

РАЗРАБОТКА МУЛЬТИАГЕНТНОГО АЛГОРИТМА ДЛЯ ПРЕОБРАЗОВАНИЯ ТЕКСТА В SQL С ИСПОЛЬЗОВАНИЕМ LLM, ВНУТРЕННЕЙ САМОКОРРЕКЦИИ И ОБРАТНОЙ СВЯЗИ ПОЛЬЗОВАТЕЛЯ

Авторы: Перов И.И. (Университет ИТМО, г. Санкт-Петербург)

Научный руководитель: доцент, старший научный сотрудник, Насонов Д.А. (Университет ИТМО, г. Санкт-Петербург), аспирант, Филатова А.А. (Университет ИТМО, г. Санкт-Петербург)

Введение. Технология Text-to-SQL позволяет преобразовывать запросы на естественном языке в SQL-запросы, что упрощает взаимодействие с базами данных. Однако такие системы сталкиваются с рядом проблем: пользователи часто формулируют неопределенные или неполные запросы, например, «Покажите продажи», не уточнив временной промежуток или регион. Это может привести к неверным результатам. Кроме того, автоматические системы подвержены ошибкам в интерпретации и генерации SQL-запросов, особенно при работе с сложными запросами.

Традиционные системы могут обнаруживать ошибки в базе данных, но часто не справляются с более сложными проблемами, такими как неправильное понимание намерений пользователя или логические ошибки. Для решения этих проблем необходимо развивать более умные системы, которые комбинируют автоматизацию и человеческую обратную связь.

Рост объема структурированных данных, который к 2024 году достигнет почти 70 зеттабайт, делает разработку таких систем особенно актуальной [2]. Исследования показывают, что компании, использующие аналитику данных, могут повысить продуктивность на 20–25%, а подходы с обратной связью от человека могут улучшить точность систем на 45% [1].

В ответ на эти вызовы предлагается метод на основе LLM для Text-to-SQL, включающий механизмы саморефайнмента и человеческой обратной связи, что позволит повысить точность и адаптивность системы.

Основная часть. Предложенная система основана на многозадачной архитектуре, которая включает несколько ключевых агентов, работающих в тесной взаимосвязи. Диаграмма ниже иллюстрирует основные компоненты системы и их взаимодействие при обработке пользовательских запросов:

1. Маршрутизатор — центральный компонент системы, который координирует рабочий процесс, направляя задачи к соответствующим компонентам.
2. Генератор SQL — принимает пользовательский запрос и преобразует его в SQL-запрос с использованием возможностей большой языковой модели (LLM).
3. Проверка SQL — выполненная в СУБД, она проверяет корректность сгенерированного SQL-запроса. В случае ошибок Checker выделяет ошибку и возвращает текст ошибки.
4. Модификатор вопроса — на основе текста ошибки от Checker, модифицирует исходный запрос пользователя, уточняя его.
5. Упаковщик истории чатов — собирает соответствующий контекст из предыдущих взаимодействий, что помогает системе поддерживать консистентность и улучшать запросы на основе прошлых данных. Этот агент также позволяет пользователю давать обратную связь.

Алгоритм обрабатывает запросы через два различных конвейера: простой конвейер и сложный конвейер.

Процесс обработки запроса начинается с предварительной обработки, на которой система классифицирует сложность запроса с помощью классификатора задач. На этом этапе также происходит расширение аббревиатур, проверка синтаксиса запроса и его соответствие схеме базы данных. Все эти шаги помогают системе лучше понять запрос и соответствующим образом подготовить его для дальнейшей обработки. После этого происходит обогащение доказательств, фильтрация схем и выбор примеров, которые направляют LLM, улучшая точность генерации SQL-запросов.

В зависимости от сложности запроса система выбирает один из двух подходов. Для простых запросов используется простой конвейер, где система генерирует SQL-запрос, очищает его, выполняет в базе данных и выбирает лучший результат. Такой подход подходит для запросов, которые имеют однозначные ответы и не требуют дополнительной логики. В случае сложных запросов система выбирает сложный конвейер, который включает разложение запроса на более мелкие компоненты. Для каждой части генерируется отдельный SQL-запрос, а затем результаты этих запросов комбинируются в один итоговый ответ.

Выводы. В заключение исследование подчеркивает три ключевых вывода. Во-первых, разработанный метод подтверждает эффективность гибридных систем, которые комбинируют автоматические процессы с человеческой обратной связью. Во-вторых, интеграция обратной связи от пользователя позволила повысить метрику на 10%, что подчеркивает её значительную роль в улучшении точности запросов и надежности системы. Наконец, было выявлено, что большие языковые модели (LLM) могут достигать еще лучших результатов при использовании техник декомпозиции запросов, что подчеркивает важность разбиения сложных запросов на более управляемые части для достижения более высокой точности.

Источники:

1. McKinsey & Company. Ops 4.0: Fueling the next 20 percent productivity rise with digital analytics. <https://www.mckinsey.de/publikationen/leading-in-a-disruptive-world/ops-40-fueling-the-next-20-percent-productivity-rise-with-digital-analytics>
2. Statista. Total enterprise data volume world wide. <https://www.statista.com/statistics/1186304/total-enterprise-data-volume-location/>

Автор _____ Перов И.И.

Научный руководитель _____ Насонов Д.А.

Научный руководитель _____ Филатова А.А.