MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 31, NO. 2, 2024
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Automatic Determination of Semantic Similarity of Student Answers With the Standard One Using Modern Models

N. S. Lagutina[1], K. V. Lagutina[1], V. N. Kopnin[1]

DOI: 10.18255/1818-1015-2024-2-194-205

[1]P.G. Demidov Yaroslavl State University, Yaroslavl, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received March 20, 2024
Revised April 11, 2024
Accepted April 17, 2024

The paper presents the results of a study of modern text models in order to identify, on their basis, the semantic similarity of English-language texts. The task of determining semantic similarity of texts is an important component of many areas of natural language processing: machine translation, information retrieval, question and answer systems, artificial intelligence in education. The authors solved the problem of classifying the proximity of student answers to the teacher's standard answer. The neural network language models BERT and GPT, previously used to determine the semantic similarity of texts, the new neural network model Mamba, as well as stylometric features of the text were chosen for the study. Experiments were carried out with two text corpora: the Text Similarity corpus from open sources and the custom corpus, collected with the help of philologists. The quality of the problem solution was assessed by precision, recall, and F-measure. All neural network language models showed a similar F-measure quality of about 86 % for the larger Text Similarity corpus and 50–56 % for the custom corpus. A completely new result was the successful application of the Mamba model. However, the most interesting achievement was the use of vectors of stylometric features of the text, which showed 80 % F-measure for the custom corpus and the same quality of problem solving as neural network models for another corpus.

Keywords: natural language processing; text similarity; text classification; neural network language models; assessing students' open responses; artificial intelligence in education

INFORMATION ABOUT THE AUTHORS

Lagutina, Nadezhda S.	ORCID iD: 0000-0002-6137-8643. E-mail: lagutinans@gmail.com PhD, associate professor
Lagutina, Ksenia V. (corresponding author)	ORCID iD: 0000-0002-1742-3240. E-mail: lagutinakv@mail.ru PhD, associate professor
Kopnin, Vladislav N.	ORCID iD: 0009-0007-5451-775X. E-mail: vlad.kopnen@mail.ru Student

Funding: Yaroslavl State University (project VIP-016).

For citation: N. S. Lagutina, K. V. Lagutina, and V. N. Kopnin, "Automatic determination of semantic similarity of student answers with the standard one using modern models", *Modeling and Analysis of Information Systems*, vol. 31, no. 2, pp. 194–205, 2024. DOI: 10.18255/1818-1015-2024-2-194-205.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 31, № 2, 2024
сайт журнала: www.mais-journal.ru

Автоматическое определение семантического сходства ответов учащихся с эталонным с помощью современных моделей

Н. С. Лагутина[1], К. В. Лагутина[1], В. Н. Копнин[1]

DOI: 10.18255/1818-1015-2024-2-194-205

[1]Ярославский государственный университет им. П.Г. Демидова, Ярославль, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 20 марта 2024 г.
После доработки 11 апреля 2024 г.
Принята к публикации 17 апреля 2024 г.

В работе представлены результаты исследования современных моделей текста с целью выявления на их основе семантической близости текстов на английском языке. Задача определения семантического сходства текстов является важной составляющей многих областей обработки естественного языка: машинного перевода, поиска информации, систем вопросов и ответов, искусственного интеллекта в образовании. Авторы решали задачу классификации близости ответов учащихся к эталонному ответу учителя. Для исследования были выбраны нейросетевые языковые модели BERT и GPT, ранее применявшиеся к определению семантического сходства текстов, новая нейросетевая модель Mamba, а так же стилометрические характеристики текста. Эксперименты проводились с двумя корпусами текстов: корпус Text Similarity из открытых источников и собственный корпус, собранный с помощью филологов. Качество решения задачи оценивалось точностью, полнотой и F-мерой. Все нейросетевые языковые модели показали близкое качество F-меры около 86 % для большего по размеру корпуса Text Similarity и 50–56 % для собственного корпуса авторов. Совсем новым результатом оказалось успешное применение модели Mamba. Однако, самым интересным достижением стало применение векторов стилометрических характеристик текста, показавшее 80 % F-меры для авторского корпуса и одинаковое с нейросетевыми моделями качество решения задачи для другого корпуса.

Ключевые слова: обработка естественного языка; сходство текстов; классификация текстов; нейросетевые языковые модели; оценка открытых ответов учащихся; искусственный интеллект в образовании

ИНФОРМАЦИЯ ОБ АВТОРАХ

Лагутина, Надежда Станиславовна	ORCID iD: 0000-0002-6137-8643. E-mail: lagutinans@gmail.com Канд. физ.-мат. наук, доцент
Лагутина, Ксения Владимировна (автор для корреспонденции)	ORCID iD: 0000-0002-1742-3240. E-mail: lagutinakv@mail.ru Канд. тех. наук, доцент
Копнин, Владислав Николаевич	ORCID iD: 0009-0007-5451-775X. E-mail: vlad.kopnen@mail.ru Студент

Финансирование: ЯрГУ (проект VIP-016).

Для цитирования: N. S. Lagutina, K. V. Lagutina, and V. N. Kopnin, “Automatic determination of semantic similarity of student answers with the standard one using modern models”, *Modeling and Analysis of Information Systems*, vol. 31, no. 2, pp. 194–205, 2024. DOI: 10.18255/1818-1015-2024-2-194-205.

Введение

Эффективный контроль знаний обучающихся в тестовых и письменных заданиях обязательно включает ответы в виде связного текста. Классический подход к проверке выполнения таких заданий преподавателем очень трудоёмкий, утомительный и субъективный. По мере развития методов обработки естественного языка и искусственного интеллекта осуществляются исследования в области автоматизации этого процесса [1].

Основной задачей проверки ответов учащихся является определение их сходства с эталонным ответом учителя. Автоматическое определение сходства текста — это, как правило, вычисление расстояния между двумя фрагментами текста, представляющего степень их близости в естественном языке [2]. Решение этой задачи можно рассматривать в двух аспектах: лексическое сходство и семантическое сходство [3].

Тексты схожи лексически, если они имеют одинаковые или однокоренные слова или последовательности слов. Автоматическое определение лексического сходства базируется на статистических методах анализа текста на символьном уровне и уровне слов. Однако в области ответов на вопросы этот аспект сильно сужает набор правильных вариантов и оставляет за его пределами фразы, отличающиеся от эталона структурой, синонимами, особенностями словоупотреблений.

Тексты обладают семантическим сходством, если они используются в одном и том же контексте, содержат одинаковую информацию и значение. Определение семантического сходства текстов представляет собой сложную задачу, так как требует учёта значений слов, контекстной информации, синтаксической структуры фраз. Огромный потенциал для её решения появляется с развитием больших языковых моделей (LLM): BERT, GPT, Mamba.

Исследователи предлагают большое количество вариантов определения сходства текстов, но часто ограничиваются решением в конкретной предметной области, строя узкоспециализированную модель текста [4]. Такой подход сильно ограничивает рамки применения разработанного решения и осложняет выбор модели в каждом конкретном случае. Поэтому, когда в ходе разработки программной системы построения языкового профиля обучающегося [5] возникла необходимость проверки открытых ответов, авторы работы поставили перед собой задачу исследовать, насколько качественно современные модели текста могут определять близость текстов на английском языке. В данной работе под сходством текстов понимается смысловая (семантическая) близость свободно сконструированного ответа учащегося с заведомо правильным, эталонным ответом преподавателя [6].

1. Обзор научных исследований

Первые методы определения сходства текста измеряли близость слов и предложений на основе оцифровки последовательности символов или строк, из которых они состоят. Полученные числовые векторы сравнивались математическими метриками расстояний: евклидово, косинусное, манхэттенское, расстояние Хэмминга [7]. Часто сходство вычислялось путем выделения самой длинной общей подстроки [8] или с использованием n-грамм [9].

Авторы работы [10] заметили, что данный подход мало учитывает семантику слов и добавили информацию о близости слов на основе WordNet. Эту же идею использовали исследователи сходства арабских текстов для оценки ответов учащихся [8]. Однако, выявление семантики текста с помощью тезаурусов и онтологий больше характерно для задач по поиску сходных данных в предметных областях [11].

Классическими методами выявления семантической информации считаются вычисление характеристик TF-IDF и латентный семантический анализ (latent semantic analysis, LSA) [11]. Молер и Михалча [12] исследовали LSA, но также предложили методы классификации путем извлечения лексических и синтаксических характеристик. Исследователи подготовили корпус Mohler's Dataset

эталонных ответов и ответов учащихся в области компьютерных наук на английском языке, размеченных по пятибальной системе.

Большая часть исследований в области оценки открытых ответов учащихся классифицирует тексты по оценкам или баллам [13]. В последнее десятилетие наиболее популярными моделями для решения этой задачи стали эмбеддинги, сначала, такие как Word2Vec и GloVe [14], потом построенные с помощью методов глубокого обучения [15]. Авторы работы [16] показали, что эмбеддинги BERT превосходят по качеству GloVe при определении сходства ответов с эталонными для корпуса Mohler's Dataset. Цамус и Филигхера [17] классифицировали набор данных SemEval-2013 из ответов учащихся, размеченных на три класса: правильный, неправильный и противоречивый. Они использовали различные модели BERT и получили F-меру от 67 % до 79 %.

Классификация ответов с точки зрения оценок важна для автоматизации контроля знаний, но непосредственное определение сходства текста с эталонным позволяет решать и другие задачи в области искусственного интеллекта в образовании (Artificial Intelligence in Education), например, организацию обратной связи или оценку профессиональных компетенций.

Современные решения по сравнению текстов используют идею векторного представления текстов и дальнейшего сравнения или классификации векторов. В работе [18] реализован гибридный подход с использованием эмбеддингов BERT в комбинации с моделью на основе нейронной сети Bi-LSTM. Авторы определяли сходство пар вопросов из набора данных Quora. Значение F-меры оказалось достаточно высоким: 91 %. Сравнение аннотаций статей оказалось намного сложнее. Витчард и др. [19] комбинировали эмбеддинги BERT и USE (Universal sentence encoder for English) и получили F-меру всего лишь 46 %. Исследователи отмечают высокую трудоёмкость применения больших языковых моделей и сильную зависимость от объёма и качества используемых корпусов текстов [20].

Сильное различие в качестве определения сходства текста побуждает исследователей строить комплексные и ансамблевые модели. Например, в работе [21] объединяются эмбеддинги FastText и n-граммы. Сравнение слов через ансамбль четырёх мер сходства осуществляют Хассан и др. [22]. Комбинацию эмбеддингов для инструмента визуализации поиска похожих текстов используется в работе [19]. Авторы исследования [23] сопоставляют ключевые слова и фразы с использованием ансамблей эмбеддингов BERT, GPT, параметров связей на основе WordNet, алгоритма сравнения строк Джаро-Винклера, и достигают максимального значения F-меры 74 %. Коллектив учёных [11] обращает внимание на важную роль синтаксической информации при сравнении текстов.

Следует заметить, что в области определения сходства текстов остаются малоисследованными классические лексические характеристики. Поэтому для своей работы авторы выбрали как языковые модели BERT, GPT, совсем новую Mamba, так и стилометрические параметры уровня символов, слов и структуры фраз и предложений.

2. Метод определения близости текстов

2.1. Основные этапы метода

Для определения близости текстов авторы применили одну из методологий, общепринятых в современных исследованиях. Это методология для работы с алгоритмами без обучения. Метод определения близости текстов включает в себя следующие этапы:

1. Векторное моделирование текста. Текст на естественном языке преобразуется в вектор чисел на основе стилометрической или предобученной языковой модели.
2. Вычисление близости векторов текстов. Для пар векторов текстов считается значение метрики близости.

Table 1. Parameters of text corpora **Таблица 1.** Параметры корпусов текстов

Корпус	Кол-во пар текстов	Близкие	Неблизкие	Уникальные	Ср. длина
Text Similarity	2142	1613	529	2597	5.18
Собственный корпус	1813	618	1195	683	6.09

3. Определение близости текстов. Для пар векторов на основе значения метрики близости дается ответ, являются ли тексты близкими (да/нет). Это осуществляется с помощью порогового значения: если метрика близости выше заданного порога, то тексты близкие, иначе нет.
4. Измерение качества метода. Ответы метода сравниваются с правильными с помощью стандартных метрик качества. Правильные ответы имеются в исходных корпусах текстов, где определены пары не только близких текстов, но и текстов, близкими не считающихся.

Близость текстов определяется для двух корпусов, состоящих из англоязычных текстов различных размеров и тематик. В следующих разделах подробно описаны корпуса, модели текстов, метрики близости и метрики качества, применяемые в методе.

2.2. Корпуса текстов

Для экспериментальной апробации векторных моделей были выбраны два корпуса англоязычных текстов: один из открытых источников и один собственный корпус.

Корпус Text Similarity[1] содержит 2142 пары текстов. Тексты представляют собой короткие фрагменты по тематике акций, взятые из источника Quovo. В корпусе содержится 1613 пар близких текстов и 529 пар неблизких.

Собственный корпус был собран авторами и их коллегами-филологами. Он содержит 1813 пар коротких текстов, которые являются ответами студентов на вопросы или эталонными ответами с точки зрения преподавателя. Вопросы задавались преподавателями-филологами с целью узнать уровень английского языка студентов по шкале CEFR. В корпусе содержится 618 пар близких текстов и 1195 пар неблизких.

В таблице 1 приведены сравнительные параметры корпусов: количество пар текстов, количество пар близких и неблизких текстов, число уникальных текстов и средняя длина текстов в словах.

2.3. Векторные модели

Каждый текст из корпуса моделируется как вектор чисел. Используемые модели делятся на две категории: стилометрические и предобученные языковые.

Стилометрические модели подразумевают подсчёт стандартных статистических характеристик стиля текста трёх уровней.

1. Уровень символов. Он включает в себя следующие характеристики:
   - количество букв;
   - количество символов;
   - количество предложений;
   - средняя длина предложения в символах;
   - частоты встречаемости букв;
   - частоты встречаемости знаков препинания.
2. Уровень слов. Он включает в себя следующие характеристики:
   - n-граммы слов, $n = 1, 2, 3$. Среди всех n-грамм в корпусе выбирается 40 самых частых для каждого значения n, а для отдельных текстов считается частота встречаемости отдельных n-грамм относительно отобранных;
   - количество слов;
   - средняя длина предложения в словах;

[1] https://www.kaggle.com/datasets/rishisankineni/text-similarity

- средняя длина предложения в символах.

3. Уровень структуры. Текст разделяется на слова, для слов вычисляются части речи, т. е. текст представляется как последовательность частей речи. Далее считаются n-граммы частей речи, $n = 1, 2, 3, 4$. Среди всех n-грамм в корпусе выбирается 40 самых частых для каждого знначения n, а для отдельных текстов считается частота встречаемости отдельных n-грамм относительно отобранных.

Описанные характеристики уже успешно показали себя в обработке текстов в предыдущих работах авторов [24].

Вычисление характеристик всех уровней происходит с использованием Python-библиотеки Stanza [25]: она разделяет текст на слова и определяет части речи. Каждый из уровней считается как одна модель текста, также уровни комбинируются попарно или в тройку конкатенацией векторов характеристик, результат считается новой моделью. Элементы в векторах характеристик размещены по заданным позициям, каждая позиция соответствует отдельной характеристике, чтобы при сопоставлении векторов числа на одних и тех же позициях соответствовали друг другу по смыслу.

Вторая категория моделей — это предобученные языковые модели, созданные на основе современных нейросетевых архитектур.

- BERT для английского языка и его вариации:
  - bert-base-cased — базовая версия BERT для английского языка, учитывающая регистр;
  - bert-base-uncased — базовая версия BERT для английского языка, не учитывающая регистр;
  - bert-large-cased — увеличенная версия BERT для английского языка, учитывающая регистр;
  - bert-large-uncased — увеличенная версия BERT для английского языка, не учитывающая регистр;
  - bert-large-cased-whole-word-masking — увеличенная версия BERT для английского языка, учитывающая регистр и маскирующая сразу все токены, соответствующие слову;
  - bert-large-uncased-whole-word-masking — увеличенная версия BERT для английского языка, не учитывающая регистр и маскирующая сразу все токены, соответствующие слову.
- GPT для английского языка и его вариации:
  - openai-community/gpt2 — базовая версия GPT-2 для английского языка с 124 миллионами параметров;
  - openai-community/gpt2-medium — средняя версия GPT-2 для английского языка с 355 миллионами параметров;
  - openai-community/gpt2-large — большая версия GPT-2 для английского языка с 774 миллионами параметров;
  - sembeddings/model_gpt_trained — версия GPT-2 для английского языка, предназначенная для задач по анализу значения текстов, в том числе близости текстов;
  - sembeddings/gptops_finetuned_mpnet_gpu_v1 — версия GPT-2 для английского языка, предназначенная для задач кластеризации и семантического поиска.
- Mamba для английского языка и её вариации:
  - Q-bert/Mamba-130M — базовая версия Mamba для английского языка с 130 миллионами параметров;
  - Q-bert/Mamba-1B — базовая версия Mamba для английского языка с миллиардом параметров.

Все языковые модели являются предобученными. Они взяты из открытого ресурса Hugging Face[2]. Модели сопровождаются собственными токенизаторами, которые преобразуют исходный

[2] https://huggingface.co/models

текст на естественном языке в последовательность токенов, подающихся на вход модели. На выходе для каждого текста формируется эмбеддинг — вектор действительных чисел фиксированного размера.

2.4. Метрики близости

Пары текстов, представленные как векторы чисел одинаковой длины, сравниваются с помощью метрик близости:

- косинусное сходство — косинус угла между векторами, принимает значения от −1 до 1;
- коэффициент корреляции Пирсона — метрика линейной зависимости между двумя величинами, принимает значения от −1 до 1;
- метрика Чебышева — максимум модуля разности компонент векторов;
- евклидово расстояние — квадратный корень из суммы квадратов разностей соответствующих компонент векторов;
- метрика Минковского — метрика-обобщение евклидова и манхэттенского расстояний.

Таким образом для корпуса текстов считается набор чисел по одной из указанных метрик, в итоге числа оказываются в одном диапазоне. Каждой паре векторов текстов сопоставляется число. Чтобы вынести вердикт, близки тексты или нет, выбирается порог значения метрики. Если значение метрики близости для пары текстов оказывается выше порога, то тексты считаются близкими, иначе — нет. Порог значения метрики является гиперпараметром описываемого метода и подбирается путём перебора.

2.5. Метрики качества

Метод в результате подсчёта метрики близости и применения порога находит для корпуса из пар текстов ответы да/нет о близости пар. Полученные ответы сравниваются с правильными с помощью следующих метрик качества:

- точность — доля пар текстов, действительно близких, относительно всех пар текстов которые метод отметил как близкие;
- полнота — доля найденных методом пар близких текстов, принадлежащих классу относительно всех пар близких текстов;
- F-мера — среднее гармоническое точности и полноты.

Метрики являются стандартными для классификационных задач.

3. Эксперименты по определению близости текстов

Метод измерения близости текстов применялся для обоих корпусов. С каждым корпусом проводились отдельные эксперименты: строились вектора текстов с помощью всех моделей, затем для пары модель+метрика близости методом перебора значений выбирался порог таким образом, чтобы результат работы метода показывал лучшую F-меру.

Оба корпуса содержат в себе как пары близких, так и пары неблизких текстов. Поэтому результаты классификации для всех моделей получились достаточно показательными. Лучшие характеристики качества приведены в таблицах 2 и 3, где для каждой модели выбирался результат с наибольшей F-мерой.

В первом столбце указывается модель по техническому имени с Hugging Face или по названию уровня стилометрических характеристик/комбинаций. Во втором столбце указываются метрики близости, для которых была достигнута наибольшая F-мера. В большинстве случаев все метрики близости позволили достичь одного и того же максимума по F-мере, для некоторых моделей такими метриками оказались косинусное сходство (cos), коэффициент корреляции Пирсона (Pearson) или метрика Чебышева (Chebyshev). В остальных столбцах приведены значения метрик качества в процентах.

Table 2. Classification results for the Text Similarity corpus

Таблица 2. Результаты классификации для корпуса Text Similarity

Модель	Метрика близости	F1-мера	Точность	Полнота
bert-base-cased	cos, Pearson	**86.01**	**76.21**	**98.70**
bert-base-uncased	все	85.91	75.30	100.00
bert-large-cased	все	85.91	75.30	100.00
bert-large-cased-whole-word-masking	все	85.91	75.30	100.00
bert-large-uncased	cos, Pearson	86.07	76.01	99.19
bert-large-uncased-whole-word-masking	все	85.91	75.30	100.00
openai-community/gpt2	все	85.91	75.30	100.00
openai-community/gpt2-medium	все	85.91	75.30	100.00
openai-community/gpt2-large	cos, Pearson	86.52	77.02	98.70
sembeddings/model_gpt_trained	cos, Pearson	87.65	80.87	95.66
sembeddings/gptops_finetuned_mpnet_gpu_v1	cos, Pearson	**87.89**	**79.91**	**97.64**
Q-bert/Mamba-130M	Pearson	**86.37**	**76.71**	**98.82**
Q-bert/Mamba-1B	Pearson	86.42	76.90	98.64
Символы	все	85.91	75.30	100.00
Структура	Pearson	86.24	76.96	98.07
Слова	Pearson	**86.33**	**76.52**	**99.01**
Символы+Структура	все	85.91	75.30	100.00
Символы+Слова	Chebyshev	85.93	75.36	99.94
Структура+Слова	все	85.91	75.30	100.00
Символы+Структура+Слова	все	85.91	75.30	100.00

На корпусе Text Similarity (2) все модели добились примерно одинакового качества результата: 85–87 % F-меры. Среди BERT-моделей можно признать лучшими bert-base-cased и bert-base-uncased, так как они при меньшем размере показали те же результаты, что и вариации bert-large. Эмбеддинги Mamba сработали аналогично. Для GPT-моделей можно отметить, что эмбеддинги от sembeddings, ориентированные именно на задачи анализа семантики и поиск близких текстов, не дали прироста качества. Стилометрические модели показали то же качество результата, что и эмбеддинги. Комбинации пар и тройки уровней стилометрических характеристик не увеличили точность, полноту и F-меру.

Для собственного корпуса (3) приведены модели, показавшие лучшее качество. Эмбеддинги показали низкое качество, среди них можно отметить только sembeddings/gptops_finetuned_mpnet_gpu_v1, превзошедшую остальные. Зато стилометрические характеристики позволили выбрать близкие тексты с качеством около 80 %. Это ниже на 5 п. п., чем на корпусе Text Similarity, зато существенно выше, чем результаты моделей на основе эмбеддингов. При этом все стилометрические характеристики показали примерно один и тот же результат, комбинация качество определения близости текстов не повысила.

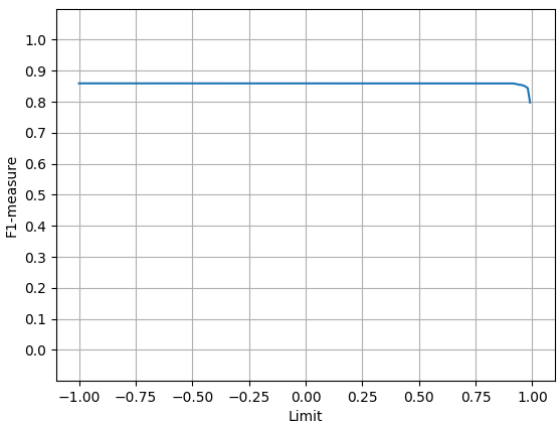
Достигнутые значения F-меры для стилометрических моделей не сильно зависят от порога, по которому происходит определение близости текстов. На рис. 1 для корпуса Text Similarity взяты стилометрические характеристики и порог от −1 до 1. Косинусная метрика на всём диапазоне позволяет достичь 85.91 % F-меры, и такая стабильная ситуация случилась для большинства пар стилометрическая модель-метрика близости. В качестве исключения можно привести косинусную метрику в комбинации с уровнем структуры (1b).

Для моделей на основе эмбеддингов значение порога имеет важную роль для обоих корпусов. На рис. 2a до порога в 0.5 показываются хорошие результаты, и при значении около 0.5 находится

Table 3. Classification results for the custom corpus

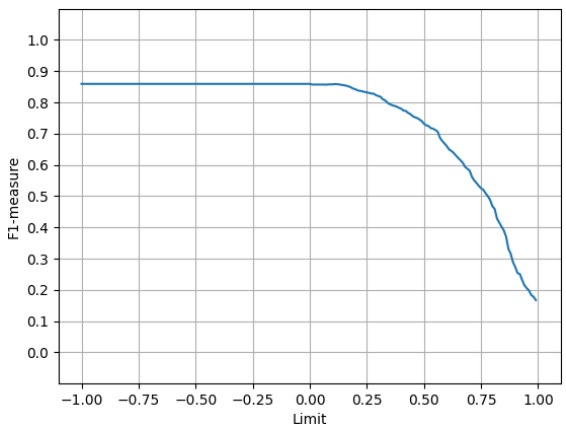
Таблица 3. Результаты классификации для собственного корпуса

Модель	Метрика близости	F1-мера	Точность	Полнота
bert-base-uncased	cos, Pearson	54.33	40.69	81.72
bert-large-uncased	cos, Pearson	51.80	36.64	88.35
bert-large-uncased-whole-word-masking	Chebyshev	51.53	36.60	87.06
openai-community/gpt2	cos, Pearson	50.86	34.11	100.00
openai-community/gpt2-large	cos	51.61	37.96	80.58
sembeddings/gptops_finetuned_mpnet_gpu_v1	cos, Pearson	**56.37**	**46.05**	**72.65**
Q-bert/Mamba-130M	Pearson	50.86	34.11	100.00
Q-bert/Mamba-1B	cos	51.75	35.45	95.79
Символы	все	**80.30**	**67.09**	**100.00**
Структура	все	80.25	67.02	100.00
Слова	все	80.25	67.02	100.00
Символы+Структура+Слова	все	80.28	67.05	100.00



a)

Fig. 1. Dependence of F=measure on the threshold for the Text Similarity corpus and stylometric features a) of all levels b) of the structure level with the cosine metric



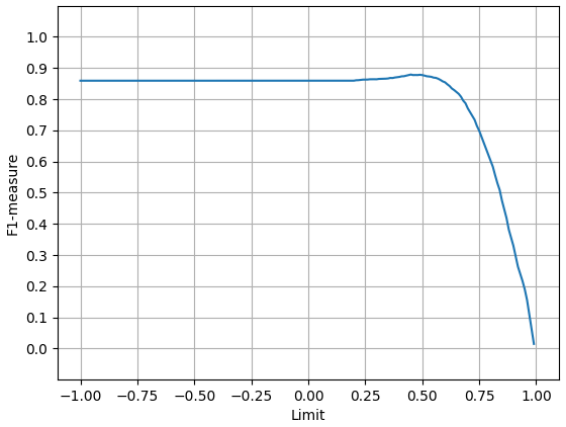
b)

Рис. 1. Зависимость F-меры от порога для корпуса Text Similarity и стилометрических характеристик a) всех уровней b) уровня структуры с косинусной метрикой

наибольшая F-мера. То же происходит и для второго корпуса, как показано на рис. 2b: лучшая F-мера достигается только при значении порога около 0.5. Стоит отметить, что порог оказывается одинаковым для различных метрик близости.

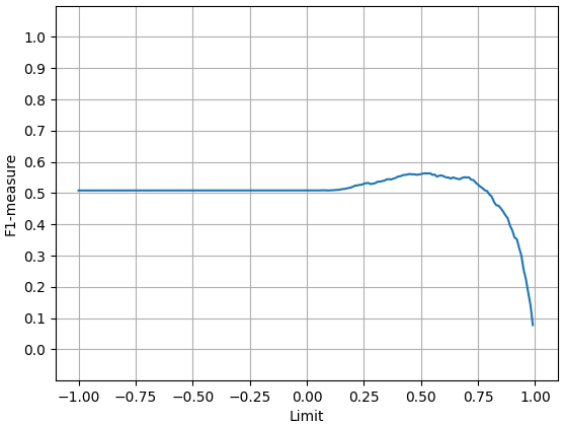
Таким образом, для собственного корпуса с короткими текстами на специфическую тематику лучше оказалось использовать стилометрические модели для определения близости. Нейросетевые модели показали для него слишком низкий результат: 51–56 % F-меры. Вероятно, такое качество работы моделей на основе эмбеддингов связано с размерами корпусов. Обычно успешные эксперименты с эмбеддингами получаются на корпусах из нескольких тысяч текстов, в то время как у собственного корпуса всего несколько сотен близких текстов.

Корпус Text Similarity также является тематическим, но в нём существенно больше близких текстов. Их поиск оказался успешнее для нейростетевых моделей: 87.89 % F-меры при максимуме в 56.37 % для этих же моделей при работе с собственным корпусом. Результат достаточно высокий и стабильный, так как все модели показали близкие результаты качества. Однако порог в 88 % F-меры не преодолела ни одна модель.



a)

Fig. 2. Dependence of F-measure on the threshold for the sembeddings/gptops_finetuned_mpnet_gpu_v1 model a) on the Text Similarity corpus with the cosine metric b) on the custom corpus with the Pearson correlation coefficient



b)

Рис. 2. Зависимость F-меры от порога для модели sembeddings/gptops_finetuned_mpnet_gpu_v1 a) на корпусе Text Similarity с косинусной метрикой b) на собственном корпусе с коэффициентом корреляции Пирсона

Следует отметить, что все модели показали высокое значение полноты определения близости текстов: 80–100 % для собственного корпуса и 98–100 % для корпуса Text Similarity. Это означает, что модели позволяют найти все близкие тексты, но при этом не справляются с некоторыми случаями отсечения неблизких текстов. Анализ ошибок и поиск усовершенствованных лингвистических моделей может повысить точность решения задачи.

Заключение

Результаты работы показывают, что задача сравнения ответа учащегося с эталонным достаточно хорошо решается методами определения сходства числовых характеристических векторов. Языковые нейросетевые модели BERT и GPT показывают лучшие результаты на корпусе большего объёма. Новизной исследования является использование модели Mamba. Качество решения задачи с её помощью практически совпадает с другими нейросетевыми моделями для корпуса Text Similarity и немного превосходит для собственного корпуса авторов. Кроме того, результаты экспериментов показывает успешность этой новой языковой модели.

Интересным результатом является высокая F-мера классификации близких текстов с помощью стилометрических характеристик. Особенно высокий результат получился для авторского корпуса. Скорее всего это обусловлено лексическими особенностями этого корпуса, так как классические нейросетевые модели показали существенно меньший результат, в то время как на другом корпусе качество практически совпадает. Этот факт открывает большие перспективы для более широкого исследования возможности применения стилометрии в области определения сходства текстов.

References

[1] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. Srinivasa, “Automatic assessment of text-based responses in post-secondary education: A systematic review”, *Computers and Education: Artificial Intelligence*, vol. 6, p. 100 206, 2024. DOI: 10.1016/j.caeai.2024.100206.

[2] J. Wang and Y. Dong, “Measurement of text similarity: A survey”, *Information*, vol. 11, no. 9, p. 421, 2020. DOI: 10.3390/info11090421.

[3] A. Rozeva and S. Zerkova, “Assessing semantic similarity of texts–methods and algorithms”, *AIP Conference Proceedings*, vol. 1910, no. 1, p. 060 012, 2017. DOI: 10.1063/1.5014006.

[4] P. D. Wibisono, A. Asad, and A. Chintan, "Short text similarity measurement methods: A review", *Soft Computing*, vol. 25, pp. 4699–4723, 2021. DOI: 10.1007/s00500-020-05479-2.

[5] N. S. Lagutina, M. V. Tihomirov, and N. K. Mastakova, "Algoritm avtomaticheskogo postroeniya yazykovogo profilya uchashchegosya", *Zametki po informatike i matematike*, no. 15, pp. 58–65, 2023, in Russian.

[6] O. B. Mishunin, A. P. Savinov, and D. I. Firstov, "Sostoyanie i uroven' razrabotok sistem avtomaticheskoj ocenki svobodnyh otvetov na estestvennom yazyke", *Modern high technologies*, no. 1, pp. 38–44, 2016, in Russian.

[7] L. Zahrotun, "Comparison Jaccard similarity, cosine similarity and combined both of the data clustering with Shared Nearest Neighbor method", *Computer Engineering and Applications Journal*, vol. 5, no. 1, pp. 11–18, 2016. DOI: 10.18495/comengapp.v5i1.160.

[8] H. A. Abdeljaber, "Automatic Arabic short answers scoring using longest common subsequence and Arabic WordNet", *IEEE Access*, vol. 9, pp. 76 433–76 445, 2021. DOI: 10.1109/ACCESS.2021.3082408.

[9] S. Sultana and I. Biskri, "Identifying similar sentences by using n-grams of characters", in *Recent Trends and Future Technology in Applied Intelligence: Proceedings of 31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, Springer, 2018, pp. 833–843. DOI: 10.1007/978-3-319-92058-0_80.

[10] S. Vij, D. Tayal, and A. Jain, "A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs", *Wireless Personal Communications*, vol. 111, pp. 1271–1282, 2020. DOI: 10.1007/s11277-019-06913-x.

[11] Y. Zhou, C. Li, G. Huang, Q. Guo, H. Li, and X. Wei, "A short-text similarity model combining semantic and syntactic information", *Electronics*, vol. 12, no. 14, p. 3126, 2023. DOI: 10.3390/electronics12143126.

[12] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading", in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 2009, pp. 567–575.

[13] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity", *Concurrency and Computation: Practice and Experience*, vol. 33, no. 5, e5971, 2021. DOI: 10.1002/cpe.5971.

[14] S. Roy, S. Dandapat, A. Nagesh, and Y. Narahari, "Wisdom of students: A consistent automatic short answer grading technique", in *Proceedings of the 13th International Conference on Natural Language Processing*, 2016, pp. 178–187.

[15] A. Ahmed, A. Joorabchi, and M. J. Hayes, "On deep learning approaches to automated assessment: Strategies for short answer grading", in *Proceedings of the 14th International Conference on Computer Supported Education*, vol. 2, 2022, pp. 85–94. DOI: 10.5220/0011082100003182.

[16] A. Ahmed, A. Joorabchi, and M. J. Hayes, "On the application of sentence transformers to automatic short answer grading in blended assessment", in *Proceedings of the 33rd Irish Signals and Systems Conference (ISSC)*, IEEE, 2022, pp. 1–6. DOI: 10.1109/ISSC55427.2022.9826194.

[17] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading", in *Proceedings of the 21st International Conference Artificial Intelligence in Education, Part II 21*, Springer, 2020, pp. 43–48. DOI: 10.1007/978-3-030-52240-7_8.

[18] D. Viji and S. Revathy, "A hybrid approach of weighted fine-tuned BERT extraction with deep Siamese Bi-LSTM model for semantic text similarity identification", *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6131–6157, 2022. DOI: 10.1007/s11042-021-11771-6.

[19] D. Witschard, I. Jusufi, R. M. Martins, K. Kucher, and A. Kerren, "Interactive optimization of embedding-based text similarity calculations", *Information Visualization*, vol. 21, no. 4, pp. 335–353, 2022. DOI: 10.1177/14738716221114372.

[20] T. Brown *et al.*, "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[21] D. Shashavali *et al.*, "Sentence similarity techniques for short vs variable length text using word embeddings", *Computación y Sistemas*, vol. 23, no. 3, pp. 999–1004, 2019. DOI: 10.13053/cys-23-3-3273.

[22] B. Hassan, S. E. Abdelrahman, R. Bahgat, and I. Farag, "UESTS: An unsupervised ensemble semantic textual similarity method", *IEEE Access*, vol. 7, pp. 85 462–85 482, 2019. DOI: 10.1109/ACCESS.2019.2925006.

[23] I. Gagliardi and M. T. Artese, "Ensemble-based short text similarity: An easy approach for multilingual datasets using transformers and wordnet in real-world scenarios", *Big Data and Cognitive Computing*, vol. 7, no. 4, p. 158, 2023. DOI: 10.3390/bdcc7040158.

[24] N. Lagutina, K. Lagutina, A. Brederman, and N. Kasatkina, "Text classification by CEFR levels using machine learning methods and BERT language model", *Modeling and Analysis of Information Systems*, vol. 30, no. 3, pp. 202–213, 2023. DOI: 10.18255/1818-1015-2023-3-202-213.

[25] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.