A Pilot Study for Chinese SQL Semantic Parsing

Contribution

- First Chinese SQL semantic parsing dataset (manually translated from the English Spider dataset)
- Additional challenges:
  - Question-to-DB matching
  - Word segmentation errors
  - Sentence pattern and zero-pronoun

What is Spider and why Spider?

- A semantic parsing dataset for text-to-SQL task.
- Why Spider?
  - Multi-domains (more than 130) rather than a single domain.
  - Different databases for training and test dataset.
  - Multiple tables in one database, more complicated SQL queries.

Chinese Spider

- Manually translate the questions in Chinese, and keep the schema information in English.
- 9691 questions with 166 databases.

		# Q	# SQL	# DB	# Table/DB
English	all	10181	5693	200	5.1
Chinese	all	9691	5263	166	5.28
	train	6831	3493	99	5.38
	dev	954	589	25	4.16
	test	1906	1193	42	5.69

Table 1: Comparisons between Spider and Chinese Spider datasets.

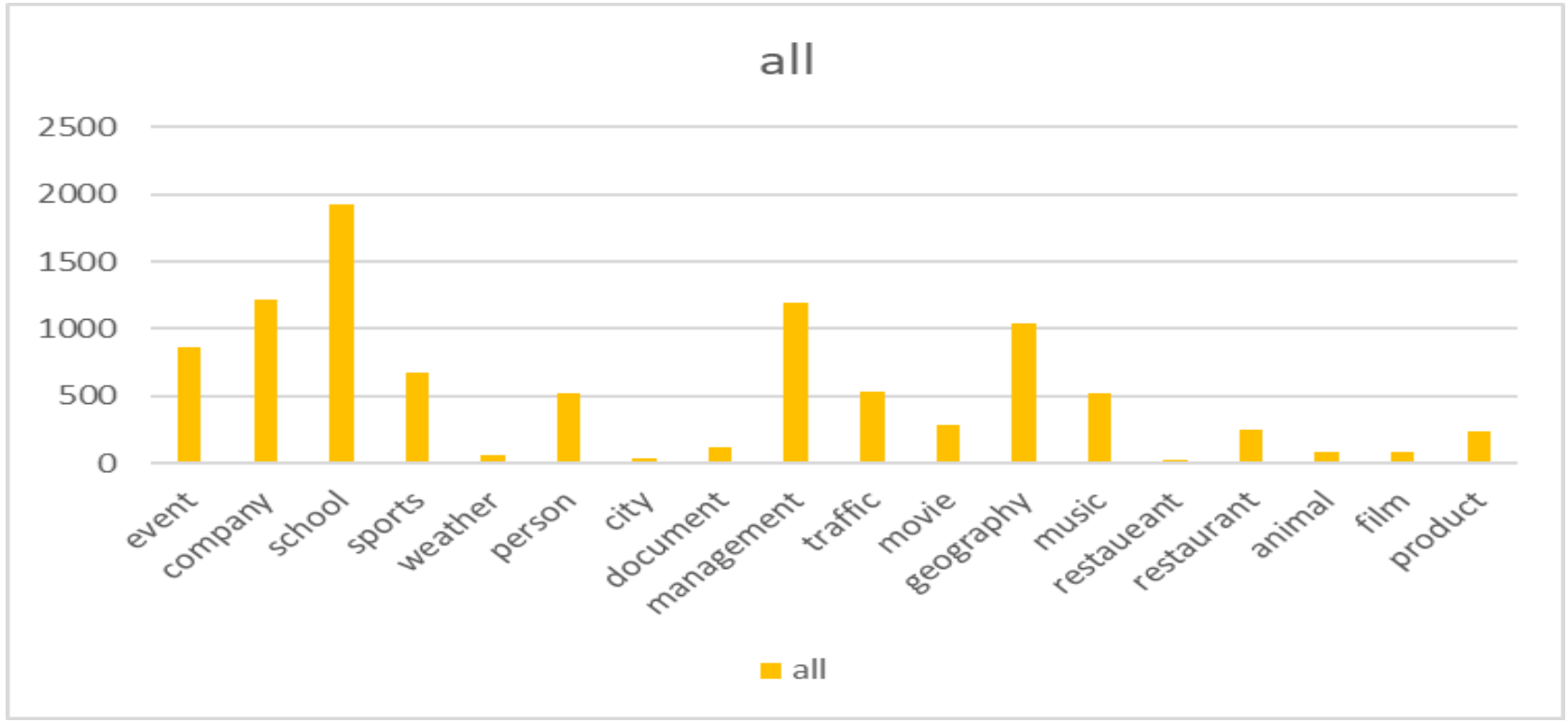
Chinese Spider

Sample 1: applying multiple tables in one database.
SQL Query: SELECT T2.star rating description FROM HOTELS AS T1 JOIN Ref Hotel Star Ratings AS T2 ON T1.star rating code = T2.star rating code WHERE T1.price range >10000;
English Question: Give me the star rating descriptions of the hotels that cost more than 10000.
Translated Chinese Question: 给出费用超过10000的酒店星级的描述。

Sample 2: with a nested SQL query.
SQL Query: SELECT T1.staff name , T1.staff id FROM Staff AS T1 JOIN Fault Log AS T2 ON T1.staff id = T2.recorded by staff id EXCEPT SELECT T3.staff name, T3.staff id FROM Staff AS T3 JOIN Engineer Visits AS T4 ON T3.staff id = T4.contact staff id";
English Question:
What is the name and ID of the staff who recorded the fault log but has not contacted any visiting engineers?
Translated Chinese Question: 那些记录了错误报告但没有联系任何到访工程师的职工的姓名和ID是什么?

Chinese Spider

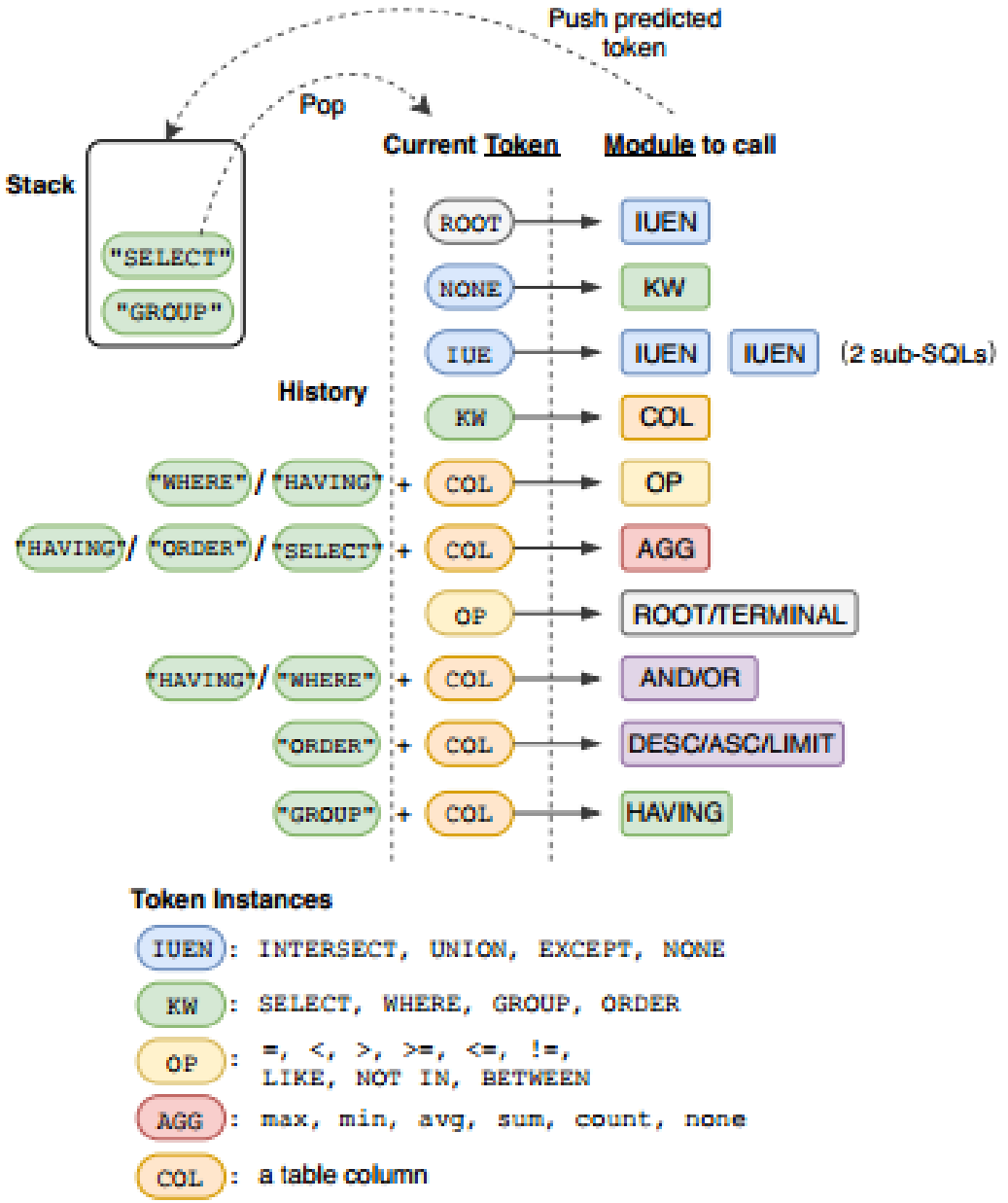
- Domain-related information



Syntaxsqlnet

- A seq-to-set method which enables to generate complicated SQL query.
- 9 modules to generate different SQL components. All these modules are trained separately and do not share parameters.
- Encoder: Using Bi-LSTM to encode question, schema and SQL history separately.
- Decoder: A syntax tree-based decoder with SQL path history and schema attention.

Syntaxsqlnet



Added challenges parsing Chinese questions into SQL

- **Question-to-DB mapping**: column in English and questions in Chinese
  - Separate sets of word embeddings
  - Multi-lingual word embeddings

Added challenges parsing Chinese questions into SQL

- Question-to-DB mapping: column in English and questions in Chinese
  - Separate sets of word embeddings
  - Multi-lingual word embeddings
- **Character-based VS Word-based**:
  - YZ segmentor
  - Jieba segmentor

Added challenges parsing Chinese questions into SQL

- Question-to-DB mapping: column in English and questions in Chinese
  - Separate sets of word embeddings
  - Multi-lingual word embeddings
- Character-based VS Word-based:
  - YZ segmentor
  - Jieba segmento
- **Sentence pattern**: incorrect predictions for many question sentences frequently

Added challenges parsing Chinese questions into SQL

- Question-to-DB mapping: column in English and questions in Chinese
  - Separate sets of word embeddings
  - Multi-lingual word embeddings
- Character-based VS Word-based:
  - YZ segmentor
  - Jieba segmentor
- Sentence pattern: incorrect predictions for many question sentences frequently
- **Zero-pronoun**: frequent for Chinese

Performance of Machine Translation

- Human-translated sentences VS Machine-translated sentences
  - 100 randomly picked machine-translated sentences, 42 translation mistakes

Evaluation Methods

- Divided by the difficulty of data
  - Easy : Containing one key word such as where or groupby;
  - Medium : Containing orderby, or two columns in select module, or aggregator;
  - Hard : containing more than three SQL keywords;
  - Extra Hard : Containing nested queries;
- Evaluation
  - Component matching
  - Exact Matching
- Do not predict values in the SQL queries.

Experiment Results

- Exact Match

		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.

Case Study

- Word Segmentation Error

Word segmentation error	Predicted query
哪些 商店 的 产品 数量 高于 平均 水平？ 把 店 名 给 我 。 Which shops' number products is above the average? Give me the shop names.	SELECT Manager_name FROM shop WHERE Number_products > (SELECT AVG(Number_products) FROM shop)
哪些 商店 的 产品 数量 高于 平均 水平？ 把 店名 给 我 。 Which shops' number products is above the average? Give me the shop names.	SELECT name FROM shop WHERE Number_products > (SELECT AVG(Number_products) FROM shop)

Figure 3: Word segmentation error.

Case Study

- Sentence Pattern

Sentence patterns	Predicted query
最 常 见 的 教师 的 家乡 是 哪里 ? What is the most common hometowns for teachers?	SELECT Hometown FROM teacher ORDER BY Age DESC LIMIT 1
列出 最 常 见 的 教师 的 家乡 。 List the most common hometown of teachers.	SELECT Hometown FROM teacher GROUP BY Hometown ORDER BY COUNT(*) DESC LIMIT 1

Figure 4: Sentence pattern.

Case Study

- Chinese zero pronoun

Chinese zero pronoun	Predicted query
代表 的 不同 党派 是 什么 ？ 显示 各 党 的 党名 和 代表 人数 。 What are the different parties of representative? Show the party name and the number of representatives.	SELECT Date , COUNT(*) FROM election GROUP BY Seats
代表 的 不同 党派 是 什么 ？ 显示 各 党 的 党名 和 各 党 的 代表 人数 。 What are the different parties of representative? Show the party name and the number of representatives in each party.	SELECT Party , COUNT(*) FROM representative GROUP BY Party

Figure 5: Chinese zero pronoun.

Thank you !