

Концепция подхода к решению задачи ответа на вопросы по табличной информации с использованием вычислительных графов.

Иван А. Поддубный¹, Никита О. Дородных¹

¹Институт динамики систем и теории управления имени В.М. Матросова Сибирского отделения Российской академии наук (ИДСТУ СО РАН), 664033, Иркутская область, г. Иркутск, ул. Лермонтова, д. 134,

Аннотация

В работе рассматривается метод предварительного обучения языковой модели для генерации ответов на табличные запросы с использованием линеаризованных вычислительных графов вместо традиционного SQL. Эксперименты с моделью BART-large на синтетическом датасете показали низкую точность на тестовых данных WikiSQL, что указывает на ограниченность предложенного подхода и необходимость дальнейших исследований в области эффективного представления табличной информации для задач автоматического ответа на вопросы.

Keywords

Ответы на вопросы по табличной информации, Вычислительный граф, Предварительное обучение, Языковая модель,

1. Введение

Табличные данные широко используются для структурированного хранения и анализа информации в различных сферах, однако их неоднородность и отсутствие явной семантики затрудняют автоматическую обработку в интеллектуальных системах [1]. Одним из перспективных направлений решения этой проблемы является Table Question Answering (TQA)[2], позволяющее формировать ответы на вопросы к табличным данным на естественном или логическом языке. Современные методы TQA, включая семантический разбор и генеративные модели, сталкиваются с трудностями в числовых рассуждениях и учёте контекста, несмотря на достигнутый прогресс. Использование графовых представлений[3] показало потенциал для повышения эффективности за счёт переноса части вычислений на внешние модули.

Попытки комбинировать различные подходы для улучшения понимания запросов на естественном языке не привели к улучшению результатов, что актуализировало задачу более глубокого понимания табличной информации языковыми моделями. Наиболее заметным решением в этой области стала модель TAPEX [4], демонстрирующая, что выполнение SQL-запросов свидетельствует о понимании таблиц. В данной работе

^{7th} International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2025), July 7–11, 2025, Irkutsk, Russia

✉ poddubnyyiv@yandex.ru (Poddubny I.); nikidorny@icc.ru (Dorodnykh N.)

🆔 0000-0002-2431-8692 (Poddubny I.); 0000-0002-1825-0097 (Dorodnykh N.)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ICCS-DE 2025 Workshop Proceedings (iccs-de.icc.ru)

DOI: 10.47350/ICCS-DE.2025.33

предлагается альтернативный подход: предварительное обучение языковой модели на графовых описаниях запросов вместо SQL, а также анализируется влияние такого представления на качество работы модели.

2. Метод генерации ответов по запросу в виде линеаризованного графа

Метод предварительного обучения генерации ответов на запросы в формате линеаризованного вычислительного графа включает следующие этапы:

1. SQL запрос преобразуется в дерево с помощью SQLGlot.
2. Из дерева формируется граф.
3. Линеаризация графа. Полученная строка состоит из узлов вычислительного графа в формате "NODE тип узла | вложенность | аргументы. Узлы расположены в направлении от листьев к корню.
4. Генерация ответа моделью.
5. Сравнение результата генерации с ответом из набора данных.

3. Набор данных

В качестве данных был взят синтетический набор, который использовался в работе [4] для предварительного обучения. Он включает в себя 5 млн. примеров: SQL-запрос, таблица, ответ. Для данного набора выполнялась предварительная обработка по экранированию частей SQL запроса для его парсинга через SQLGlot, а так же фильтрация по превышению рамера входной последовательности токенов над максимально допустимым для модели. В результате набор был уменьшен до 3.5 млн.

4. Детали реализации

В качестве языковой модели был использован BART-large Предварительное обучение выполнялось на 3 GPU NVIDIA A100 80 GB PCIe с батчем 16 на каждую карту и коэффициентом аккумуляирования градиента 5, что дает общий батч равный 240. При этом $\text{seed} = 544$, скорость обучения = $3e-5$. Контрольные точки (checkpoints) сохранялись в конце каждой эпохи. Обучение на подготовленном наборе данных TAPEX длилось 3 эпохи и заняло 15 дней. В результате такого обучения была получена модель, которая генерирует ответ по запросу в виде линеаризованного графа.

5. Результаты

Оценка проводилась на тестовом и валидационном наборе WiKiSQL [5] после тонкой настройки в течении 20 эпох с аналогичными предварительному обучению параметрами

но с батчем 120 тренировочном наборе WiKiSQL. В качестве метрики оценки использовалась точность денотации (DA):

$$DA = P/N,$$

где P – количество точно совпадающих пар «вопрос-ответ»; N – общее количество пар «вопрос-ответ». Получены оценки для тестового и валидационного набора 12% и 14% соответственно для чекпоинта с минимальным лоссом на валидационном наборе. Полученные значения говорят о провальном обучении и деградации базовой модели.

6. Заключение

В ходе проведённого исследования был предложен и реализован метод предварительного обучения языковой модели для генерации ответов по табличным запросам в формате линеаризованного вычислительного графа. Несмотря на высокий потенциал использования вычислительных графов в аналогичных работах и использование современных инструментов (BART-large, SQLGlot, синтетический датасет TAPEX), результаты экспериментов показали низкую точность денотации (12% на тестовом и 14% на валидационном наборе WiKiSQL), что свидетельствует о неэффективности предложенного метода и деградации базовой модели. Полученные данные указывают на необходимость дальнейшего поиска и разработки более эффективных стратегий представления и обработки табличной информации для задач Table Question Answering.

Благодарности

Работа выполнена при поддержке государственного задания Министерства науки и высшего образования Российской Федерации (тема № 1023110300006-9).

Список литературы

- [1] S. Zhang, K. Balog, Web table extraction, retrieval, and augmentation: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020) 1–35.
- [2] N. Jin, J. Siebert, D. Li, Q. Chen, A survey on table question answering: recent advances, in: China Conference on Knowledge Graph and Semantic Computing, Springer, 2022, pp. 174–186.
- [3] R. Mouravieff, B. Piwowarski, S. Lamprier, Training table question answering via sql query decomposition, arXiv preprint arXiv:2402.13288 (2024).
- [4] Q. Liu, B. Chen, J. Guo, M. Ziyadi, Z. Lin, W. Chen, J.-G. Lou, Tapex: Table pre-training via learning a neural sql executor, arXiv preprint arXiv:2107.07653 (2021).
- [5] S. Yavuz, I. Gür, Y. Su, X. Yan, What it takes to achieve 100% condition accuracy on wikisql, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1702–1711.