

ЗАСТОСУВАННЯ МОВНОЇ МОДЕЛІ T5 ДЛЯ ПЕРЕТВОРЕННЯ ПРИРОДНОЇ МОВИ В SQL-ЗАПИТИ

Вінницький Національний Технічний Університет

Анотація

У цій роботі досліджується процес перетворення запитів природною мовою у SQL-запити, використовуючи модель T5. Метою є розробка системи, яка надає можливість користувачам без технічних знань формувати запити природною мовою та отримувати необхідну інформацію з баз даних. Дослідження показує, що використання моделі T5 забезпечує високу точність і ефективність у перетворенні запитів, значно спрощуючи доступ до даних та роботу з ними.

Ключові слова: штучний інтелект, машинне навчання, трансформери, T5, великі мовні моделі, LLM, SQL, природна мова, NLP.

Abstract

This paper explores the process of converting natural language queries into SQL queries using the T5 model. The goal is to develop a system that enables users without technical expertise to formulate queries in natural language and retrieve the necessary information from databases. The study demonstrates that utilizing the T5 model ensures high accuracy and efficiency in query transformation, significantly simplifying data access and interaction.

Keywords: artificial intelligence, machine learning, transformers, T5, large language models, LLM, SQL, natural language, NLP.

Вступ

Зростаючий обсяг даних, що постійно накопичуються, підвищує потребу у рішеннях для їх ефективного зберігання та швидкого доступу. Незважаючи на широкий вибір сучасних технологій управління даними, їх запит і обробка часто вимагають залучення спеціалістів, що ускладнює роботу для користувачів без технічних знань [1]. Оптимальним вирішенням є створення систем, які дозволяють формувати запити природною мовою та отримувати необхідну інформацію безпосередньо з баз даних.

Дослідження у сфері перетворення природної мови в SQL активно розвивається, використовуючи семантичний аналіз та нейронні мережі. Сучасні моделі, зокрема попередньо навчені трансформери, дозволяють автоматично генерувати SQL-запити з текстових запитів користувачів. У цій роботі розглянуто підхід до побудови такої системи, який базується на використанні моделі T5, спеціально адаптованої для задачі *NL-to-SQL* (перетворення природної мови в SQL запити) [2, 3].

Зростаючий обсяг даних підвищує інтерес до рішень, що забезпечують ефективне та спрощене отримання інформації. Останні досягнення в глибокому навчанні та обробці природної мови активно використовуються для задач перетворення тексту в SQL, що дозволяє створювати зручні інтерфейси між природною мовою та базами даних.

Це дослідження зосереджено на розвитку моделей, що перетворюють запит користувача описаний природною мовою в SQL-запити для автоматизованого доступу до баз даних. Використання таких рішень дає змогу перетворювати звичайні запитання, наприклад: «Скільки користувачів відвідала веб-сайт?», у відповідні SQL-запити. Це значно спрощує процес отримання необхідних даних без залучення технічних спеціалістів.

Мовна модель T5

У цьому дослідженні для вирішення завдання перетворення тексту в SQL було використано навчання моделі T5 на спеціально підготовленому наборі даних. T5 — це попередньо навчена модель на основі трансформерів, яка забезпечує єдину структуру для перетворення тексту з одного формату в інший [4]. Основний принцип її роботи полягає в отриманні текстового запиту на вході та генерації відповідного текстового виходу після навчання.

Існує кілька варіацій моделі T5, проте в експериментах було обрано T5-base як базову модель через її менший розмір, що дозволяє зменшити обчислювальні витрати. Навчання проводилося на спеціально створеному наборі даних, а такі параметри, як кількість епох, розмір пакета даних, швидкість навчання тощо, були налаштовані для досягнення оптимальних результатів.

Навчальні та тестові набори даних

В якості набору даних було взяти відкритий набір даних *WikiSQL* з платформи Kaggle. *WikiSQL* - великий набір даних, зібраний з різних джерел, для розробки інтерфейсів природної мови для реляційних баз даних [5].

У наборах даних містилися відповідні SQL-запити для кожного запиту природною мовою. Спеціальні навчальні набори були створені так, щоб у процесі навчання модель T5 охоплювала всі можливі варіанти використання колонок і таблиць. Це дозволяло їй «ознайомитися» з ними ще на етапі тренування. Тестові набори були розроблені аналогічним чином, але вже з метою оцінки точності роботи моделі в реальних умовах.

Навчальні набори даних складалися з комбінації набору даних *WikiSQL* та спеціально створених наборів. Набір *WikiSQL* допомагав навчання загальних правил синтаксису SQL-запитів, тоді як спеціальні набори забезпечували знайомство моделі з конкретними назвами колонок і таблиць, необхідними для правильної генерації запитів.

Процес навчання моделі здійснювався на стаціонарному комп'ютері з відеокартою *NVIDIA RTX4080* та центральним процесором *Intel Core i9-13th*. Типовий цикл навчання займав приблизно 3 години.

Корекція SQL запитів

Після того як навчена модель T5 генерує SQL-запит на основі введеного текстового запиту, застосовується додатковий етап, спрямований на виправлення можливих помилок відповідно до схеми бази даних. Цей процес аналізує згенерований SQL-запит і перевіряє коректність назв таблиць і колонок. У разі виявлення невідповідностей вони автоматично виправляються згідно зі схемою бази даних. Наприклад, якщо модель помилково використовує назву "user", а в схемі бази даних коректна назва "users", алгоритм виправить цю помилку в згенерованій моделлю SQL запиті.

Виправлення здійснюється методом по символічного послідовного порівняння з урахуванням позицій у тексті. Алгоритм замінює невірні назви таблиць або колонок на правильні, ґрунтуючись на доступній інформації про структуру бази даних.

Тестування та аналіз результатів роботи

Для оцінки якості роботи моделі були використані дві основні метрики: точність логічної форми (*logical form accuracy*) та точність виконання (*execution accuracy*) [6].

Точність логічної форми визначається як відсоток згенерованих SQL-запитів, які точно відповідають правильному SQL-запиту.

Точність виконання показує відсоток SQL-запитів, які успішно виконуються в базі даних і повертають правильний результат.

Слід зазначити, що два різні SQL-запити можуть давати однаковий результат, але якщо оцінювати лише точність логічної форми, правильно виконаний запит із відмінною від очікуваної структурою може бути визнаний помилковим. Тому для повноцінної оцінки моделі слід враховувати обидві метрики [6].

У цьому дослідженні основним критерієм оцінки була точність повного збігу (*exact match accuracy*), яка доповнювалася ручним аналізом результатів тестування. Ми вважаємо, що з

урахуванням ручного аналізу точність повного збігу та точність виконання в нашому випадку є ідентичними.

Навчання моделі T5 здійснювалося ітеративно, з постійним коригуванням навчальних наборів даних. Для оцінки продуктивності використовувалися два тестові набори даних. За результатами тестування модель досягла 72,9% точності повного збігу (exact match accuracy) на наборі даних 1 і 83,7% для набору даних 2 для SQL-запитів.

Висновки

Результати дослідження демонструють, що використання попередньо навченої моделі T5 для перетворення запитів природною мовою в SQL є ефективним підходом, який дозволяє значно спростити доступ до баз даних без необхідності залучення спеціалістів. Отримані показники точності підтверджують, що модель здатна з високою достовірністю генерувати SQL-запити, які коректно виконуються в системі.

Застосування автоматичної корекції SQL-запитів додатково підвищує якість роботи моделі, усуваючи можливі помилки в іменах таблиць і колонок. Використання спеціальних навчальних наборів даних дозволило адаптувати модель до специфіки запитів, що використовуються в реальних умовах.

Подальші дослідження можуть бути спрямовані на вдосконалення архітектури моделі, використання більш потужних варіантів T5-Large, T5-3B, T5-11B або інших сучасних трансформерних моделей. Крім того, важливим напрямком є розробка методів автоматичної генерації навчальних запитів на основі схеми бази даних, що дозволить значно прискорити процес адаптації моделі до нових середовищ. Це відкриває перспективи для подальшого розвитку систем автоматизованого запиту до баз даних і створення ще більш доступних і точних інтерфейсів для користувачів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Chaudhari, M.S., Hire, M.A., Mandale, M.B., & Vanjari, M.S. Structural Query Language Question Creation by using Inverse Way, 2021.
2. 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 RASSE 2021 - IEEE International Conference on Recent Advances in Systems Science and Engineering, Proceedings, 2021.
3. K. Affolter, K. Stockinger, and A. Bernstein, "A comparative survey of recent natural language interfaces for databases," VLDB Journal, vol. 28, no. 5, pp. 793–819, 10 2019.
4. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 10 2019.
5. Офіційна сторінка набору даних WikiSQL [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/shahrukhkhan/wikisql>
6. V. Zhong, C. Xiong, and R. Socher, Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning, pp 8, 2017.

Борисюк Володимир Миколайович – аспірант кафедри комп'ютерних наук, Вінницький Національний Технічний Університет, email: volodymyr.borysiuk0@gmail.com.

Козловський Андрій Володимирович – доцент кафедри комп'ютерних наук, Вінницький Національний Технічний Університет, email: akozlovskyi@vntu.edu.ua.

Borysiuk Volodymyr M. – Post-Graduate Student of the Computer Sciences Chair; Vinnytsia National Technical University, email: volodymyr.borysiuk0@gmail.com.

Kozlovskyi Andrii V. – Cand. Sc., Associate Professor of the Department of Computer Sciences, Vinnytsia National Technical University, email: akozlovskyi@vntu.edu.ua.