Dataset Creation and Modeling for Context Dependent Database Question Answering

Shreya Dixit
Advisors: Professor Dragomir Radev, Graduate Student Tao Yu

Summary

The project attempts to tackle the problem of allowing users of any background to interact with relational databases via natural language. The LILY team is taking on two approaches to achieve this: sequential question answering (SQA) and a dialogue system. Previous text to SQL models such as Salesforce’s Seq2Sql, Shanghai Jiao Tong University’s SQLNet, and TypeSQL have achieved their results on simple SQL datasets and did not perform to the same par on the complex SQL dataset, Spider, compiled by Yale in 2018. The idea of these two new approaches is to see whether a deep learning approach and converting natural language queries to logic forms first and then to SQL improves performance on complex SQL datasets. My project comprises of two parts: (1) dataset creation for the SQA to SQL model and Dialog to SQL model and (2) construction of a model that modifies previous text to SQL models that have done well on simpler datasets to fit our text to SQL task.

Dataset Construction for SQA and Dialogue system

Goal

The major challenge of creating natural language to SQL system is the lack of data due to the proprietary nature of many databases, and simply the lack of collected “normal” natural language translations to SQL queries. Moreover, the dataset collection is even more essential in light of the approaches suggested above: SQA and Dialog System, as the form of the data becomes crucial.

Previous Work

Several Semantic Parsing Datasets exist, but few satisfy the goals of emerging semantic parsing models that apply neural network approaches, which require expanded dataset sizes, and multiple databases. Previous datasets such as Restaurants, Yelp and IMDB, and Scholar are large, but operate within a singular domain, forcing existing models to use the same database for both training and testing. In 2017, Zhong et al. published the WikiSQL, currently the most popular dataset, generated 80,654 natural language sequence-sql pairs on 24,241 different Wikipedia tables. However, the dataset was constructed at the expense of using only basic SQL labels, such as SELECT etc. Yale released a smaller SPIDER dataset that consists of 10,181
questions, but it consists of unseen queries and unseen databases, forcing any model to have to adjust to domain changes as well. Conducting a similar process as that for the Spider Dataset, this project will develop a semantic parsing dataset for SQA to SQL and Dialog System to SQL that satisfies the size, complexity, and multiple domain requirements.

Method

The Sequential Question Answering dataset will be created by Yale students by manually decomposing SQL queries into natural language questions that encompass different transformative methods: refinement, theme-entity, theme-property, paraphrase, coreference. For the Dialog System, the project will attempt to use random participants via Amazon’s Mechanical Turk to construct sample dialogs to be used as a training dataset. But in order to proceed to the collection phase, the project will require developing a user-interface to properly incentivise random ‘Turkers’ to complete a dialog with a member of the research team. The project hopes to create a high quality, large, complex, and cross domain student-annotated dataset for Sequential Question Answering.

Model for Text-to-SQL

I will replicate the model that achieves the Text-to-SQL model the best result on the WikiSQL dataset and attempt to modify it to fit on Yale’s SPIDER dataset, in order to assess the robustness of the model on a more complex dataset. Specifically, I will begin by replicating Naver’s SQLova that currently achieves a 90% accuracy on the WikiSQL dataset. Naver’s SQLova model consists of three parts: BERT based table- and context-aware word embedding, a sequence-to SQL model developed upon previous works: SEQ2SQL and SQLNet, and Execution guided decoding. The project will require a careful study of the published code, and appropriate adjustments such that the new input can be fed to the model.

Timeline

1. **Phase 1: Feb 6 - Feb 15:**
   a. Annotate Databases for SQA to SQL:
      i. (400 questions, total 14 hours, 2 weeks + review grammar after 2/15)
   b. Create familiarity with Dialogue to SQL and present suggestions about the user interface
2. **Phase 2: Feb 15 - April 15:**
   a. Finish to Dialogue-to-SQL Annotation (live with Amazon Mechanical Turk)
   b. Contribute to Text-to-SQL code
3. **Phase 3: April 15- May 2**
   a. Finish up loose ends, perhaps assist in writing conference papers (ACL, EMNLP)
4. **May 2: Submit Senior Project**
References:
4. Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Minjoon Seo. 2019. From Achieving 90% accuracy in WikiSQL.
5. Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, Dragomir R. Radev. 2018. From Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task.