

A Model for Automated Business Writing Assessment

D. D. Zafievsky¹, N. S. Lagutina¹, O. A. Melnikova¹, A. Y. Poletaev¹

DOI: [10.18255/1818-1015-2022-4-348-365](https://doi.org/10.18255/1818-1015-2022-4-348-365)

¹P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50

Research article

Full text in Russian

Received September 27, 2022

After revision November 14, 2022

Accepted November 16, 2022

This study is aimed at building an automated model for business writing assessment, based on 14 rubrics that integrate EFL teacher assessment frameworks and identify expected performance against various criteria (including language, task fulfillment, content knowledge, register, format, and cohesion). We developed algorithms for determining the corresponding numerical features using methods and tools for automatic text analysis. The algorithms are based on a syntactic analysis with the use of dictionaries. The model performance was subsequently evaluated on a corpus of 20 teacher-assessed business letters. Heat maps and UMAP results represent comparison between teachers' and automated score reports. Results showed no significant discrepancies between teachers' and automated score reports, yet detected bias in teachers' reports. Findings suggest that the developed model has proved to be an efficient tool for natural language processing with high interpretability of the results, the roadmap for further improvement and a valid and unbiased alternative to teachers' assessment. The results may lay the groundwork for developing an automatic students' language profile. Although the model was specifically designed for business letter assessment, it can be easily adapted for assessing other writing tasks, e.g. by replacing dictionaries.

Keywords: natural language processing; text features; automated essay scoring; business letter

INFORMATION ABOUT THE AUTHORS

Daniil Dmitrievich Zafievsky | orcid.org/0000-0001-8266-2283. E-mail: zafievsky@mail.ru
student.

Nadezhda Stanislavona Lagutina | orcid.org/0000-0002-6137-8643. E-mail: lagutinans@gmail.com
correspondence author | PhD, associate professor.

Oksana Andreyevna Melnikova | orcid.org/0000-0001-8814-7696. E-mail: oam8108@gmail.com
associate professor.

Anatoliy Yurievich Poletaev | orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru
post-graduate student.

Funding: The reported study was funded by YarSU Program according to the research project No. P2-GM5-2021.

For citation: D. D. Zafievsky, N. S. Lagutina, O. A. Melnikova, and A. Y. Poletaev, "A Model for Automated Business Writing Assessment", *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 348-365, 2022.

Модель текста для автоматической оценки делового письма на заданную тему

Д. Д. Зафиевский¹, Н. С. Лагутина¹, О. А. Мельникова¹, А. Ю. Полетаев¹

DOI: [10.18255/1818-1015-2022-4-348-365](https://doi.org/10.18255/1818-1015-2022-4-348-365)

¹Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912

Научная статья

Полный текст на русском языке

Получена 27 сентября 2022 г.

После доработки 14 ноября 2022 г.

Принята к публикации 16 ноября 2022 г.

В статье описана модель текста, предназначенная для автоматической оценки связного текста в виде письма на заданную тему. Параметры оценки сформулированы и формализованы в виде 14 критериев при помощи экспертов в области обучения английскому языку. Критерии включают параметры, относящиеся к анализу лексики, включая особенности предметной области, тематики текста, стилю и формату письма, средствам логической связи предложений. Авторами разработаны алгоритмы определения соответствующих числовых характеристик с использованием методов и инструментов автоматического анализа текстов. Алгоритмы основаны на анализе состава и структуры предложений, для чего используются, в том числе данные специализированных словарей. Характеристики ориентированы на проверку электронного делового письма, но могут быть адаптированы к анализу других письменных текстов, например, с помощью замены словарей. На основе разработанных алгоритмов создана система автоматической оценки текстов. Проведён эксперимент по анализу результатов работы этой системы на корпусе из 20 текстов, предварительно размеченных преподавателями английского языка. 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, кроме того, автоматическая оценка оказалась более объективной. Таким образом, разработанная модель успешно справилась с поставленной задачей и может применяться для оценки текстов, написанных человеком. Результаты будут использованы в проекте автоматического построения языкового профиля учащегося. Достоинствами модели являются хорошая интерпретируемость получаемых результатов, объективность, перспективы развития.

Ключевые слова: автоматическая обработка текста; параметры текста; автоматизированная оценка эссе; деловое письмо

ИНФОРМАЦИЯ ОБ АВТОРАХ

Даниил Дмитриевич Зафиевский	orcid.org/0000-0001-8266-2283 . E-mail: zafievsky@mail.ru студент.
Надежда Станиславовна Лагутина автор для корреспонденции	orcid.org/0000-0002-6137-8643 . E-mail: lagutinans@gmail.com канд. физ.-мат. наук, доцент.
Оксана Андреевна Мельникова	orcid.org/0000-0001-8814-7696 . E-mail: oam8108@gmail.com доцент.
Анатолий Юрьевич Полетаев	orcid.org/0000-0003-0116-4739 . E-mail: anatoliy-poletaev@mail.ru аспирант.

Финансирование: Исследование выполнено в рамках Программы развития ЯрГУ, проект № П2-ГМ5-2021.

Для цитирования: D. D. Zafievsky, N. S. Lagutina, O. A. Melnikova, and A. Y. Poletaev, "A Model for Automated Business Writing Assessment", *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 348-365, 2022.

Введение

Изменение процесса межличностного взаимодействия вследствие процессов глобализации и цифровизации повлекли за собой постановку ряда новых задач перед сообществом преподавателей, и в первую очередь перед сообществом преподавателей иностранного языка. Одной из таких задач сегодня является обучение эффективной онлайн-коммуникации как преобладающему способу взаимодействия в межкультурном пространстве. Актуальность данной задачи подтверждается включением компетенций и стратегий онлайн-коммуникации в обновленную версию стандарта для описания уровней владения иностранным языком Common European Framework of Reference (CEFR). Обучение такому виду коммуникации на практике представляет особую сложность вследствие наличия объективных различий в стратегиях для достижения эффективности при онлайн- и оффлайн-взаимодействии, которые часто не находят отражения в общепринятых методических подходах к обучению иностранному языку [1].

Важнейшее место в методике преподавания иностранного языка и, в частности, онлайн-коммуникации на иностранном языке, занимает обучение продуктивным видам речевой деятельности, ключевым аспектом которых является овладение навыком письма [2]. Поскольку письменная речь предполагает одновременное владение коммуникативными, лингвистическими и экстралингвистическими компетенциями на достаточном для эффективной коммуникации уровне [3], обучение навыкам письма и особенно их оценка нередко сопровождаются рядом проблем, которые усугубились в ходе пандемии COVID-19 [1, 4]. Оценка письменной речи является одним из способов измерения знаний, практических умений, логического и критического мышления и творческих способностей [5]. Чаще всего текст, написанный учащимся, квалифицируется человеком-экспертом на основе некоторых критериев. Однако ручная оценка требует много времени, особенно при большом количестве экзаменуемых, и достаточно необъективна, так как зависит от множества субъективных факторов: строгости оценщика, его последовательности, усталости [6]. Решением проблемы может стать система автоматизированной оценки письменной речи (Automated Essay Scoring, AES). AES — современная, актуальная задача в области обработки естественного языка [7]. Следует отметить, что термин «эссе (essay)» в области автоматической обработки текста имеет более широкую трактовку по сравнению с классической лингвистикой. В качестве эссе рассматривается короткий связный письменный текст на естественном языке.

Целью систем AES является имитация человека-оценщика. Для её достижения таких систем использует методы и инструменты обработки естественного языка и машинного обучения. При этом основой любого решения служит построение модели текста, т. е. выбор числовых параметров и формирование соответствующего вектора характеристик. Условно можно выделить две группы параметров: определяемые экспертами-людьми и получаемые автоматически с помощью методов глубокого обучения. Методы глубокого обучения показывают хорошие результаты в различных задачах, связанных с классификацией текстов. Однако следует отметить ряд проблем: хороший результат получается только при наличии больших объёмов данных для обучения, мало исследованы оценки тематических текстов в рамках предметной области, результаты работы не поддаются интерпретации, что затрудняет анализ ошибок [8]. Часть проблем даёт возможность решить группа характеристик, определяемых экспертами. Эти параметры включают меры, определяющие лексическую сложность и разнообразие текста, его синтаксические особенности, шаблоны, величины для кодирования аспектов, связанных с письменной речью. Этот подход может работать и с небольшими корпусами текстов, а также учитывать особенности предметной области, позволяет интерпретировать результаты и проводить анализ ошибок [9].

Поэтому, когда перед авторами работы встала задача автоматизации построения профиля языковых компетенций обучающихся английскому языку, в рамках которой потребовалась оценка делового письма на заданную тему, было принято решение построить и исследовать модель текста

письма на основе параметров, предложенных экспертами-лингвистами. Построение модели текста, описываемое в данной статье, заключается в разработке алгоритмов оценки уровня владения компетенциями языкового профиля и формировании соответствующего вектора числовых характеристик, а также анализе корпуса писем с применением разработанной модели.

1. Обзор связанных работ

Алгоритмы автоматической оценки писем разрабатываются с шестидесятих годов прошлого века. В настоящий момент наиболее популярны подходы на основе нейронных сетей.

Одна из первых моделей на основе рекуррентной нейронной сети (RNN) для решения задачи AES была предложена в работе [10]. Многоуровневая нейронная сеть обучалась на размеченном корпусе текстов и предсказывала оценку эссе из диапазона $[0, 1]$. Характерной работой в рамках обсуждаемого подхода является [11]. Её авторы предложили двухуровневую двунаправленную нейронную сеть LSTM для оценки эссе. Моделью текста послужили функции, извлеченные с помощью word2vec для обучения LSTM, точность решения в среднем составила 0,87.

Однако другие исследователи показали [12], что качество работы таких методов сильно зависит от квалификации и согласованности экспертов, оценивающих эссе, из которых формируется обучающий корпус. Чтобы решить эту проблему, авторы предложили двухэтапную архитектуру, которая объединила вероятностную психологическую модель оценки текстов и нейронную сеть DNN-AES. Эксперименты показали, что такой метод хорошо подходит для ситуаций, когда сложно подготовить качественные обучающие данные. Другим слабым местом моделей на основе RNN является оценка связности текста. Для решения этой задачи было предложено несколько моделей DNN-AES, учитывающих особенности когерентности отдельных частей эссе [13, 14]. Проблема решается дополнением в базовые нейронные сети элементов, моделирующих связность и семантику текста. Авторы работы [15] сформулировали и решили вспомогательную задачу извлечения семантических признаков на уровне отдельных предложений и на уровне документа в целом. Эти параметры были добавлены к другим характеристикам для получения целостной оценки письма и привели к росту качества решения задачи.

Одной из самых популярных современных моделей текста на основе глубоких нейронных сетей является BERT, выпущенная командой Google AI Language в 2018 году. BERT успешно применялся для AES и показал хорошие результаты [16–18].

К сожалению, методические и числовые особенности оценки эссе очень разнообразны и затрудняют использование каждой разработанной системы для других наборов данных или условий оцениваемой задачи. На эту проблему указывают авторы работы [19] и решают её с помощью структуры, объединяющей несколько моделей AES с учетом характеристик каждой модели с использованием теории оценки ответов. В этой структуре несколько моделей AES сначала обучаются независимо, а затем используются для получения интегрированной оценки целевых эссе. Для оценки эссе по заданному стандарту учёные [20] предложили мультимодальный подход к машинному обучению. Они подсчитывают грамматические и статистические параметры текста с помощью библиотеки spaCy и дополняют их эмбедингами слов GloVe. Подготовленные таким образом векторы предложений передают в нейронную сеть LSTM для нахождения окончательного результата. Bi-LSTM проверяет каждое предложение в двух направлениях, чтобы найти семантику.

Описанные системы на основе нейронных сетей моделируют текст автоматически, с помощью эмбедингов, однако сложные экспертные лингвистические характеристики тоже играют важную роль. В работе [21] предлагается использовать объединение нечёткой онтологии с методами латентно-семантического анализа (LSA). При этом извлекаются два типа функций: синтаксические и морфологические. На основе дерева синтаксического анализа, поиска подобия слов и предложений получают параметры, характеризующие семантику текста. Синтаксические и морфологические

особенности объединяются для получения окончательной оценки. Похожая модель, агрегирующая простые и сложные параметры текста, используется в работе [22].

Авторы работы [23] замечают, что многословные выражения связаны с владением языком, но игнорируются в работах по автоматической оценке обучающихся. Поэтому в своей статье авторы оценивают их важность для автоматизированной оценки эссе и сравнивают с другими классическими функциями. Хотя эксперименты показали, что для автоматического оценивания классические функции более значимы, чем многословные выражения, тем не менее, предложенные на основе многословных выражений параметры позволяют анализировать текст более детально.

Поддавляющее большинство работ над AES сосредоточено на комплексной оценке, которая обобщает качество эссе с помощью единой оценки. Корпуса эссе, аннотированные вручную, с такими оценками являются общедоступными, что облегчает разработку систем AES, особенно на основе методов машинного обучения. Эти системы дают возможность автоматизировать оценку миллионов эссе, написанных для стандартных тестов, что экономит много усилий по ручному оцениванию. Однако главным недостатком такого подхода является невозможность интерпретации результата и, соответственно, отсутствие обратной связи с обучающимся. Возвращение единственной оценки не говорит о том, какие аспекты письма способствовали этой оценке и тому, как ее можно улучшить [24].

Кроме этого в области автоматической оценки письма остаются вопросы: какие параметры в наибольшей степени способствуют точности оценки, можно ли обобщить эти характеристики на другие наборы данных с другой схемой оценивания и другими целевыми группами учащихся, какие конкретные лингвистические свойства письменной речи учащегося способствуют получению более высокой оценки владения языком. Исследуемые в данной статье характеристики, сформированные с помощью экспертов-лингвистов, дают возможность решать эти вопросы и строить образовательные системы с обратной связью.

2. Критерии оценки компетенций языкового профиля

Для полной, всесторонней и объективной оценки предложенного учащемуся письменного задания экспертами была создана матрица критериальных шкал для оценивания, включившая в себя как опыт зарубежных коллег, отраженный в системе уровней владения иностранным языком CEFR и критериях оценки международных экзаменов IELTS Cambridge Assessment English, так и опыт отечественных методистов, нашедший отражение в критериях оценки письменных заданий ЕГЭ и ОГЭ. Составленная матрица включает в себя четыре блока критериев: решение коммуникативной задачи (Communicative Achievement), организация речи (Organisation), лексика (Vocabulary), грамматика (Grammar).

Блок «Решение коммуникативной задачи» предназначен для контроля уровня сформированности коммуникативной компетенции, а именно оценки эффективности решения коммуникативной задачи: написание письма начальнику с просьбой о назначении на определенную должность. Для этого был разработан и выделен ряд критериев, нацеленных на определение способности автора письма решать коммуникативные микро-задачи, т. е. проверка наличия указанной в задании информации в письме. Данный блок также нацелен на контроль сформированности медиативной компетенции, в данном случае — умения вежливо и эффективно разрешить потенциально конфликтную ситуацию, для которого был выделен отдельный критерий оценки использованных в письме норм вежливости и умения эффективно и доброжелательно выразить свою точку зрения. Блок также включает в себя проверку сформированности лингвистической и экстралингвистической компетенций: стилистического оформления речи, необходимого для решения поставленной задачи, что предполагает наличие определенных фоновых знаний по затронутой проблеме.

Блок «Организация речи» включает в себя критерии проверки когезии — внутренней, лексико-грамматической связности текста, и когерентности — логичности изложения, связности дискурса,

а также правильности структурирования текста и его соответствия заявленному формату электронного письма.

Блок «Лексика» оценивает словарный запас и разнообразие используемой лексики, а также соответствие употребляемой лексики уровню по общеевропейской шкале CEFR. Данный блок также включает в себя критерии оценки точности выбора лексических единиц и словосочетаний, корректности их употребления и соответствия теме и ситуации общения.

Блок «Грамматика» предназначен для контроля уровня сформированности грамматических навыков: разнообразия употребляемых грамматических структур и явлений, точности и корректности их употребления и соответствия употребляемых структур и явлений теме и ситуации общения.

Авторами работы была предложена модель текста для проверки деловых писем. Основой модели служит набор формализованных числовых параметров (критериев), приведенный в таблице 1. Каждый из критериев оценки делового письма принимает значение один или ноль.

Table 1. Criteria used in evaluating texts

Таблица 1. Критерии, применяемые при оценке текстов

Критерий	Описание критерия	Метод оценки
K1	Письмо содержит информацию про презентацию и связанные с ней факты	Алгоритм определения схожести слов
K2	Письмо содержит информацию про коллегу и связанные с ним факты	Алгоритм поиска слова
K3	Письмо содержит информацию про приложение и связанные с ним факты	Алгоритм определения схожести слов
K4	Письмо содержит упоминание про навыки публичных выступлений автора	Алгоритм анализа темы предложения
K5	Письмо содержит упоминание про навыки публичных выступлений коллеги	Алгоритм анализа темы предложения
K6	Письмо содержит обращение к начальнику	Алгоритм поиска слова
K7	Письмо содержит просьбу о назначении автора	Работа со словарём
K8	Стиль письма соответствует заявленному формату задания	Работа со словарём
K9	Присутствует обращение	Регулярное выражение
K10	Присутствует разделение на абзацы	Регулярное выражение
K11	Присутствуют клишированные выражения	Работа со словарём
K12	Присутствует заключение (подпись)	Регулярное выражение
K13	Присутствуют связующие элементы и средства логической связи	Работа со словарём
K14	Наблюдается разнообразие корректно подобранных связующих элементов и средств логической связи	Работа со словарём

В первую очередь, письменное задание оценивается с точки зрения успешности выполнения коммуникативной задачи, то есть соответствия содержания письма заявленной теме и ситуации. Контроль решения коммуникативной задачи осуществлялся с помощью критериев K1–K7. В данном случае главной целью учащегося является обращение к начальнику с просьбой о назначении автора выступающим на презентации проекта. Для выполнения этой цели в письме требуется:

- упомянуть предстоящую презентацию (K1);
- упомянуть коллегу (K2) и навыки его публичных выступлений (K5);
- упомянуть приложение (K3);
- упомянуть навыки своих публичных выступлений (K4);
- обратиться к начальнику (K6) с просьбой о назначении автора выступающим на презентации проекта (K7).

Критерий считается выполненным, если учащимся было упомянуто слово или словосочетание, обозначающее представленные в описании критерия явления, людей и предметы, или использована грамматическая структура, на них указывающая.

Другим важным критерием успешности решения коммуникативной задачи является отсутствие в тексте письма фактических ошибок и/или информации, противоречащей заданию или заявленной ситуации общения.

Ещё одним критерием успешности решения коммуникативной задачи стало стилевое оформление (K8). Поскольку письмо, которое требуется написать, является примером деловой коммуникации, для решения поставленной коммуникативной задачи учащемуся необходимо продемонстрировать владение деловым стилем общения, в частности, умением вести деловую переписку, используя специальные стилистически маркированные языковые средства. Для проверки соответствия стилевому оформлению нами был составлен список таких средств, т. е. наиболее часто употребляемой лексики, а также наиболее употребимых грамматических явлений, на базе ведущих и наиболее популярных учебно-методических комплексов по бизнес-английскому и деловой переписке, например, Oxford English for Emails. Критерий считался удовлетворённым, если учащийся не только употребил указанные в списке средства, но и не использовал средств, характерных для иных функциональных стилей речи.

Блок «Организация речи» содержит критерии, проверяющие соответствие заявленному формату задания, т. е. электронному письму начальнику (K9–K13). Такое письмо должно быть разделено на абзацы, в нём должны присутствовать обращение и заключение, характерные для электронной деловой переписки. Учащемуся также необходимо продемонстрировать владение соответствующими стилю клишированными выражениями.

С помощью критерия K14 осуществлялась оценка когезии текста, т. е. наличия связующих элементов и средств логической связи. Учащемуся важно не просто использовать разнообразные средства, но и употребить их корректно в соответствии с речевой ситуацией. Формализация этого критерия — наиболее сложная задача, в данной работе для её решения использовались специально составленные словари фраз и словосочетаний. Лексика и грамматика оцениваются опосредованно во всех критериях.

3. Алгоритмы определения параметров текста

Для оценки соответствия текста критериям K1–K6 были разработаны следующие алгоритмы: поиска слов в тексте, поиска сходных слов в тексте и анализа темы предложения.

Оценка соответствия текста критериям K9, K10 и K12 производится при помощи регулярных выражений, поскольку в соответствующих критериях можно явно выделить конструкции, которые должны присутствовать в тексте. Так для критерия K10 такой конструкцией является пустая строка между параграфами (два символа перехода на новую строку в строковом представлении текста). Для получения балла за критерий K9 необходимо иметь в первом абзаце письма формальное обращение к начальнику, такое как «Good morning, Dear Mrs. Addams» или «Good afternoon, Mrs. Addams» и подобные им. При оценке критерия K12 применяется регулярное выражение, отслеживающее наличие в подписи полного имени человека, например, Ivanov Ivan Ivanovich или Mrs. Addams. Следует отметить, что система обнаруживает и более сложные имена, такие как Ludwig Heinrich Edler von Mises, Muhammad ibn Musa al-Khwarizmi и множество других.

Для критериев K7, K8, K11, K13, K14 при проверке соответствия им текста используются специально составленные словари слов-связок, клишированных и формальных выражений.

3.1. Алгоритм поиска слов

Алгоритм выполняет последовательный поиск необходимого слова с начала текста. Например, при применении для определения соответствия текста критерию K2 происходит поиск упоминания

информации о коллеге отслеживанием использования слова «colleague» или любой из допустимых форм полного имени коллеги автора письма, а при проверке соответствия критерию K6 ищется обращение автора письма к начальнику. При оценке данных критериев достаточно самого использования слова в письме без учета контекста, а само их применение явно продиктовано постановкой задания.

3.2. Алгоритм поиска сходных слов

Алгоритм поиска сходных слов используется в случаях, когда необходимо определить, есть ли в тексте слово, максимально приближенное к заданному. На вход алгоритм получает слово для поиска, множество лемматизированных слов из письма без стоп-слов для английского языка (получаемых из библиотеки NLTK), а также множество слов-исключений.

Для решения этой задачи формируется множество, содержащее как само исходное слово, так и его гипонимы и гиперонимы. При необходимости есть возможность указания специальных слов-исключений, которые являются контекстуальными синонимами искомого слова. Они тоже будут включены в это множество. Наличие такой возможности обусловлено тем, что используемая библиотека не учитывает особенности предметной области. Таким образом получается обойти данную проблему. После формирования множества ищется его пересечение с множеством слов текста, очищенного от стоп-слов из библиотеки NLTK. Если пересечение не пусто — ставим за критерий один балл и прекращаем работу алгоритма. В противном случае необходимо выделить из искомого слова корень и начать искать однокоренное слово в тексте. При обнаружении — за критерий ставится один балл, а иначе ставится ноль баллов.

В алгоритме для оценки соответствия критерию K1 используется слово «presentation», а множество слов-исключений пустое. В алгоритме для оценки соответствия критерию K3 применяется слово «application», а множество слов-исключений содержит только слово «product».

3.3. Алгоритм анализа темы предложения

Алгоритм 1 анализа темы предложения применяется для поиска в тексте упоминаний о навыках публичного выступления автора текста и его коллеги. Подходящие предложения из всех предложений письма определяются с помощью ключевого глагола, существительного и прилагательного, а также их гипонимов и гиперонимов, кроме того, может учитываться присутствие в предложении отрицания. По наличию ключевых слов или сходных с ними слов в тексте принимается решение о соответствии предложения требуемой теме. Поскольку одну мысль возможно выразить множеством способов, данный алгоритм может быть применен для одного критерия несколько раз для охвата большего количества возможных комбинаций с различными наборами ключевых слов. При работе с прилагательными наличие сравнительных прилагательных или прилагательных превосходной степени (JJR и JJS соответственно) в тексте следует учитывать как альтернативу ключевому прилагательному или сходному к нему слову. Возможность учета отрицания в предложении необходима по причине того, что негативная характеристика навыков публичного предложения коллеги может быть описана не только при помощи негативного прилагательного, но и положительного с применением отрицания в тексте. Таким образом алгоритм учитывает большее количество возможных описаний негативной характеристики коллеги.

При помощи NLTK_mark_pos_tags происходит токенизация предложения и определение PoS-тега для каждого слова.

При помощи NLTK_build_lemmatized_sentence_word_set происходит составление лемматизированного множества слов текущего предложения.

При реализации обеих функций используется библиотека NLTK.

При проверке соответствия критерию K4 используются следующие данные: «speaker», «represent», «adequate» — ключевые слова, «I», «myself» — местоимения. Отрицание не ищется.

Algorithm 1. Sentence topic analysis algorithm**Algorithm 1.** Алгоритм анализа
темы предложения**Data:** *sentence_list*, *noun_verb_word_set*, *adj_word_set*, *pronoun_list*, *check_negation***for** *sentence* in *sentence_list* **do** *sentence_pos_tags* $\leftarrow$ *NLTK_mark_pos_tags*(*sentence*); *words_in_sentence_set* $\leftarrow$ *NLTK_build_lemmatized_sentence_word_set*(*sentence_pos_tags*); *intersection_size* $\leftarrow$ *len*(*words_in_sentence_set.intersection*(*noun_verb_word_set*)); *pronoun_contained* $\leftarrow$ *adjective_contained* $\leftarrow$ *negation_contained* $\leftarrow$ *False*; **for** *current_pos* in *sentence_pos_tags* **do** **if** *current_pos*[0].*lower*() in *pronoun_list* **then** *pronoun_contained* $\leftarrow$ *True*; **end** **if** *current_pos*[1] in ['JJ', 'JJR'] or *current_pos*[0].*lower*() in *adj_word_set* **then** *adjective_contained* $\leftarrow$ *True*; **end** **if** *check_negation* and *current_pos*[0].*lower*() in ['n't', 'not', 'no', 'never'] **then** *negation_contained* $\leftarrow$ *True*; **end** **end** *sentence_suitability* $\leftarrow$ *adjective_contained* or *negation_contained*; *sentence_suitability* $\leftarrow$ *pronoun_contained* and *sentence_suitability*; **if** *intersection_size* $\neq$ 0 and *sentence_suitability* **then** **return** 1; **end****end****return** 0;

При проверке соответствия критерию K5 используется два набора данных, а оценка ставится как максимум из ответов. Первый набор данных: «speaker», «represent», «adequate» — ключевые слова, «he», «she», «his», «her», «himself», «herself», «john» — набор местоимений. Отрицание ищется. Второй набор данных: «speaker», «represent», «incompetent» — ключевые слова, «he», «she», «his», «her», «himself», «herself», «john» — набор местоимений. Отрицание не ищется.

4. Результаты экспериментов

Разработанные алгоритмы легли в основу автоматической системы оценивания связного текста. С использованием этой системы была проведена экспериментальная оценка двадцати писем на английском языке, размеченных экспертами-лингвистами по указанным критериям. Результаты оценки эксперта-лингвиста 1 и системы 2 представлены в виде тепловых карт, на которых насыщенность цвета прямо пропорциональна величине оценки по указанному критерию. Оценки эксперта более тонкие за счет возможности выставления половинных баллов; система же выставляет по каждому критерию только один или ноль баллов.

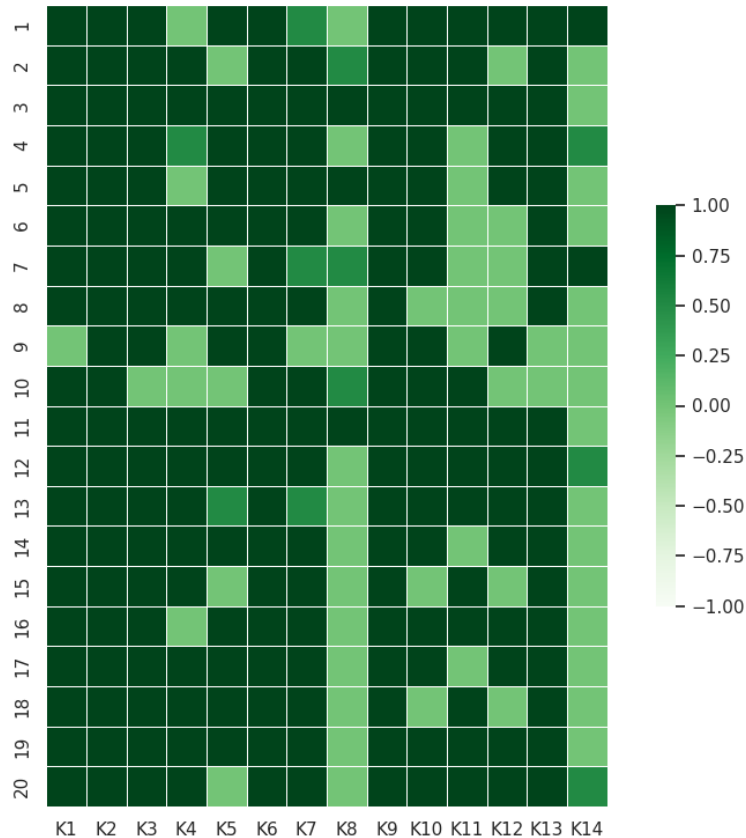


Fig. 1. Heat map of the lingvist's estimates

Рис. 1. Тепловая карта оценок эксперта-лингвиста

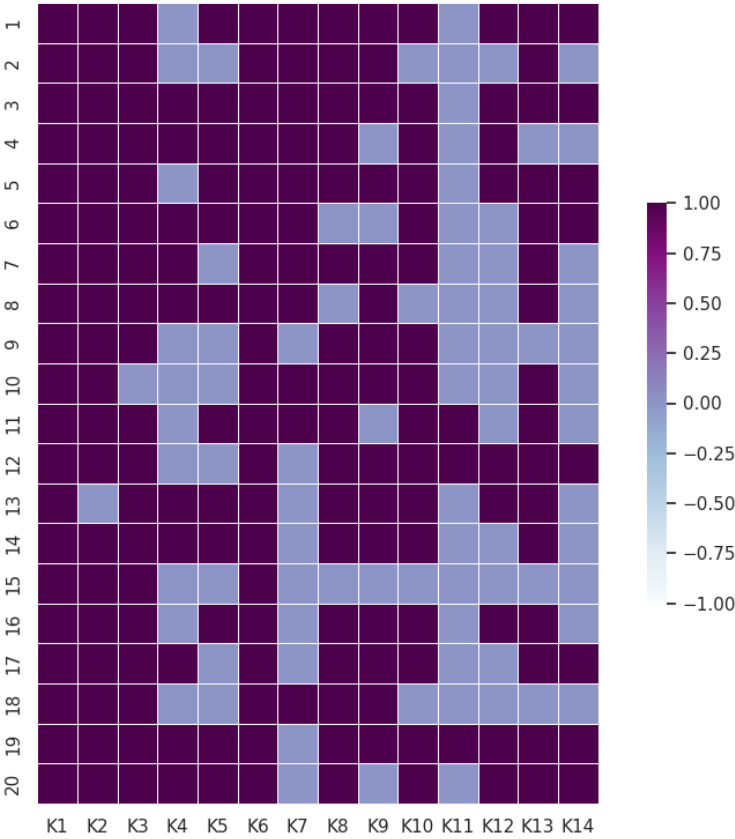


Fig. 2. Heat map of estimates of the developed system

Рис. 2. Тепловая карта оценок разработанной системы

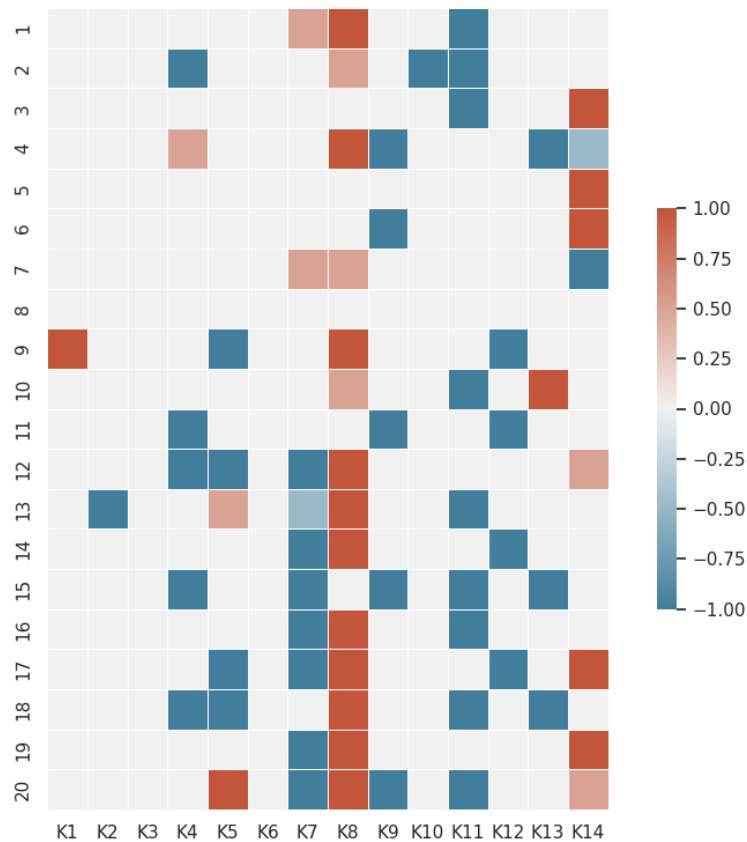


Fig. 3. Estimation difference heat map

Рис. 3. Тепловая карта разницы оценок

Различия между оценками системы и эксперта также представлены в виде соответствующей тепловой карты 3. На ней красными тонами показаны ложно-положительные ошибки системы (система поставила балл за критерий, а эксперт — нет), а синими тонами показаны ложно-негативные (система не поставила балл, а эксперт — поставил). Белые ячейки обозначают совпадение оценок. По данной карте видно, что присутствуют как хорошо оцениваемые критерии: K1, K2, K3, K6 и K10, так и достаточно проблемные: K7, K8, K11, K14.

Был также проведён кластерный анализ текстов, на которых проводились эксперименты. Для представления четырнадцатимерных векторов оценок на двухмерном пространстве применялся алгоритм нелинейного снижения размерности UMAP. Поскольку алгоритм использует стартовые точки для выполнения работы, то построение графиков выполнялось неоднократно с различными значениями. Благодаря этому были получены представленные графики 4, 5, позволяющие отобразить расхождения оценки системы и экспертов на деловых письмах.

Как видно из обоих графиков, система автоматически разделила письма на два кластера. Первый кластер (письма 2, 8, 15, 18), несмотря на явные лингвистические различия в самих текстах, имеет ряд сходств, а именно одинаковые оценки по ряду критериев. Так, все четыре работы получили 1 экспертный балл за критерии K1–K4. Три работы из четырёх (2, 15, 18) получили 0 баллов за крите-

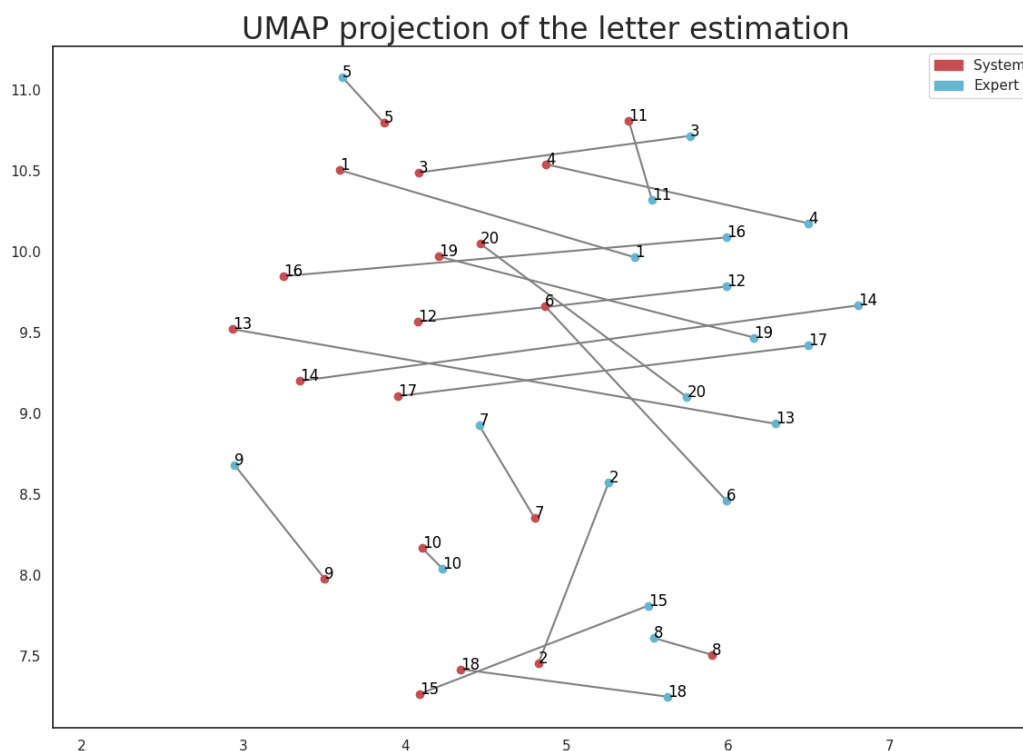


Fig. 4. Two-dimensional representation of score vectors using UMAP 1

Рис. 4. Двумерное представление векторов оценок при помощи UMAP 1

рии K4 и K14 при экспертной оценке в 1 балл. Все четыре письма не удовлетворяют критерию K8, но удовлетворяют критерию K9. В 8, 15 и 18 письмах отсутствует разделение на абзацы (критерий K0). В трех из четырех случаев (2, 15, 18) присутствует достаточное количество клишированных выражений (критерий K11). Во всех 4 письмах также отсутствует подпись (критерий K12), присутствуют логические средства связи (критерий K13), но они употреблены либо некорректно, либо в недостаточном количестве (критерий K14).

Во втором кластере (письма 1, 3, 4, 6, 11, 12, 13, 14, 16, 17, 19, 20) также прослеживается сходство. Во всех работах критерии K1–K3 и критерий K13 были оценены экспертом в 1 балл, в 10 случаях из 12 (кроме 1 и 16 писем) — так же был оценен критерий K4. В 10 случаях (кроме 3 и 11 писем) критерий K8 был обнулен экспертом, при этом в 9 из 10 таких писем по данному критерию был автоматически ошибочно выставлен 1 балл. Во всех письмах находилась информация, удовлетворяющая критериям K6 и K7, однако при этом в 7 из 12 писем по критерию K7 было автоматически выставлено 0 баллов. В 10 из 12 писем (кроме 6 и 20) присутствуют заключение и подпись (критерий K12), но отсутствует разнообразие корректно подобранных связующих элементов — в данном случае такое явление наблюдается во всех письмах кроме писем 4 и 12.

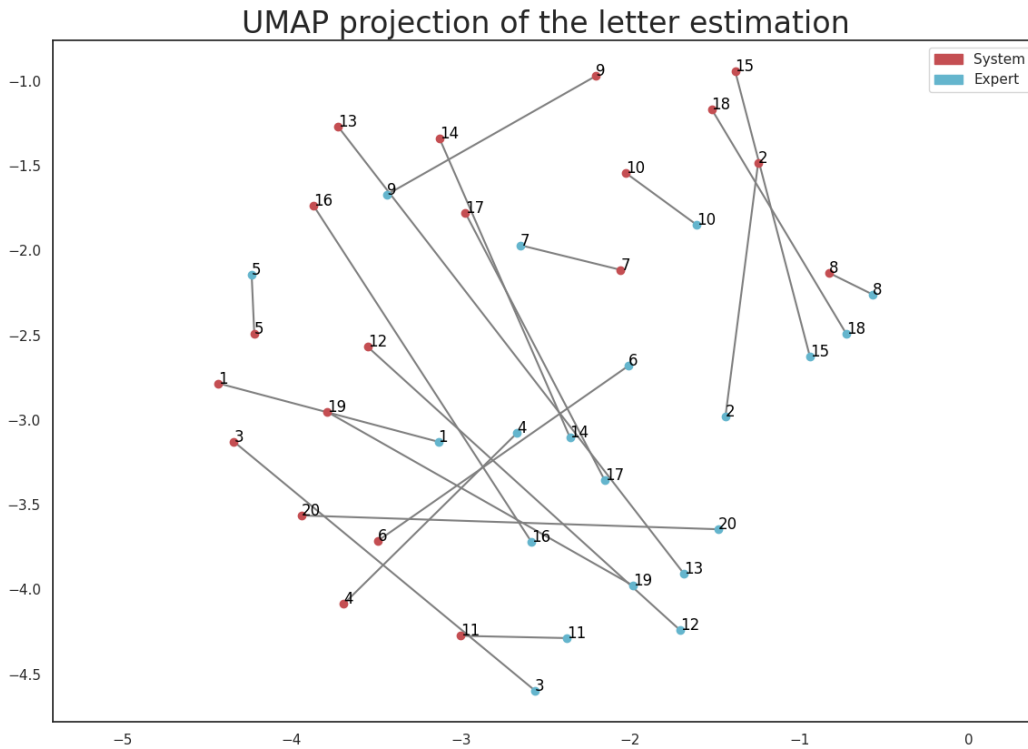


Fig. 5. Two-dimensional representation of score vectors using UMAP 2

Рис. 5. Двумерное представление векторов оценок при помощи UMAP 2

5. Обсуждение результатов

Как видно из полученных результатов, в части «Решение коммуникативной задачи» наибольшее количество проблем при автоматической оценке вызвали критерии K4, K5, K7 и K8. Это может быть связано с тем, что для решения данных задач учащиеся, как правило, использовали не ключевые слова и выражения, а языковые средства, которые не называли напрямую, но указывали на необходимые факты или явления, т. е. были логически связаны с искомыми фактами или явлениями. Так, в частности, при оценке соответствия критерию K4 (письмо содержит упоминание про навыки публичных выступлений автора) в письмах 2, 11, 12, 15 и 18 были использованы выражения и слова, которые отражают причину или следствие обладания навыками публичных выступлений (здесь и далее все ошибки сохранены): «The presentation will sound more dignified on my behalf» (письмо 11, следствие того, что у автора есть все необходимые для этого навыки), «I have excellent knowledge of these skills and can present our product» (письмо 18, причина и следствие + оценочное прилагательное excellent). В ряде случаев, как, например, в письме 11, решению задачи способствует исключительно контекст, когерентность и использование оценочной лексики: «I am well acquainted with the features of our application, because I was directly involved in its creation», которое позиционно находится сразу после упоминания навыков публичных выступлений коллеги и просьбы о своем назначении. Все вышеупомянутые случаи крайне трудно поддаются формализации, поскольку, во-первых, не содержат в себе каких-либо постоянных ключевых слов или выражений, которые можно было бы выделить в отдельную группу, а во-вторых, зависят от позиционирования и когерентности, которая является крайне субъективным и нечетким явлением по своей природе.

Похожая картина видна и при анализе результатов автоматической проверки критерия K5 (письмо содержит упоминание про навыки публичных выступлений коллеги). Успешности автоматической проверки в данном случае также препятствовало решение учащимися данной задачи с помощью логической связи, контекстуализации и когерентности: письмо 18, «he does not have the skills to speak in public» подразумевается, что он не сможет выступить должным образом; письмо 17: «if I make a presentation of the application, there are more chances to impress customers», следовательно, у коллеги шансов меньше.

Ошибки наблюдаются также и при автоматической оценке критерия K7 (письмо содержит просьбу о назначении автора). Как и при оценке критериев K4 и K5, возможной причиной этому явлению послужила невозможность оценить системой имплицитную логическую связь. Письмо 17: «Please, do not find me rude but if I make a presentation of the application, there are more chances to impress customers. I hope you will do a right choice», просьба о своем назначении имплицитно присутствует и выражается с помощью прилагательного *many* в сравнительной степени и выражения «make the right choice», где учащимся была допущена ошибка в употреблении коллокаций. В ряде случаев учащимися были использованы слова и выражения, которые могут быть отнесены к синонимам, антонимам, гипонимам и гиперонимам ключевых выражений или являться самими ключевыми выражениями, но не удовлетворять условиям иных критериев. Письмо 15: «I want to ask you assign me as the application presenter». Технически учащийся решает коммуникативную задачу, используя выражения, прямо называющие, а не указывающие на явление. Однако в данном случае при составлении этого предложения учащимся был допущен ряд других ошибок, что могло стать возможной причиной того, что система не смогла оценить соответствие данному критерию.

Наибольшее количество ошибок в блоке «Решение коммуникативной задачи» было выявлено при автоматической оценке критерия K8 (стиль письма соответствует заявленному формату задания): в 14 из 20 писем. Это может объясняться, во-первых, тем, что принадлежность текста к тому или иному функциональному стилю проявляется с помощью языковых единиц разных уровней языковой системы, из чего следует, что проверка такой принадлежности должна представлять собой сложную, разветвленную систему ряда параметров. Реализация такой системы на практике крайне затруднительна и, помимо этого, будет осложняться тем, что общепринятых утвержденных подробных списков и словарей таких единиц на настоящий момент не существует. Во-вторых, даже при условии использования учащимся стилистически корректных единиц, совершенные им/ей ошибки, как правило, нивелируют все корректно употребленные единицы. Так, автором письма 12 были корректно употреблено как и клишированное выражение «Dear Mrs Addams, I am writing to you about. . . », так и лексические средства, которые чаще употребляются именно в текстах официально-делового стиля, однако использования сокращений: «That's why», неуместной оценочной лексики: «bad public speaker, was a total failure» привело к обнулению экспертом баллов по данному критерию.

Наибольшие затруднения при автоматической проверке вызвал блок «Организация речи». Так, в данном блоке больше всего ошибок при автоматической проверке писем наблюдалось в критериях K9, K11, K12, K13 и K14. В частности, ошибки при проверке критериев K9 (присутствует обращение) и K12 (присутствует заключение или подпись), которые проверялись с помощью регулярных выражений, в первую очередь связаны с тем, что содержащиеся в письмах обращения и подписи могли либо не иметь отступа, либо содержать в себе стилистические и/или лексические ошибки. Так, например, в письме 6 формально присутствует обращение «Hello Mrs Addams», однако учащимся была допущена стилистическая и пунктуационная ошибка. В ряде случаев в качестве заключения учащиеся использовали гиперонимы вместо конкретного имени, что формально считается подписью, хоть и не способствует решению коммуникативной задачи: в письме 11 «Yours sincerely, An employee of your company».

Интересно, что анализ расхождений оценок экспертов и автоматической системы сработал и в обратную сторону: выявил недостаточную объективность человека. При проверке критерия K11 причиной «ошибочной оценки» стал исключительно человеческий фактор: во всех 9 письмах, в которых за данный критерий было автоматически выставлено 0 баллов, действительно наблюдается недостаток клишированных выражений, однако эксперты излишне лояльно отнеслись к работам.

Многочисленные ошибки при автоматической оценке критериев K13 и K14 (присутствуют связующие элементы и средства логической связи и наблюдается разнообразие корректно подобранных связующих элементов и средств логической связи) вполне логичны и могут объясняться самой природой механизмов когезии. Унификация требований к таким механизмам крайне сложна в силу объективных причин, поскольку, во-первых, связность текста обеспечивается за счет языковых средств разного уровня, а не только лексического, а во-вторых, сама корректность употребления таких средств будет всегда зависеть от контекста и от их позиции в тексте, а также от стилистической принадлежности таких средств к тому или иному функциональному стилю. В частности, в письме 19 автором было употреблено «And early I was a public speaker a few times so I understand that I must do». Само употребление «and» в данном случае не является ошибкой и вполне допустимо в официально-деловой переписке, однако в данном случае его позиция в начале предложения некорректна.

Однако в целом можно с уверенностью утверждать, что в подавляющем большинстве случаев система справилась с оценкой письма по выделенным критериям. Выявленные расхождения в оценке системой и в экспертной оценке объясняются комплексной, многоаспектной природой проверяемых явлений и сложностью их унификации, а также человеческим фактором, т. е. более лояльным отношением эксперта и более строгой и объективной автоматической оценкой.

Заключение

Модель текста, описанная в статье, включает набор числовых параметров, формализующих детали оценки письма на заданную тему. Основная часть этих параметров легко адаптируется к проверке связных текстов в других предметных областях. Проведенные эксперименты показали, что в большинстве случаев значимых различий между оценкой эксперта и оценкой системы выявлено не было. Это свидетельствует о том, что автоматическая проверка письменного задания является весьма эффективным инструментом оценивания, обладающим огромным и неоспоримым потенциалом в обучении иностранным языкам. Она также более объективна: система строго придерживается заранее заданных параметров, тем самым полностью исключая человеческий фактор. Те критерии, в которых наблюдались расхождения в оценках эксперта и системы сложны для алгоритмизации. Работа над формализацией таких критериев представляет собой перспективу для дальнейшего исследования как с точки зрения анализа естественного языка, так и с точки зрения методики преподавания иностранных языков и лингвистики.

Интересно, что данное исследование также поставило ряд актуальных вопросов перед преподавательским сообществом и выявило необходимость в дальнейшем пересмотре принятой системы оценивания письменных заданий экспертом, в частности, выделения более четких и строгих критериев оценки письма, а также объективности самого оценивания. В перспективе данного исследования также лежит анализ ошибок, допущенных экспертами, а также исследование возможных причин их возникновения. Это связано с особенностями методов обработки естественного языка, основанных на правилах. Результаты работы этих методов могут быть интерпретированы и проанализированы с разных точек зрения. Перспективой развития такого анализа является построение автоматических систем обучения с обратной связью не только для учеников, но и преподавателей.

References

- [1] A. Al-Bargi, “Exploring Online Writing Assessment Amid Covid-19: Challenges and Opportunities from Teachers’ Perspectives”, *Arab World English Journal*, pp. 3–21, 2022.
- [2] N. P. Soboleva and M. A. Nilova, “Obuchenie pis’mu studentov gumanitarnykh special’nostej s ispol’zovaniem sovremennykh obrazovatel’nykh tekhnologij”, *Kazanskij vestnik molodyh uchyonyh*, vol. 2, no. 5 (8), pp. 57–59, 2018.
- [3] M. Fareed, A. Ashraf, and M. Bilal, “ESL learners’ writing skills: Problems, factors and suggestions”, *Journal of education and social sciences*, vol. 4, no. 2, pp. 81–92, 2016.
- [4] K. N. A. Al-Mwzaiji and A. A. F. Alzubi, “Online self-evaluation: the EFL writing skills in focus”, *Asian-Pacific Journal of Second and Foreign Language Education*, vol. 7, no. 1, pp. 1–16, 2022.
- [5] M. A. Hussein, H. Hassan, and M. Nassef, “Automated language essay scoring systems: A literature review”, *PeerJ Computer Science*, vol. 5, e208, 2019.
- [6] H. John Bernardin, S. Thomason, M. Ronald Buckley, and J. S. Kane, “Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability”, *Human Resource Management*, vol. 55, no. 2, pp. 321–340, 2016.
- [7] Z. Ke and V. Ng, “Automated Essay Scoring: A Survey of the State of the Art.”, in *IJCAI*, vol. 19, 2019, pp. 6300–6308.
- [8] M. Uto, “A review of deep-neural automated essay scoring models”, *Behaviormetrika*, vol. 48, no. 2, pp. 459–484, 2021.
- [9] S. Vajjala, “Automated assessment of non-native learner essays: Investigating the role of linguistic features”, *International Journal of Artificial Intelligence in Education*, vol. 28, no. 1, pp. 79–105, 2018.
- [10] K. Taghipour and H. T. Ng, “A neural approach to automated essay scoring”, in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1882–1891.
- [11] L. Xia, J. Liu, and Z. Zhang, “Automatic essay scoring model based on two-layer bi-directional long-short term memory network”, in *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, 2019, pp. 133–137.
- [12] M. Uto and M. Okano, “Robust neural automated essay scoring using item response theory”, in *International Conference on Artificial Intelligence in Education*, Springer, 2020, pp. 549–561.
- [13] Y. Tay, M. Phan, L. A. Tuan, and S. C. Hui, “Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring”, in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018, pp. 5948–5955.
- [14] Y. Farag, H. Yannakoudakis, and T. Briscoe, “Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input”, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics (ACL), 2018, pp. 263–271.
- [15] Y. Yang and J. Zhong, “Automated essay scoring via example-based learning”, in *International Conference on Web Engineering*, Springer, 2021, pp. 201–208.
- [16] E. Mayfield and A. W. Black, “Should you fine-tune BERT for automated essay scoring?”, in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 151–162.
- [17] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, “Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking”, in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1560–1569.

- [18] M. Uto, Y. Xie, and M. Ueno, “Neural automated essay scoring incorporating handcrafted features”, in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6077–6088.
- [19] I. Aomi, E. Tsutsumi, M. Uto, and M. Ueno, “Integration of automated essay scoring models using item response theory”, in *International Conference on Artificial Intelligence in Education*, Springer, 2021, pp. 54–59.
- [20] W. Zhu and Y. Sun, “Automated Essay Scoring System using Multi-Model Machine Learning”, in *CS & IT Conference Proceedings*, CS & IT Conference Proceedings, vol. 10, 2020, pp. 109–117.
- [21] S. M. Darwish and S. K. Mohamed, “Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis”, in *International Conference on Advanced Machine Learning Technologies and Applications*, Springer, 2019, pp. 566–575.
- [22] Y. Salim, V. Stevanus, E. Barlian, A. C. Sari, and D. Suhartono, “Automated English Digital Essay Grader Using Machine Learning”, in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, IEEE, 2019, pp. 1–6.
- [23] R. Wilkens, D. Seibert, X. Wang, and T. François, “MWE for Essay Scoring English as a Foreign Language”, in *2nd Workshop on Tools and Resources for READING Difficulties (READI)*, 2022, pp. 62–69.
- [24] D. Ramesh and S. K. Sanampudi, “An automated essay scoring systems: a systematic literature review”, *Artificial Intelligence Review*, pp. 1–33, 2021.