金融NL2SQL评测任务

## 任务描述

随着金融机构数智化转型，企业内部数据正急剧增长。一方面，这些数据作为金融机构核心资产，能大大提高企业的运作效率；另一方面，数据也正在成为幸福的烦恼，在金融领域，投研人员、投资经理、运营人员和投资者都苦于如何快速精准地查找需要的数据。传统匹配方法在覆盖度、准确性、可迁移性等方面难以满足需求。

自然语言转SQL（NL2SQL）方法设想自动将自然语言问句转化为SQL查询，从而实现精准数据问答。现有NL2SQL数据和方法主要关注“封闭场景指定库/表”设定，这很难满足业务范围动态发展的需求。从领域特性来看，金融数据多为时间序列，包括日频行情、季频财报、年度GDP、不定期股票质押解质押等，这种复杂性无疑会增大问题转SQL难度。考虑到垂直领域知识可能有助于构建可迁移系统，我们将开放金融股票、基金、宏观等场景结构化表格及相关知识，希望和社区共同努力推动垂直领域下NL2SQL技术发展。

## 数据描述

本次赛题将提供有标注数据训练集5000条和验证集1000条，2000条无标签数据作为测试集。除标注数据外，还包括数据库基础信息（db\_info.json）、数据库表（.sql文件和.sqlite文件）、领域知识（fin\_kb.json）等。

**以训练数据train.json为例，每一个标注数据形如**

 {

 "db": "stock",

 "q\_id": 0, # 第几个问题

"question": "浙江省的上市公司有多少家？",

"query": "SELECT COUNT(\*) FROM LC\_StockArchives WHERE State = '浙江省'",

 "sql": {

 "from": {

 "table\_units": [

 [

 "LC\_StockArchives", # 表名

 0 # 表在db中的id

 ]

 ],

 "conds": [] # 表关联关系，形如[[[表名,字段名,字段id],[ 表名,字段名,字段id]], ...]，其中[[表名,字段名,字段id],[ 表名,字段名,字段id]]表示一组关联条件

 },

 "select": [

 false, # 是否进行distinct, true/false

 [

 [

 "count", # 字段agg操作 (null, 'max', 'min', 'count', 'sum', 'avg')

 "LC\_StockArchives", # 字段来源表名，如果是表关联字段，取该字段任一关联表均正确

 "\*", # 字段名称

 0, # 该字段在该表对应的column id

 false # 是否进行distinct

 ]

 ]

 ],

 "where": [ # [[表名,字段名,字段id,where\_op,约束值], and/or, ...]

 [

 "LC\_StockArchives", # 表名

 "State", # 字段名

 3, # 字段id

 "=", # where\_op包括('between', '=', '>', '<', '>=', '<=', '!=', 'like')

 "浙江省", # 约束值

 null # 对datetime字段取时间操作，包括(null, 'year', 'month', 'day', 'quarter')

 ]

 ],

 "groupBy": [], # [[表名,字段名,字段id], [表名, 字段名, 字段id], ...]

 "having": [], # [[agg操作, is\_distinct,表名,字段名,字段id, where\_op,约束值], and/or, ..]

 "orderBy": [], # [[asc/desc/none, agg操作, is\_distinct, 表名, 字段名, 字段id],...]

 "limit": null # null 或者数值

 }

 }

**数据库基础信息db\_info.json记录数据库-表-字段关系、表关联关系、字段类型等信息**。

[

 {

 "db\_name": "stock", # 股票db的信息

 "tables": [ # 该db对应的所有表

 "LC\_StockArchives",

 "LC\_ExecutivesHoldings"

 ],

 "column\_info": [ # 每个表对应的字段信息

 {

 "table": "LC\_StockArchives",

 "columns": [ # 表对应的字段集合

 "\*",

 "CompanyCode",

 "ChiName",

 "State",

 "RegAddr",

 "EstablishmentDate"

 ],

 "column\_chiName": [ # 字段中文名

 null,

 "公司代码",

 "公司名称",

 "省份",

 "公司注册地址",

 "公司成立日期"

 ],

 "column\_type": [ # 字段对应的值类型

 null,

 "int",

 "text",

 "text",

 "text",

 "datetime"

 ]

 },

 …

 "table\_rel": [ # 表关联关系

 [

 [

 "LC\_StockArchives", # 表名

 "CompanyCode" # 字段名

 ],

 [

 "LC\_ExecutivesHoldings", # 表名

 "CompanyCode" # 字段名

 ]

 ]

 ]

 ]

 }

]

**领域知识fin\_kb.json主要维护金融领域重要实体对应的知识，包括实体类型、属性、属性值类型、属性别称和属性枚举值。**

[

 {

 "entityID": 1,

 "entityName": "上市公司",

 "entityType": "Company",

 "entityAttribute": [

 {

 "attrID": 1,

 "attrName": "成立日期",

 "attrType": "Date",

 "attrAlias": [

 "成立时间",

 "建立时间",

 "创立日期"

 ],

 "attrEnum": []

 },

 {

 "attrID": 2,

 "attrName": "A股代码",

 "attrType": "Number",

 "attrAlias": [

 "证券代码",

 "股票代码",

 "股票"

 ],

 "attrEnum": [

 "000021",

 "000035",

 "000040"

 ]

 },

 {

 "attrID": 3,

 "attrName": "总经理",

 "attrType": "Person",

 "attrAlias": [

 "经理"

 ],

 "attrEnum": []

 }

 ]

 }

]

## 评价指标

本次评测任务的评分标准包括：

Exact Set Match（EM）: 预测结构体中sql完全匹配的准确率，当所有元素（from, select, where, groupBy, having, orderBy, limit）都匹配才算作该样本预测结果正确，在每个元素匹配过程中，在不影响执行结果的情况下内部顺序打乱也算作正确（例如select对象可交换）；

Execution Accuracy (EA): 预测SQL语句与真实标注SQL执行结果一致。

最终评测指标为：（EM + EA） / 2，评测指标越高，排名越高。

## 任务提交方式

为构建端到端NL2SQL系统，本次任务的测试将提供未标注的question（不指定库、表），其他db\_info.json、fin\_kb.json和数据库表文件与训练保持一致。选手需要基于搭建的NL2SQL模型推理得到每个测试question对应的结果（与训练标注数据样式相同），文件命名为results.json。

最终提交结果包括results.json、相关代码及说明和方案文档，供指标评测、创新性评估使用。

## 奖励设置

恒生提供30000元奖金，奖励评测结果优异的团队，具体设置如下：

第一名：15000元

第二名：10000元

第三名：5000元

创新奖：5000元 （没有名次限制，由学会提供）

## 时间安排

评测任务发布：4月6日

报名时间：4月6日—7月25日

训练及验证数据发布：4月25日

测试数据发布：7月25日

提交测试结果：7月31日

评测论文提交：8月12日

CCKS会议日期(评测报告及颁奖)：8月25日—28日

## 组织者信息

林金曙：linjs13607@hundsun.com

娄东方：loudf24454@hundsun.com

张少杰：zhangsj41876@hundsun.com