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

Содержание

1 Описание предметной области и постановка задачи **2**
1.1 Введение в предметную область 2
1.2 Постановка задачи 2

2 Обзор литературы **3**

3 Описание датасета **4**
3.1 Структура датасета 4
3.2 Измерение качества 4

4 Эксплоративный анализ **5**
4.1 Распределение Score 5
4.2 Схожести текстов 6

5 Бейслайн решение **7**

6 Описание выбранной базовой архитектуры **7**

7 Описание полученных результатов **8**

8 Направления дальнейшей работы **8**

9 Источники **9**

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* на английском языке. Необходимо обучить модель, которая предсказываеи STR двух предложений. Доступны тренировчный и тестовый датасеты на английском, Leaderboard.

2 Обзор литературы

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

3 Описание датасета

3.1 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.

Датасет имеет следующую структуру:

- **PairID** - уникальный id пары предложений
- **Text** - тексты, записанные подряд через ' \n'
- **Score** - оценка семантической близости предложений

Тренировочный датасет - 5500 объектов; тестовый датасет - 250.

	PairID	Text_1	Text_2	Score
0	ENG-train-0000	It that happens, just pull the plug.	if that ever happens, just pull the plug.	1.0
1	ENG-train-0001	A black dog running through water.	A black dog is running through some water.	1.0
2	ENG-train-0002	I've been searchingthe entire abbey for you.	I'm looking for you all over the abbey.	1.0
3	ENG-train-0003	If he is good looking and has a good personali...	If he's good looking, and a good personality, ...	1.0
4	ENG-train-0004	She does not hate you, she is just annoyed wit...	She doesn't hate you, she is just annoyed.	1.0

Таблица 1: Пример выборки (тексты были разбиты на пары)

3.2 Измерение качества

Официальной метрикой оценки для этого задания является коэффициент ранговой корреляции Спирмена.

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{(R(X))}\sigma_{(R(Y))}}$$

- ρ обозначает коэффициент корреляции Пирсона, но применяется к ранговым переменным
- $cov(R(X), R(Y))$ является ковариацией ранговых переменных
- $\sigma_{(R(X))}$ и $\sigma_{(R(Y))}$ являются стандартными отклонениями ранговых переменных

Коэффициент отражает, насколько хорошо предсказанные системой скоры согласуются с суждениями человека.

4 Эксплоративный анализ

Для предварительного анализа и имплементации бейслайна пары предложений в датасете были лемматизированны и очищены от стоп-слов. Сам датасет был разбит на тестовую и валидационную выборки в отношении $0.8/0.2$.

Мы удостоверились, что датасет чист от выбросов и пробелов.

4.1 Распределение Score

Посмотрим на распределение скоров по частоте и в зависимости от модуля разности длин пары предложений. Для

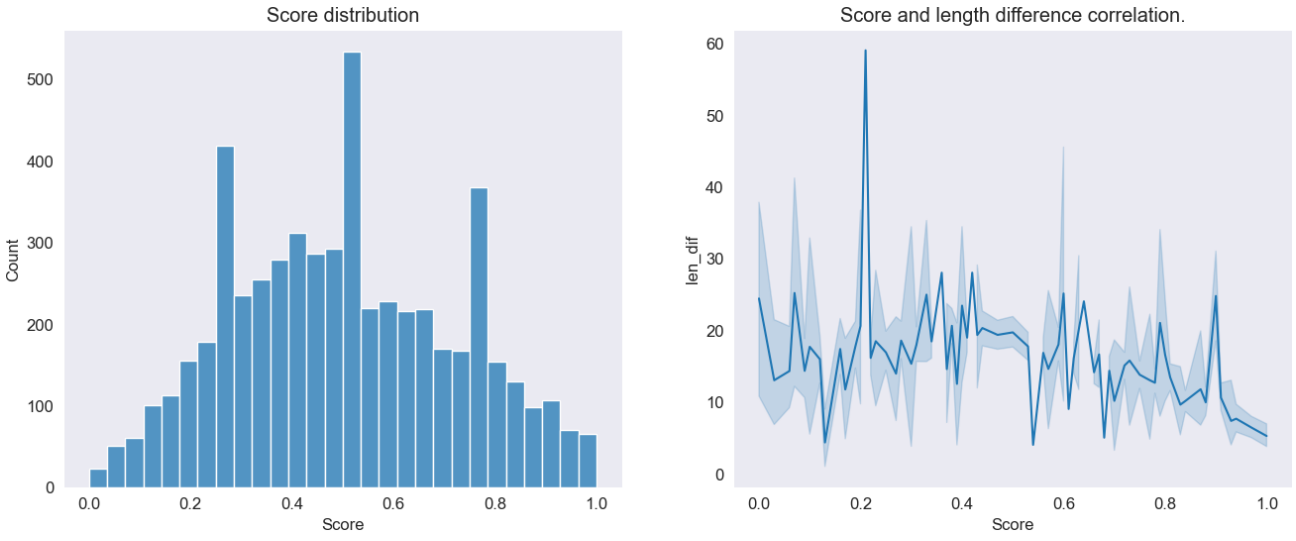

Рис. 1: Распределение Score

Распределение по целевой функции похоже на нормальное, есть концентрация плотности в значении 0.5.

Также видим, что скор обратно пропорционален разности длин текстов, что соответствует интуиции: длинные предложения обычно имеют смысл отличные от коротких.

4.2 Схожести текстов

Дополнительно взглянем на близости текстов в датасете. В качестве меры близости будем использовать WER на лемматизированных предложениях:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

где:

- *S* – количество замен,
- *D* – количество делеций,
- *I* — количество вставок,
- *C* – количество правильных слов,
- *N* — количество слов в ссылке $(N = S + D + C)$

Рассмотрим распределение WER по частоте и в зависимости от скора.

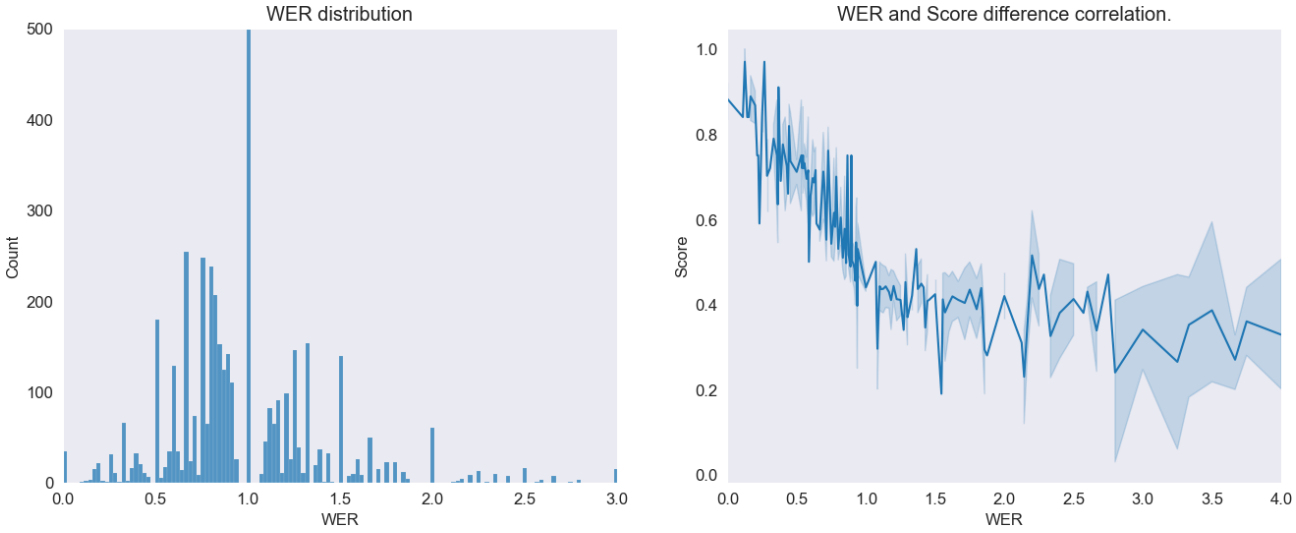

Рис. 2: Распределение WER

Мы видим, что на лемматизированных текстах WER в среднем равен 1. Это означает, что пары текстов различны морфологически в основном (то есть наборами, представляющих их слов).

Ожидаемо, чем ниже WER - тем лучше Score.

5 Бейслайн решение

В качестве бейслайна авторы предложили посчитать долю общих слов в предложениях.

Оба предложения токенизируются по словам и знакам препинания с помощью регулярного выражения, и составляются множества токенов каждого из этих предложений. После этого предсказанная степень сходства вычисляется как мощность пересечения этих множеств, делённая на суммарную мощность обоих множеств. На лемматированных текстах такой подход даёт корреляцию Пирсона 0.58.

Collab с авторским решением.

В качестве дополнительного подхода мы обучили TF_IDF модель на тестовой выборке. Над векторами был обучен RandomForestClassifier. Такой метод дал следующие скоры:

- MSE: 0.04
- Spearman correlation: 0.435
- Pearson correlation: 0.433

6 Описание выбранной базовой архитектуры

В рамках данной части проекта ставилась задача выдвинуть архитектуру модели, которая превзойдёт базовое решение организаторов по целевой метрике качества.

Для достижения этой цели использовался следующий метод.

- Для каждого предложения в выборке вычислялся его BERT-эмбеддинг как усреднённые по всем токенам в предложении выходные эмбеддинги последнего слоя модели bert-base-uncased.
- Семантическое сходство пары предложений оценивалось как косинусное сходство BERT-эмбеддингов этих предложений (косинус угла между эмбеддингами).

Использовался предобученный bert-base-uncased.

Данный метод был выбран из-за простоты реализации, а также потому, что он развивает концепцию анализа слов предложений из базового решения организаторов на более сложные признаки предложений, такие как BERT-эмбеддинги. Таким образом, можно чётко отследить вклад в качество решения, который даёт переход от анализа совместных слов к анализу BERT-эмбеддигов предложений.

Также важно отметить, что данное решение фактически относится к unsupervised-подходу, т.к. не использует целевую переменную при обучении. С одной стороны, это является недостатком из-за потери информации от целевой переменной. Однако такой подход избавляет от проблемы переобучения сложных моделей под выборку, часто возникающей в задачах с маленькой обучающей выборкой (в данном случае её размер составляет 5500 пар предложений).

В будущей работе можно попробовать улучшить качество решения путём перехода к supervised-подходу с помощью методов регуляризации.

7 Описание полученных результатов

Чтобы оценить предложенное решение, в тестирующую систему соревнования нашей командой было отправлено как наше решение, так и базовое решение организаторов для сравнения значений функционала качества.

Результаты получились следующими. Базовое решение организаторов набрало score 0.63, а решение, предложенное нами - 0.67 (значения округлены до сотых). Таким образом, наше решение превзошло базовое решение организаторов по значению целевого функционалу качества. Ссылка на посылку в контест.

Ссылка на репозиторий проекта.

8 Направления дальнейшей работы

В качестве направления дальнейшей работы предлагается дообучить Transformer-based модель на Score. Дополнительно стоит рассмотреть поведение модели при инференсе на разных ground-truth скорах - а именно, где она чаще ошибается.

Отдельным направлением хочется выделить сбор и разметку данных. Тренировочная выборка в 5500 элементов мала. Также хочется посмотреть как соотносятся предсказания реальных людей с предложенной в датасете разметкой. Эту проблему можно решить с помощью краудсорсинга, например Toloka.

9 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”