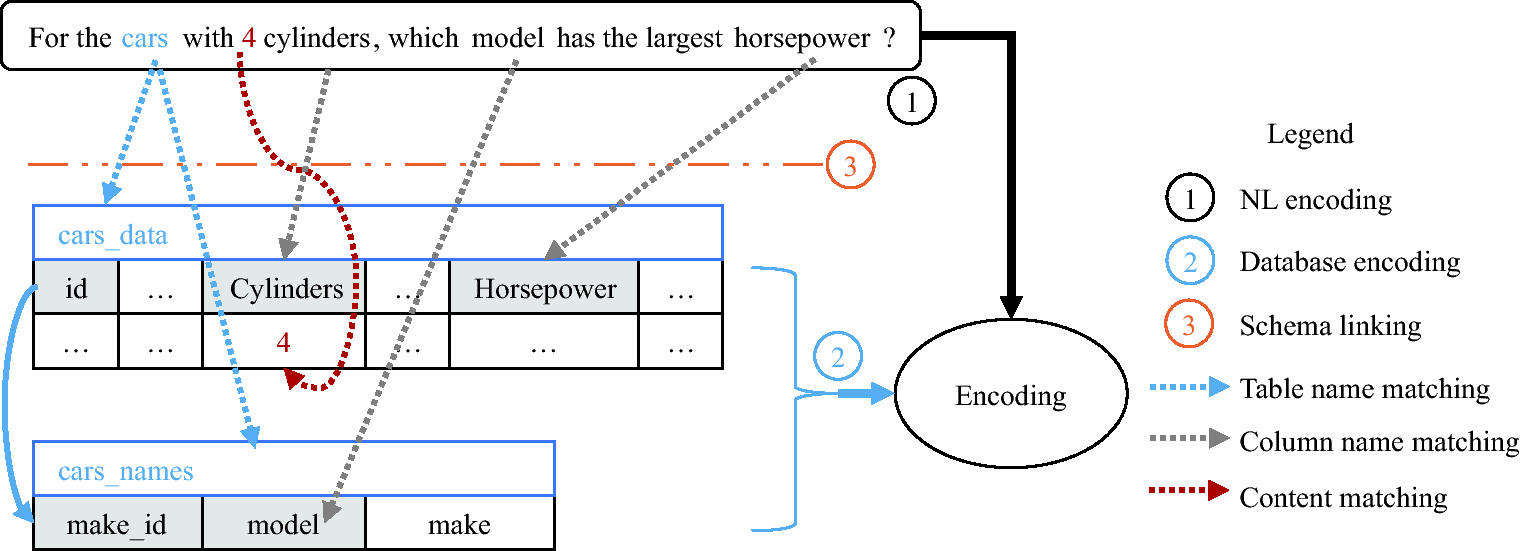
According to the complexity of SQL sentences, like some simple scenarios, SQL sentences usually just include “select”, “from”, “in”, “where” and so on. However, in some complex scenarios, SQL sentences can include “ground by”, “order by” and some other complicated keywords so that they can be used in more situations.

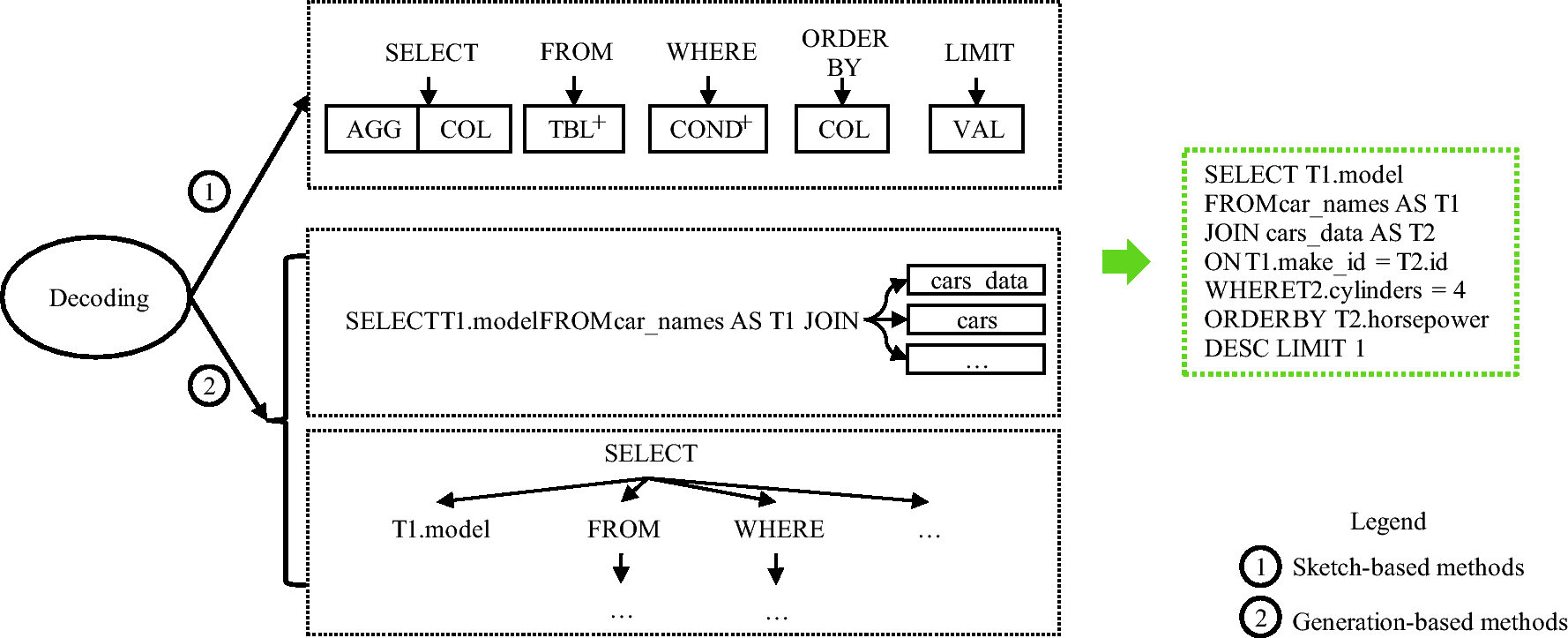
And we also have single-round SQL generation scenarios and multi-round SQL generation scenarios. Single-round SQL can only use to fix out some simple sentences, but it’s much more easy compared with multi-round SQL. Multi-round SQL can be used to some very complicated situations like some long subordinate clauses or the sentene which cannot be describe clearly at one time.

Most SQL-generated datasets only involve a single database for a particular domain since it’s the easiest way to train the SQL. These kind of datasets contain multiple data tables from a sigle database, all the training and testing process are generated from this database.

There are aslo some datases involve several databases, so the SQL generation model is required to generalize to database which has not existed auring the training process. So this kind of SQL model is much more complex with first one.

1. Model Structure



This is an example of coding process. Number 1 is the natural language encoding part, Number 2 is the database encoding part and Number 3 is schema linking part.

This is an example of decoding process.

The first is based on the template interpolation method, the black box is the content to be inserted, the AGG indicates the aggregation operation symbol, COL indicates the query column, TBL+ indicates the query table, which can be joined by multiple tables, VAL indicates the specific value type.

The second way is based on the generation method, which can be specifically divided into the sequence-based method and the syntax-based method