We’R
Проектный практикум

Предсказание качества распознавания речи

27 мая 2024

Описание датасета

Состав:

- "model_annotation": расшифровка аудио моделью ASR
- "human_markup": ручная расшифровка аудио
- "audio_path": пути к .wav файлам
- "label": метки класса (0 - верно, 1 - ошибка)

Характеристики:

- 1 дубликат, нет пропущенных значений

Дисбаланс классов: 60% класс 0, 40% класс 1

Длина фраз (медиана):

- Класс 0: 6 слов
- Класс 1: 5 слов
- Средняя длина слова:
- Разметка моделью: 4.82 символа
- Разметка человеком: 4.73 символа

Методы подготовки данных

Токенизация текста:

- Разбиение текста из столбца "model_annotation" на отдельные слова (токены)
- Текст уже приведен к нижнему регистру и очищен от знаков препинания
- Токенизация необходима для подачи текста на вход языковых моделей

Векторизация текста:

- Преобразование токенизированного текста в числовые векторы
- Использование предобученных моделей (rubert-base-cased, rubert-base-cased-conversational, rubert-tiny2)
- Модели преобразуют каждый токен в векторное представление (embedding)

Разделение данных:

- Разбиение датасета на обучающую, валидационную и тестовую выборки
- Соотношение 85/15
- Необходимо для обучения и оценки качества моделей

Варианты моделей

Модели на основе архитектуры BERT:

- rubert-base-cased: русскоязычная модель BERT, обученная на большом корпусе текстов, чувствительна к регистру
- rubert-base-cased-conversational: русскоязычная модель BERT, дополнительно обученная на диалогах, лучше подходит для разговорной речи
- rubert-tiny2: облегченная версия rubert, меньше параметров, быстрее в обучении и инференсе, но потенциально менее точная

Дообучение (fine-tuning) моделей:

- Использование предобученных моделей как основы
- Дообучение последнего слоя (или нескольких последних слоев) на данных проекта
- Позволяет адаптировать модели к специфике задачи и данных

Гиперпараметры моделей:

- Размер батча: влияет на скорость обучения и качество модели
- Скорость обучения (learning rate): определяет величину обновления весов на каждом шаге
- Количество эпох: сколько раз модель проходит по всему датасету во время обучения
- Размер эмбеддингов: размерность векторных представлений слов

Промежуточные результаты

Модель	AUC-ROC
rubert-base-cased	0.77
rubert-base-cased-conversational	0.82
rubert-tiny2	0.73

Итоговый результат

Лучшая модель отобрана для проверки на фрагменте отложенной выборки.

Модель	AUC-ROC
rubert-base-cased-conversational	0.805

Команда

Артем Смирнов

Екатерина Кубракова

Елизавета Лилиом

Александр Ильиных

Андрей Гриценко

Эдуард Антонов