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 Датасет

IMDB Dataset - датасет отзывов к фильмам, классифицированным по тональности. По 25000 положительных и отрицательных объектов.
Признаки:

- **review** - текст отзыва
- **sentiment** - тональность текста

2.3 Предобработка

Следующие обработки были применены перед применением классических алгоритмов NLP.

1. **Удаление следов html разметки.** Было замечено, что почти в каждом объекте встречаются теги по типу *

*. От них избавляемся с помощью BeautifulSoup.
2. **Токенизация.** Предложения приводились в нижний регистр и разбивались по словам из английских символов.
3. **Лемматизация.** Слова классифицировались по части речи и приводились в инфинитив с помощью WordNet.

3 Эксплоративный анализ

1. Найдите топ-300 слов по частоте без учета стоп-слов.
2. Найдите топ слов, характеризующих каждую тональность отдельно.

[бонус] Найдите еще что-то интересное в корпусе (что-то специфичное для данной темы)

Самые популярные слова:

word	count
movie	103234
film	95844
one	55427
make	46122
like	44296
…	…

Для нахождения топ слов для разных классов оценим разделяющую способность каждого слова:

$$P = \{\text{text} \mid \text{word} \in \text{text},\ \text{text}_{\text{label}} = \text{positive}\}$$

$$N = \{\text{text} \mid \text{word} \in \text{text},\ \text{text}_{\text{label}} = \text{negative}\}$$

$$\text{word}^{\text{positive}}_{\text{separating power}} = \frac{|P|}{|P| + |N| + \alpha}$$

$$\text{word}^{\text{negative}}_{\text{separating power}} = \frac{|N|}{|P| + |N| + \alpha}$$

α - коэффициент значимости, выбрал равным 100. Таким образом слова, которые встретились всего пару раз в одном классе текстов не будут иметь высокую разделяющую способность.

word	positive	negative
wonderful	0.803922	0.165913
excellent	0.798810	0.177381
super	0.796998	0.131523
beautifully	0.776181	0.121150
fantastic	0.758344	0.179852
…	…	…

Таблица 1: Топ по позитивной разделяющей способности

word	positive	negative
waste	0.084324	0.893250
awful	0.085755	0.886037
terrible	0.125291	0.845571
poorly	0.089933	0.842953
horrible	0.129672	0.832189
…	…	…

Таблица 2: Топ по негативной разделяющей способности

[бонус] Дополнительно определим количество слов с высокой разделяющей способностью (> 0.7) для тональностей. Таких 25 для положтельного и 48 для отрицательного класса. Как видим, мы имеем дело с негативным комьюнити, которое знает больше разных негативных слов, нежели позитивных. Хотя, например, слово 'zombie' чаще встречается в негативных отзывах, из чего можно сделать выводы, что фильмы про зомби так себе.

Слов с высокой разделяющей способностью мало относительно общего числа уникальных слов (85949). Значит, мы потеряем слишком много информации, если ограничим словарь только на такие слова.

Длины отзывов имеют схожие распределения. Это значит, что нету смысла вводить признак длины отзыва.

4 Описание выбранной архитектуры

5 Описание полученных результатов

6 Направления дальнейшей работы

7 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”