Dialogue2SQL: Modeling Conversation with a SQL Database Backend

Davey Proctor
Yale University / New Haven, CT 06511
david.proctor@yale.edu

Abstract

The Spider dataset (Yu et al., 2018) was the largest text to complex SQL dataset to date. Joining the team, I look to augment the dataset in a promising new research direction, that of sequential question answering (SQA) and dialogue. I will also model these problems by adapting baselines to these so-called SQA2SQL and Dialogue2SQL tasks, as well as designing and implementing novel models.

1 Introduction

The problem of translating natural language to a computer-understandable encoding has been studied for a long time. This need has increased recently due to the exponential explosion of recorded data in structured databases query-able by SQL. At the same time, few people have adequate understanding of SQL to leverage this data; even the ones who do find it tedious to express their query in such a rigid syntax.

There has been much work towards automating the “sequence to SQL” translation process, including the generation of datasets such as ATIS (Price, 1990) (Dahl et al., 1994), GeoQuery (Zelle and Mooney, 1996), and WikiSQL (Zhong et al., 2017). Each of these provided valuable datasets and models; however, the queries were not nearly complex enough to extract powerful information from databases. Further problems include that the datasets have evaluation unfairness, such as randomly splitting highly-redundant examples into train and test.

The Spider dataset (Yu et al., 2018) strove towards complex queries. Its complexity is represented in transcending simple “SELECT ... WHERE,” queries; instead using “GROUP BY,” nested, “HAVING” and “JOIN”-ing over many tables. See figure 1 for the contributions of Spider.

2 Proposed Contributions

My work will fall into three categories on the two problems, SQA2SQL and Dialogue2SQL, as follows.

2.1 Dataset creation

We seek to augment the Spider dataset to allow for sequential question answering. This involves taking the complex queries and asking largely orthogonal sequential questions that contextualize later questions. The system is thus asked to learn representations of the prior questions in order to answer the more simply stated complex-logic questions asked later. This is meant to both service human sequential questions, and model the sequential contextualization of information. Figure 2 shows one example of a sequential question in this dataset.
1. Find the the name of the oldest director.
SELECT name FROM director ORDER BY age DESC LIMIT 1
Q: what is the average age of all directors?
SELECT avg(age) FROM director
Q: how about the closest?
SELECT age FROM director ORDER BY age DESC LIMIT 1
Q: who is it?
SELECT name FROM director ORDER BY age DESC LIMIT 1

Figure 2: Example list of sequential questions in the SQA2SQL dataset.

<table>
<thead>
<tr>
<th>User</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where would you go to eat in the south part of town?</td>
<td>inform(area=south)</td>
</tr>
<tr>
<td>System actions:</td>
<td>request(food), request(price-range)</td>
</tr>
<tr>
<td>I just want to eat at a cheap restaurant in the south part of town. What food types are available, can you also provide some phone numbers?</td>
<td>inform(price-range=cheap), inform(area=south), request(price-range), request(phone)</td>
</tr>
<tr>
<td>I found two restaurants serving cheap food. Would you prefer Portuguese or Chinese food?</td>
<td>request(food)</td>
</tr>
<tr>
<td>Either is fine, can I have the phone number please?</td>
<td>request(phone)</td>
</tr>
<tr>
<td>The lucky star is at 0123444277 and Nanos is at 012332708.</td>
<td></td>
</tr>
<tr>
<td>Thank you very much.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Proposed user-system interaction with information slots in the Dialogue2SQL dataset.

2.2 Baselining
I will replicate the baseline presented in (Zhong et al., 2018) to the new datasets for Dialogue2SQL. This model uses global estimators to share parameters for the so-called “slots” of dialogue states, and then local modules to fill in those slots. This models the user goals within and between queries. Figure 3, from that paper, shows the additional slot-learning procedure of a model user-system dialogue.

2.3 Modeling
I will work with Tao Yu to build the best neural model for Dialogue2SQL. This will work off of the model presented by (Zhong et al., 2018) but will extend it.

3 Timeline
1. Feb 1 - Feb 15. Annotate 400 SQA2SQL questions, across 12 databases (3 databases completed).
3. March 3 (?). Submission to ACL for SQA2SQL.
5. April 15 - May 2. Finish paper for EMNLP on Dialogue2SQL.
6. May 2. Submit senior project.

References


Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongyu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In EMNLP.


Victor Zhong, Caiming Xiong, and Richard Socher.  