Translating Natural Language Queries to SQL Using the T5 Model

Albert Wong
Mathematics and Statistics
Langara College
Vancouver, Canada
0000-0002-0669-4352

Lien Pham
Mathematics and Statistics
Langara College
Vancouver, Canada
Email: honglien.pham@gmail.com

Young Lee
Mathematics and Statistics
Okanagan College
Kelowna, Canada
Email: yolee3112@gmail.com

Shek Chan
Mathematics and Statistics
Langara College
Vancouver, Canada
0000-0002-5932-5390

Razel Sadaya
Mathematics and Statistics
Langara College
Vancouver, Canada
Email: razelsadaya@gmail.com

Youry Khmelevsky
Computer Science
Okanagan College
Kelowna, Canada
0000-0002-6837-3490

Mathias Clement
Computer Science
Okanagan College
Kelowna, Canada
0000-0001-8206-307X

Flor1 22085a3atistics

The main contributions of the research study are the development of an effective data warehouse and the insights and machine learning models [28], [35], [36], [43]–[46], [49]–[51] to leverage contextual information.

The availability of large datasets such as Spider, WikiSQL, or SparC, has enabled researchers to fine-tune the model for text-to-SQL tasks. For example, Shaw et al. [52] showed competitive results from fine-tuning the Text-to-Text Transfer Transformer (T5) model [53] without relational structures. Authors of UnifiedSKG [54] achieved state-of-the-art results using the T5 model for various semantic parsing tasks including text-to-SQL.

Using pre-trained language models, PICARD [55] attempted to constrain the auto-regressive decoder of language models through incremental parsing. This method can be operated directly on the output of a pre-trained language model such as T5. The authors claimed to have significantly improved performance on the Spider dataset. RASAT [56] also tried to improve the performance of pre-trained language models in text-to-SQL tasks by incorporating relational structures such as schema linking and schema encoding while still inheriting the pre-trained parameters from the T5 model effectively.

4) Spider and Other Public Data Sets for Development: A number of data sets, such as Spider, WikiSQL, and Sparc, have been compiled to support the development of the NL to SQL models and for establishing benchmarks in comparison of accuracy for these models. The availability of these data sets has also enabled researchers to fine-tune their models. We have chosen to use the Spider data set as part of the development efforts in this research.

III. BUILDING THE NL TO SQL MODEL FOR A UTILITYO-SQL.

ensured that the model learned the necessary “vocabulary” (name of columns and tables) so that the model could provide effective translation for queries from the natural language to SQL.

The development was completed on both Kaggle cloud computing environment using GPU P100 and on our own Ubuntu server with an NVIDIA GPU. A typical training run of the model takes about 4 hours.

C. SQL Correction

After using the trained T5 model to translate the natural language query to SQL, we implemented a simple post-processing method to correct the SQL query with reference to the database schema.

This process scans the output SQL for incorrect names of columns and tables and, if necessary, corrects them according to the database schema. For example, it will change “meter” in the SQL query to “meters” if the correct name for the column in the schema is “meters”. It is done using a character-by-character, sequential match that is based on positions. It replaces tokens with the correct tables or column names based on the information from the database schema.

D. Performance Metrics

For model evaluation, Zhong proposed two evaluation metrics from different perspectives: logical form accuracy and execution accuracy [30].

Logical form accuracy is defined as the percentage of generated queries that are converted correctly from the actual query. On the other hand, execution accuracy stands for the percentage of generated queries that can be executed against the database and produce the correct results.

Zhong mentioned that two database queries can produce the same result. If only the logical form accuracy is utilized, some generated queries with a correctly executed result but not in the same syntax would be treated as incorrect queries. It is suggested that both of the metrics should be considered to evaluate the performance of models [30].

For this research, we used the exact match accuracy with manual inspection of the testing results to measure performance. We believe that with manual inspection and adjustments, the exact match accuracy and execution accuracy are the same in this case.

[8] A. Wong, S. Whang, E. Sagre, N. Sachin, G. Dutra, Y.-W. Lim, G. Hains, Y. Khmelevsky, and F. Chang Zhang, “Short-Term Stock Price Forecasting using Exogenous Variables and Machine Learning Algorithms,” in *2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 2023, pp. 260–265.

[9] A. Wong, L. Pham, Y. Lee, S. Chan, R. Sadaya, Y. Khmelevsky, M. Clement, F. W. Y. Cheng, J. Mahony, and M. Ferri, “Translating Natural Language Queries to SQL Using the T5 Model,” 2023.

[10] Y. Khmelevsky, A. Wong, N. Ebadifard, F. Zhang, G. Bhangu, and G. Hains, “Pan-Institutional Applied Research within Undergraduate and Post-Degree Diploma Teaching Programs,” in *Proceedings of the 25th Western Canadian Conference on Computing Education*, ser. WCCCE ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3593342.3593353

[11] A. Wong, J. Figini, A. Raheem, G. Hains, Y. Khmelevsky, and P. C. Chu, “Forecasting of Stock Prices Using Machine Learning Models,” in *2023 IEEE International Systems Conference (SysCon)*, 2023, pp. 1–7.

[12] D. Joiner, A. Vezeau, A. Wong, G. Hains, and Y. Khmelevsky, “Algorithmic Trading and Short-term Forecast for Financial Time Series with Machine Learning Models; State of the Art and Perspectives,” in *2022 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 2022, pp. 1–9.

[13] A. Wong, P. Unni, A. L. K. P. Henrique, T. A. Nguyen, C. Chiu, Y. Khmelevsky, and J. Mahony, “Machine Learning Models Application in Daily Forecasting of Hourly Electricity Usage,” in *2022 IEEE International Systems Conference (SysCon)*, 2022, pp. 1–5.

[14] A. Wong, C. Chiu, A. Abdulgapul, M. N. Beg, Y. Khmelevsky, and J. Mahony, “Estimation of Hourly Utility Usage Using Machine Learning,” in *2022 IEEE International Systems Conference (SysCon)*, 2022, pp. 1–5.

[15] A. Wong, C. Chiu, G. Hains, J. Behnke, Y. Khmelevsky, and C. Mazur, “Modelling Network Latency and Online Video Gamers’ Satisfaction with Machine Learning,” in *2021 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 2021, pp. 1–5.

[16] A. Wong, C. Chiu, G. Hains, J. Behnke, Y. Khmelevsky, and T. Sutherland, “Network Latency Classification for Computer Games,” in *2021 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 2021, pp. 1–6.

[17] A. Wong, D. Joiner, C. Chiu, M. Elsayed, K. Pereira, Y. Khmelevsky, and J. Mahony, “A Survey of Natural Language Processing Implementation for Data Query Systems,” in *2021 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 2021, pp. 1–8.

[18] A. Wong, C. Y. Chiu, G. Hains, J. Humphrey, H. Fuhrmann, Y. Khmelevsky, and C. Mazur, “Gamers Private Network Performance Forecasting. From Raw Data to the Data Warehouse with Machine Learning and Neural Nets,” 2021.

[19] J. Mazur Chris
*g*and Ayers, H. Jack, H. Gaétan, and K. Youry, “Machine Learning Prediction of Gamer’s Private Networks (GPN®S),” in *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2*, S. Arai Kohei
*g*and Kapoor and B. Rahul, Eds. Cham: Springer International Publishing, 2021, pp. 107–123.

[20] Y. Khmelevsky, “Ten Years of Capstone Projects at Okanagan College: A Retrospective Analysis,” in *Proceedings of the 21st Western Canadian Conference on Computing Education*. New York, NY, USA: ACM, 2016, pp. 7:1–7:6. [Online]. Available: http://doi.acm.org/10.1145/2910925.2910949

[21] Y. Khmelevsky, V. Ustimenko, G. Hains, C. Kluka, E. Ozan, and D. Syrotovsky, “International collaboration in SW engineering research projects,” in *Proceedings of the 16th Western Canadian Conference on Computing Education*, 2011, pp. 52–56.

[22] Y. Khmelevsky, “Research and teaching strategies integration at post-secondary programs,” in *Proceedings of the 16th Western Canadian Conference on Computing Education*, ser. WCCCE ’11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 57–60. [Online]. Available: https://doi.org/10.1145/1989622.1989638

[23] G. Hains, C. Li, N. Wilkinson, J. Redly, and Y. Khmelevsky, “Performance analysis of the parallel code execution for an algorithmic trading system, generated from UML models by end users,” in *2015 National*