В. Н. Копнин, Н. С. Лагутина
Ярославский государственный университет, г. Ярославль
УДК 004.912

ОПРЕДЕЛЕНИЕ СЕМАНТИЧЕСКОГО СХОДСТВА ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКОВЫХ МОДЕЛЕЙ НА ОСНОВЕ ТРАНСФОРМЕРОВ

Аннотация. В работе рассматриваются результаты применения трех современных языковых моделей на основе нейросетей-трансформеров BERT для задачи определения семантического сходства текста ответа учащегося и эталонного ответа учителя. Эксперимент проведен на собственном корпусе текстов, состоящем из 1812 пар схожих фраз и словосочетаний. Качество решения задачи оценивалось с помощью пяти статистических метрик: точность, полнота, F-мера, P4, MCC.

Ключевые слова: автоматический анализ текстов, семантическое сходство текстов, языковые модели, BERT, автоматическая оценка ответов учащихся.

Введение. В рамках взаимодействия двух факультетов ЯрГУ, информатики и вычислительной техники и филологии и коммуникаций, разрабатывается проект автоматического определения языкового профиля изучающих английский язык [1]. Он заключается в оценке знаний и умений учащегося в соответствии с уровнями шкалы общеевропейских компетенций владения иностранным языком CEFR. Основной частью проекта является автоматическая проверка заданий, среди которых значительную часть составляют вопросы открытого типа, требующие ответа в виде связного текста. Оценка правильности такого ответа требует сравнения эталонного текста и текста, написанного учащимся. В области обработки естественного языка эта задача решается методами определения семантического сходства текста [2].

Оценка сходства между текстовыми данными является одной из актуальных проблем в области обработки естественного языка. Наибольшее количество работ посвящено решению этой задачи в сфере машинного перевода и информационно-поисковых систем [3]. Первые подходы к определению сходства текстов опирались на статистические методы и методы, основанные на правилах. Такие способы решения задачи не утратили свою актуальность и используются исследователями, но, как правило, в комбинации с другими. Например, в работе [4] сочетается классический подход определения схожести документов TF-IDF с семантической информацией из онтологии предметной области. Однако в настоящее время наиболее популярны и успешны методы машинного обучения и нейронные сети [5]. Для прогнозирования сходства текстов ученые разрабатывают собственную архитектуру нейронных сетей [6] или используют предобученные языковые модели и классические нейронные сети для получения числовых характеристик текста — эмбеддингов [7]. Качество решения задачи очень сильно колеблется от значения F-меры 0,46 при сравнении аннотаций к статьям с помощью ансамблей эмбеддингов [8] до 0,91 для исследования семантического сходства пар вопросов Quora с помощью эмбеддингов BERT и модели Bi-LSTM [7].

Проблема исследования. Определение сходства текста ответа ученика с эталонным ответом учителя является очень актуальной для любых естественных языков, так как классическая экспертная проверка выполнения заданий требует много времени и усилий и зависит от человеческого фактора, например, усталости [9]. Однако исследований в этой области мало даже для английского языка. Авторы статей часто сосредотачиваются на классификации ответов по оценкам, распознавании рукописных текстов и не уделяют внимания исследованию моделей семантического сходства текстов [10].

Поэтому авторы работы поставили задачу исследовать современные языковые модели применительно к оценке сходства текстов на естественном языке для определения правильности ответов учащихся на задания в области знания английского языка.

Материалы и методы. Основой решения является представление текста в виде числового вектора. Для сравнения векторов введем метрическое пространство (X, ρ). Пусть задано число ε. Векторы $x, y \in X$ назовем схожими, если значение метрики $\rho(x, y) < \varepsilon$. Два текста будем считать семантически схожими, если соответствующие им числовые векторы характеристик схожи.

В рамках взаимодействия двух факультетов ЯрГУ (информатики и вычислительной техники и филологии и коммуникаций) был составлен корпус для проведения исследования. Он состоит из 1812 уникальных пар коротких фраз и словосочетаний. Представлены как семантически схожие пары, так и не схожие.

Для формирования числовых характеристик текста, эмбеддинго, были выбраны три языковые модели на основе нейросетей трансформеров BERT.

1. bert-base-cased. Модель имеет размер в 109 миллионов параметров, 768-мерные эмбеддинги, чувствительна к регистру, обучена на книжном корпусе и корпусе википедии.
2. bert-large-cased. Модель имеет размер в 336 миллионов параметров, 768-мерные эмбеддинги, 24 слоя, 1024 скрытых измерения, чувствительна к регистру, обучена на книжном корпусе и корпусе википедии.
3. sembeddings/model_gpt_trained. Построена на базе модели DistilBERT [11], имеет размер в 65,8 миллионов параметров, 768-мерные эмбеддинги, чувствительна к регистру, обучена на книжном корпусе и корпусе википедии и на корпусе автора.

Для сравнения эмбеддингов были выбраны следующие метрики из пространства R^n:

– косинусное сходство — косинус угла между векторами, принимает значения от -1 до 1;

– коэффициент корреляции Пирсона — указывает линейную зависимость между величинами, принимает значения от -1 до 1;

– метрика Чебышева — максимум модуля разности координат векторов, принимает значения от 0 до $+\infty$;

• Евклидово расстояние — квадратный корень из суммы квадратов координат векторов, принимает значения от 0 до $+\infty$. Sanh V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter //arXiv preprint arXiv:1910.01108. — 2019;

• метрика Минковского — обобщение Евклидова расстояния и Манхэттенской метрики, принимает значения от 0 до $+\infty$.

Чтобы результаты сравнения эмбеддингов были репрезентативными, все расстояния были отображены из луча $[0; +\infty)$ в отрезок $[0; 1]$ с помощью функции $f(\rho) = (1 + \rho)^{-1}$.

Для оценки качества бинарной классификации пар текстов моделями были выбраны следующие метрики:

• полнота — доля обнаруженных моделью схожих пар текстов относительно всех семантически схожих пар;

• точность — доля пар фраз, действительно являющихся семантически схожими, относительно всех пар, классифицированных моделью как схожие;

• F-мера — среднее гармоническое полноты и точности;

- P4 — расширение F-меры, обладает свойством симметрии относительно инверсии классов;
- MCC — коэффициент корреляции Мэтьюса — коэффициент корреляции между фактическими и предсказанными моделью бинарными классификациями. Является сбалансированной мерой, применима даже в условиях сильного дисбаланса классов.

Результаты.

В результате исследования был получен корпус из 1812 уникальных пар фраз, оценены потенциалы некоторых нейросетевых классификаторов. Результаты экспериментов представлены в табл. 1.

Лучшая F-мера (0.549) оказалась у модели sembeddings/model_gpt_trained. Лучшие результаты по метрикам качества точность (0.474), P4 (0.604) и MCC (0.243) показала модель bert-base-cased при мерах. Модель bert-large-cased обнаружила лучшую полноту (0.851). Все модели показали лучшие результаты при мерах «косинусное сходство» и «коэффициент корреляции Пирсона».

Таблица 1

Оценка качества предсказания сходства текстов

Название модели	*Метрика близости*	*F-мера*	*Точность*	*Полнота*	*P4*	*MCC*	*ε*
bert-base-cased	косинусное сходство	0.527	0.474	0.594	0.604	0.243	0.13
	коэффициент корреляции Пирсона	0.527	0.474	0.594	0.604	0.243	0.13
	метрика Чебышева	0.508	0.341	1.000	0.000	0.000	1.00
	Евклидово расстояние	0.508	0.341	1.000	0.000	0.000	1.00
	метрика Минковского	0.508	0.341	1.000	0.000	0.000	1.00
bert-large-cased	косинусное сходство	0.538	0.393	0.851	0.495	0.184	0.39
	коэффициент корреляции Пирсона	0.538	0.393	0.851	0.495	0.184	0.39
	метрика Чебышева	0.508	0.341	1.000	0.000	0.000	1.00
	Евклидово расстояние	0.508	0.341	1.000	0.000	0.000	1.00
	метрика Минковского	0.508	0.341	1.000	0.000	0.000	1.00
sembeddings/ model_gpt _trained	косинусное сходство	0.549	0.424	0.780	0.563	0.227	0.54
	коэффициент корреляции Пирсона	0.549	0.423	0.780	0.563	0.227	0.54
	метрика Чебышева	0.508	0.341	1.000	0.000	0.000	1.00
	Евклидово расстояние	0.508	0.341	1.000	0.000	0.000	1.00
	метрика Минковского	0.508	0.341	1.000	0.000	0.000	1.00

Заключение. Результаты экспериментов показывают, что языковые модели на основе нейронных сетей трансформеров могут применяться для определения сходства текстов. Полученные значения метрик качества означают, что для полноценного применения в системах анализа ответов обучающихся требуется дополнительная доработка этих моделей. Перспективными направлениями могут стать комбинации эмбеддингов и дополнительных характеристик текста с учетом предметной области, применение ансамблей классификаторов, увеличение объемов обучающих корпусов текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Лагутина Н.С., Тихомиров М.В., Мастакова Н.К. Алгоритм автоматического построения языкового профиля учащегося // Заметки по информатике и математике. — 2023. — С. 58-65.
2. Qurashi A.W., Holmes V., Johnson A.P. Document processing: Methods for semantic text similarity analysis // 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). — IEEE, 2020. — С. 1-6.
3. Chandrasekaran D., Mago V. Evolution of semantic similarity—a survey // ACM Computing Surveys (CSUR). — 2021. — Т. 54, № 2. — С. 1-37.
4. Fei L. Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method [J] // Advances in Multimedia. — 2022. — С. 1-11.
5. Minaee S. et al. Deep learning--based text classification: a comprehensive review // ACM computing surveys (CSUR). — 2021. — Т. 54, № 3. — С. 1-40.
6. Wang K. et al. Comparison between calculation methods for semantic text similarity based on siamese networks // 2021 4th International Conference on Data Science and Information Technology. — 2021. — С. 389-395.
7. Viji D., Revathy S. A hybrid approach of Weighted Fine-Tuned BERT extraction with deep Siamese Bi–LSTM model for semantic text similarity identification // Multimedia Tools and Applications. — 2022. — Т. 81, № 5. — С. 6131-6157.
8. Witschard D. et al. Interactive optimization of embedding-based text similarity calculations // Information Visualization. — 2022. — Т. 21, № 4. — С. 335-353.
9. John Bernardin H. et al. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability // Human Resource Management. — 2016. — Vol. 55, № 2. — P. 321-340.
10. Sanuvala G., Fatima S. S. A study of automated evaluation of student’s examination paper using machine learning techniques // 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). — IEEE, 2021. — С. 1049-1054.
11. Sanh V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // arXiv preprint arXiv:1910.01108. — 2019.