MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 32, NO. 1, 2025
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

A Survey of Models for Automatic Assessment of Similarity of Student's Answer to the Reference Answer

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

DOI: 10.18255/1818-1015-2025-1-42-65

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

MSC2020: 68T50
Research article
Full text in Russian

Received February 10, 2025
Revised February 20, 2025
Accepted February 26, 2025

The development of automatic assessment systems is a relevant task designed to simplify the routine work of a teacher and speed up feedback for a student. The survey is devoted to research in the field of automatic assessment of student answers based on a teacher's reference answer. The authors of the work analyzed text models used for the tasks of automatic assessment of short answers (ASAG) and automated essay assessment (AES). Several approaches were also taken into account for the task of determining the text similarity, since it is a close task, and the methods for solving it can also be useful for analyzing student answers.

Text models can be divided into several large categories. The first is linguistic models based on various stylometric features, both simple ones like a bag of words and n-grams, and complex ones like syntactic and semantic ones. The authors attributed neural network models based on various embeddings to the second category. It highlights large language models as universal, popular and high-quality modeling methods. The third category includes combined models that unite both linguistic features and neural network embeddings. A comparison of modern studies on models, methods and quality metrics showed that the trends in the subject area coincide with the trends in computational linguistics in general. A large number of authors choose large language models to solve their problems, but standard features remain in demand. It is impossible to single out a universal approach; each subtask requires a separate choice of method and adjustment of its parameters. Combined and ensemble approaches allow achieving higher quality than other methods. The vast majority of studies examine texts in English. However, successful results for national languages are also found. It can be concluded that the development and adaptation of methods for assessing students' answers in national languages is a relevant and promising task.

Keywords: natural language processing; text similarity; text classification; neural network language models; assessing students' answers; 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

Funding: This work was supported by a grant from the Russian Science Foundation (Project no. 25-21-00196).

For citation: N. S. Lagutina and K. V. Lagutina, "A survey of models for automatic assessment of similarity of student's answer to the reference answer", *Modeling and Analysis of Information Systems*, vol. 32, no. 1, pp. 42–65, 2025. DOI: 10.18255/1818-1015-2025-1-42-65.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 32, № 1, 2025
сайт журнала: www.mais-journal.ru

Обзор моделей автоматической оценки сходства ответа учащегося с эталонным ответом

Н. С. Лагутина[1], К. В. Лагутина[1]

DOI: 10.18255/1818-1015-2025-1-42-65

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

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 10 февраля 2025 г.
После доработки 20 февраля 2025 г.
Принята к публикации 26 февраля 2025 г.

Разработка систем автоматического оценивания является актуальной задачей, призванной упростить рутинный труд учителя и ускорить обратную связь для учащегося. Обзор посвящён исследованиям в области автоматической оценки ответов учащихся на основе эталонного ответа учителя. Авторы работы проанализировали модели текстов, применяемые для задач автоматической оценки коротких ответов (ASAG) и автоматизированной оценки эссе (AES). Также принималось во внимание несколько подходов для задачи определения близости текстов, так как она является аналогичной задачей, и методы её решения могут быть полезны и для анализа ответов студентов.

Модели текста можно разделить на несколько больших категорий. Первая — это лингвистические модели, основанные на разнообразных стилометрических характеристиках, как простых вроде мешка слов и n-грамм, так и сложных вроде синтаксических и семантических. Ко второй категории авторы отнесли нейросетевые модели, основанные на разнообразных эмбеддингах. В ней выделяются большие языковые модели как универсальные, популярные и качественные методы моделирования. Третья категория включает в себя комбинированные модели, которые объединяют в себе как лингвистические характеристики, так и нейросетевые эмбеддинги. Сравнение современных исследований по моделям, методам и метрикам качества показало, что тренды в предметной области совпадают с трендами в компьютерной лингвистике в целом. Большое количество авторов выбирают для решения своих задач большие языковые модели, но и стандартные характеристики остаются востребованными. Универсальный подход выделить нельзя, каждая подзадача требует отдельного выбора метода и настройки его параметров. Комбинированные и ансамблевые подходы позволяют достичь более высокого качества, чем остальные методы. В подавляющем большинстве работ исследуются тексты на английском языке. Однако успешные результаты для национальных языков также встречаются. Можно сделать вывод, что разработка и адаптация методов оценки ответов студентов на национальных языках является актуальной и перспективной задачей.

Ключевые слова: обработка естественного языка; сходство текстов; классификация текстов; нейросетевые языковые модели; оценка 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

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 25-21-00196.

Для цитирования: N. S. Lagutina and K. V. Lagutina, “A survey of models for automatic assessment of similarity of student’s answer to the reference answer”, *Modeling and Analysis of Information Systems*, vol. 32, no. 1, pp. 42–65, 2025. DOI: 10.18255/1818-1015-2025-1-42-65.

Введение

Определение и оценка сходства ответа учащегося с эталонным ответом учителя является основой эффективного контроля знаний. Классический подход к проверке преподавателем выполненных заданий с развернутым ответом очень трудоёмкий, утомительный и субъективный [1]. Автоматическая оценка работ позволяет упростить и ускорить этот процесс, основываясь на объективных критериях, что способствует более справедливому отношению к учащимся и повышению мотивации к обучению [2].

В области применения методов искусственного интеллекта к оценке развернутых ответов на естественном языке можно условно обозначить два направления: автоматическая оценка коротких ответов (automatic short answer grading, ASAG) и автоматизированная оценка эссе (automated essay scoring, AES). Типичным критерием разделения является длина текста. В обоих случаях оценка сосредоточена на семантике, а не на синтаксисе ответа [3, 4]. Конкретные результаты работы систем автоматической оценки ответов могут сильно отличаться в зависимости от постановки задачи или цели системы. Для направления ASAG характерно прогнозирование бинарной оценки правильности ответа или оценки по заданной шкале, AES также может иметь целью получить определённые баллы, но часто предлагаются качественные критерии, вычисляемые на основе автоматической обработки текста [5]. Во многих случаях основную роль играет сравнение с эталонным ответом или ответами.

В исторической ретроспективе методы оценки развернутых ответов развиваются в соответствии с методами обработки естественного языка (natural language processing, NLP). Ранние системы автоматической оценки письма использовали простейшие характеристики текста, такие как частоты символов и слов, средняя длина предложения и т. п. С появлением более мощных компьютеров и методов NLP, анализирующих структуру слов и предложений, появились более сложные характеристики, отражающие контекст употребления слов и семантику текста [6]. Машинное обучение и большие языковые модели в последнее десятилетие позволили вычислять и отбирать релевантные признаки текста для прогнозирования оценки и показали высокое качество при обучении на больших корпусах данных, качественно размеченных экспертами [7].

Прорыв в развитии языковых моделей, позволяющих более эффективно находить смысловую составляющую текстов, вызвал существенное увеличение количества исследований по оценке открытых ответов и необходимость систематизации результатов. Поэтому авторы данной работы поставили целью анализ современных достижений в этой области. Основной задачей была выбрана автоматическая оценка ответов учащихся на основе близости к эталону. Формальное понятие близости или сходства определяется через расстояние между числовыми характеристическими векторами текстов или как результат классификации таких векторов в соответствии с выбранной шкалой оценок. Все исследования рассматривались с точки зрения используемых моделей.

Когда автоматическая оценка ответов учащихся ставится как задача классификации, это может быть либо бинарная классификация (правильный/неправильный ответ), либо мультиклассификация в двух вариантах: классификация по категории правильности (правильный/частично правильный/неправильный ответ/ответ не по теме и т. д.) и классификация по оценке, когда за ответ выставляется балл, например, по пятибалльной шкале. Последняя задача также может решаться как задача регрессии, но, так как баллы обычно ставятся либо целые, либо с долей 0.5, набор итоговых баллов строго ограничен, и задача сводится к классификационной.

В качестве исходных данных для автоматической оценки ответа учащегося выступают сам текстовый ответ, эталонный ответ и заданный вопрос. Ответ учащегося — это основной моделируемый текст. Эталонный ответ применяется в первую очередь для сравнения с ответом учащегося, но в некоторых случаях он может быть использован как источник для моделирования предметной области в целом, например, создания тезауруса. Заданный вопрос чаще всего не используется

совсем, но в некоторых исследованиях его текст применяется для определения темы текстов или моделирования предметной области.

Качество работы методов для автоматической оценки ответа учащегося определяется при помощи стандартных метрик для классификации и регрессии. К первым относятся доля правильных ответов (accuracy), точность, полнота и F-мера. Ко вторым относится среднеквадратичная ошибка (RMSE), средняя абсолютная ошибка (MAE) и квадратично-взвешенная каппа (QWK). В некоторых работах также встречаются каппа Коэна и коэффициент корреляции Пирсона. Часть исследователей ограничиваются экспертными оценками, когда не проводят масштабных экспериментов.

Обзор структурирован следующим образом. В разделе 1 рассматриваются модели текста на основе лингвистических характеристик, как стандартных стилометрических, так и сложных структурных и семантических. В разделе 2 обсуждаются нейросетевые модели текста. Раздел 3 посвящён моделям, комбинирующим нейросетевые эмбеддинги и лингвистические характеристики. В заключении делается вывод о состоянии предметной области.

1. Лингвистические модели

1.1. Модели уровня слов

Лингвистические характеристики входят в состав самых первых моделей текста и не теряют актуальности сейчас. Числовые характеристики уровня слов без учёта их порядка (модель «мешок слов»), хотя и почти не учитывают контекст и синтаксис естественного языка, тем не менее несут важную информацию о содержании текста, которую необходимо принимать во внимание при оценке ответов учащихся.

Hsu и др. [8] разработали автоматизированную систему оценки кратких ответов. Авторы работали со студентами своего университета, собрав собственный англоязычный корпус ответов, представляющих собой текстовые описания программ. Метод оценки использовал логистическую регрессию для биграмм и числовых характеристик на основе мешка слов, показав долю правильных ответов около 90 %. Исследователи отмечают, что в их системе возникло существенное количество ложноотрицательных срабатываний, что пришлось компенсировать механизмом ручных апелляций от студентов.

Suzen и др. [9] оценивали англоязычные короткие тексты студентов из университета Северного Техаса. Корпус состоял из 29 ответов, оценки за ответы учащихся выставлялись по шкале от 0 до 5. Распределение по классам (оценкам за ответ) оказалось несбалансированным: большинство ответов получили оценку 5, второй по популярности была оценка 2. Авторы сравнивали ответ студента с эталонным, базируясь на количестве общих слов. До эталонного ответа вычислялось расстояние Хэмминга, которое затем подставлялось в авторскую формулу расстояния. Полученные ответы сопоставлялись с оценками двух преподавателей, которые были получены вручную. Среднеквадратичное отклонение между результатом работы автоматического метода и средним значением двух преподавательских оценок составило 0.17, доля правильных ответов — 17 %. Авторы отмечают, что их модель следует усложнить, например, с применением синонимов.

Классический вариант моделирования текста в рамках подхода «мешок слов» применён в работе [10]. Векторизация осуществлялась на основе частот слов и биграмм и метрики TF-IDF. Для классификации по пяти категориям (полностью правильный или аналогичный эталонному ответ; частично правильный ответ; противоречащий эталонному ответ; неуместный ответ в рамках предметной области; ответ, не относящийся к предметной области и демонстрирующий отсутствие знаний) использовались методы опорных векторов, случайного леса, k-ближайших соседей и искусственной нейронной сети. Для экспериментов был взят англоязычный набор ответов на открытые вопросы SciEntsBank. В данном корпусе около 40 % правильных ответов, около 27 % частично правильных,

около 22 % неуместных, около 10 % противоречащих и 1 % ответов не из предметной области. Лучший результат по доле правильных ответов оказался 70 %.

Poguda и Tape [11] разработали систему автоматического оценивания англоязычных ответов учащихся по пятибалльной системе. Основой оценки явился расчет схожести между студенческим ответом и эталонным текстом с использованием алгоритма Джаро — Левенштейна, учитывающего как порядок символов, так и их совпадения. Дополнительно оценивались оригинальность, полнота и корректность текстов. Для окончательной оценки использовались алгоритмы машинного обучения. К сожалению, нет описания корпуса, на котором проводились эксперименты, и нет статистических оценок качества. Среди положительных качеств системы является возможность обнаружения плагиата в студенческих работах.

Pribadi и др. [12] оценивали короткие ответы из корпуса университета Северного Техаса. Авторы дополнили в нем набор эталонных ответов с помощью алгоритма суммаризации максимальной предельной релевантности на основе студенческих ответов и имеющихся эталонных ответов. Затем сгенерированные тексты были доработаны вручную. Такой метод сделал эталонные ответы более разнообразными. Измерение близости между студенческими и эталонными ответами производилось по методу GAN-LCS, предложенному авторами. Данная формула подсчета близости двух текстов опирается на самую длинную общую подпоследовательность символов между двумя текстами и на длины этих текстов. Для каждого студенческого ответа считалась метрика GAN-LCS до каждого эталонного ответа, затем бралось максимальное значение. Среднее значение среднеквадратичной ошибки составило 0.88, шкала оценок за ответы учащихся была от 0 до 5.

В работе [13] автоматически оценивались короткие описания программного кода, данные студентами. Авторы вычисляли простые стилометрические характеристики текстов: длину, долю самостоятельных частей речи, метрику читабельности. Также авторы применяли стандартные методы измерения близости текстов: ChrF и METEOR, основанные на n-граммах, и BERTScore, основанную на языковой модели BERT. Авторы экспериментировали с корпусом из 3019 пар ответ — эталон, подсчитывали метрики близости, но не вычисляли общие метрики качества. Эксперименты с анализом статистики у полученных характеристик и метрик близости показали, что поверхностные стилометрические характеристики текстов не позволяют отличить правильные ответы от неправильных. Только метрика METEOR позволила отделить правильные ответы от неправильных по близости к эталону.

Meccawy и др. [14] использовали для моделирования коротких ответов из корпуса AR-ASAG «мешок слов» и синонимы из арабского WordNet, а для вычисления близости текстов — косинусную метрику. Оценки учащимся выставлялись по шкале от 0 до 5. Значение среднеквадратичной ошибки, равное 1.12, достаточно хорошее, но использование нейросетевой языковой модели позволило авторам улучшить результат, как показано в следующем разделе.

В работе [15] для слов из коротких ответов искались синонимы в арабском WordNet. Слова из ответа и их синонимы сопоставлялись со словами из эталонного ответа. Для измерения близости ответа к эталону использовался алгоритм LCS, основанный на самой длинной общей подпоследовательности символов. Оценка за ответ выставлялась по шкале от 0 до 10. Эксперименты на собственном корпусе из 330 ответов студентов показали среднеквадратичную ошибку, равную 0.81. F-мера варьировалась от 60 % до 96 % для разных вопросов.

Buditjahjanto и др. [16] использовали «мешок слов», TF-IDF и косинусную метрику для определения близости эссе к эталонному. Далее для определения итоговой оценки применялась нейронная сеть с обратным распространением ошибки. Эксперименты с собственным корпусом из эссе на индонезийском языке показали MAE, равную 0.06, и F-меру, равную 88.57 %.

Непосредственное отношение к анализу ответов учащихся имеет задача определения сходства текстов. В этом случае под сходством текстов понимается смысловая (семантическая) близость

свободно сконструированного ответа учащегося с заведомо правильным, эталонным ответом преподавателя [17].

Авторы работы [18] проанализировали наборы лексических характеристик уровня символов, слов и структуры для определения близости текстов на английском языке. Текст на естественном языке преобразовывался в вектор чисел на основе стилометрической модели, для пар векторов текстов рассчитывались метрики близости пяти видов: косинусное сходство, коэффициент корреляции Пирсона, метрика Чебышева, евклидово расстояние, метрика Минковского. Если метрика близости оказывалась выше заданного порога, то тексты классифицировались как близкие, иначе нет. Ответы метода сравнивались с эталонными с помощью стандартных метрик качества. Для экспериментальной апробации векторных моделей были выбраны два корпуса англоязычных текстов: Text Similarity[1], содержащий 1613 пар близких и 529 пар неблизких текстов; собственный корпус из ответов студентов и эталонных, содержащий 618 пар близких и 1195 пар неблизких текстов. На корпусе Text Similarity лучшее качество показала модель на основе характеристик уровня слов: F-мера 86 %. Для собственного корпуса векторное представление уровня символов показало лучшее значение F-меры, равное 80 %.

Крюкова А. В. [19] решала проблему оценки семантической близости текстов на русском языке с помощью открытого программного инструмента DKPro Similarity [20]. В нём реализованы 15 различных строковых метрик близости. Эксперименты были проведены с текстами из четырёх наборов: аннотации научных статей, новости СПбГУ, переводы романа Набокова, заголовки новостных статей. Из каждой группы случайным образом было выбрано несколько текстов: пять аннотаций; по пять сообщений из двух частей новостного корпуса; пять соответствующих друг другу отрывков из трех переводов; пять пар заголовков. Эталонное сходство текстов определялось экспертами по шкале «0–1–2», где «0» — небольшая степень схожести, «1» — средняя степень, «2» — сильная степень схожести. Для задач классификации эксперименты проводились с несколькими моделями: логистическая регрессия, гребневый классификатор, классификатор SGD, пассивно-агрессивный классификатор, перцептрон. Объём обучающей выборки составлял примерно две трети (23 текста), а тестовой, соответственно, треть (12 объектов). Лучшую F-меру 63–100 % для разных групп текстов показали мера включения n-грамм, косинусная метрика и классификатор на основе логистической регрессии.

Все подходы в рассмотренных исследованиях объединяет простота реализации методов и возможность применения для корпусов текста ограниченного размера. Однако заметно, что оценка качества сильно колеблется даже в рамках одной научной работы. Анализ результатов показывает большое количество ложных результатов. Для преодоления недостатков часто используется работа экспертов, а также предлагается использование более сложных методов.

1.2. Модели на основе лингвистических правил и синтаксиса

Более сложные лингвистические характеристики связаны в первую очередь с синтаксисом фраз и предложений естественного языка. С одной стороны, современные инструменты NLP уже содержат алгоритмы определения лемм, морфологических характеристик, ключевых слов, параметров структуры предложений и дают возможность быстро реализовывать сложные правила проверки, с другой стороны, остаются трудоёмкими.

Nandini и Uma Maheswari [21] разработали систему, которая использует метод извлечения признаков на основе синтаксических отношений для автоматической оценки ответов описательного типа. Для моделирования ответов применялись тип ответа, определяемый из текста вопроса, n-граммы, ключевые слова, отфильтрованные по алгоритму разрешения лексической многозначности Lesk, а также общие слова. Для определения близости к эталону использовалась комплексная

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

метрика, учитывающая косинусную метрику, коэффициент Жаккара и количество общих слов. Корпус состоял из 50 вопросов, на которые дали ответы 10 студентов. Качество работы алгоритма достигло 95 % точности и 94 % полноты.

Авторы исследования [22] ставят задачу разработать систему автоматизированной проверки ответов учащихся на русском языке на открытые вопросы. Предложения в ответе студента могут быть любой сложности. Эталонный ответ представляет собой набор коротких односложных предложений, смысл которых должен присутствовать в ответе. Разработанная система анализирует ответы и на основе сравнения с эталонным и выставляет предварительную оценку. При сравнении учитываются полностью совпадающие слова, синонимы, антонимы, различные формы слов, написания дат и имен собственных, а также использование местоимений. Метод решения задачи комбинирует сравнение деревьев, построенных на основе графа синтаксического разбора, и извлечение фактов и формирует числовую метрику текста ответа. Система, руководствуясь заданными преподавателем порогами, относит ответ учащегося к одному из трех классов: правильный, неправильный или неопределённый, требующий дополнительной ручной проверки преподавателем. Эксперимент с 456 текстами показал F-меру 70 %. Интересно, что отсутствовали ошибки первого и второго рода, но 57 ответов, оценённых преподавателем как верные, и 77, оценённых как неверные, попали в класс неопределённых.

Зафиевский и др. [23] рассматривают модель, предназначенную для автоматической оценки делового письма на заданную тему на английском языке. Параметры оценки сформулированы и формализованы в виде 14 критериев при помощи экспертов в области обучения английскому языку. Критерии включают анализ лексики, особенности предметной области, тематики текста, стиля и формата письма, наличие средств логической связи предложений. Алгоритмы определения критериев основаны на анализе состава и структуры предложений, используют специализированные словари и регулярные выражения. Проведён эксперимент по анализу результатов работы этой системы на корпусе из 20 текстов. Автоматическая оценка и оценка экспертов сравнивались с помощью тепловых карт и технологии двумерного представления векторов UMAP, применённой к характеристическим векторам текстов. В большинстве случаев не было выявлено значимых различий между этими оценками.

Программный модуль для оценки открытых ответов на русском языке построен на основе системы формальных грамматик и правил и описан в работе [24]. Автор вводит индивидуальные концептуальные грамматики для синтаксического анализа текста в условиях, определённых контекстом вопроса. В рамках грамматик формулируются правила использования лексем, отношений между ними, полноты ответа. В тексте выделяются смысловые сегменты, которые анализируются на соответствие правилам, процесс продолжается рекурсивно до завершения текста ответа. В результате вычисляются семь числовых характеристик, объединённых в вектор ситуации. Эксперимент проведён для 20 ответов, в ходе которого вектора этих текстов сравнивались с векторами эталонных. В описании работы приведён анализ ошибочных ситуаций и сделан вывод о возможной коррекции правил.

Анализ синтаксических структур предложений лежит в основе определения сходства русскоязычных текстов в исследовании [25]. Авторы предлагают пять числовых критериев, агрегация которых показывает долю схожести текстов: доля покрытия предложения-эталона предложением сопоставляемого текста на основе метрики TF-IDF; оценка информационной значимости слов предложения-эталона в предложении сравниваемого текста; оценка сходства предложения-эталона и предложения сравниваемого текста на основе совпадения синтаксических структур; доля совпавших семантических значений у словоупотреблений в предложении эталона и в предложении сопоставляемого текста; оценка совпадающих семантических связей на основе роли слов в предложении. Однако в работе отсутствует описание экспериментов с корпусом текстов.

Относительно небольшое количество исследований, применяющих сложные лингвистические характеристики, можно объяснить несоответствием между приростом качества решения задачи и трудоёмкости методов. Однако, этот подход может оказаться эффективным для отражения специфики предметных областей, особенно в случае малого размера обучающего корпуса текстов. Кроме того, рассматриваемые характеристики текста могут быть частью комплексных методов определения сходства текстов и оценки ответов.

1.3. Модели на основе семантики

Самой сложной задачей любой области NLP является выявление смысловой составляющей текста. Семантическую близость понятий часто определяет отношение синонимии. Некоторые подходы предоставляют методы извлечения информации, тезаурусы и онтологии. Высококачественное определение семантики текста еще впереди, однако и существующие подходы успешно используются для оценки ответов.

Ouahrani и Bennouar [26] собрали корпус арабских текстов AR-ASAG для измерения близости короткого ответа к эталону. Для демонстрации работы с корпусом был поставлен эксперимент по определению близости текстов косинусной метрикой. Ответы моделировались параметрами «мешка слов», причём слова выбирались из семантического пространства, предварительно составленного для корпуса методом COALS. Характеристикой слова для вектора текста выступала NTFlog. Среднеквадратичная ошибка данного метода составила в среднем 0.80.

Хорошие результаты показала система автоматического оценивания ответов на открытые вопросы на русском языке на правильные и неправильные, описанная в статье [27]. Из ответа учащегося с помощью Томита-парсера извлекается набор фактов, которые сравниваются с эталонным ответом. Грамматические правила, позволяющие извлекать факты, строятся отдельно для каждого конкретного вопроса, кроме того, вводятся весовые коэффициенты значимости. Такой подход приводит к высокой F-мере 90 %, но не позволяет легко применять метод для аналогичных задач из других предметных областей.

В работе [28] предлагается решение задачи определения семантической близости команды, выданной на естественном языке, с эталонным текстом. Сравниваемые тексты относятся к одной предметной области и содержат термины из определённого набора. Моделирование текста происходит на основе тезаурусного графа основных и второстепенных терминов предметной области со связями трёх видов: «включает в себя», «обеспечивает результат» и «соответствует». Степень семантической близости определяется по количеству совпавших понятий, а также расстоянию между соответствующими вершинами в графе. К сожалению, описания корпуса текстов и экспериментов нет.

Maharjan и Rus [29] предложили оценивать ответы студентов с помощью концептуальной карты для предметной области. Концептуальная карта представляет собой базу знаний, сформированную экспертами. В процессе оценивания ответов авторы сопоставляли слова и другие фрагменты ответов с данной картой, измеряя близость фрагментов к концептам из карты. Полученные вектора характеристик для студенческих ответов и для эталонных ответов конкатенировались и использовались для классификации студенческого ответа как корректного/некорректного/неполного/спорного. Бинарная классификация на корректный/некорректный осуществлялась с помощью косинусной метрики с эмпирически установленным порогом, F-мера на собственном корпусе достигла 80.2 %. Для мультиклассификации использовался классификатор LSTM, F-мера на корпусе DT-Grade достигла 62.0 %.

Автор работы [30] обращает внимание на то, что проблема получения новых эффективных алгоритмов для вычисления семантической близости предложений остается актуальной, поскольку современные методы часто не дают достоверных результатов. Он предлагает алгоритм определения схожести предложений на основе вычисления семантической близости биграмм и триграмм. На основе корпуса текстов предметной области строятся бинарные деревья понятий с иерархиче-

скими связями. Семантическая близость двух понятий определяется длиной пути наследования, сложность которого вычисляется разностью бинарных иерархических чисел [31]. Эксперимент был проведён с семью новостными предложениями на политическую тематику на русском и английском языках. Программная реализация алгоритма позволила построить матрицу попарного сходства, которая была проанализирована экспертом, сделавшим качественный вывод о пригодности предложенного подхода к решению поставленной задачи.

1.4. Сравнение лингвистических моделей

Исследования, опирающиеся на лингвистические модели текстов, приведены в таблице 1. Для исследований указаны тип задачи, язык текстов, модель текста, метод сравнения или классификации, лучшее качество результата или диапазон лучших значений, если корпусов было несколько.

Типы задач по анализу студенческих текстов указаны англоязычными аббревиатурами (ASAG — automatic short answer grading — автоматическая оценка кратких ответов, AES — automated essay scoring — автоматизированная оценка эссе), отдельно указаны исследования по определению близости других текстов. Наиболее популярной задачей является автоматическая оценка кратких ответов, она встречается чаще всего. Автоматизированная оценка эссе также распространена, но когда исследователям нужно сравнить эссе с эталонным, они прибегают не к лингвистическим моделям, что будет видно в следующих разделах.

Наиболее хорошо и часто изучаются тексты на английском языке. Однако следует отметить, что для русского и арабского языков область активно развивается. Для арабского языка это отчасти связано с появлением арабской версии WordNet и открытого корпуса AR-ASAG. Других открытых корпусов для разных языков существует немного, но это не является существенным препятствием: очень многие авторы собирают собственные корпуса текстов.

Среди лингвистических моделей наиболее часто используются стандартные стилометрические характеристики: «мешок слов», n-граммы, синонимы из WordNet. Однако сами по себе они редко дают высокое качество результата, поэтому авторы их применяют в комплексе с синтаксическими и семантическими характеристиками текстов, что повышает статистические метрики оценки решения.

Метод для сравнения и бинарной классификации векторных моделей — это в большинстве случаев косинусная метрика с пороговым значением, подобранным вручную. Для мультиклассовой классификации, то есть выставления оценки студенческому ответу, обычно используются или регрессионные алгоритмы, или нейронные сети.

Общая оценка качества анализа студенческих текстов часто дается с помощью стандартных для классификации метрик: доли правильных ответов (accuracy в таблице), точности, полноты и F-меры. В таблице приводится в основном F-мера, как наиболее сбалансированная и показательная метрика или доля правильных ответов, если F-меру исследователи не вычисляли. Помимо данных метрик качества, в задачах сравнения близости текстов часто измеряется среднеквадратичная ошибка, которая должна быть как можно меньше. Вместе со значением среднеквадратичной ошибки приводится шкала, по которой выставлялась оценка учащимся, чтобы сделать поправку на диапазон значений: чем больше шкала оценок, тем большее значение среднеквадратичной ошибки можно считать хорошим. Значения метрик качества сильно варьируются: от 62 до 100 % F-меры, от 17 до 90 % доли правильных ответов. Только у среднеквадратичной ошибки можно указать достаточно часто встречающийся уровень около 0.8. Некоторые авторы вообще не приводят данных об общих метриках качества, что означает, что модель и метод решения подбираются под конкретные условия и оцениваются вручную. Таким образом, среди лингвистических моделей текста нет универсального подхода, дающего стабильно высокие результаты.

Table 1. Research where the tasks of analASAG: Automatic Short Answer Gradingyzing a text answer are solved using linguistic models

Таблица 1. Исследования, где задачи анализа текстового ответа решаются с помощью лингвистических моделей

Авторы	Задача	Язык	Модель	Метод	Качество
Hsu и др. [8]	ASAG	Английский	Мешок слов	Логистическая регрессия	Accuracy 90 %
Suzen и др. [9]	ASAG	Английский	Мешок слов	Расстояние Хэмминга	Accuracy 17 %
Кербенева [10]	ASAG	Английский	Мешок слов	Нейронная сеть	Accuracy 70 %
Poguda и Tape [11]	ASAG	Английский	Мешок слов	Алгоритм Джаро-Левенштейна	-
Pribadi и др. [12]	ASAG	Английский	GAN-LCS	GAN-LCS	RMSE 0.88 шкала 0–5
Lekshmi-Narayanan и др. [13]	ASAG	Английский	n-граммы	METEOR Экспертная оценка	-
Meccawy и др. [14]	ASAG	Арабский	WordNet	Косинусная метрика	RMSE 1.12 шкала 0–5
Abdeljaber и др. [15]	ASAG	Арабский	WordNet	LCS	F-мера 60–96 % RMSE 0.81 шкала 0–10
Buditjahjanto и др. [16]	AES	Индонезийский	Мешок слов	Косинусная метрика BPNN	MAE 0.06 F-мера 89 %
Лагутина и др. [18]	Близость текстов	Английский	Мешок слов	Корелляция Пирсона	F-мера 80–86 %
Крюкова А. В. [19]	Близость текстов	Русский	Косинусная метрика, мера включения n-грамм	Логистическая регрессия	F-мера 63–100 %
Nandini и др. [21]	ASAG	Английский	Тип ответа n-граммы Ключевые слова Общие слова	Косинусная метрика Коэффициент Жаккара	Точность 95 % Полнота 94 %
Леонов и др. [22]	ASAG	Русский	Граф синтаксических связей	Косинусная метрика	F-мера 70 %
Зафиевский и др. [23]	AES	Английский	Правила	Косинусная метрика	-
Прокопьев [24]	Близость текстов	Русский	Грамматики и правила	Экспертная оценка	-
Хорошилов и др. [25]	Близость текстов	Русский	Синтаксические характеристики	Экспертная оценка	-
Ouahrani и Bennouar [26]	ASAG	Арабский	Мешок слов	Косинусная метрика	RMSE 0.80 шкала 0–5
Кожевников и др. [27]	ASAG	Русский	Грамматические правила	Доля правильности	F-мера 90 %
Гадасин и др. [28]	ASAG	Русский	Тезаурус	Экспертная оценка	-
Maharjan и Rus [29]	ASAG	Английский	Концептуальная карта	Косинусная метрика LSTM	F-мера 80 % F-мера 62 %
Каширин [30]	Близость текстов	Русский Английский	Бинарные иерархические числа	Экспертная оценка	-

ASAG — automatic short answer grading, автоматическая оценка кратких ответов.
AES — automated essay scoring, автоматизированная оценка эссе.
RMSE — root mean square error, среднеквадратичная ошибка.

2. Нейросетевые модели

2.1. Большие языковые модели

Огромное количество методов, так или иначе связанных с нейронными сетями, отражает современное состояние NLP. С точки зрения моделирования текста, следует выделить уже ставшие

классическими эмбеддинги и их комбинации с другими характеристиками. При моделировании текстов только с помощью эмбеддингов значительная часть исследователей напрямую использует числовые параметры больших языковых моделей для каждого текста. Отдельно хотелось бы выделить подход, агрегирующий эмбеддинги ответов и эталона, он будет рассмотрен в следующем разделе. Этот раздел посвящен большим языковым моделям, построенными на основе нейронных сетей, которые учитывают многие особенности устройства естественного языка. Эмбеддинги становятся характеристическими векторами текстов, а сравнение с эталоном трактуется как определение близости соответствующих векторов.

В работе [14], описанной ранее, авторы использовали для моделирования коротких ответов из корпуса AR-ASAG не только синонимы из WordNet, но и эмбеддинги word2vec и AraBERT. Вектора эталонных и проверяемых ответов сравнивались по косинусной метрике. Языковая модель AraBERT позволила достичь лучших результатов, чем модель на основе лингвистических характеристик: среднеквадратичная ошибка 1.00 против 1.12.

В работе [32] рассматривалось применение векторных моделей для анализа ответов студентов, сформулированных в свободной форме на русском языке. В качестве моделей представления слов и документов были выбраны word2vec, doc2vec, BERT. Сравнение ответа, данного обучающимся, и эталонного с помощью косинусной меры показало, что более качественно модели выявляют неверные ответы. Такой же вывод можно сделать из экспериментов бинарной классификации на правильные и неправильные ответы, где использовались эмбеддинги word2vec и классификатор Случайный лес. F-мера оказалась 74 % для определения класса неверных ответов и 55 % для верных. Для повышения качества авторы предлагают проводить дополнительные этапы проверки, такие как анализ ключевых слов и экспертная проверка.

В продолжении исследований [33] предложен метод оценивания ответов через нахождение косинусного сходства полученных векторов и уточнение оценок проверкой ключевых слов. Векторное представление текста строилось на основе модели BERT. Отдельно оценивались ответы на каждый из нескольких вопросов по теме «Компьютерная графика». Максимальная доля автоматического определения правильности и неправильности в соответствии с экспертным мнением составила 90 %, средняя: 77 %.

Ndukwe и др. [34] применили Sentence-BERT, чтобы поставить 228 ответам студентов оценку от 0 до 5. Эталонные ответы преподавателей получили оценку 5 и использовались для обучения нейронной сети. Квадратично-взвешенная каппа достигла 0.70. Авторы очень коротко описали постановку и результаты экспериментов и значения других метрик в работе не привели.

Fateen и Mine [35] оценивали короткие ответы на арабском языке с помощью нескольких подходов. В первом подходе моделировались вопрос, эталонный ответ и ответ студента с помощью языковой модели AraBERT, далее нейронная сеть выступала классификатором этих троек, определяя оценку ответа студента. Во втором подходе моделировались эталонный ответ и ответ студента также с AraBERT, а вместо классификатора использовалась сиамская нейронная сеть, которая оценивала близость ответа к эталону с помощью косинусной метрики. QWK для разных вопросов варьировалась очень сильно: от 0.02 до 0.80, так что оба подхода нельзя назвать стабильными.

Gaddipati и др. [36] применили предобученные языковые модели ELMo, BERT, GPT и GPT-2 для автоматической оценки кратких ответов. Ответы разделялись на слова, слова токенизировались, сумма эмбеддингов слов выступала эмбеддингом ответа. Для пар эмбеддингов эталонного и студенческого ответа вычислялась косинусная метрика близости, порог решения о том, близки тексты или нет, определялся на тренировочной выборке с помощью изотонической, линейной и гребневой регрессии. Эксперименты проводились на корпусе Mohler, состоящем из 2273 ответов, оцениваемых по шкале от 0 до 5. Среднеквадратичная ошибка составила 1.00 для ELMo и греб-

невой регрессии. Авторы провели сравнение с предыдущими работами, где экспериментировали с корпусом Mohler, и показали, что их результаты выше на 0.11.

Schneider и др. [37] оценивают корректность коротких ответов по близости к эталону: корректный ответ/некорректный ответ/не определено. Корпус вопросов-ответов состоит из 10 миллионов ответов на 990000 вопросов на 14 языках, собранных с сайта Classtime. Языки включают в себя в основном языки европейских стран, в том числе русский, а также турецкий и тайский. В корпусе больше правильных ответов, чем неправильных: 58 %, но для экспериментов авторы искусственно уравняли размеры классов, добавляя ответы на другие вопросы в качестве неправильных. Тексты на всех языках моделировались одинаково с помощью мультиязыковых моделей BERT и LaBSE, причём эмбеддинги ответа строились на основе текстов и ответа, и соответствующего вопроса. Близость между эмбеддингами ответов оценивалась с помощью формулы, основанной на косинусной метрике. Доля правильных ответов метода достигла 87 % для модели LaBSE, для BERT качество оказалось немногим ниже. Авторы отмечают, что для полноценной замены работы преподавателя автоматическим оценщиком данного уровня качества недостаточно, так как алгоритм совершает слишком много ошибок.

Hendre и др. [38] оценивают качество эссе с помощью эмбеддингов GSE, ELMo и GloVe. Близость ответа студента к эталонному считается с помощью косинусной метрики. В корпусе ASAP, использовавшемся для экспериментов, нет эталонных ответов, поэтому авторы выбрали в качестве эталонных эссе с наибольшей оценкой для каждого вопроса. Если эссе с высшей оценкой для вопроса было несколько, то выбиралось эссе с большей длиной. В качестве метрики для сравнения результатов метода с оценками двух преподавателей считалась корреляция Пирсона. Она достигла 0.52–0.69 для различных вопросов, лучшие результаты были достигнуты моделью GSE.

Doewes и др. [39] сравнивали эссе с их перефразированными версиями и измеряли близость между ними. Моделирование производилось несколькими способами: с помощью «мешка слов», TF-IDF, эмбеддингов Sentence-BERT и LASER. Метрикой близости выступала косинусная метрика, итоговая оценка за эссе выставлялась при помощи алгоритмов регрессии: гребневой, случайного леса или повышения градиента. На экспериментах с собственным корпусом QWK для эмбеддингов превысила QWK для стилометрических моделей и составило 0.72–0.77.

В работе [40] сравнивались 50 эссе студентов, изучающих английский, с образцовыми эссе от экспертов. В качестве моделей использовались эмбеддинги из открытых программных библиотек: InferSent, spaCy, ADW, DKPro, SEMILAR, LSA, скомбинированные с косинусной метрикой для измерения близости. Корреляция Пирсона с оценками экспертов составила 0.82 для эмбеддингов InferSent.

Dhini и др. [41] оценивали 1028 эссе на индонезийском языке, собранные из LMS университета Terbuka. Эссе моделировались как эмбеддинги предложений с помощью Sentence-BERT и как эмбеддинги ключевых слов с помощью KeyBERT. Близость между образцовыми текстами и эссе студентов считалась по косинусной метрике между данными эмбеддингами. Значение MAE достигло 0.71.

Работа [42] посвящена применению генеративных моделей, взаимодействующих с пользователем посредством чат-ботов ChatGPT и PerplexityAI на базе модели OpenAI GPT 3.5, для оценки студенческих эссе, написанных в формате стандартизированного экзамена по английскому языку. Эссе были проверены преподавателем по 9-балльной шкале IELTS, баллы были целочисленными или кратными 0.5. Оценивались тексты в целом и аспекты: решение коммуникативной задачи, связанность, лексический ресурс, грамматические навыки. Для экспериментов было использовано 19 эссе. Выставленные баллы были сопоставлены с оценкой преподавателя и двух чат-ботов путем вычисления коэффициентов согласованности (альфа Кронбаха) и межэкспертного согласия (каппы Коэна и Флейса). Доля оценок, отличавшихся не более чем на балл, была высока и варьировалась от 79 до 100 %, доля полностью идентичных оценок была мала, наибольшее согласие прослеживалось

у двух чат-ботов, что вполне объясняется общей базовой моделью языка. В результате качественного анализа выявлены особенности обратной связи от чат-ботов, такие как периодическое игнорирование инструкций в запросе, тенденция к нахождению несуществующих ошибок, выставление разных баллов одной и той же работе при последовательных запросах.

Преподаватель английского языка для русскоязычных студентов описывает опыт применения чат-бота Mistral AI для оценки эссе на заданную тему длиной не более 250 слов [43]. В статье приведён результат анализа одного эссе, который содержит баллы по отдельным критериям. Автор обращает внимание, что качество результата зависит от формулировки запроса к чат-боту.

В работе [44] рассматриваются результаты применения трех языковых моделей на основе BERT и GPT для задачи определения семантического сходства текста ответа учащегося и эталонного ответа учителя. Эксперимент проводился на собственном корпусе текстов, состоящем из 1812 пар схожих фраз и словосочетаний. Тексты классифицировались на похожие и не похожие на эталонный ответ на основе шести метрик сходства. Наиболее высокие результаты оказались для косинусной метрики и коэффициента корреляции Пирсона. Лучшая F-мера 55 % оказалась у модели sembeddings/model_gpt_trained. Лучшую точность 47 % показала модель bert-base-cased. Модель bert-large-cased обнаружила лучшую полноту 85 %.

В уже упоминавшейся работе [18] по определению сходных текстов сравнивались различные варианты больших языковых моделей BERT, GPT и Mamba. На корпусе Text Similarity все модели добились примерно одинакового качества результата: 85–87 % F-меры. Среди BERT-моделей лучшими оказались bert-base-cased и bert-base-uncased. Эмбеддинги GPT-моделей и Mamba сработали аналогично. Интересно, что стилометрические модели показали то же качество. На собственном корпусе эмбеддинги языковых моделей показали низкое качество: 50–56 % F-меры, среди них можно отметить только sembeddings/gptops_finetuned_mpnet_ gpu_v1. Их существенно опередили стилометрические характеристики, позволившие выбрать близкие тексты с качеством около 80 %. Вероятно, такой результат связан с меньшим размером собственного корпуса.

Таким образом, на текущий момент большие языковые модели становятся очень популярными для автоматизации оценок текстов как на английском, так и на других языках. Однако, требуется большое количество экспериментов для объективной оценки качества их работы с разных сторон: аспектного анализа языковых компетенций, особенностей предметных областей, анализа ошибок.

Можно заметить, что прямое использование эмбеддингов больших языковых моделей лучше большинства лингвистических характеристик, но недостаточно для построения по-настоящему эффективных систем оценок. Повысить качество удаётся дообучением или дополнением моделей.

2.2. Модели, использующие эталонный ответ для последующей классификации

В этом разделе рассматривается вариант решения задачи, когда характеристический вектор ответа формируется как модификация эмбеддингов исходного текста и эталона. После построения модели прогнозирование оценки ответа решается как мультиклассификационная задача.

Gomaa и Fahmy [45] доработали стандартную модель word2vec и представили Ans2vec, адаптированную для проверки коротких ответов. Разность векторов эталонного и студенческого ответа представляла собой итоговый вектор ответа. Оценка ответа вычислялась при помощи логистической регрессии. Лучшие результаты на корпусе университета Северного Техаса достигли среднеквадратичной ошибки 0.91, на корпусе Cairo — 0.92, на корпусе SCIENTSBANK — 0.46. Данные результаты близки к результатам других исследователей на этих же корпусах, но не превышают их, как показывают авторы при анализе результатов.

Weegar и Idestam-Almquis [46] оценивали 230 коротких ответов студентов, написанных на шведском и английском языках. Корпус с помощью алгоритмов кластеризации k-средних и GMM делился на тренировочную и тестовую выборки. Тексты моделировались как эмбеддинги BERT, также в вектор ответа добавлялось значение косинусной метрики близости между ответом студента и эталоном.

Классификаторами служили метод случайного леса, SVM и наивный байесовский классификатор, первый позволил достичь лучших результатов. F-мера оказалась 57 %, доля правильных ответов — 84 %.

Рассматриваемый подход к классификации ответов является стандартным с точки зрения методов машинного обучения. Невысокая F-мера, скорее всего, обусловлена слабым учётом особенностей предметной области.

2.3. Модели на основе ансамблей и собственных архитектур

Одним из способов учёта особенностей предметной области и повышения качества решения задачи является построение ансамблевых моделей и собственных оригинальных решений.

Li и др. [47] предложили семантическую сеть SFRN, основанную на связях из базы знаний о вопросах, справочных ответов или рубрик и маркированных ответов студентов. SFRN применялась для автоматической оценки кратких ответов. С её помощью для каждого ответа строились три вектора: для текста самого ответа, эталонного ответа и вопроса. Векторизация текстов производилась с помощью моделей LSTM и BERT. Три вектора объединялись в один посредством семантической нейронной сети. Корпусом для экспериментов выступал SemEval-2013, автоматически увеличенный с помощью переводов ответов студентов с английского на французский или китайский и обратно. Задача определения близости ставилась как классификационная задача: ответ правильный, неправильный, правильный частично, противоречивый, нерелевантный, не по теме. Классификатором выступил многослойный персептрон. Качество работы алгоритма оценивалось с помощью QWK и составило 0.79.

Tan и др. [48] оценивали короткие ответы студентов с помощью графовой свёрточной сети (GCN). Граф содержал два типа вершин. Первый тип — вершины на уровне предложений, соответствующие эталонным ответам, ответам, оценённым человеком, и ответам студентов, которые должны быть оценены. Второй тип — вершины на уровне слов/биграмм, соответствующие словам и биграммам из эталонных и студенческих. Вершины были соединены рёбрами в соответствии с их отношениями включения или совместной встречаемости. Вес ребра соответствовал TF-IDF или PMI (точечной взаимной информации). Подобный граф моделирует весь корпус, что обуславливает его способность решать классификационную задачу по оценке ответа студента. Эксперименты на корпусе SemEval-2013 позволили достичь F-меры 73 %. Эксперименты с китайским языком и двумя собственными тематическими корпусами текстов показали F-меру, равную 77 %, для корпуса по математической теме (17248 ответов) и 65 % для литературной темы (10104 ответов). Можно отметить, что алгоритм сработал хуже для корпуса меньшего размера.

Xie и др. [49] применяют для оценки эссе собственную модель NPCR. Она основывается на принципе того, что для двух заданных эссе расстояние между похожими эссе из одной и той же категории должно быть небольшим, в то время как расстояние между разнородными эссе должно быть большим, и семантическая связь может быть отражена путем измерения расстояния. В качестве эмбеддингов для входных данных NPCR брался BERT, который обеспечил QWK 0.82 на корпусе ASAP.

Garg и др. [50] составили ансамбль из двух методов оценивания короткого ответа. Первая часть ансамбля считала косинусную метрику близости между эмбеддингами ответа и эталона. Для подсчета эмбеддингов бралась версия BERT, предназначенная для вопросно-ответных систем. Вторая часть ансамбля объединяла ответ и эталон в общий текст и строила для них один BERT-эмбеддинг. Далее с помощью BERT-регрессора вычислялась оценка за ответ. Оценки из обеих частей ансамбля отправлялись в финальный нейросетевой классификатор, который предсказывал итоговый ответ. Качество работы на корпусе Mohler достигло среднеквадратичной ошибки, равной 0.73.

Комбинация эмбеддингов USE и BERT для инструмента визуализации поиска похожих текстов используется в работе [51]. Эмбеддинги вычисляются параллельно и конкатенируются, близость

Table 2. Research where text answers analysis tasks are solved using neural network models

Таблица 2. Исследования, где задачи анализа текстового ответа решаются с помощью нейросетевых моделей

Авторы	Задача	Язык	Модель	Метод	Качество
Meccawy и др. [14]	ASAG	Арабский	word2vec AraBERT	Косинусная метрика	RMSE 1.00 шкала 0–5
Миннегалиева и др. [32]	ASAG	Русский	word2vec	Косинусная метрика	F-мера 55–74 %
Миннегалиева и др. [33]	ASAG	Русский	BERT Косинусная метрика	Случайный лес	Accuracy 77 %
Ndukwe и др. [34]	ASAG	Английский	SBERT	Нейронная сеть	QWK 0.70
Fateen и Mine [35]	ASAG	Арабский	AraBERT	Siamese NN Косинусная метрика Нейронная сеть	QWK 0.02–0.80
Gaddipati и др. [36]	ASAG	Английский	ELMo, BERT, GPT, GPT-2	Косинусная метрика Гребневая регрессия	RMSE 1.00 шкала 0–5
Schneider и др. [37]	ASAG	14 языков	BERT, LaBSE	Косинусная метрика	Accuracy 87 %
Hendre и др. [38]	AES	Английский	GSE, ELMo GloVe	Косинусная метрика	Пирсон 0.52–0.69
Doewes и др. [39]	AES	Английский	SBERT, LASER	Косинусная метрика Повышение градиента	QWK 0.72–0.77
Wang [40]	AES	Английский	InferSent, spaCy, ADW, DKPro, SEMILAR, LSA	Косинусная метрика	Пирсон 0.82
Dhini и др. [41]	AES	Индонезий-ский	SBERT KeyBERT	Косинусная метрика	MAE 0.71
Боголерова и др. [42]	AES	Английский	ChatGPT PerplexityAI	Экспертная оценка	Accuracy 79–100 %
Евстигнеев [43]	AES	Английский	Mistral AI	Экспертная оценка	-
Копнин и др. [44]	Близость текстов	Английский	BERT, GPT	Косинусная метрика	F-мера 55 %
Лагутина и др. [18]	Близость текстов	Английский	Mamba BERT, GPT	Косинусная метрика	F-мера 85–87 %
Gomaa и Fahmy [45]	ASAG	Английский	Ans2vec	Линейный классификатор	RMSE 0.46–0.91 шкала 0–5
Weegar и др. [46]	ASAG	Шведский Английский	BERT Косинусная метрика	Случайный лес	F-мера 57 %
Li и др. [47]	ASAG	Английский	SFRN	Нейронная сеть	QWK 0.79
Tan и др. [48]	ASAG	Английский Китайский	GCN	GCN	F-мера 72.5 % F-мера 65–77 %
Xie и др. [49]	AES	Английский	NPCR, BERT	Нейронная сеть	QWK 0.82
Garg и др. [50]	ASAG	Английский	BERT	BERT-регрессор Косинусная метрика Нейронная сеть	RMSE 0.73 шкала 0–5
Witschard и др. [51]	Близость текстов	Английский	USE, BERT	Косинусная метрика	F-мера 55 %

ASAG — automatic short answer grading, автоматическая оценка кратких ответов.
AES — automated essay scoring, автоматизированная оценка эссе.
RMSE — root mean square error, среднеквадратичная ошибка.
QWK — quadratic weighted kappa, квадратично-взвешенная каппа.
MAE — mean absolute error, средняя абсолютная ошибка.

текстов измеряется по косинусной метрике. F-мера на корпусе IEEE VIS достигает не очень высокого значения 55 %.

2.4. Сравнение нейросетевых моделей

Исследования, опирающиеся на нейросетевые модели текстов, приведены в таблице 2. Таблица по структуре аналогична таблице из раздела 1.4.

Наиболее популярной задачей снова является автоматическая оценка кратких ответов, однако автоматизированная оценка эссе тоже часто решается при помощи нейросетей.

Мультиязычные языковые модели и языковые модели для конкретных языков позволили исследователям достигать хороших результатов не только для английского, но и многих других национальных языков, а также оценивать мультиязычные корпуса текстов. Тем не менее наиболее часто изучаются англоязычные тексты, особенно из открытого корпуса Mohler. Однако авторы собирают собственные корпуса и нередко проводят большое количество экспериментов и на собственном корпусе, и на Mohler в рамках одного исследования.

Тексты моделируются в основном с помощью языковых моделей, популярных и в других задачах компьютерной лингвистики, — BERT и GPT, а также их вариаций. Именно они обеспечивают или лучшие результаты, или хотя бы близкие к лучшим. Прочие эмбеддинги таких стабильных результатов не показывают.

Метод для сравнения и классификации векторных моделей — это или косинусная метрика с пороговым значением, подобранным вручную, или простая нейронная сеть. В нескольких случаях используются сиамские нейронные сети, которые предназначены для сравнения пар текстов.

Общая оценка качества анализа студенческих текстов дается с помощью достаточно разнообразных метрик. Помимо доли правильных ответов (accuracy в таблице), F-меры и среднеквадратичной ошибки, исследователи вычисляют QWK — квадратично-взвешенную каппу и коэффициент корреляции Пирсона. Значения метрик качества нельзя назвать высокими и стабильными: в каждом исследовании оказывается свой уровень или диапазон значений. Тем не менее большинство авторов успешно решают или свою основную задачу, или одну из подзадач. Они достигают высоких значений F-меры, доли правильных ответов и QWK и низких значений метрик, описывающих ошибки, либо на корпусе текстов в целом, либо на его части. Таким образом, для каждого предложенного метода находится своя область, где он оказывается полезен.

3. Комбинированные модели

Прямое применение больших языковых моделей обычно не даёт желаемого качества решения задачи, так как мало учитывает предметную область, а для дообучения модели недостаточно материалов, поскольку экспертная разметка текстов требует значительных временных и человеческих ресурсов. Один из путей решения этих проблем находится в комбинировании эмбеддингов и лингвистических и других характеристик, отражающих особенности проверки правильности ответов.

Tulu и др. [52] используют в методе оценки коротких ответов синонимы из WordNet, переведённые в эмбеддинги методом SemSpace, и модель MaLSTM, основанную на Манхэттенском расстоянии и позволяющую работать с парами эмбеддингов эталонных ответов и студенческих ответов. Эксперименты с корпусом Mohler из 87 ответов показали среднеквадратичную ошибку 0.04 и среднюю абсолютную ошибку (MAE) 0.23. Эксперименты с собственным корпусом CU-NLP из 171 ответа показали MAE, равную 0.02. Следует отметить, что хорошие результаты метод показал на авторском корпусе.

Zhang и др. [53] скомбинировали для оценки коротких ответов модели правильных ответов, ответа студента, вопроса и уровня знаний студента. Для моделирования пар «правильный ответ» — «студенческий ответ» использовался мешок слов, IDF, LSA, косинусная метрика, разница в количестве слов, метрика близости с учетом предметной области, общая метрика близости. Для модели-

рования вопроса определялась его тема и уровень сложности. Для моделирования уровня знаний студента применялась скрытая марковская модель с двумя состояниями, базирующаяся на предварительном тестировании студента по темам, которым посвящены вопросы. В качестве классификатора правильный/неправильный ответ использовалась нейронная сеть DBN. F-мера достигла 83 % на корпусе данных из системы Cordillera, состоящем из ответов 158 студентов на 482 вопроса.

Lubis и др. [54] скомбинировали эмбеддинги word2vec, обученные на индонезийской Википедии, и лингвистические характеристики, основанные на частях речи и синтаксической структуре предложения. На основе синтаксических связей между словами в коротком ответе эмбеддинги слов word2vec объединялись в эмбеддинг текста. Близость ответа к эталону измерялась с помощью косинусной метрики. Эксперименты на собственном корпусе индонезийских коротких ответов показали MAE, равную 0.70 .

Shashavali и др. [55] разбивали предложения на n-граммы и вычисляли для них эмбеддинги FastText. Эмбеддинги ответов сравнивались с эталонными по косинусной метрике. Эксперименты на собственном корпусе предложений различной длины показали F-меру, равную 93 % .

Prabhudesai и Duong [56] оценивают короткие ответы студентов при помощи комбинации эмбеддингов GloVe и стилометрических характеристик: количества слов, длины ответа, количества уникальных слов. Вектора характеристик строятся для студенческого и эталонного ответов и подаются на вход сиамской нейронной сети на основе LSTM. Эксперименты на корпусе Mohler позволили достичь MAE, равной 0.62, и среднеквадратичной ошибки, равной 0.89.

Lakshmi и др. [57] моделировали каждый короткий ответ как набор метрик близости между ним и эталоном. Метрики близости включали в себя статистическую близость по длинам текстов, количествам всех слов и уникальных слов, близость по общим словам, близость по ключевым, уникальным и лемматизированным словам, близость по мешку слов и TF-IDF, близость LSA, семантическую близость по эмбеддингам Infersent, близость по сгенерированным аннотациям ответов. Вектора из метрик близости классифицировались по оценкам с помощью стандартных методов машинного обучения: KNN, байесовский классификатор, SVM, дерево решений, случайный лес, XGBoost. Эксперименты проводились на корпусе текстов, составленном из известных открытых корпусов для оценки коротких текстов. Качество классификации достигло F-меры в среднем 75 % для метода случайного леса.

Сравнение текстов через ансамбль четырёх мер сходства осуществляют Hassan и др. [58]. Ансамбль включает в себя две характеристики близости на основе мешка слов и синонимов из BabelNet, эмбеддинги DSM, учитывающие контекст слов и характеристику близости по расстоянию Левенштейна. Эксперименты на собственном корпусе и на корпусах SemEval показывают коэффициент корреляции Пирсона 0.70–0.85.

Del Gobbo и др. [59] моделировали короткие ответы с помощью TF-IDF и эмбеддингов BERT. В процессе работы метод использует для измерения близости эталонных ответов и ответов студентов регрессор на основе метода опорных векторов (SVR). Для экспериментов были объединены открытые корпуса ASAP, SciEntBank и Cu-NLP, для корпуса ASAP ответы с лучшими оценками были размечены как эталонные, ответы оценивались по шкале о 0 до 5. Качество работы метода достигло среднеквадратичной ошибки в среднем 0.73 среди 83 вопросов.

Zhao и др. [60] построили модель анализа семантической целостности студенческих эссе на основе BERT. Авторы скомбинировали эту модель с характеристиками на основе ключевых моментов эссе: особенности стиля, содержание абзацев и темы абзацев. В качестве классификатора применялся ансамбль из сиамской нейронной сети и ESIM, который оценивал пары студенческое эссе — эталонное эссе. Эксперименты на собственном корпусе показали сильный разброс коэффициента каппы Коэна от 0.50 до 0.85 для 8 различных вопросов для эссе.

Table 3. Research where text answer analysis tasks are solved using combined models

Таблица 3. Исследования, где задачи анализа текстового ответа решаются с помощью комбинированных моделей

Авторы	Задача	Язык	Модель	Метод	Качество
Tulu и др. [52]	ASAG	Английский	WordNet, SemSpace MaLSTM	MaLSTM	RMSE 0.04 шкала 0–5 MAE 0.02–0.23
Zhang и др. [53]	ASAG	Английский	Стилометрия Метрики близости Скрытая марковская модель	DBN	F-мера 83 %
Lubis и др. [54]	ASAG	Индонезийский	word2vec Синтаксические характеристики	Косинусная метрика	MAE 0.70
Shashavali и др. [55]	Близость текстов	Английский	*n*-граммы FastText	Косинусная метрика	F-мера 93 %
Prabhudesai и др. [56]	ASAG	Английский	GloVe Стилометрия	Siamese LSTM	MAE 0.62 RMSE 0.89 шкала 0–5
Lakshmi и др. [57]	ASAG	Английский	Метрики близости	Случайный лес	F-мера 75 %
Hassan и др. [58]	Близость текстов	Английский Арабский	Эмбеддинги DSM Мешок слов BabelNet	Косинусная метрика	Пирсон 0.70–0.85
Del Gobbo и др. [59]	ASAG	Английский	TF-IDF, BERT	SVR	RMSE 0.73 шкала 0–5
Zhao и др. [60]	AES	Английский	BERT Стилометрия	Siamese NN ESIM	Каппа 0.50–0.85
Faseeh и др. [61]	AES	Английский	RoBERTa Лингвистические характеристики	Нейронная сеть	QWK 0.93 RMSE 0.18 шкала 0–5
Gagliardi и др. [62]	Близость текстов	Английский Французский	BERT, GPT WordNet	Косинусная метрика	F-мера 74 %
Beseiso и др. [63]	AES	Английский	BERT, word2vec Стилометрия	Нейронная сеть	Accuracy 75 % Каппа 0.77

ASAG — automatic short answer grading, автоматическая оценка кратких ответов.
AES — automated essay scoring, автоматизированная оценка эссе.
RMSE — root mean square error, среднеквадратичная ошибка.
QWK — quadratic weighted kappa, квадратично-взвешенная каппа.
MAE — mean absolute error, средняя абсолютная ошибка.

Faseeh и др. [61] применили гибридный подход для автоматической оценки эссе, скомбинировав языковую модель RoBERTa и лингвистические характеристики. Лингвистические характеристики включают в себя грамматические ошибки, близость между эталонным и студенческим эссе, рассчитанную как косинусную метрику между BERT-эмбеддингами, длины абзацев, тональность, частоту встречаемости частей речи, появление вспомогательных слов и словосочетаний, маркирующих структуру эссе, читабельность, богатство словаря. Эксперименты на корпусе ASAP показали QWK 0.93 и среднеквадратичную ошибку 0.18.

Авторы исследования [62] сопоставляют ключевые слова и фразы с использованием ансамблей эмбеддингов BERT, GPT, параметров связей на основе WordNet, алгоритма сравнения строк Джаро-Винклера. Эксперименты на мультиязычных корпусах с текстами на английском и французском языках показывают максимальное значение F-меры 74 %.

Beseiso и Alzahrani [63] оценивали эссе из корпуса ASAP. Они сконкатенировали три векторные репрезентации текстов: word2vec, BERT и стилометрические характеристики. Последние включали в себя количество именованных сущностей, слов, знаков пунктуации, частей речи, предложений,

а также значения метрик близости до 8 эссе, выбранных авторами вручную как образцовые. Доля правильных ответов достигла 75 %, каппа Коэна — 0.77.

Исследования, опирающиеся на комбинированные модели текстов, приведены в таблице 3. Таблица по структуре аналогична таблице из раздела 1.4.

Комбинированные модели встречаются существенно реже, чем только нейросетевые или только лингвистические. В основном они предназначены для английского языка.

В качестве эмбеддингов для моделирования текстов чаще берутся BERT и word2vec. Они конкатенируются с разнообразными лингвистическими моделями как на основе стандартных характеристик уровня слов, так и с более сложными характеристиками уровней структуры предложения или семантики.

Методы для сравнения и классификации комбинированных моделей совпадают с аналогичными методами, применяемыми для нейросетевых моделей.

Значения метрик качества достаточно хорошие. F-мера часто оказывается выше 74 %, среднеквадратичная ошибка и MAE в отдельных случаях близки к нулю.

Заключение

Анализ исследований в области оценки ответов учащихся на естественном языке показывает, что применяемые модели в целом соответствуют общей тенденции развития методов NLP. Среди стандартных лингвистических параметров популярны n-граммы, метрика TF-IDF, ключевые слова, синонимы. Эмбеддинги больших языковых моделей занимают главное место в работах последнего десятилетия. Однако не существует универсального подхода, обеспечивающего хорошее качество прогнозирования оценки ответа, даже в случаях бинарной классификации на правильные и неправильные. Явное преимущество показывают ансамблевые и комбинированные модели. Их комплексный подход позволяет учесть особенности поставленной задачи, но неизбежно ведёт к специализации разрабатываемого метода.

Можно отметить практически полное отсутствие исследований зависимости проверки текстов ответов от тематики задаваемых вопросов и их типов. В первую очередь это связано с отсутствием корпусов размеченных данных. Подавляющее большинство учёных формирует собственные корпуса, отсутствующие в открытом доступе. Такая ситуация сильно ограничивает объективность результатов. Интересно, что это характерно даже для английского языка.

Работы с английским языком сильно преобладают, несмотря на то, что хорошие языковые модели существуют для большинства естественных языков. Это открывает широкое поле для исследователей. Результаты автоматической проверки англоязычных ответов показывают перспективность разработки подобных национальных систем.

References

[1] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. 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] N. A. Medvedeva, N. G. Malkov, and M. L. Prozorova, “Professional and public accreditation as an assessment of agricultural educational program quality in Russia”, *Asian Journal of University Education*, vol. 17, no. 1, pp. 100–111, 2021. DOI: 10.24191/ajue.v17i1.12611.

[3] S. Bonthu, S. Rama Sree, and M. Krishna Prasad, “Automated short answer grading using deep learning: A survey”, in *Proceedings of the 5th International Cross-Domain Conference Machine Learning and Knowledge Extraction*, Springer, 2021, pp. 61–78. DOI: 10.1007/978-3-030-84060-0_5.

[4] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review", *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022. DOI: 10.1007/s10462-021-10068-2.

[5] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading", *International journal of artificial intelligence in education*, vol. 25, pp. 60–117, 2015. DOI: 10.1007/s40593-014-0026-8.

[6] L. Parra G and X. Calero S, "Automated writing evaluation tools in the improvement of the writing skill.", *International Journal of Instruction*, vol. 12, no. 2, pp. 209–226, 2019. DOI: 10.29333/iji.2019.12214a.

[7] A. Mizumoto and M. Eguchi, "Exploring the potential of using an ai language model for automated essay scoring", *Research Methods in Applied Linguistics*, vol. 2, no. 2, p. 100 050, 2023. DOI: 10.1016/j.rmal.2023.100050.

[8] S. Hsu, T. W. Li, Z. Zhang, M. Fowler, C. Zilles, and K. Karahalios, "Attitudes surrounding an imperfect AI autograder", in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15. DOI: 10.1145/3411764.3445424.

[9] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods", *Procedia Computer Science*, vol. 169, pp. 726–743, 2020. DOI: 10.1016/j.procs.2020.02.171.

[10] A. Y. Kerbeneva, "Razrabotka procedur intellektual'noj ocenki znanij na osnove semanticheskoj obrabotki otvetov pol'zovatelej na estestvennom yazyke", in *Informacionnye sistemy i tekhnologii IST-2021*, in Russian, 2021, pp. 187–191.

[11] A. A. Poguda and J. Tape, "Development of an algorithm and module for automatic evaluation of student papers based on semantic analysis of text", *Open Education*, vol. 28, no. 3, pp. 46–55, 2024, in Russian. DOI: 10.21686/1818-4243-2024-3-46-55.

[12] F. S. Pribadi, A. E. Permanasari, and T. B. Adji, "Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS)", *Education and Information Technologies*, vol. 23, pp. 2855–2866, 2018. DOI: 10.1007/s10639-018-9745-z.

[13] A.-B. Lekshmi-Narayanan and P. Brusilvosky, "Evaluating correctness of student code explanations: Challenges and solutions", in *Proceedings of 8th Educational Data Mining in Computer Science Education Workshop*, 2024, p. 1079.

[14] M. Meccawy, A. A. Bayazed, B. Al-Abdullah, and H. Algamdi, "Automatic essay scoring for Arabic short answer questions using text mining techniques", *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[15] 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.

[16] I. Buditjahjanto, M. Idhom, M. Munoto, and M. Samani, "An automated essay scoring based on neural networks to predict and classify competence of examinees in community academy.", *TEM Journal*, vol. 11, no. 4, 2022. DOI: 10.18421/TEM114-34.

[17] 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.

[18] 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, in Russian. DOI: 10.18255/1818-1015-2024-2-194-205.

[19] A. V. Kryukova, "Computing semantic similarity of russian texts by means of DKPro similarity tool", in *Trudy ob"edinyonnoj nauchnoj konferencii ¡¡Internet i sovremennoe obshchestvo¿¿*, 2017, pp. 87–97. DOI: 10.17586/2541-9781-2017-1-87-97.

[20] D. Bär, T. Zesch, and I. Gurevych, "Dkpro similarity: An open source framework for text similarity", in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2013, pp. 121–126.

[21] V. Nandini and P. Uma Maheswari, "Automatic assessment of descriptive answers in online examination system using semantic relational features", *The Journal of Supercomputing*, vol. 76, no. 6, pp. 4430–4448, 2020. DOI: 10.1007/s11227-018-2381-y.

[22] A. G. Leonov, N. S. Martynov, K. A. Mashchenko, A. A. Kholkina, and A. V. Shlyakhov, "Automation of semantic analysis for textual responses of students in a digital educational platform", *Software & Systems*, vol. 37, no. 3, pp. 440–452, 2024, in Russian. DOI: 10.15827/0236-235X.142.440-452.

[23] D. D. Zafievsky, N. S. Lagutina, O. A. Melnikova, and A. Y. Poletaev, "Text model for the automatic scoring of business letter writing", *Automatic Control and Computer Sciences*, vol. 57, no. 7, pp. 828–840, 2023. DOI: 10.3103/S0146411623070167.

[24] N. A. Prokopyev, "Automatic grading of answers in knowledge control for «definition» and «description» question types", *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, vol. 166, no. 4, pp. 580–593, 2024, in Russian. DOI: 10.26907/2541-7746.2024.4.580-593.

[25] A. Khoroshilov, A. Kan, E. Evdokimova, and S. Pitskhelauri, "Establishing similarities between text documents", *Modelling and Data Analysis*, vol. 13, no. 4, pp. 45–58, 2023, in Russian. DOI: 10.17759/mda.2023130403.

[26] L. Ouahrani and D. Bennouar, "AR-ASAG an Arabic dataset for automatic short answer grading evaluation", in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 2634–2643.

[27] V. A. Kozhevnikov and O. Y. Sabinin, "System of automatic verification of answers to open questions in Russian", *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems*, vol. 11, no. 3, pp. 57–72, 2018, in Russian. DOI: 10.18721/JCSTCS.11306.

[28] D. V. Gadasin, A. V. Shvedov, and I. S. Vakurin, "Opredelenie semanticheskoj blizosti tekstov s ispol'zovaniem algoritma sravneniya sushchnosti grafov", *REDS: Telekommunikacionnye ustrojstva i sistemy*, vol. 12, no. 4, pp. 11–19, 2022, in Russian.

[29] N. Maharjan and V. Rus, "A concept map based assessment of free student answers in tutorial dialogues", in *Artificial Intelligence in Education: 20th International Conference, AIED*, Springer, 2019, pp. 244–257. DOI: 10.1007/978-3-030-23204-7_21.

[30] I. Y. Kashirin, "Binarnye ierarhicheskie chisla dlya vychisleniya semanticheskoj blizosti predlozhenij estestvennogo yazyka", *Vestnik of RSREU*, no. 86, pp. 110–120, 2023, in Russian. DOI: 10.21667/1995-4565-2023-86-110-121.

[31] I. Y. Kashirin, "Ierarhicheskie chisla dlya proektirovaniya ICF-taksonomij iskusstvennogo intellekta", *Vestnik of RSREU*, no. 71, pp. 71–82, 2020, in Russian. DOI: 10.21667/1995-4565-2020-71-71-82.

[32] C. Minnegalieva, G. Sabitova, and A. Gayaliev, "Method of pre-assessment of students' answers based on the vector model of documents", *Russian Digital Libraries Journal*, vol. 26, no. 3, pp. 324–339, 2023, in Russian. DOI: 10.26907/1562-5419-2023-26-3-324-339.

[33] C. Minnegalieva, I. Kashapov, and O. Morozova, "Automated students' short answers grading using language models", *Russian Digital Libraries Journal*, vol. 27, no. 3, pp. 278–293, 2024, in Russian. DOI: 10.26907/1562-5419-2024-27-3-278-293.

[34] I. G. Ndukwe, C. E. Amadi, L. M. Nkomo, and B. K. Daniel, "Automatic grading system using Sentence-BERT network", in *Artificial Intelligence in Education*, 2020, pp. 224–227.

[35] M. Fateen and T. Mine, "In-context meta-learning vs. semantic score-based similarity: A comparative study in Arabic short answer grading", in *Proceedings of ArabicNLP 2023*, 2023, pp. 350–358. DOI: 10.18653/v1/2023.arabicnlp-1.28.

[36] S. K. Gaddipati, D. Nair, and P. G. Plöger, *Comparative evaluation of pretrained transfer learning models on automatic short answer grading*, 2020. arXiv: 2009.01303 [cs.CL].

[37] J. Schneider, R. Richner, and M. Riser, "Towards trustworthy autograding of short, multi-lingual, multi-type answers", *International Journal of Artificial Intelligence in Education*, vol. 33, no. 1, pp. 88–118, 2023.

[38] M. Hendre, P. Mukherjee, R. Preet, and M. Godse, "Efficacy of deep neural embeddings based semantic similarity in automatic essay evaluation", *International Journal of Computing and Digital Systems*, vol. 9, pp. 1–11, 2020.

[39] A. Doewes, A. Saxena, Y. Pei, and M. Pechenizkiy, "Individual fairness evaluation for automated essay scoring system", in *Proceedings of the 15th International Conference on Educational Data Mining*, 2022, pp. 206–216. DOI: 10.5281/zenodo.685315.

[40] Q. Wang, "The use of semantic similarity tools in automated content scoring of fact-based essays written by EFL learners", *Education and Information Technologies*, vol. 27, no. 9, pp. 13 021–13 049, 2022.

[41] B. F. Dhini, A. S. Girsang, U. U. Sufandi, and H. Kurniawati, "Automatic essay scoring for discussion forum in online learning based on semantic and keyword similarities", *Asian Association of Open Universities Journal*, vol. 18, no. 3, pp. 262–278, 2023. DOI: 10.1108/aaouj-02-2023-0027.

[42] S. V. Bogolepova and M. G. Zharkova, "Researching the potential of generative language models for essay evaluation and feedback provision", *Domestic and Foreign Pedagogy*, vol. 1, no. 5, pp. 123–137, 2024, in Russian. DOI: 10.24412/2224fi??0772fi??2024fi??101fi??123fi??137.

[43] M. N. Evstigneev, "Thematic control and criteria-based assessment of foreign language writing skills using artificial intelligence technologies", *Tambov University Review. Series: Humanities*, vol. 29, no. 4, pp. 913–926, 2024, in Russian. DOI: 10.20310/1810-0201-2024-29-4-913-926.

[44] V. N. Kopnin and N. S. Lagutina, "Opredelenie semanticheskogo skhodstva tekstov s ispol'zovaniem yazykovyh modelej na osnove transformerov", in *Matematicheskoe i informacionnoe modelirovanie : materialy Vserossijskoj konferencii molodyh uchenyh, Tyumen'*, in Russian, Tyumen' : TyumGU-Press, 2024, pp. 143–146.

[45] W. H. Gomaa and A. A. Fahmy, "Ans2vec: A scoring system for short answers", in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4*, Springer, 2020, pp. 586–595.

[46] R. Weegar and P. Idestam-Almquist, "Reducing workload in short answer grading using machine learning", *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 247–273, 2024. DOI: 10.1007/s40593-022-00322-1.

[47] Z. Li, Y. Tomar, and R. J. Passonneau, "A semantic feature-wise transformation relation network for automatic short answer grading", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6030–6040. DOI: 10.18653/v1/2021.emnlp-main.487.

[48] H. Tan, C. Wang, Q. Duan, Y. Lu, H. Zhang, and R. Li, "Automatic short answer grading by encoding student responses via a graph convolutional network", *Interactive Learning Environments*, vol. 31, no. 3, pp. 1636–1650, 2023. DOI: 10.1080/10494820.2020.1855207.

[49] J. Xie, K. Cai, L. Kong, J. Zhou, and W. Qu, "Automated essay scoring via pairwise contrastive regression", in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 2724–2733.

[50] J. Garg, J. Papreja, K. Apurva, and G. Jain, "Domain-specific hybrid BERT based system for automatic short answer grading", in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, IEEE, 2022, pp. 1–6. DOI: 10.1109/CONIT55038.2022.9847754.

[51] 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.

[52] C. N. Tulu, O. Ozkaya, and U. Orhan, "Automatic short answer grading with SemSpace sense vectors and MaLSTM", *IEEE Access*, vol. 9, pp. 19 270–19 280, 2021.

[53] Y. Zhang, C. Lin, and M. Chi, "Going deeper: Automatic short-answer grading by combining student and question models", *User modeling and user-adapted interaction*, vol. 30, no. 1, pp. 51–80, 2020.

[54] F. F. Lubis, A. Putri, D. Waskita, T. Sulistyaningtyas, A. A. Arman, Y. Rosmansyah, *et al.*, "Automated short-answer grading using semantic similarity based on word embedding", *International Journal of Technology*, vol. 12, no. 3, pp. 571–581, 2021. DOI: 10.14716/ijtech.v12i3.4651.

[55] 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.

[56] A. Prabhudesai and T. N. Duong, "Automatic short answer grading using Siamese bidirectional LSTM based regression", in *Proceedings of the IEEE International Conference on Engineering, Technology and Education (TALE)*, IEEE, 2019, pp. 1–6. DOI: 10.1109/TALE48000.2019.9226026.

[57] P. S. Lakshmi, J. Simha, and R. Ranjan, "Empowering educators: Automated short answer grading with inconsistency check and feedback integration using machine learning", *SN Computer Science*, vol. 5, no. 5, p. 653, 2024. DOI: 10.1007/s42979-024-02954-7.

[58] 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.

[59] E. Del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "GradeAid: A framework for automatic short answers grading in educational contexts—design, implementation and evaluation", *Knowledge and Information Systems*, vol. 65, no. 10, pp. 4295–4334, 2023.

[60] J. Zhao, Y. Li, and W. Feng, "Investigating the validity and reliability of a comprehensive essay evaluation model of integrating manual feedback and intelligent assistance", *International Journal of Emerging Technologies in Learning*, vol. 18, no. 4, 2023. DOI: 10.3991/ijet.v18i04.38241.

[61] M. Faseeh *et al.*, "Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy", *Mathematics*, vol. 12, no. 21, p. 3416, 2024. DOI: 10.3390/math12213416.

[62] 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.

[63] M. Beseiso and S. Alzahrani, "An empirical analysis of BERT embedding for automated essay scoring", *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020. DOI: 10.14569/ijacsa.2020.0111027.