

Содержание

1	Краткое описание контрольной точки 1	2
2	Новая модель	3
3	Результаты на соревновании	4
4	Итоги	5

1 Краткое описание контрольной точки 1

Давайте немного вспомним какая наша цель.

Мы участвуем в соревновании по оценке семантической текстовой связи между предложениями. В датасете есть два предложения и оценка их семантического сходства, полученная с помощью ассессоров. Метрикой качества является коэффициент ранговой корреляции Спирмена.

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

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

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

Для бейслайн решения мы использовали TF-IDF вместе с RandomForestClassifier. Получили корреляцию 0.435.

Авторы соревнования предложили использовать в качестве бейслайна долю общих слов в предложениях. Такой подход дал корреляцию 0.63.

Наша основная модель это Bert. Мы вычисляли для каждого предложения его Bert-эмбединги, затем оценивали схожесть между предложениями как косинусное сходство этих эмбедингов. Его выбрали из-за простоты реализации, а также потому, что он улучшает анализ слов относительно базового решения организаторов.

В качестве улучшений планировалось перебрать другие архитектуры предобученных Bert-base моделей.

2 Новая модель

Для улучшения метрики качества мы стали пробовать другие Bert-base модели. Рассмотрим подробнее каждую из них.

- **ALBERT**

Модель ALBERT была предварительно обучена на [BookCorpus](#), наборе данных, состоящем из 11 038 неопубликованных книг и [English Wikipedia](#) (исключая списки, таблицы и заголовки).

Модель выдала корреляцию 0.08%.

- **RoBERTa**

Модель RoBERTa была предварительно обучена на объединении пяти наборов данных:

- [BookCorpus](#) — набор данных, состоящий из 11 038 неопубликованных книг;
- [English Wikipedia](#) (за исключением списков, таблиц и заголовков);
- [CC-News](#) - набор данных, содержащий 63 миллиона новостных статей на английском языке, просканированных в период с сентября 2016 года по февраль 2019 года;
- [OpenWebText](#), воссоздание набора данных WebText с открытыми источниками, используемый для обучения GPT-2;
- [Stories](#), содержащий подмножество данных CommonCrawl.

Модель показала корреляцию 0.07%.

- **DistilBERT**

Как и ALBERT, был предобучен на BookCorpus и English Wikipedia.

Данная модель показала качество 0.797%.

По итогу мы использовали модель DistilBert из-за маленькой выборки входных данных. Она является облегченной версией Bert (число слоёв encoder снижено с 12 до 6), которая, практически, не теряет своей эффективности, но ощутимо увеличивает скорость.

Использование небольшой модели снизило переобучение, которое наблюдалось при использовании крупных Bert-base сетей.

3 Результаты на соревновании

CodaLab - Competition

Search Competitions My Competitions Help Sign Up Sign In

#	User	Entries	Date of Last Entry	Team Name	Spearman Correlation ▲	Detailed Results
1	jgrnmnx	1	12/08/23		0.844 (1)	View
2	aazhangmiaoran	5	11/18/23		0.826 (2)	View
3	S_231110043	7	11/09/23		0.823 (3)	View
4	Ayon	7	12/10/23	Unknown	0.815 (4)	View
5	Syavaprd	5	11/11/23		0.814 (5)	View
6	S_231110608	3	10/19/23		0.808 (6)	View
7	Sls	15	12/05/23		0.806 (7)	View
8	spoker	1	12/10/23		0.797 (8)	View
9	janko0329	4	11/28/23	DDDDDDDD	0.781 (9)	View
10	S_231110606	7	11/02/23		0.753 (10)	View
11	ivan_belyaev	2	11/08/23		0.671 (11)	View
12	Pollgorka	1	10/31/23		0.469 (12)	View
13	Shilpa	1	10/17/23	SemanticLM	0.394 (13)	View

#	Username	Score
1	S_231110608	0.675

Join us on Github for contact & bug reports About Privacy and Terms v1.6

Поиск

12:11 11.12.2023

Рис. 1: Leaderboard

Выделение синим - наш текущий результат 0.797 (User poker)

Наш прошлый результат - 0.671 (User ivan belyaev)

4 Итоги

В конечном итоге у нас получилось достичь около 80% точности ответа. Как ни странно, облегчение модели помогло нам улучшить результат.

Большинство участников, которые нас превзошли, получили результаты буквально на пару процентов выше. От лучшего результата мы отстаем не более чем на 5%.