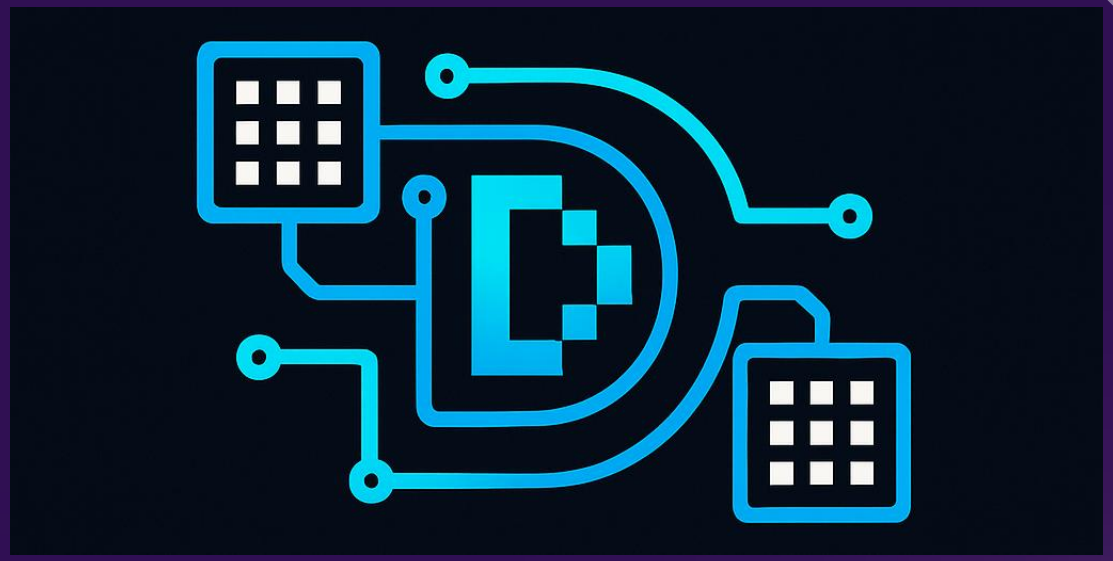
10



Duo Byte

Задача 10 Сервис выделения сущностей из поискового запроса клиента в мобильном приложении торговой сети «Пятерочка»

КОМАНДА «НАЗВАНИЕ»

ПРОЕКТ МЭРА МОСКВЫ | ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ | РАЗВИТИЕ ЧЕЛОВЕЧЕСКОГО КАПИТАЛА | ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

О команде

- Белгород
- 2
- Писаренко Антон

Наименование задачи:

Сервис выделения сущностей из поискового запроса клиента в мобильном приложении торговой сети «Пятерочка»

Как вы планируете дальше использовать или развивать ваше решение:

Проект может быть масштабирован за счёт расширения и очистки датасета: тщательная фильтрация и ручная проверка аннотаций позволят снизить количество ошибок и повысить качество. На следующем этапе возможно обучение более крупной модели для достижения ещё более высокой точности, однако текущая компактная версия уже оптимальна для мобильных приложений. Решение может быть интегрировано в поисковую систему сети «Пятёрочка» или адаптировано под другие сценарии работы с клиентскими запросами.

Описание решения:

Суть и цель проекта:

Создание сервиса для выделения сущностей в поисковых запросах клиентов «Пятёрочки» с целью точного определения типа товара, бренда, числовых и процентных характеристик.

Технологии и подходы:

Базовая модель rubert-tiny2 с CRF-слоем для согласованной BIO-разметки, дополненная синтетическими данными и кастомной подготовкой входных текстов.

Уникальность решения:

Компактная модель объёмом 114 Мб эффективно работает без GPU, что делает её оптимальной для мобильных приложений при сохранении высокого качества распознавания.

Команда «Duo Byte»

Антон Писаренко

- Сбор данных, машинное обучение
- Telegram: @antonSHBK
- Номер телефона: 8-980-320-56-16

Игорь Шмаль

- Backend, разработка web-сервиса
- Telegram: @mathd0g
- Номер телефона: 8-999-700-21-94

Команда «Duo Byte»

Краткая история команды:

Команда сформировалась после объявления о хакатоне: один из нас предложил другому попробовать свои силы в конкурсе. Ранее в хакатонах не участвовали, но давно дружим и вместе развиваемся в IT направлении. Один из занимается машинным обучением и искусственным интеллектом, второй — веб-разработкой (full-stack). Интересный факт: мы живём в Белгородской области, в городе Шебекино, у самой границы с Украиной.

Почему вы выбрали именно эту задачу из предложенных на хакатоне?

Эта задача показалась нам наиболее интересной с точки зрения реализации и практической пользы. У нас уже есть опыт работы с моделями-трансформерами, поэтому подход на базе NER выглядел понятным и выполнимым. Кроме того, проект напрямую связан с реальным бизнес-кейсом и даёт возможность применить современные методы обработки естественного языка в задаче, которая действительно востребована.

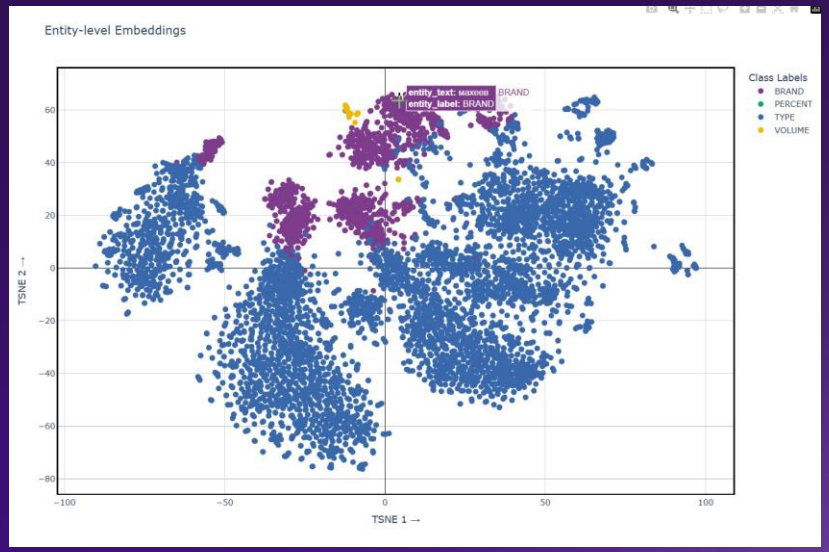
С какими основными сложностями или вызовами вы столкнулись и как их преодолели?

- Одной из проблем с которой столкнулись это была исходные данные. Объём в ~27 тысяч примеров оказался относительно мал для обучения модели. Мы использовали аугментацию и сгенерировали дополнительные синтетические примеры с помощью крупных языковых моделей
- Вторым вызовом оказалась низкая чистота разметки. Уже на этапе валидации выяснилось, что в датасете присутствуют ошибки и противоречия: бренд размечался как тип товара и наоборот, одинаковые конструкции в одних случаях получали метку сущности, а в других — метку «O». Мы провели ручной анализ, очистили данные.

Техническая проработка решения

В качестве базовой архитектуры использована компактная модель rubert-tiny2 (114 Мб), поверх которой добавлен CRF-слой для согласованных предсказаний в BIO-формате.

Данные включали исходный тренировочный корпус и синтетические примеры. Мы проводили пошаговую обработку и тестирование: запускали модель, анализировали ошибки и на их основе генерировали дополнительные контрпримеры, чтобы повысить устойчивость и качество распознавания.



Для обучения применялся оптимизатор AdamW, косинусный шедулер с разогревом (10% шагов), число эпох — 6. Обучение выполнялось на ноутбуке с 12 ядрами и 16 Гб оперативной памяти, полный цикл из 6 эпох занимал около 12 минут. Функция потерь объединяла CRF-лосс и взвешенную кросс-энтропию в соотношении 0.5 : 0.5.

Разработанный сервис развёрнут на сервере с 2 ядрами и 4 Гб оперативной памяти. Этого достаточно, чтобы обрабатывать тестовый набор из 5000 запросов.

Автоматизация фильтрации BIO разметки моделью

Идея

Для повышения качества данных и контроля уверенности предсказаний используется методика на основе энтропии предсказанных распределений.

Как это работает

- Для каждой сущности рассчитывается энтропия по эмбеддингам токенов.
- Если значение энтропии выше заданного порога — предсказание считается неуверенным и требует ручной проверки.
- Если энтропия ниже порога — предсказание признаётся надёжным и может автоматически приниматься.

Применение

Автоматическая фильтрация — система сама отбирает надёжные примеры для дообучения или использования.

Анализ датасета — существующие данные можно прогнать через модель:

- если уверенное предсказание не совпадает с разметкой - высокая вероятность ошибки в данных;
- помогает выявлять шум, некорректные метки и улучшать датасет.

Результат

Метод позволяет автоматизировать контроль качества BIO-разметки, уменьшить влияние ошибок исходных данных и сделать обучение более устойчивым.

	text	entity_text	true_entities	pred_entity	entropy	threshold	correct
113	йикорий	йикорий	[(0, 7, B-TYPE)]	B-TYPE	0.039521	0.157873	1
2824	кофейник vites	кофейник	[(9, 14, B-BRAND), (0, 8, B-TYPE)]	B-TYPE	0.031297	0.157873	1
1685	йогурт детский	йогурт	[(7, 14, I-TYPE), (0, 6, B-TYPE)]	B-TYPE	0.019398	0.157873	1
1482	копчёности	копчёности	[(0, 10, B-TYPE)]	B-TYPE	0.030618	0.157873	1
2556	рис красно	рис	[(4, 10, I-TYPE), (0, 3, B-TYPE)]	B-TYPE	0.024860	0.157873	1
1886	окорок слово мясника	слово	[(7, 12, I-TYPE), (13, 20, I-TYPE), (0, 6, B-TYPE)]	B-BRAND	0.107561	0.157873	0
2090	lind	lind	[(0, 4, B-BRAND)]	B-BRAND	0.106300	0.157873	1
209	колбаса окраина брауншвейгская	окраина	[(16, 30, I-BRAND), (0, 7, B-TYPE), (8, 15, B-BRAND)]	B-BRAND	0.074999	0.157873	1
2198	булоча	булоча	[(0, 6, B-TYPE)]	B-TYPE	0.020616	0.157873	1
2296	artfruit нектари	нектари	[(9, 16, B-TYPE), (0, 8, B-BRAND)]	B-TYPE	0.118901	0.157873	1

Уверенные предсказания

	text	entity_text	true_entities	pred_entity	entropy	threshold	correct
766	маслом	маслом	[(0, 6, B-TYPE)]	B-TYPE	0.203328	0.157873	1
230	пахлава petr	пахлава	[(0, 7, B-TYPE), (8, 12, B-BRAND)]	B-TYPE	0.341908	0.157873	1
883	бомбар	бомбар	[(0, 6, B-BRAND)]	B-TYPE	0.289808	0.157873	0
543	aquaftesh	aquaftesh	[(0, 9, B-BRAND)]	B-BRAND	0.418888	0.157873	1
1068	сыр красная	красная	[(4, 11, I-TYPE), (0, 3, B-TYPE)]	I-TYPE	0.455666	0.157873	1
852	эинкали	эинкали	[(0, 7, B-TYPE)]	B-TYPE	0.223967	0.157873	1
913	хагис	хагис	[(0, 5, B-BRAND)]	B-TYPE	0.536469	0.157873	0
162	snak	snak	[(0, 4, B-TYPE)]	B-TYPE	0.662781	0.157873	1
4	печенье с цукатами	с	[(10, 18, I-TYPE), (0, 7, B-TYPE), (8, 9, I-TYPE)]	I-TYPE	0.480432	0.157873	1
437	хлеб.	хлеб.	[(0, 5, B-TYPE)]	B-TYPE	0.167756	0.157873	1

Неуверенные предсказания