Выполнили: Успенский Д. А.
Беляев И. А.
Карбаев С. А.
Группа: 208

Содержание

Выполнили: Успенский Д. А.
Беляев И. А.
Карбаев С. А.
Группа: 208

1 Описание предметной области и постановка задачи

В качестве проекта мы выбрали *Task 1: Semantic Textual Relatedness (STR)*, где предлагается решить задачу по автоматическому определению степени семантической связи между парами предложений на различных языках.

1.1 Введение в предметную область

Semantic Textual Relatedness (семантическая текстовая родственность) показывает степень, в которой две единицы языка, такие как слова, фразы или предложения, близки по своему значению. Это мера того, насколько семантически схожи или родственны две лингвистические единицы друг другу. Задача измерения этой связи возникает естественным образом при работе с текстовыми данными.

Применения меры семантического сходства:

- **Лингвистические**
  - изучение отношения между различными языковыми единицами
  - измерение тесктовой связности
  - сравнение стиля текстов на различных языках
- **NLP**
  - поиск информации в тексте
  - обобщение документов
  - оценка методологий: инструмент сравнения с groud truth текстами
  - создание датасетов: наборы данных семантической текстовой связанности, таких как STR-2022 (Mohamed Abdalla et. al. , 2023)

1.2 Постановка задачи

На соревновании мы выбрали трек *A: Supervised*. Необходимо обучить модель, которая предсказываеи STR двух предложений. Доступны тренировчный и тестовый датасеты на английском.

2 Описание датасета

2.1 Структура датасета

Каждый объект в датасете представляет собой пару предложений и оценку, отражающую степень семантической текстовой связи между двумя предложениями. Баллы могут варьироваться от 0 (максимально не связаны) до 1 (максимально связаны). Эти оценки были определены путем ручного аннотирования. В частности, использовался метод *comparative annotation*, чтобы избежать известных ограничений традиционных методов оценки сходства.

пример аннотации

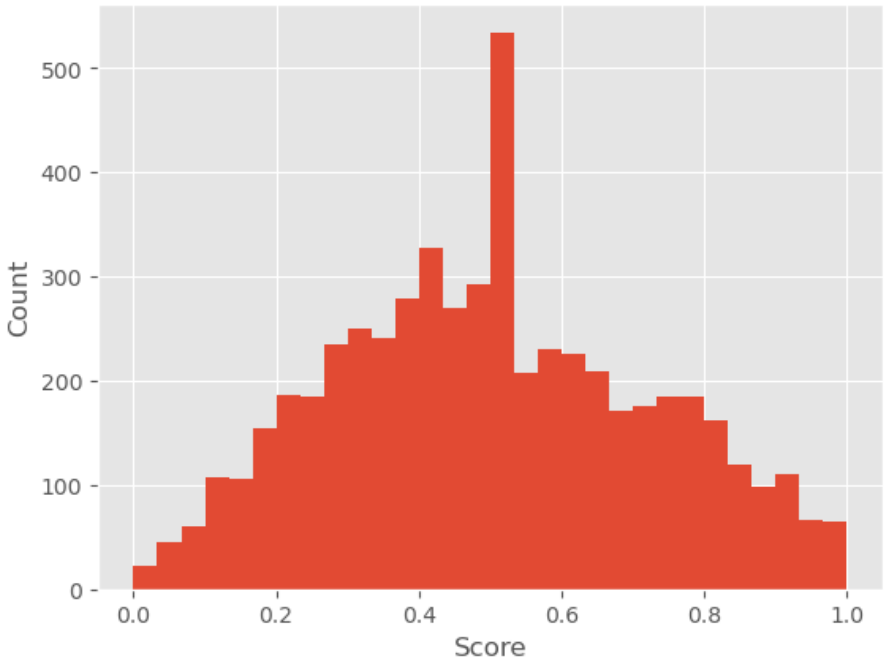
2.2 Измерение качества

Официальной метрикой оценки для этого задания является коэффициент ранговой корреляции Спирмена, который отражает, насколько хорошо предсказанные системой рейтинги тестовых экземпляров согласуются с суждениями человека.

3 Предварительный анализ данных

3.1 График распределения

Посмотрим на график распределения целевой переменной



Получили что-то немного похожее на нормальное распределение

3.2 Базовое решение

Сначала попробуем что-то простое. Составим словарь слов каждого из 2 предложений и посмотрим сколько слов в них пересекаются. Затем, для получения целевой переменной, поделим найденное число на общее количество слов в каждом предложении.

Удалим из наших предложений самые распространенные слова (если ’the’ есть в каждом предложении, более похожими они не станут) и ненужные символы.

В итоге мы получили результат

4 Обзор литературы

1. A SEMANTIC SIMILARITY MEASURE BETWEEN SENTENCES
2. обобщение документов
3. оценка методологий: инструмент сравнения с groud

5 Описание выбранной архитектуры

6 Описание полученных результатов

7 Направления дальнейшей работы

8 Источники

- Nedjma Ousidhoum et. al. SemEval 2024 Task 1: Semantic Textual Relatedness (STR) Source
- Mohamed Abdalla et. al. , 2023 ”What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study”