Enhancing Natural Language to SQL Translation with Advanced NLP Techniques

Ashwin Daswani
Boston University
ashwind@bu.edu

Harshil Gandhi
Boston University
harshilg@bu.edu

Abstract

This paper explores the application of advanced NLP techniques in translating natural language queries into SQL queries. We focus on integrating schema information with GPT models, applying retrieval augmentation using sentence transformers, and fine-tuning LLaMA 7B using Low-Rank Adaptation (LoRA). These methods aim to improve query accuracy and provide user-friendly database interactions.

1 Introduction

The translation of natural language into SQL queries is a key step in enhancing user interaction with databases. Our research leverages cutting-edge language models to refine this translation process, aiming to make databases more accessible to users without SQL expertise.

2 Methodology

Our methodology employs a combination of NLP methods to enhance natural language to SQL translation. We utilize the Spider dataset, a large-scale, complex, and cross-domain semantic parsing and text-to-SQL dataset. It comprises real and complex questions paired with SQL queries across various databases, making it an ideal benchmark for our experiments.

Execution Match as a Metric: We employ the 'execution match' metric to evaluate our models. Unlike the 'exact match' metric, which requires the predicted SQL query to match the reference query exactly, 'execution match' assesses whether the execution result of the predicted query matches that of the reference. This metric better reflects the practical effectiveness of the models in a real-world setting, as it allows for variations in SQL syntax while still ensuring correct query outcomes.

2.1 Schema Integration with GPT Models

In our first experiment, we enhanced GPT models by appending relevant database schema information to the natural language queries. This integration aimed to provide a more comprehensive context for the model, thereby improving the accuracy of the generated SQL queries.

2.2 Retrieval Augmentation Using Sentence Transformers

For our second experiment, we employed retrieval augmentation. Using the 'all-MiniLM-L6-v2' sentence transformer, we generated embeddings for our dataset. These embeddings were then used to create an index, facilitating the retrieval of semantically similar SQL queries based on cosine similarity. This process aimed to enhance the model's ability to provide relevant and accurate SQL translations.

2.3 Fine-Tuning LLaMA 7B with LoRA

In our third experiment, we focused on fine-tuning the LLaMA 7B model with Low-Rank Adaptation (LoRA). LoRA allows for efficient fine-tuning of large language models with minimal changes to their parameters. We conducted this fine-tuning with and without Retrieval-Augmented Generation (RAG), training each model for 30 epochs. This fine-tuning aimed to tailor the LLaMA model's responses more precisely to our Text-to-SQL translation tasks.

3 Results and Discussion

In our experimental evaluation on the Spider dataset, we observed distinct performance variations across different models and methodologies.

GPT Model Performance: The GPT model with schema augmentation exhibited a significant increase in performance, achieving an execution match of 54.35

In comparison, the implementation of Retrieval-Augmented Generation (RAG) with the GPT model further boosted performance, reaching an execution match of 59.73

Comparative Analysis with State-of-the-Art: It's noteworthy that the current state-of-the-art on the Spider dataset has achieved an execution match of 77.1 % (source: https://paperswithcode.com/sota/text-to-sql-on-spider). While our results show promising improvements, particularly with the GPT models, they also highlight the gap that remains to reach top-tier performance levels. This comparison serves as a valuable benchmark for future development and optimization of our models.

Challenges with LLaMA Model: The LLaMA 7B model presented unique challenges, primarily characterized by hallucinatory behavior. Despite fine-tuning efforts using LoRA and incorporating RAG, the model frequently generated queries based on non-existent or imagined inputs. This behavior complicated the measurement of the execution match metric due to the need for extensive cleanup and interpretation of the outputs. The lack of a discernible pattern in these hallucinations further compounded the issue, making it difficult to systematically address or correct through model adjustments. Consequently, quantifying the LLaMA model's performance in terms of execution match proved problematic.

4 Future Directions

The experimentation with RASAT, a novel approach aiming to augment transformer models with relation-aware weights, encountered technical hurdles, particularly CUDA out of memory errors. Overcoming these implementation challenges remains a key objective for future research. Successfully integrating our retrieval augmentation technique with the RASAT architecture could potentially elevate performance, aligning it closer to state-of-the-art levels.

Exploring larger variants of the LLaMA model, such as 13B or 70B, may provide a solution to the hallucination issues we encountered. Additionally, our experience suggests that further refinement in prompt engineering and advanced fine-tuning techniques could be crucial in enhancing the model's ability to accurately translate natural language into SQL queries. This direction holds promise for significant advancements in natural language processing applications, particularly in the domain of database query generation.

5 Code and Data Repository

The code and related materials for our project are available in our GitHub repository.

GitHub Repository: https://github.com/ashwindaswanibu/Text-to-SQL

Please note that while the repository contains comprehensive code, large datasets, such as the complete Spider dataset used for training and evaluation, are not included due to their size. However, links to these datasets and instructions for accessing them are provided within the repository.

6 Contributions

The successful completion of this project was the result of collaborative efforts by both team members, Ashwin Daswani and Harshil Gandhi, who contributed equally to various aspects of the research.

Ashwin Daswani: Ashwin played a crucial role in the initial setup and configuration of the advanced NLP models. He was instrumental in integrating the schema information with the GPT models and fine-tuning the LLaMA 7B model using Low-Rank Adaptation (LoRA). Ashwin also took the lead in coding the initial versions of our model scripts and setting up the project repository on GitHub.

Harshil Gandhi: Harshil focused on the implementation of the retrieval augmentation using sentence transformers. He was responsible for generating embeddings, creating the index for the dataset, and implementing the cosine similarity-based retrieval system. Additionally, Harshil contributed to refining and debugging the model scripts and played a significant role in analyzing the experimental results and metrics.

Both team members actively participated in designing the experiments, discussing and interpreting the results, and drafting the final research paper. Their combined efforts ensured a comprehensive approach to tackling the challenges of natural language to SQL translation using advanced NLP techniques.

References

1. Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Ro-

man, S., Zhang, Z., & Radev, D. (2018). Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. ArXiv. https://arxiv.org/abs/1809.08887

2. Qi, J., Tang, J., He, Z., Wan, X., Cheng, Y., Zhou, C., Wang, X., Zhang, Q., & Lin, Z. (2022). RASAT: Integrating Relational Structures into Pretrained Seq2Seq Model for Text-to-SQL. ArXiv. https://arxiv.org/abs/2205.06983

3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv. https://arxiv.org/abs/2005.11401

4. Hu, E. J., Shen, Y., Wallis, P., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv. https://arxiv.org/abs/2106.09685

5. Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022). A Survey on In-context Learning. ArXiv. https://arxiv.org/abs/2301.00234

6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., et al. (2020). Language Models are Few-Shot Learners. ArXiv. https://arxiv.org/abs/2005.14165

7. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv. https://arxiv.org/abs/2307.09288