Автоматизированный анализ коротких ответов с использованием векторизации текста

Ч. Б. Миннегалиева

Институт вычислительной математики и информационных технологий, Казанский федеральный университет

mchulpan@gmail.com

С. Ф. Зиятдинова

Казанский национальный исследовательский технологический университет

ibr.sumbel@yandex.ru

Аннотация. **Автоматизированное оценивание коротких ответов, полученных в ходе тестирования знаний, остаётся актуальной проблемой. В данной работе приведены результаты анализа ответов на открытые вопросы. Первоначальные оценки получены при помощи косинусного сходства и языковых моделей. Предлагается алгоритм уточнения оценок с использованием бинарной классификации и инструментов объяснимого искусственного интеллекта. Проверены возможности современных чат-ботов на основе больших языковых моделей для анализа ответов студентов, сформулированных в свободной форме.**

Ключевые слова: языковая модель, BERT, косинусное сходство, автоматизированное оценивание коротких ответов, объяснимый искусственный интеллект

I. Введение

При тестировании знаний вопросы, которые требуют от обучающегося произвольного, свободного ответа (открытые вопросы), позволяют повысить объективность оценивания. Они помогают получить обратную связь о понимании студентами пройденного материала. Такие вопросы рекомендуется использовать в тестах, заданиях, проектах, чтобы преподаватель мог проанализировать и улучшить стратегии преподавания [1]. Открытые вопросы являются эффективным инструментом оценки для измерения навыков критического мышления учащихся разных возрастов, уровня усвоения учебного материала, характеристики когнитивного развития учащихся [2]. Но несмотря на то, что использование открытых вопросов потенциально приносит более богатую информацию, оно создает проблемы с точки зрения извлечения полезной информации, при этом значительно увеличивая время анализа. Так как в большинстве случаев проверку проводит преподаватель [3]. Поэтому автоматическая оценка коротких ответов является важной задачей. Для оценки ответов студентов можно использовать косинусную меру сходства, модели классификации на основе нейронной сети [4]. В следующей работе представлено исследование проблемы автоматической классификации коротких связных текстов (эссе) на английском языке по уровням международной шкалы CEFR. Определение уровня текста на естественном языке является важным компонентом оценки знаний обучающегося. Для решения этой задачи рассматриваются векторные модели текста на основе стилометрических признаков. Полученные векторы классифицируются стандартными классификаторами машинного обучения [5]. Часто встречаются открытые вопросы, требующие классификации по нескольким меткам, т. е. которым присвоено несколько кодов. В статье основное внимание уделяется классификации по нескольким меткам текстовых ответов на открытые вопросы. Авторы оценивают производительность модели BERT для немецкого языка по сравнению с традиционными алгоритмами [6]. В следующей работе представлена новая гибридная модель на основе BERT. Используется настраиваемый механизм, адаптированный с BERT, который позволяет надежно идентифицировать семантические зависимости между словами в предложении и, следовательно, способствует эффективности и надежности системы подсчета баллов [7]. Изучена эффективность различных моделей на основе BERT для извлечения данных из арабских текстов, полученных во время опросов в образовательных учреждениях. Разработана система, которая автоматизирует процесс анализа настроений на основе аспектов [8]. Часто семантическое сходство между многомерными объектами находится путем применения косинусного сходства, но авторы предупреждают о возможных ошибках [9]. Бинарная классификация определяет принадлежность элементов к одной из двух возможных категорий. Существуют различные метрики для оценки таких систем классификации. В статье авторы представили новый математический подход для точной оценки метрик бинарной классификации [10]. Хотя в последнее время произошел взрыв исследований по объяснимому искусственному интеллекту на моделях, которые работают с изображениями и табличными данными, текстовые наборы данных представляют новые проблемы. В последнее время были разработаны методы для решения этих проблем и предоставления объяснений по моделям обработки естественного языка [11]. Метод объяснимого искусственного интеллекта, такой как LIME, демонстрирует работу модели. Использование LIME повышает точность анализа настроений; объяснимость помогает лучше понять модель и ее можно применять для улучшения производительности модели, что, в свою очередь, повышает надежность и достоверность систем искусственного интеллекта [12].

Таким образом, подходы к автоматизированному анализу ответов студентов, сформулированных в свободной форме, разносторонне изучаются

специалистами. Ранее нами были представлены результаты оценивания ответов на открытые вопросы с использованием языковых моделей [13]. В этой работе будут описаны возможности уточнения полученных оценок.

II. Данные

Мы рассматриваем базу, состоящую из 1580 ответов, полученных от студентов. Ответы были даны во время занятий в аудиториях, студенты не имели возможности пользоваться сетью Интернет и справочной литературой. Оценивались ответы на 42 вопроса по темам, относящимся к компьютерной графике. Ранее мы проводили эксперимент, в ходе которого перед тестированием студентам была предоставлена возможность повторить теоретический материал. В этом случае увеличиваются как количество правильных ответов, так и доля корректно оцененных ответов. Так как обучающиеся используют преимущественно слова, которые были в тексте для повторения [14].

После анализа всех ответов мы пришли к выводу, что формулировку части вопросов необходимо скорректировать. Некоторые студенты недостаточно четко поняли задание. При автоматизированной проверке у обучающихся и у преподавателя нет возможности задать уточняющий вопрос, и данное обстоятельство необходимо учитывать при составлении вопросов. Например, вопрос «Приведите примеры градационных преобразований изображений» может быть уточнен как «Приведите все примеры градационных преобразований изображений, изученных во время прохождения курса» или «Приведите любые два примера градационных преобразований изображений». Были составлены рекомендации для преподавателей с учетом всех вопросов, возникших у обучающихся.

Всем ответам были даны экспертные оценки. Экспертами были преподаватели и магистранты, обучающиеся по направлению «Информационные системы и технологии». Так как задавались вопросы, требующие определения одного понятия или объяснения одного процесса, они оценивались как верные (0 баллов) и неверные (1 балл). В некоторых случаях, например, когда требовалось перечисление понятий, оценка могла быть выражена дробным числом.

Для ясности приведем примеры вопросов и некоторых ответов, полученных от студентов. На вопрос «Что определяет цветовое пространство?» были получены ответы: «определяет реально достижимые цвета в рамках выбранной цветовой модели» (верный); «модель представления цвета» (неверный); «координаты цвета» (неверный); «реально получаемые цвета в рамках выбранной модели» (верный). На задание «Unity. Напишите назначение объекта Terrain.» были получены ответы «Выбрав GameObjec – Create Others – Terrain из меню» (неверный); «Функция ландшафта – это инструмент для создания реалистичной и захватывающей среды для игр. Она позволяет создавать большие и сплошные ландшафты с горами, долинами и др.» (верный); «не в курсе» (неверный); «создание ландшафта» (верный); «представляет собой компонент для создания и редактирования больших, открытых ландшафтов. Основное назначение: моделирование местности с холмами, долинами и растительностью» (верный); «для рисования» (неверный); «создание ландшафта» (верный). Таким образом, ответы студентов разные, могли состоять из одного или нескольких слов, были ответы длиной до трех предложений.

III. Методы

Все ответы были проверены при помощи косинусного сходства [15], [16]. Если в тест добавляется новый вопрос, ответ студента сравнивается с шаблонным ответом, сформулированным преподавателем. То есть мы находим значение косинуса между вектором ответа студента и вектором корректного ответа. Когда в базе достаточное количество ответов студентов, вычисляется значение косинуса между вектором ответа студента и вектором, являющимся средним векторов верных ответов, полученных от студентов. Если абсолютное значение разности между вычисленным значением косинуса и оценкой эксперта меньше, чем 0.5, мы считаем, что оценка, полученная при помощи языковой модели, и оценка эксперта совпадают. Ранее мы пришли к выводу, что если в ответе больше специальных терминов, то можно применить модель word2vec, обученную на специальных текстах. Для вопросов и ответов, в которых больше общеупотребительных слов, лучше применить большие языковые модели.

Была исследована модель бинарной классификации для вопросов, ответы на которые были разделены на 2 класса – верные и неверные. В случаях, когда количество ответов было недостаточным, была проведена аугментация данных с использованием словаря синонимов, чат-ботов на основе больших языковых моделей. Аугментированные данные (сгенерированные ответы) были проверены и скорректированы экспертом (преподавателем). Мы также изучили возможности инструмента LIME для интерпретации результатов классификации. В ходе работе были использованы язык программирования Python с необходимыми библиотеками, возможности Google Colab.

IV. Результаты

Доля корректно оцененных ответов с использованием модели на основе BERT для разных вопросов меняется от 71,4 % до 100 % (рис. 1). На рисунке по горизонтальной оси отмечены номера вопросов. Рис. 2 показывает эти же данные, сгруппированные по интервалам.

В табл. 1 приведены доли ошибочно оцененных ответов. В данном случае вычислялось значение косинуса между вектором ответа студента и вектором, являющимся средним векторов верных ответов, полученных от студентов. Как видно из таблицы, оценка эксперта и оценка, полученная при помощи косинусного сходства, полностью совпали только для двух вопросов (вопросы 26 и 36). Поэтому необходимо уточнение и в качестве дополнительного этапа оценивания предлагаем провести бинарную классификацию. Мы рассматривали отдельно каждый вопрос. Как было сказано выше, точность результатов оценивания разная для разных вопросов.

Предлагаются следующие шаги алгоритма автоматизированного оценивания ответов. 1). Вычисление косинусного сходства. 2). Результат предсказания с использованием бинарной классификации. 3). Визуальная интерпретация классификации для принятия окончательного решения преподавателем.

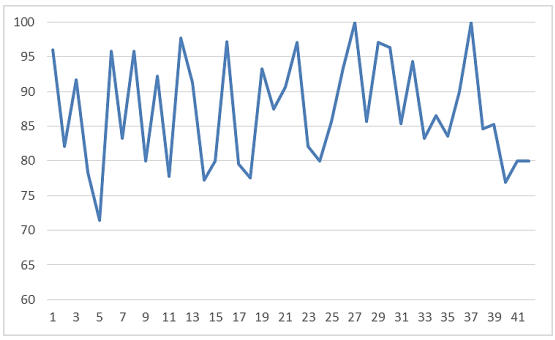

Рис. 1. Доля корректно оцененных ответов (в процентах)

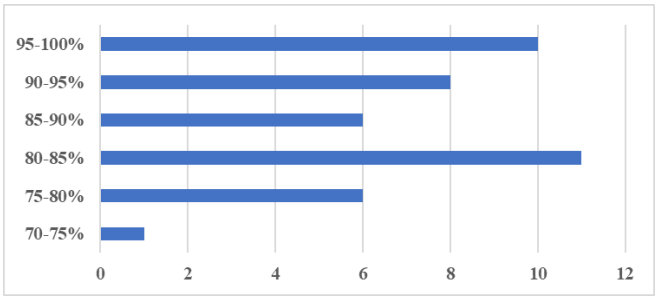

Рис. 2. Корректно оцененные ответы

ТАБЛИЦА I.

Номер вопроса	**Доля ошибочно оцененных ответов (в %)**	**Номер вопроса**	**Доля ошибочно оцененных ответов (в %)**
1	4	**22**	17.9
2	17.9	**23**	20
3	8.3	**24**	14.3
4	21.7	**25**	6.6
5	28.6	**26**	0
6	4.2	**27**	14.3
7	16.7	**28**	2.9
8	4.2	**29**	3.6
9	20	**30**	18.2
10	7.8	**31**	5.6
11	22.2	**32**	16.7
12	2.3	**33**	13.5
13	8.7	**34**	16.4
14	22.8	**35**	10
15	2.8	**36**	0
16	20.4	**37**	15.4
17	22.4	**38**	14.7
18	6.7	**39**	23.1
19	12.5	**40**	20
20	9.3	**41**	20
21	2.9	**42**	3.6

Рассмотрим пример 1. Вопрос: «Какие преобразования называются градационными преобразованиями изображений?». Ответ студента: «Это преобразования, при которых для вычисления интенсивности точки используется интенсивность самой точки» (оценка эксперта – 1). Косинусное сходство равно 0.877. Результат предсказания с использованием бинарной классификации – 1 (ответ верный).

При помощи инструмента LIME (Local Interpretationable Model-Agnostic Explains) можно понять, какие слова повлияли на решение модели. Представим слова с их коэффициентами, показывающими влияние данного слова на решение (метод опорных векторов в текстовой классификации с использованием TfidfVectorizer): «точки» – 0.4884, «самой» – 0.1056, «преобразования» – -0.0929, «интенсивности» – 0.0787, «интенсивность» – 0.0682, «используется» – 0.0598. Интерпретация классификации представлена на рис. 3. Такие визуализации помогут преподавателю корректно оценить ответ.

Пример 2. Вопрос: «Какие преобразования называются градационными преобразованиями изображений?». Ответ студента: «те, где используется только интенсивность соответствующей исходной точки» (оценка эксперта – 1). Косинусное сходство равно 0.687. Результат предсказания с использованием бинарной классификации – 1 (ответ верный). Слова с их коэффициентами, показывающими влияние данного слова на решение: «точки» – 0.1528, «только» – 0.1089, «исходной» – 0.072, «интенсивность» – 0.0403, «используется» – 0.031, «соответствующей» – - 0.0104. Интерпретация классификации представлена на рис. 4.

Инструменты объяснимого искусственного интеллекта находятся в развитии. Поэтому данная интерпретация может выступать только как дополнительный этап оценивания.

Также дополнительным этапом оценивания возможно использование чат-ботов на основе LLM. В табл. 2 представлены результаты оценивания некоторых ответов, полученные при помощи ChatGPT. При обращении к данному чат-боту возможно составить промпт так, чтобы получить также обоснования оценок. Например, при рассмотрении вопроса «Что определяет цветовое пространство?» для ответа «Реально достижимые цвета» объяснение было таким: «Этот ответ ближе к правильному, поскольку цветовое пространство действительно определяется реально достижимыми цветами. Однако, это определение не совсем полное, поскольку не учитывает другие факторы, такие как модель цвета и доступные цвета.»

Оценки, определенные ChatGPT не всегда совпадают с оценками экспертов, их можно уточнить при помощи дополнительных промптов. Мы использовали разные промпты и сервисы (ChatGPT, DeepSeek). Результаты различаются для разных вопросов. Ни по одному вопросу мы не получили 100 процентного совпадения с оценками эксперта. Конечно, чат-боты могут заметно облегчить проверку студенческих ответов. Но преподаватель должен будет эти результаты перепроверить.

Возможно использование чат-ботов на базе больших языковых моделей для составления вопросов, для генерирования ответов. Сгенерированные ответы могут помочь при недостаточном количестве данных для классификации. Но, как показала практика, и в этом случае преподаватель должен перепроверить результаты.

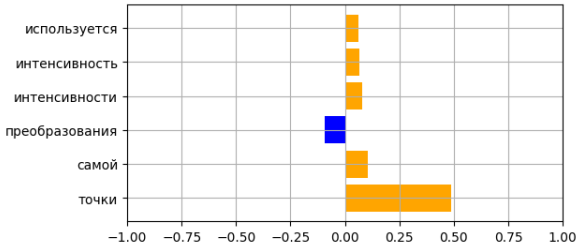

Рис. 3. Визуальная интерпретация для примера 1

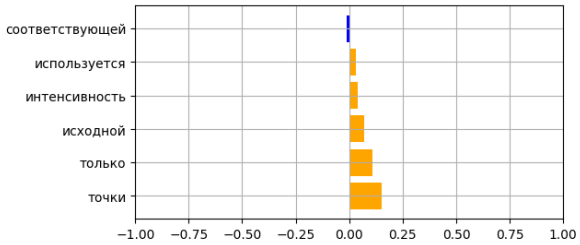

Рис. 4. Визуальная интерпретация для примера 2

ТАБЛИЦА II.

Ответ студента	**Оценка**
Характеристика цвета: яркость, насыщенность, контраст и т.д.	0.2
Реально достижимые цвета.	0.8
Доступные цвета в выбранной модели.	0.6
Свет.	0.3

Отметим, что все примеры ответов, рассмотренные выше, приведены в том виде, в котором получены от студентов. Ранее мы получили, что влияние исправления орфографических, пунктуационных и других ошибок на результат оценивания незначительно.

V. Заключение

Автоматизированное оценивание коротких ответов, сформулированных в свободной форме, остаётся актуальной проблемой. С одной стороны, языковые модели, современные инструменты обработки естественного языка предоставляют большие возможности для оптимизации работы преподавателей. С другой стороны, как показал анализ, ошибочно оцененные ответы остаются при использовании любого подхода. Поэтому необходимо методы применять комплексно.

Точность оценивания ответов на открытые вопросы зависит от предметной области, от объема заполненной базы данных. Также необходимо обратить внимание на корректную формулировку вопроса. Вопросы, которые могут быть поняты студентами неоднозначно, необходимо переформулировать.

В данной работе представлен алгоритм автоматизированной проверки коротких ответов, состоящих из 1-3 предложений. Шагами алгоритма были определены вычисление косинусного сходства, использование моделей бинарной классификации. Для определения косинусного сходства сначала находятся векторные представления ответа студента и корректного (шаблонного) ответа. Вычисляется значение косинуса между вектором ответа студента и вектором корректного ответа или вектором, являющимся средним векторов всех верных ответов в базе. До наполнения базы ответами вычисленное значение косинуса округляется до целого значения и предъявляется преподавателю как рекомендуемая оценка. Преподаватель перепроверит эти оценки. После добавления в базу достаточного количества ответов, преподаватель увидит оценку, полученную при помощи косинусного сходства, результаты предсказания и интерпретацию по бинарной классификации.

Работа по изучению подходов к автоматизированному оцениванию ответов будут нами продолжена. С увеличением количества ответов в базе точность оценивания повысится. Проблемой является то, что невозможно получить большое количество ответов, например, выполнив, парсинг данных с сайтов, с помощью которых проводится тестирование знаний. Так как когда проводится онлайн тестирование, не всегда применяется прокторинг и часто студенты могут пользоваться сетью Интернет, справочными системами, дополнительной литературой. Поэтому в данной работе были проанализированы ответы, полученные в аудитории.

Список литературы

[1] H. Husain, B. Baisb, A.Hussainb, S. A. Samadb. How to Construct Open Ended Questions // Procedia - Social and Behavioral Sciences, vol. 60, pp. 456-462, 2012.

[2] Sarwanto L.E.W. Fajari Chumdari. Open-Ended Questions to Assess CriticalThinking Skills in Indonesian Elementary School // International Journal of Instruction, vol. 14(1), pp. 615-630, 2021.

[3] V. Baburajan, J.d.A.e Silva, F.C. Pereira. Open-Ended Versus Closed-Ended Responses: A Comparison Study Using Topic Modeling and Factor Analysis // IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 4, pp. 2123-2132, April 2021.

[4] M. Putnikovic, J. Jovanovic. Embeddings for Automatic Short Answer Grading: A Scoping Review // IEEE Transactions on Learning Technologies, vol. 16, no. 2, pp. 219-231, 1 April 2023.

[5] N.S. Lagutina, K.V. Lagutina, A.M. Brederman, N.N. Kasatkina. Text Classification by CEFR Levels Using Machine Learning Methods and the BERT Language Model // Automatic Control and Computer Sciences, vol. 58, pp. 869-878, 2024.

[6] M. Schonlau, J. Weiß, J. Marquardt. Multi-label classification of open-ended questions with BERT // 2023 Big Data Meets Survey Science (BigSurv), Quito, Ecuador, 2023, pp. 1-8.

[7] M. Kaya, I. Cicekli. A Hybrid Approach for Automated Short Answer Grading // IEEE Access, vol. 12, pp. 96332-96341, 2024.

[8] K.A. Alshaikh, O.A. Almatrafi, Y.B. Abushark. BERT-Based Model for Aspect-Based Sentiment Analysis for Analyzing Arabic Open-Ended Survey Responses: A Case Study // IEEE Access, vol. 12, pp. 2288-2302, 2024.

[9] H. Steck, C. Ekanadham, N. Kallus. Is Cosine-Similarity of Embeddings Really About Similarity? // Companion Proceedings of the ACM Web Conference 2024 (WWW '24). Association for Computing Machinery, New York, NY, USA, pp. 887-890, 2024.

[10] M. Shirdel, M. Di Mauro, A. Liotta. Worthiness Benchmark: A novel concept for analyzing binary classification evaluation metrics // Information Sciences, vol. 678, 120882, September 2024.

[11] J.E. Zini, M. Awad. On the Explainability of Natural Language Processing Deep Models // ACM Computing Surveys, vol. 55, pp. 1-31, 2022.

[12] V. Bidve, P.M. Shafi, P. Sarasu, A.A. Pavate, A. Shaikh, S. Borde, V.B. Pratap Singh, R. Raut. Use of explainable AI to interpret the results of NLP models for sentimental analysis // Indonesian Journal of Electrical Engineering and Computer Science. vol. 35, no. 1, pp. 511-519, July 2024.

[13] C. Minnegalieva. Using Text Vectorization to Analyze Students' Answers Formulated in Different Languages // 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, pp. 1262-1266, 2024.

[14] Миннегалиева Ч.Б., Кашапов И.И., Морозова О.Д. Автоматизированное оценивание коротких ответов обучающихся с использованием языковых моделей // Электронные библиотеки. 2024. Т. 27, № 3. С. 278-293.

[15] Миннегалиева Ч.Б., Сабитова Г.А., Гаялиев А.М. Метод предварительной оценки ответов обучающихся на основе векторной модели документов // Электронные библиотеки. 2023. Т. 26, № 3. С. 324-339.

[16] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) Hong Kong, China Association for Computational Linguistics, 2019, 3982–3992. DOI: 10.48550/arXiv.1908.10084