ПЕРЕВОД РУССКОГО ЕСТЕСТВЕННОГО ЯЗЫКА НА SQL

Полонский М.И.[1], Савелов Д.Ю.[2], Губин Е.И.[3]
[1]Томский политехнический университет, ИШИТР, студент гр. 8К13, e-mail: novs965@mail.ru
[2]Санкт-Петербургский государственный университет телекоммуникаций (СПбГУТ), ИСиТ, e-mail: savasavelck@mail.ru
[3]Томский политехнический университет, ИШИТР, доцент, e-mail: gubine@tpu.ru

Введение

Модель машинного обучения для перевода естественного языка на SQL может обрабатывать как простые, так и сложные запросы. Решение упростит работу людям, которые не умеют составлять запросы SQL.

Разработанное решение позволяет сократить время при принятии решений, ускорит бизнес-процессы в компаниях, а также позволит их командам быстрее обучаться работе с данными и сократит нагрузку на дата-аналитиков.

Описание алгоритма

В решении задачи для перевода естественного языка использовалась модель машинного обучения (архитектуры трансформер [1]) T5-base, которая была дообучена для данной цели.

Для этого был использован датасет Spider [2], состоящий из 10181 вопросов и 5693 уникальных SQL запросов (в том числе и очень сложных) по 200 базам данных с несколькими таблицами, охватывающими 138 различных областей.

Easy

What is the number of cars with more than 4 cylinders?

```
SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
```

Meidum

For each stadium, how many concerts are there?

```
SELECT T2.name, COUNT(*)
FROM concert AS T1 JOIN stadium AS T2
ON T1.stadium_id = T2.stadium_id
GROUP BY T1.stadium_id
```

Hard

Which countries in Europe have at least 3 car manufacturers?

```
SELECT T1.country_name
FROM countries AS T1 JOIN continents
AS T2 ON T1.continent = T2.cont_id
JOIN car_makers AS T3 ON
T1.country_id = T3.country
WHERE T2.continent = 'Europe'
GROUP BY T1.country_name
HAVING COUNT(*) >= 3
```

Рис. 1. Примеры запросов из датасета Spider

Spider является англоязычным датасетом, поэтому для перевода русских вопросов на английский использовалась модель от Facebook WMT19.

Финальное решение позволяет формировать запросы SQL разной сложности по запросу пользователя на русском. На тестовой выборке модель T5 показывает результаты, представленные в таблице 1.

Таблица 1

Точность модели на тестовых данных

Метрика	Значение
Точное совпадение	32,4%
BLEU [4]	32,9

Учитывая специфику датасета Spider, можно сказать, что модель показывает достойный результат для использования в решении поставленной задачи.

Заключение

Проект был разработан в рамках хакатона «Цифровой прорыв. Сезон: искусственный интеллект». В задаче перевода русского естественного языка на SQL команда заняла первое место.

В планах команды перевести вопросы на естественном языке датасета Spider с помощью модели WMT19 на русский, получив новый датасет, и обучить T5-base или схожую модель уже на нем.

Список использованных источников

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J. Attention Is All You Need // [Электронный ресурс]. – URL: https://arxiv.org/pdf/1706.03762.pdf (дата обращения 18.02.2023).
2. Yu T., Zhang R., Yang K., Yasunaga M. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. [Электронный ресурс]. – URL: https://arxiv.org/pdf/1809.08887v5.pdf
3. Spider 1.0. Yale Semantic Parsing and Text-to-SQL Challenge. [Электронный ресурс]. – URL: https://yale-lily.github.io/spider
4. Понимание оценки BLEU в кастомизированном машинном переводе. [Электронный ресурс]. – URL: https://habr.com/ru/post/661377/ (дата обращения 18.02.2023).
5. Raffel C., Shazeer N., Roberts A., Lee K. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research 21, 2020. – 1–67 с.
6. Git-репозиторий [Электронный ресурс]. – URL: https://github.com/mathewpolonsky/NLSQL