УДК 004.891

АДАПТАЦИЯ МОДЕЛИ TEXT-TO-SQL ДЛЯ ЛОКАЛЬНОГО ИСПОЛЬЗОВАНИЯ

Курочка Константин Сергеевич
кандидат технических наук, зав. кафедрой «Информационные технологии»
Гомельский государственный технический университет имени П.О. Сухого,
Республика Беларусь, г. Гомель
E-mail: kurochka@gstu.by

Житко Алина Сергеевна*
студент 4 года обучения
Гомельский государственный технический университет имени П.О. Сухого,
Республика Беларусь, г. Гомель
E-mail: zhitkoalina@gmail.com

Предлагается вариант реализации задачи преобразования текста на естественном языке в SQL-запросы (Text-to-SQL) с использованием локальной языковой модели, что позволяет использовать данный подход без передачи дополнительных запросов на сторонние сервисы. Такой подход обеспечивает более высокий уровень безопасности и приватности пользователей, повышает автономность программных продуктов, снижает нагрузку на сетевые ресурсы.

Ключевые слова: трансферное обучение, машинное обучение, Text-to-SQL, языковые модели

Введение. Задача преобразования текстов на естественном языке в SQL-запросы (Text-to-SQL) представляет собой процесс, при котором система преобразует пользовательские запросы, сформулированные на естественном языке, в соответствующие SQL-запросы для получения данных из базы данных. Например, для запроса пользователя «Вывести список пациентов с искривлением позвонков вдоль вертикальной оси» [1] система должна сгенерировать соответствующий SQL-запрос, который позволит извлечь информацию о пациентах из базы данных (рисунок 1).

Рисунок 1 – Принцип работы Text-to-SQL модели

Разработка программного обеспечения с применением моделей Text-to-SQL необходима для создания пользовательского интерфейса с элементами искусственного интеллекта, где пользователям предоставляется возможность взаимодействия с программным продуктом на естественном языке. Такие интерфейсы значительно расширяют диапазон пользователей, поскольку они позволяют людям без специализированных технических знаний эффективно взаимодействовать с системами доступа к базам данных.

Адаптация модели Text-to-SQL для локального использования становится особенно актуальной задачей в условиях, когда обработка данных не может быть передана на внешние серверы по соображениям безопасности и конфиденциальности, а также из-за ограниченных ресурсов сети. Также использование мощных языковых моделей, таких как GPT, через API может оказаться финансово невыгодным, поскольку большинство таких сервисов предоставляют услуги на платной основе [2]. Это может стать значительной статьей расходов для организаций, особенно при обработке больших объемов данных.

Таким образом, разработка автономных, локально работающих решений становится важной

задачей для организаций, стремящихся снизить затраты и одновременно обеспечить независимость от внешних сервисов. Такие подходы позволяют максимально адаптировать модели для внутренних баз данных и запросов, поддерживая стабильную и безопасную работу системы.

Архитектура предлагаемой системы. Существует множество парадигм для решения задачи преобразования естественного языка в SQL-запросы [3–5]. Ключевыми среди них являются single-turn и multi-turn подходы, а также парадигма предварительного обучения (pre-training), каждая из которых обеспечивает различные методы обработки и генерации запросов.

Парадигма предварительного обучения (pre-training) значительно повысила эффективность разработки Text-to-SQL моделей, позволяя им извлекать знания из обширных корпусов данных и адаптироваться к разнообразным языковым структурам. Эти модели проходят этап предварительного обучения на больших объемах текстов, что улучшает их способность обрабатывать сложные языковые конструкции и формировать корректные SQL-запросы. Для данного проекта была выбрана архитектура T5 [6], которая представляет собой мощную модель типа «последовательность в последовательность» (seq-to-seq).

Архитектура T5 (Text-to-Text Transfer Transformer) основана на модели трансформера, состоящей из энкодера и декодера. Энкодер отвечает за обработку входного текста, преобразуя его в контекстные представления, которые затем используются декодером для генерации выходного текста. Этот подход обеспечивает гибкость модели, позволяя ей работать с различными задачами, связанными с обработкой естественного языка, путем формулирования всех задач в виде текстовых пар: входного запроса и соответствующего выходного ответа.

В T5, как показано на рисунке 2, энкодер состоит из множества слоев, каждый из которых включает механизм самовнимания, позволяющий каждому токену взаимодействовать с другими токенами во входной последовательности. Это взаимодействие осуществляется с помощью многоголового внимания, что дает возможность модели захватывать различные аспекты взаимосвязей между токенами. В дополнение к этому, каждый слой содержит полносвязную нейронную сеть с активацией ReLU и применением нормализации и остаточных соединений, что способствует эффективному обучению и стабилизации процесса. Декодер имеет аналогичную структуру, однако он включает в себя механизмы маскирования, которые предотвращают доступ к будущим позициям в выходной последовательности, обеспечивая тем самым корректность генерации.

Отличительной чертой архитектуры T5 является использование различных масок внимания в зависимости от специфики задачи. Например, в стандартной архитектуре энкодера-декодера используется полностью видимая маска для энкодера, что позволяет ему обращать внимание на всю входную последовательность. В декодере, однако, применяется причинная маска, предотвращающая зависимость выходного элемента от будущих входных элементов, что критически важно для задач последовательной генерации текста.

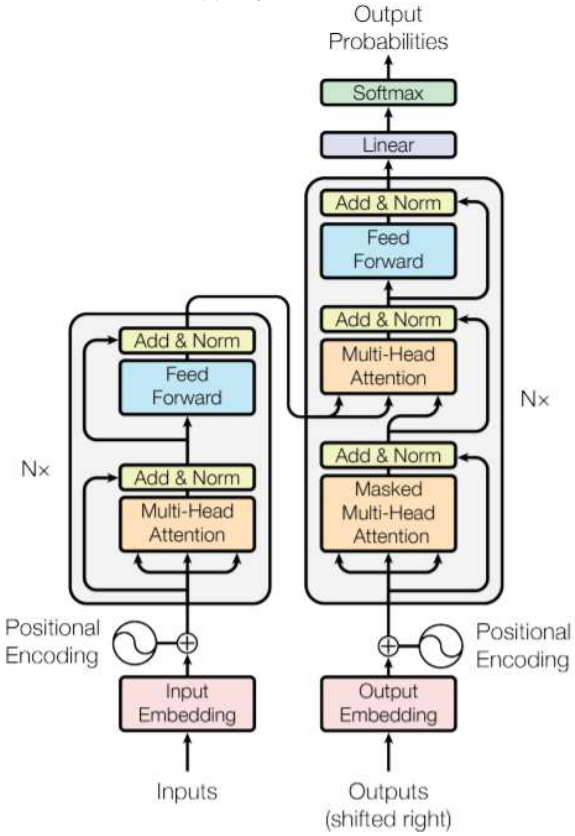

Рисунок 2 – Архитектура трансформера

Особенностью T5 является масштабируемость: архитектура представлена в нескольких вариантах с различным количеством параметров, что позволяет адаптировать модель под конкретные задачи и вычислительные ограничения. В рамках данного проекта, учитывая баланс между производительностью и ограниченными вычислительными ресурсами, была выбрана модификация T5-small, содержащая 60 миллионов параметров. Легковесность данной версии делает ее оптимальным выбором для локального использования, обеспечивая при этом достаточную точность в генерации SQL-запросов для задач средней сложности.

Этапы обучения. В то время как для англоязычных систем уже проведено множество исследований и созданы крупные наборы данных, такие как WikiSQL [7] и Spider [8], для русскоязычного сегмента подобных ресурсов практически нет. Это существенно ограничивает возможности для разработки и применения Text-to-SQL систем на русском языке. Очевидно, что для создания полноценной системы преобразования текстов на естественном русском языке в SQL необходимо либо разрабатывать собственные русскоязычные датасеты, либо адаптировать существующие англоязычные решения через трансферное обучение.

Поскольку русскоязычные ресурсы для Text-to-SQL отсутствуют, для прототипирования системы был использован англоязычный датасет Spider. Этот датасет содержит более 10,000 примеров разнообразных и сложных SQL-запросов, что позволяет оценить производительность модели в различных сценариях и условиях. Благодаря большому количеству записей, включающих различные структуры данных и типы запросов, использование Spider обеспечивает надежную базу для тестирования способностей модели обрабатывать сложные запросы.

Процесс обучения модели T5-small, начинается с загрузки предобученной версии модели. На этом этапе модель уже обучена на большом корпусе данных, что позволяет ей иметь общее понимание языка. Затем T5-small дообучается на специфичных тестовых и валидационных данных, соответствующих конкретной задаче. На этапе дообучения происходит адаптация модели к новым данным, в ходе которой она настраивается на специфику поставленной задачи. Этот процесс включает в себя оптимизацию параметров модели для повышения ее производительности на тестовых наборах. В процессе обучения используются методы градиентного спуска, которые помогают модели минимизировать функцию потерь, что в свою очередь позволяет улучшить точность предсказаний.

Для оценки эффективности модели данные делятся на тренировочный и валидационный наборы в соотношении 70 на 30. Валидационный набор используется для промежуточной оценки производительности модели, что помогает предотвратить переобучение и настроить гиперпараметры. Этот подход обеспечивает обобщенные результаты и позволяет гарантировать, что модель успешно справляется с новыми данными при окончательной проверке на тестовом наборе.

Результаты обучения модели можно проиллюстрировать графиком (рисунок 3), который отображает изменение функции потерь по мере итераций.

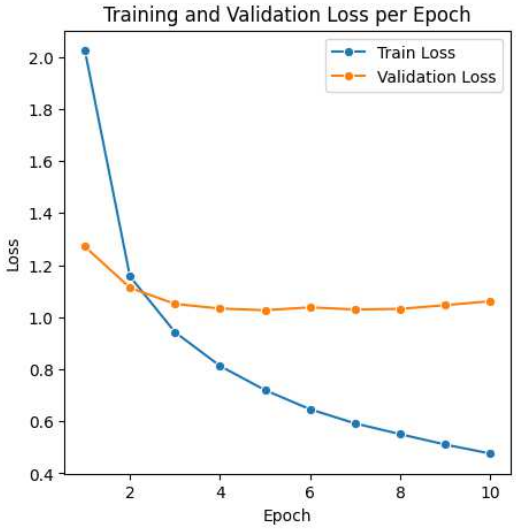

Рисунок 3 – График обучения модели T5-small

На графике обучения видно, как значение функции потерь уменьшается с каждой эпохой, что указывает на успешное дообучение модели на тестовых и валидационных данных. В начале процесса обучение сопровождается резким снижением функции потерь, что свидетельствует о том, что модель эффективно осваивает новые паттерны в данных. Со временем снижение функции потерь становится менее выраженным, что может указывать на приближение к плато, где дальнейшее улучшение становится минимальным.

Оценка разработанного решения. Оценка качества разработанного решения проводилась с использованием метрики BLEU (Bilingual Evaluation Understudy), которая широко применяется для оценки качества машинного перевода и генерации

текста. BLEU вычисляется на основе совпадений n-грамм между сгенерированным текстом и эталонными фразами, а также учитывает степень их порядкового соответствия, что позволяет оценить не только точность перевода, но и его стилистическое качество.

Конкретное значение BLEU на валидационной выборке составило 0,26, что свидетельствует о том, что модель демонстрирует степень совпадения с эталонными результатами, близкую к значениям, представленным в [6, табл. 14]. Эти результаты свидетельствуют о способности предложенного решения эффективно преобразовывать естественный текст в SQL-запросы, что делает его подходящим для практического применения в задачах машинного перевода и других областях обработки естественного языка.

Дополнительно стоит упомянуть, что размер разработанной модели составляет 500 МБ, что делает ее эффективной для интеграции в различные системы. Такой размер позволяет использовать модель на устройствах с ограниченными вычислительными ресурсами без значительной потери производительности [9].

Заключение. Выбранная архитектура T5 успешно решает задачу преобразования текстов на естественном языке в SQL-запросы. Модель продемонстрировала высокую точность генерации запросов и эффективное использование ресурсов, что подтверждает её потенциал для локальной реализации. Однако для дальнейшего расширения системы и повышения её точности на русскоязычных данных необходимо разработать специализированный русскоязычный датасет.

Список используемой литературы

1 Validity and reliability of a computer-assisted system method to measure axial vertebral rotation / Hurtado-Avilés J. [et al.] // Quantitative Imaging in Medicine and Surgery. – 2022. – Т. 12. – №. 3. – P. 1706.

2 OpenAI [Electronic resource] : Pricing . – Mode of access: https://openai.com/api/pricing/. – Date of access: 18.10.2024.

3 **Deng, N.** Recent Advances in Text-to-SQL: A Survey of What We Have and What We Expect / N. Deng, Y. Chen, Y. Zhang // Proceedings of the 29th International Conference on Computational Linguistics / Ed. N. Calzolari [et al.]. – Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022. – P. 2166–2187.

4 A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions [Electronic resource] / B. Qin [et al.] – 2022. – Mode of access: https://arxiv.org/abs/2208.13629. – Date of access: 18.10.2024.

5 **Katsogiannis-Meimarakis, G. A** Survey on Deep Learning Approaches for Text-to-SQL / G. Katsogiannis-Meimarakis, G. Koutrika // The VLDB Journal. – 2023. – Vol. 32. – P. 905–936.

6 Exploring the limits of transfer learning with a unified text-to-text transformer [Electronic resource] / C. Raffel [et al.] – 2019. – Mode of access: https://arxiv.org/pdf/1910.10683. – Date of access: 18.10.2024.

7 Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task / T. Yu [et al.]. // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing / Ed. E. Riloff [et al.]. – Brussels: Association for Computational Linguistics, 2018. – P. 3911–3921.

8 **Zhong, V.** Seq2SQL: Generating Structured Queries from Natural Language Using Reinforcement Learning [Electronic resource] / V. Zhong, C. Xiong, R. Socher – 2017. – Mode of access: https://arxiv.org/abs/1709.00103. – Date of access: 18.10.2024.

9 **Курочка, К. С.** Локализация позвонков на КТ-изображениях в условиях ограниченности вычислительных ресурсов / К. С. Курочка, К. А. Панарин // Новые горизонты - 2021 : сборник материалов VIII Белорусско-Китайского молодежного инновационного форума, 11-12 ноября 2021 года / Белорусский национальный технический университет. – Минск : БНТУ, 2021. – Т. 1. – С. 177-179.