# A Pilot Study for Chinese SQL Semantic Parsing

Qingkai Min, Yuefeng Shi, Yue Zhang
School of Engineering, Westlake University, China
Institute of Advanced Technology, Westlake Institute for Advanced Study

## Highlights

- We build a Spider dataset with multiple tables and complicated queries in Chinese, called **CSpider**, which is a low-resource data source in this area.
- We evaluate a baseline model using this dataset with different sets of embeddings and segmentors. Results show that segmentation accuracy and cross-lingual embeddings are useful for this task.
- We start a competition, which is released at **https://github.com/taolusi/chisp**.

## CSpider

We translate all English questions from Spider dataset into Chinese. The work is undertaken by 2 NLP researchers and 1 computer science student.

**Challenges:**

- The schema data (i.e. table names and column names) in CSpider has not been translated to simulate the industry situations, which brings further challenges on **question-to-DB mapping**.
- The basic sematic unit for denoting columns or cells can be words, but **word segmentation** can be erroneous.
- Linguistic characteristics of Chinese, such as zero-pronoun, can influence on its SQL parsing.

## Examples

*Easy*

**What is the number of cars with more than 4 cylinders?**
**有多少辆车的汽缸超过4个?**

```
SELECT COUNT(*) FROM cars_data WHERE cylinders > 4
```

*Medium*

**For each stadium, how many concerts are there?**
**每个体育馆开过多少演唱会?**

```
SELECT T2.name, COUNT(*) FROM concert AS T1 JOIN
stadium AS T2 ON T1.stadium_id = T2.stadium_id
GROUP BY T1.stadium_id
```

*Hard*

**Which countries in Europe have at least 3 car manufacturers?**
**欧洲有哪些国家至少有3家汽车制造商?**

```
SELECT T1.country_name FROM countries AS T1 JOIN
continents AS T2 ON T1.continent = T2.cont_id JOIN
car_makers AS T3 ON T1.country_id = T3.country
WHERE T2.continent = 'Europe' GROUP BY
T1.country_name HAVING COUNT(*) > 3
```

*Extra Hard*

**What is the average life expectancy in the countries where English is not the official language?**
**在那些英语不是官方语言的国家，平均预期寿命是多少?**

```
SELECT AVG (life_expectancy) FROM country
WHERE name NOT IN
  (SELECT T1.name FROM country AS T1 JOIN
  country_language AS T2 ON T1.code = T2.
  country_code WHERE T2.language = "English" AND
  T2.is_official = "T")
```

## Experiments

| | | Easy | Medium | Hard | Extra Hard | All |
|---|---|---|---|---|---|---|
| | ENG | 31.8% | 11.3% | 9.5% | 2.7% | 14.1% |
| HT | C-ML | **27.3%** | **9.9%** | 7.5% | **2.3%** | **12.1%** |
| | C-S | 23.1% | 7.7% | 6.2% | 1.7% | 9.9% |
| | WY-ML | 21.4% | 8.1% | **8.0%** | 1.7% | 10.0% |
| | WY-S | 20.2% | 6.4% | 6.7% | 2.0% | 8.9% |
| | WJ-ML | 19.8% | 8.6% | 5.0% | 1.3% | 9.2% |
| | WJ-S | 20.1% | 5.0% | 5.7% | 1.7% | 8.2% |
| MT | C-ML | 18.1% | 4.6% | 5.2% | 0.3% | 7.9% |
| | WY-ML | 17.9% | 4.7% | 4.5% | 0.3% | 7.6% |

Table 3: Accuracy of Exact Matching on test set.

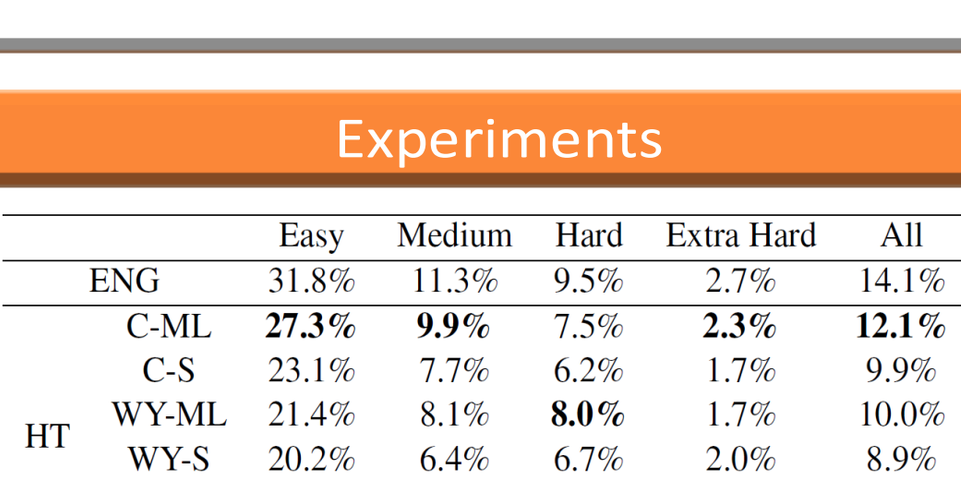
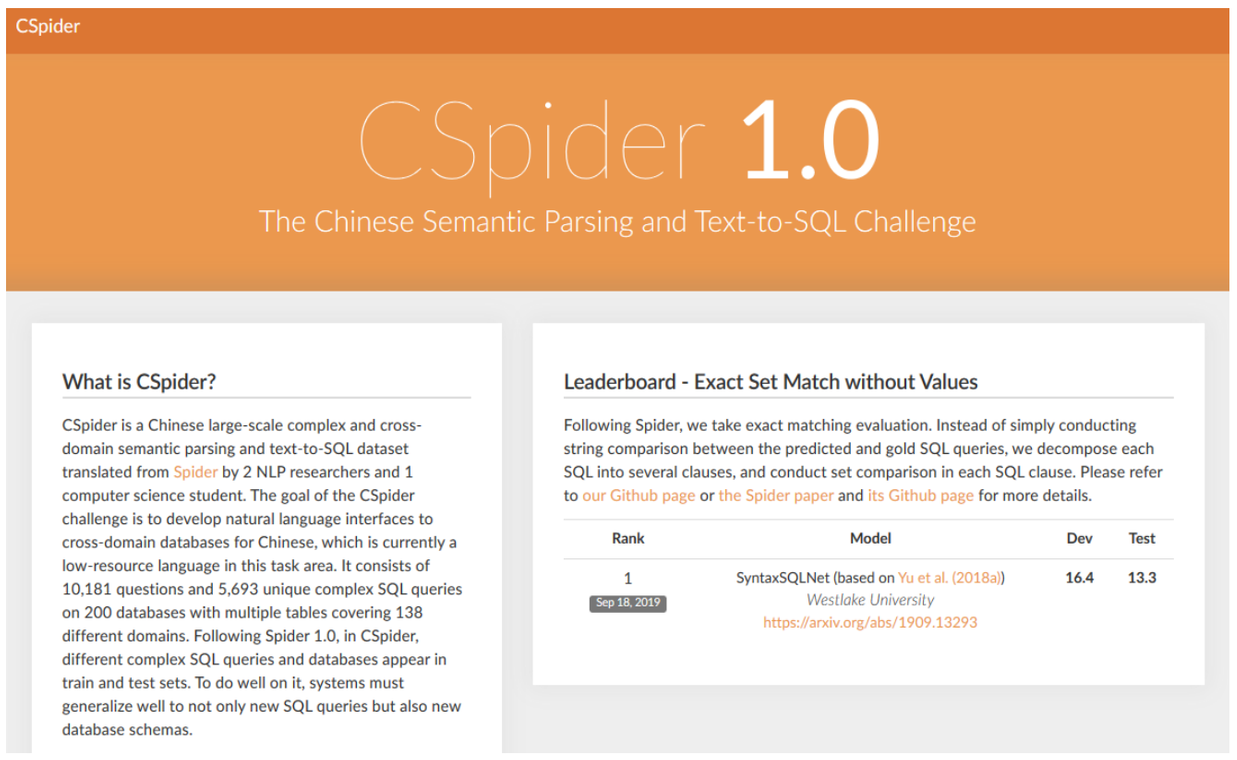



Figure 2: Component Matching Comparisons.

## Call For Participant




Our github page:
https://github.com/taolusi/chisp

To participate our competition:
https://taolusi.github.io/CSpider-explorer