A Pilot Study for Chinese SQL Semantic Parsing

Qingkai Min, Yuefeng Shi and Yue Zhang
School of Engineering, Westlake University, China
Institute of Advanced Technology, Westlake Institute for Advanced Study
{minqingkai, shiyuefeng, zhangyue}@westlake.edu.cn

Abstract

The task of semantic parsing is highly useful for dialogue and question answering systems. Many datasets have been proposed to map natural language text into SQL, among which the recent Spider dataset provides cross-domain samples with multiple tables and complex queries. We build a Spider dataset for Chinese, which is currently a low-resource language in this task area. Interesting research questions arise from the uniqueness of the language, which requires word segmentation, and also from the fact that SQL keywords and columns of DB tables are typically written in English. We compare character- and word-based encoders for a semantic parser, and different embedding schemes. Results show that word-based semantic parser is subject to segmentation errors and cross-lingual word embeddings are useful for text-to-SQL.

1 Introduction

The task of semantic parsing is highly useful for tasks such as dialogue (Chen et al., 2013; Gupta et al., 2018; Einolghozati et al., 2019) and question answering (Gildea and Jurafsky, 2002; Yih et al., 2015; Reddy et al., 2016). Among a wide range of possible semantic representations, SQL offers a standardized interface to knowledge bases across tasks (Astrova, 2009; Xu et al., 2017; Dong and Lapata, 2018; Lee et al., 2011). Recently, Yu et al. (2018b) released a manually labelled dataset for parsing natural language questions into complex SQL, which facilitates related research.

Yu et al. (2018b)’s dataset is exclusive for English questions. Intuitively, the same semantic parsing task can be applied cross-lingual, since SQL is a universal semantic representation and database interface. However, for languages other than English, there can be added difficulties parsing into SQL. Take Chinese for example, the additional challenges can be at least two-fold. First, structures of relational databases, in particular names and column names of DB tables, are typically represented in English. This adds to the challenges to question-to-DB mapping. Second, the basic semantic unit for denoting columns or cells can be words, but word segmentation can be erroneous. It is also interesting to study the influence of other linguistic characteristics of Chinese, such as zero-pronoun, on its SQL parsing.

We investigate parsing Chinese questions to SQL by creating a first dataset, and empirically evaluating a strong baseline model on the dataset. In particular, we translate the Spider (Yu et al., 2018b) dataset into Chinese. Using the model of Yu et al. (2018a), we compare several key model configurations.

Results show that our human-translated dataset is significantly more reliable compared to a dataset composed of machine-translated questions. In addition, the overall accuracy for Chinese SQL semantic parsing can be comparable to that for English. We found that cross-lingual word embeddings are useful for matching Chinese questions with English table columns and keywords and that language characteristics have a significant influence on parsing results. We release our dataset named CSpider and code at https://github.com/taolusi/chisp.

2 Related Work

Existing datasets for semantic parsing can be classified into two major categories. The first uses logic for semantic representation, including ATIS (Price, 1990; Dahl et al., 1994) and GeroQuery (Zelle and Mooney, 1996). The second and dominant category of datasets uses SQL, which includes Restaurants (Tang and Mooney, 2001; Popescu et al., 2003), Academic (Iyer et al., 2017), Yelp and IMDB (Yaghmazadeh et al., 2017), Ad-
vising (Finegan-Dollak et al., 2018) and the recently proposed WikiSQL (Zhong et al., 2017) and Spider (Yu et al., 2018b). One salient difference between Spider and prior work is that Spider uses different databases across domains for training and testing, which can verify the generalization power of a semantic parsing model. Compared with WikiSQL, Spider further has multiple tables in each database and correspondingly more complex queries. We thus consider Spider for sourcing our dataset. Existing semantic parsing datasets for Chinese include a small corpus for assigning semantic roles (Sun and Jurafsky, 2004) and SemEval-2016 Task 9 for Chinese semantic dependency parsing (Che et al., 2012), but these data are not related to SQL. To our knowledge, we are the first to release a Chinese SQL semantic parsing dataset.

There has been a line of work improving the model of Yu et al. (2018a) since the release of the Spider dataset (Guo et al., 2019; Bogin et al., 2019; Lin et al., 2019). At the time of our investigation, however, the models are not published. We thus chose the model of Yu et al. (2018a) as our baseline. The choice of more different neural models is orthogonal to our dataset contribution, but can empirically give more insights about the conclusions.

3 Dataset

We translate all English questions in the Spider dataset into Chinese.\(^1\) The work is undertaken by 2 NLP researchers and 1 computer science student. Each question is first translated by one annotator, and then cross-checked and corrected by a second annotator. Finally, a third annotator verifies the original and corrected versions. Statistics of the dataset are shown in Table 1. There are originally 10181 questions from Spider, but only 9691 for the training and development sets are publicly available. We thus translated these sentences only. Following the database split setting of Yu et al. (2018b), we make training, development and test sets split in a way that no database overlaps in them as shown in Table 1.

The translation work is performed on a database to database basis. For each database, the same translator translates relevant inquiries sentence by sentence. The translator is asked to read the original question as well as the SQL query before making its Chinese translation. If the literal translation is possible, the translator is asked to stick to the original sentence style as much as feasible. For complex questions, the translator is allowed to rephrase the English question, so that the most natural Chinese translation is made. In addition, we keep the diversity of style in the English dataset by matching different English expressions to different Chinese expressions. A sample of our dataset is shown in Table 2. Our dataset is named CSpider.

4 Model

We use the neural semantic parsing method of Yu et al. (2018a) as the baseline model, which can be regarded as a sequence-to-tree model. In particular, the input question is encoded using an LSTM sequence encoder, and the output is a SQL query in its syntactic tree form. The tree is generated incrementally top-down, in a pre-order traversal sequence. Tree nodes include keyword nodes (e.g., SELECT, WHERE, EXCEPT) and table column name nodes (e.g., ID, City, Surname, which are defined in specific tables), which are represented in respective embedding spaces. Each keyword or column is generated by attention to the embedding space using the question representation as a key. A stack is used for incremental decoding, where

---

\(^1\)Note that we do not translate the database schema (i.e., column names) into Chinese because in practice databases have English schema and Chinese contents in the industry.
the whole output history is leveraged as a feature for deciding the next term. This method gives the current released state-of-the-art results while submitting this paper. We provide a visualization of the model in Figure 1.

5 Experiments

We focus on comparing different word segmenta-
tion methods and different embedding representa-
tions. As discussed above, column names are se-
lected by attention over column embeddings using
sentence representation as a key. Hence there must be
a link between the embeddings of columns and
those of the questions. Since columns are written
in English and questions in Chinese, we consider
two embedding methods. The first method is to
use two separate sets of embeddings for Chinese
and English, respectively. We use Glove (Pen-
nington et al., 2014)\(^2\) for embeddings of English
keywords, column names etc., and Tencent em-
beddings (Song et al., 2018)\(^3\) for Chinese. The
second method is to directly use the cross-lingual
word embeddings. To this end, the Tencent multi-
lingual embeddings are chosen, which contain
both Chinese and English words in a multi-lingual
embedding matrix.

Evaluation Metrics. We follow Yu et al.
(2018b), evaluating the results using two major
types of metrics. The first is exact matching accu-
racly, namely the percentage of questions that have
exactly the same SQL output as its reference. The
second is component matching F1, namely the F1
scores for SELECT, WHERE, GROUP BY, ORDER
BY and all keywords, respectively.

Hyperparameters. Our hyperparameters are
mostly taken from Yu et al. (2018a), but tuned
on the Chinese Spider development set. We use
column names and word embeddings from Tencent em-
bedding; both of them are not fine-tuned during
model training. Embedding sizes are set to 200
for both characters and words. For the different
choices of keywords and column names embed-
dings, sizes are set to 200 and 300, respectively.
Adam (Kingma and Ba, 2014) is used for opti-
mization, with a learning rate of 1e-4. Dropout
is used for the output of LSTM with a rate of 0.5.

For word-based models, segmentation is neces-
sary. We take two segmentors with different per-
formances, including the Jieba segmentor and the
model of Yang et al. (2017), which we name Jieba
and YZ, respectively. To verify their accuracy, we
manually segment the first 100 sentences from the
test set. Jieba and YZ give F1 scores of 89.8% and
91.7%, respectively.

5.1 Overall Results

The overall exact matching results are shown in
Table 3. In this table, ENG represents the re-
sults of Yu et al. (2018a)’s model on their En-

---

\(^2\)https://nlp.stanford.edu/projects/glove/

\(^3\)https://ai.tencent.com/ailab/nlp/embedding.html

---
sentences, 42 have translation mistakes such as semantic changes (28 sentences) and grammar errors (14 sentences). Both of these facts indicate that data by machine-translation is not reliable for semantic parsing research.

Second, comparisons among C-ML, WY-ML and WJ-ML, and among C-S, WY-S and WJ-S show that multi-lingual embeddings give superior results compared to monolingual embeddings, which is likely because they bring a better connection between natural language questions and database columns.

Third, comparisons between WY-ML and WJ-ML, and WY-S and WJ-S indicate that better segmentation accuracy has a significant influence on question parsing. Word-based methods are subject to segmentation errors.

Moreover, with the current segmentation accuracy of 92%, a word-based model underperforms a character-based model. Intuitively, since words carry more direct semantic information as compared with database columns and keywords, improved segmentation may allow a word-based model to outperform a character-based model.

Finally, for easy questions, the character-based model shows strong advantages over the word-based models. However, for medium to extremely hard questions, the trend becomes less obvious, which is likely because the intrinsic semantic complexity overwhelms the encoding differences.

Our best Chinese system gives an overall accuracy of 12.1%, 4 which is less but comparable to the English results. This shows that Chinese semantic parsing may not be significantly more challenging compared to English with text to SQL.

**Component matching.** Figure 2 shows F1 scores of several typical components, including SELN (SELECT NO AGGREGATOR), WHEN (WHERE NO OPERATOR) and GBN (GROUP BY NO HAVING), applying the superior multi-lingual embeddings. The trends are consistent with the overall results.

The detailed results are shown in Table 4. Specifically, the char-based methods achieve around 41% on SELN and SEL (SELECT), which are about 5% higher compared to the word-based methods. This result may be due to the fact that word-based models are sensitive to the OOV words (Zhang and Yang, 2018; Li et al., 2019). Unlike other components, SEL and SELN are confronted with more severe OOV challenges caused by recognizing the unseen schema during testing.

In addition, the models using multi-lingual embedding overperform the models using separate embeddings on both WHEN and OB (ORDERBY), which further demonstrates that embeddings in the same dimension distribution benefit to strengthen the connection between the question and the schema.

Contrary to the overall result, the models employing the jieba segmentor perform better than those using the YZ segmentor on OB. The reason is that the jieba segmentor has different word segmentation results in terms of the superlative of adjectives. For example, the word “最高” (the highest) is segmented as “最” (most) and “高” (high) by YZ segmentor but “最” in jieba segmentor. This again demonstrates the influence of word segmentation. Finally, for GB (GROUPBY) there is not a regular contrast pattern between different models, which can be likely because of the lack of sufficient training data.

5.2 Case study

Figure 3 shows the negative influence of segmentation errors. In particular, the incorrect segmentation of the word “店名” (shop name) leads to

---

4 Note that the results are lower than those reported by Yu et al. (2018a) under their split due to different training/test splits. Our split has less training data and more test instances in the “Hard” category and less in “Easy” and “Medium”.

Table 3: Accuracy of Exact Matching on test set.

<table>
<thead>
<tr>
<th></th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
<th>Extra Hard</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG</td>
<td>31.8%</td>
<td>11.3%</td>
<td>9.5%</td>
<td>2.7%</td>
<td>14.1%</td>
</tr>
<tr>
<td>C-ML</td>
<td>27.3%</td>
<td>9.9%</td>
<td>7.5%</td>
<td>2.3%</td>
<td>12.1%</td>
</tr>
<tr>
<td>C-S</td>
<td>23.1%</td>
<td>7.7%</td>
<td>6.2%</td>
<td>1.7%</td>
<td>9.9%</td>
</tr>
<tr>
<td>WY-ML</td>
<td>21.4%</td>
<td>8.1%</td>
<td><strong>8.0%</strong></td>
<td>1.7%</td>
<td>10.0%</td>
</tr>
<tr>
<td>WY-S</td>
<td>20.2%</td>
<td>6.4%</td>
<td>6.7%</td>
<td>2.0%</td>
<td>8.9%</td>
</tr>
<tr>
<td>WJ-ML</td>
<td>19.8%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>1.7%</td>
<td>9.2%</td>
</tr>
<tr>
<td>WJ-S</td>
<td>20.1%</td>
<td>5.0%</td>
<td>5.7%</td>
<td>1.7%</td>
<td>8.2%</td>
</tr>
<tr>
<td>MT C-ML</td>
<td>18.1%</td>
<td>4.6%</td>
<td>5.2%</td>
<td>0.3%</td>
<td>7.9%</td>
</tr>
<tr>
<td>WY-ML</td>
<td>17.9%</td>
<td>4.7%</td>
<td>4.5%</td>
<td>0.3%</td>
<td>7.6%</td>
</tr>
</tbody>
</table>

Figure 2: Component Matching Comparisons.
<table>
<thead>
<tr>
<th>SEL</th>
<th>SELN</th>
<th>WHE</th>
<th>WHEN</th>
<th>GB</th>
<th>GBN</th>
<th>OB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG</td>
<td>47.3%</td>
<td>48.2%</td>
<td>19.9%</td>
<td>24.4%</td>
<td>35.0%</td>
<td>40.6%</td>
</tr>
<tr>
<td>HT</td>
<td>C-ML</td>
<td>40.7%</td>
<td>41.2%</td>
<td>19.9%</td>
<td>23.6%</td>
<td>33.6%</td>
</tr>
<tr>
<td></td>
<td>C-S</td>
<td>40.6%</td>
<td>41.0%</td>
<td>15.3%</td>
<td>17.3%</td>
<td>29.2%</td>
</tr>
<tr>
<td>WY-ML</td>
<td>34.8%</td>
<td>35.6%</td>
<td>18.1%</td>
<td>21.4%</td>
<td>26.7%</td>
<td>30.9%</td>
</tr>
<tr>
<td>WY-S</td>
<td>34.5%</td>
<td>35.6%</td>
<td>16.5%</td>
<td>19.8%</td>
<td>30.2%</td>
<td>34.6%</td>
</tr>
<tr>
<td>WJ-ML</td>
<td>34.7%</td>
<td>35.4%</td>
<td>15.8%</td>
<td>19.2%</td>
<td>27.9%</td>
<td>31.4%</td>
</tr>
<tr>
<td>WJ-S</td>
<td>35.7%</td>
<td>36.8%</td>
<td>15.9%</td>
<td>19.6%</td>
<td>24.4%</td>
<td>26.8%</td>
</tr>
<tr>
<td>MT</td>
<td>C-ML</td>
<td>36.5%</td>
<td>37.2%</td>
<td>11.3%</td>
<td>14.2%</td>
<td>29.1%</td>
</tr>
<tr>
<td>WY-ML</td>
<td>32.1%</td>
<td>32.8%</td>
<td>11.3%</td>
<td>13.4%</td>
<td>24.8%</td>
<td>27.5%</td>
</tr>
</tbody>
</table>

Table 4: F1 scores of Component Matching on test set.

<table>
<thead>
<tr>
<th>Word segmentation error</th>
<th>Predicted query</th>
<th>Sentence patterns</th>
<th>Predicted query</th>
</tr>
</thead>
<tbody>
<tr>
<td>哪些商店的产品数量高于平均水平？</td>
<td>SELECT Manager_name FROM shop WHERE Number_products &gt; (SELECT AVG(Number_products) FROM shop)</td>
<td>SELECT Hometown FROM teacher ORDER BY Age DESC LIMIT 1</td>
<td>SELECT Date, COUNT(*) FROM election GROUP BY Seats</td>
</tr>
<tr>
<td>哪些商店的产品数量高于平均水平？</td>
<td>SELECT name FROM shop WHERE Number_products &gt; (SELECT AVG(Number_products) FROM shop)</td>
<td>SELECT Hometown FROM teacher GROUP BY Hometown ORDER BY COUNT(*) DESC LIMIT 1</td>
<td>SELECT Party, COUNT(*) FROM representative GROUP BY Party</td>
</tr>
</tbody>
</table>

Figure 3: Word segmentation error.

Figure 4: Sentence pattern.

incorrect SQL for the whole sentence, since the character “店” (shop) can typically be associated with “店长” (shop manager).

Figure 4 shows the sensitivity of our model to sentence patterns. In particular, the word-based model gives incorrect predictions for many question sentences frequently. As shown in the first row, the word “where” confuses the system for making a choice between “ORDER BY” and “GROUP BY”. When we manually change the sentence pattern into “List the most common hometown of teachers”, the parser gives the correct keyword. In contrast, the character-based model is less sensitive to question sentences, which is likely because characters are less sparse compared with words. More training data or contextualized embeddings may alleviate the issue for the word-based method, which we leave for future work.

Figure 5 shows the sensitivity of the model to Chinese linguistic patterns. In particular, the first sentence has a zero pronoun “各党” (in each party), which is omitted later. As a result, a semantic parser cannot tell the correct database columns from the sentence. We manually add the correct entity for the zero pronoun, resulting in the second sentence. The parser can correctly identify both the column name and the table name for this corrected sentence. Since zero-pronouns are frequent for Chinese (Chen and Ng, 2016), they give added difficulty for its semantic parsing.

6 Conclusion

We constructed a first resource named CSpider for Chinese sentence to SQL, evaluating the performance of a strong English model on this dataset. Results show that the input representation, embedding forms and linguistic factors all have the influence on the Chinese-specific task. Our dataset can serve as a starting point for further research on this task, which can be beneficial to the investigation of Chinese QA and dialogue models.

Acknowledgments

We thank the anonymous reviewers for their detailed and constructive comments. Yue Zhang is the corresponding author.
References


Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In Proceedings of the 57th Conference of the Association for Computational Linguistics, pages 3242–3252.


