PROJECT PROPOSAL

Text-to-SQL Semantic Parser: A Novel Approach for Generating SQL Queries from Text

(by Shreyashri, Naman, Sudarshan, Nisharg)

The primary objective of this project is to design and implement a Text-to-SQL semantic parser with long short-term memory (LSTM) that effectively converts natural language queries into SQL queries. The goal is to enhance the efficiency and accessibility of database interactions, enabling users to interact with databases using plain language queries.

1. ***Datasets*** : We will be using WikiSQL, a crowd-sourced dataset. The dataset is designed for training and evaluating models that aim to convert human-readable questions into SQL queries, facilitating the development of systems capable of understanding and executing database queries based on natural language input. It is a dataset comprising questions, corresponding SQL queries, and SQL tables.
2. ***Scope*** : Simple Select-From-Where query. If time permits we will extend the scope.
3. ***Methodology*** :
    a. **Inputs:** The inputs for our program will be a dataset with Natural Language Query and SQL schema (under perusal)
    b. **Word embedding**: These are dense, low-dimensional vector representations of words designed to capture their semantic and syntactic relationships. Since our input is a natural language query which is a variable input, we need to apply word embeddings like Word2Vec, GloVe or BERT etc.
    c. **Encoder:** The output of word embedding is then input encoded into a machine-readable format that the model can understand.
    d. **Model:** The encoded input is then processed by a neural network, which is a machine learning model that has been trained on a massive amount of text data.
    Long Short Term Memory (LSTM) / Bi-LSTM based translation models are long used for text translation due to its robustness towards vanishing/exploding gradient problems. Also, pre-trained models like SEQ2SQL and SQLNet perform well for text translation tasks. We aim to build these models exploring how they perform with SQL queries focusing on SELECT clause, AGG, OP functions. Given the time constraints we will be implementing LSTM architecture instead of LLM models like HuggingFace-StarChat as mentioned in feedback
    e. **Decoder**: The generated output is then decoded from the machine-readable (vectors) format back into human-readable SQL Query.
4. ***References:***

- Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. CoRR, abs/1709.00103, 2017.
- Katsogiannis-Meimarakis, G., & Koutrika, G. (2023). A survey on deep learning approaches for text-to-SQL. *The VLDB Journal*, 1-32.
- Kumar, A., Nagarkar, P., Nalhe, P., & Vijayakumar, S. (2022). Deep Learning Driven Natural Languages Text to SQL Query Conversion: A Survey. *arXiv preprint arXiv:2208.04415*.
- Xu, X., Liu, C., & Song, D. (2017). Sqlnet: Generating structured queries from natural language without reinforcement learning. *arXiv preprint arXiv:1711.04436*.
- Xu, K., Wang, Y., Wang, Y., Wen, Z., & Dong, Y. (2021). Sead: End-to-end text-to-sql generation with schema-aware denoising. *arXiv preprint arXiv:2105.07911*.