НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2024 Том 24 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2024 Vol. 24 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-6-999-1006
УДК 81'33

Сравнительный анализ сгенерированных и оригинальных аннотаций научных статей по филологической тематике

Мария Владимировна Хохлова[1]✉, Михаил Витальевич Корышев[2]

[1,2] Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация

[1] m.khokhlova@spbu.ru✉, https://orcid.org/0000-0001-9085-0284
[2] m.koryshev@spbu.ru, https://orcid.org/0000-0001-8946-4431

Аннотация

Введение. Появление систем генеративного искусственного интеллекта оказало значительное влияние на задачи, имеющие отношение к обработке естественного языка: машинный перевод, сентимент-анализ, генерация и суммаризация текстов и т. п. Цель работы заключалась в определении особенностей автоматически сгенерированных научных текстов по сравнению с текстами, созданными авторами, а также в оценке возможностей разных методов применительно к задаче их классификации. **Метод.** Выполнен анализ аннотаций двух типов: собранные из научных журналов по компьютерной лингвистике и по германистике, сгенерированные по заголовкам соответствующих научных статей при помощи Generative Pre-trained Transformer (ChatGPT-4o mini). Общий объем данных составил 60 единиц. Выбор тематики работ обусловлен тем, что тексты относятся к одной предметной области, но отличаются по своей структуре. Первая группа, в которую собраны оригинальные тексты по компьютерной лингвистике, схожа с аннотациями научных работ по информационным технологиям, и содержит большое количество англоязычной терминологии. Вторая группа содержит тексты по германистике и носит более описательно-нарративный характер. Проведен анализ отличий аннотаций двух типов, выполнена их классификация по двум типам с привлечением экспертов, трех систем-детекторов для определения участия искусственного интеллекта при создании текстов (Smodin, ZeroGPT и GPTZero), а также самой системой ChatGPT-4o mini. **Основные результаты.** Проведенный анализ показал, что сгенерированные тексты отличаются четкой формальной структурой и соблюдением правил построения научных текстов в соответствии с IMRAD (наличием введения, методов, результатов и заключения). Содержательно они носят поверхностный характер, в них не всегда соблюдается научный стиль, присутствуют повторы конструкций и перефразирование названий статей, что не встречается в аннотациях, написанных авторами без привлечения искусственного интеллекта. Автоматически сгенерированные аннотации нуждаются не только в дальнейшей редакторской правке (поскольку в ряде случаев нарушены лексическая и синтаксическая сочетаемость, присутствует неоднозначность), но и в проверке упоминаемых фактов и терминов. Среди систем-детекторов наиболее высокие показатели по метрикам precision, accuracy и F1-score достигаются системой Smodin, в то время как по критерию Recall лучшие результаты демонстрирует система ZeroGPT. Наиболее низкие результаты при оценке аннотаций при сравнении с другими инструментами были достигнуты системой ChatGPT-4o mini. Классификация с привлечением экспертов показала наиболее высокие результаты в случае аннотаций по германистике. **Обсуждение.** Полученные результаты могут быть полезны исследователям при работе с научными текстами по лингвистике, а также для дальнейшего дообучения нейросетевых моделей.

Ключевые слова

ChatGPT, генерация текстов, искусственный интеллект, аннотации, научные статьи

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 24-28-00937, https://rscf.ru/project/24-28-00937/.

Ссылка для цитирования: Хохлова М.В., Корышев М.В. Сравнительный анализ сгенерированных и оригинальных аннотаций научных статей по филологической тематике // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 999–1006. doi: 10.17586/2226-1494-2024-24-6-999-1006

Comparative analysis of AI-generated and original abstracts of academic articles on philology

Maria V. Khokhlova[1]✉, Mikhail V. Koryshev[2]

[1,2] St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation

[1] m.khokhlova@spbu.ru✉, https://orcid.org/0000-0001-9085-0284

[2] m.koryshev@spbu.ru, https://orcid.org/0000-0001-8946-4431

Abstract

Generative artificial intelligence systems have a significant impact on tasks related to natural language processing: machine translation, sentiment analysis, text generation, and summarisation, etc. The aim of the presented work was to determine the features of automatically generated academic texts in comparison with texts created by authors, and to evaluate the capabilities of different methods in relation to the task of their classification. The paper analyses two types of abstracts: collected from academic journals on computational linguistics and Germanic studies and generated from the titles of the corresponding articles using ChatGPT-4o mini. The total amount of data was 60 items. The choice of article topics is due to the fact that the texts belong to the same subject area but differ in their structure. The first group which contains original texts on computational linguistics, is similar to the abstracts of academic articles on computer science, and contains a large amount of English terminology. The second group contains texts on Germanic studies and is more descriptive-narrative in their nature. We analyzed the differences between the two types of abstracts and classified them into two categories with the help of experts, three detector systems to determine the involvement of artificial intelligence in the creation of texts (Smodin, ZeroGPT and GPTZero), as well as the ChatGPT system itself. The analysis showed that the generated texts are characterized by a clear formal structure and adherence to the rules of academic text construction in accordance with IMRAD (Introduction, Methods, Results and Discussion). They are superficial in content and they do not always follow the scientific style; there are repetitions of constructions and paraphrasing of article titles, which is not found in the abstracts written by the authors without artificial intelligence. Automatically generated abstracts need not only further editing (because in some cases lexical and syntactic coherence is broken and ambiguity is present), but also verification of the facts and terms mentioned. Among the detector systems, the highest scores in Precision, Accuracy and F1-score are achieved by Smodin tools, while the best results in Recall are achieved by ZeroGPT. The lowest results in abstract evaluation when compared with other tools were achieved by the ChatGPT system itself. Expert-assisted classification showed the highest results in the case of Germanic abstracts. The results may be useful for researchers when working with academic texts on linguistics as well as for further fine-tuning of neural network models.

Keywords

ChatGPT, text generation, artificial intelligence, abstracts, academic articles

Acknowledgements

The study was supported by the Russian Science Foundation, Project No. 24-28-00937, https://rscf.ru/en/project/24-28-00937/.

For citation: Khokhlova M.V., Koryshev M.V. Comparative analysis of AI-generated and original abstracts of academic articles on philology. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 999–1006 (in Russian). doi: 10.17586/2226-1494-2024-24-6-999-1006

Введение

Технологии искусственного интеллекта (ИИ) находят все большее применение у широкой аудитории, заменяя обращение к специализированным системам. Последнее десятилетие ознаменовалось бурным ростом информационных технологий, которые оказали влияние на область, посвященную автоматической обработке языка. Достижения в области разработки моделей глубокого обучения и больших языковых моделей привели к повышению качества результатов при решении таких прикладных задач компьютерной лингвистики как машинный перевод, генерация текстов, сентимент-анализ, извлечение информации и другие. В настоящее время исследовательский интерес обращен к нейросетевым моделям, которые стали применяться к задачам языкового моделирования с начала 2000-х. В 2013 году появилась модель обучения векторных представлений слов word2vec [1, 2], совершившая прорыв в области компьютерных технологий. В 2017 году была представлена новая архитектура глубоких нейронных сетей под названием трансформеры, которая пришла на смену рекуррентным нейронным сетям. В основе новой архитектуры лежат технологии внимания и самовнимания, которые позволяют сделать акцент на самом важном токене для данного контекста.

Отдельного внимания заслуживают системы генеративного ИИ, которые служат для целей порождения текстов, изображений или мультимодальных данных. Примерами таких систем являются: Generative Pre-trained Transformer (ChatGPT) (компании OpenAI, США); Bard (компании Google, США); BingAI (Sydney) (компании Microsoft, США) или YandexGPT (компании «Яндекс», Россия). В настоящей работе использована система ChatGPT — одна из наиболее популярных систем генеративного ИИ. Она была представлена компанией OpenAI в конце 2022 года и стала первой системой такого рода, основанной на трансформерах. Система создает тексты как результат обработки промптов — специальных запросов для общения с ней. Промпты формулируются максимально формальным образом. ChatGPT является примером системы, которая способна переводить [3, 4] и генерировать контент, что повлияло на решение разных задач, связанных с созданием текстов. Прежде всего, это коснулось учебных и научных текстов, новостных статей, постов для блогов, однако

также в настоящее время есть возможность написания псевдокода или программного кода на определенном языке.

Становится очевидным, что ИИ может оказывать влияние на повседневную жизнь. Возникает вопрос о том, насколько сопоставимы результаты, которые получены при его использовании, с теми, что могли быть представлены человеком. Цель работы заключается в определении особенностей автоматически сгенерированных научных текстов по сравнению с текстами, созданными авторами, а также в оценке возможностей разных инструментов применительно к задаче их классификации по двум типам (созданные при участии систем ИИ или написанные людьми).

Обзор исследований

Способность отвечать на поставленные вопросы является одной из ключевых характеристик систем, стремящихся к пониманию естественного языка. Однако использование ИИ для получения ответов может быть связано с некоторыми сложностями.

Во-первых, требуется так называемый фактчекинг, предполагающий проверку сведений, полученных от ИИ, поскольку тексты могут быть недостоверными по своему содержанию. В исследованиях обсуждается проблема «галлюцинации» систем ИИ — явления, при котором система генерирует содержание, имеющее ошибочную или весьма правдоподобную (но ничем не обоснованную) информацию. В работе [5] изучен вопрос о том, что системы могут делать ничем не подкрепленные выводы в тех случаях, когда верный ответ не содержится эксплицитно в тексте (для тестирования системы были предложены отрывки, по которым необходимо было ответить на поставленные вопросы). Авторы представили усовершенствованную коллекцию вопросов SQuADRUn на основе Стенфордского вопросно-ответного набора данных (он является бенчмарком при обучении и тестировании систем понимания текстов), в котором содержится более 50 тыс. вопросов, на которые не существует ответов. Наиболее высокие результаты (точность 66,3 % по мере F1-score) по распознаванию подобных единиц показывает модель DocQA + ELMo, однако она уступает ручной оценке (точность 89,5 % по мере F1-score) [5]. Работа [6] посвящена оценке ChatGPT как вопросно-ответной системы применительно к английскому и персидскому языкам с привлечением SQuADRUn. Полученные результаты показали, что система справляется лучше с ответами, касающимися конкретных фактов, чем с вопросами «как» и «почему», однако в целом результаты ниже, чем у специализированных вопросно-ответных систем. Склонность к «галлюцинациям» проявляется в ответах на те вопросы, для которых нет ответов в представленном контексте. В [7] рассмотрен вопрос о применении ChatGPT при написании научных статей на медицинские темы, обращаясь к проблеме галлюцинаций. Некоторые из фактов, сгенерированные системой, были правдивыми, однако уточнение остальных фактов привело к выяснению следующего: упомянутые позиции в списках литературы оказались полностью выдуманными, а идентификаторы были связаны с другими работами. Однако в работе [7] отмечено, что ChatGPT может быть полезен при создании когерентного текста на основе разрозненных фрагментов из разных источников, а также при организации цитирования литературы в самом тексте статьи (во избежание цитирования одного и того же источника дважды). А также предложено редколлегиям использовать данный инструмент для выявления подобного контента.

Во-вторых, возникает вопрос авторства созданного текста, особо актуальный при работе с учебными или научными текстами. ChatGPT указан как автор в ряде статей по медицине: обсуждению этого вопроса посвящены работы [8, 9], в которых перечислены те критерии, которым должен удовлетворять автор научной статьи. Согласно им, ChatGPT не может выступать в качестве полноценного автора, однако в разделе, посвященном благодарностям, должно быть указано его использование, а также в этом случае авторам следует взять на себя ответственность (ввиду последствий необъективно поданной или ложной информации) [9, 10]. Отметим, что крупные издательства (например, Elsevier, Nature, Springer) обновляют свою политику и эксплицитно прописывают в правилах, что системы ИИ не могут быть указаны в качестве авторов [11]. Сравнительный анализ новых правил известных издательств приведен в [12].

Вопрос авторства носит не только этический характер, но и традиционно связан с задачей определения плагиата. В компьютерной лингвистике существует отдельное направление, посвященное стилеметрическому анализу. В качестве наиболее часто используемых методов обсуждаются разнообразные метрики сложности текстов, в том числе применительно к автоматически сгенерированным текстам [13].

Сравнению научных статей, написанных авторами-людьми и сгенерированными автоматически, в последнее время уделяется большое внимание. В работе [14] выполнена оценка двух наборов аннотаций как экспертами, так и автоматическими системами, выявляющими ИИ-контент. Первый был взят из научных статей, опубликованных в медицинских журналах, в то время как вторые были сгенерированы автоматически по предложенным заголовкам. Результаты показали, что 86 % из первой группы были ошибочно определены как созданные ботом, в то время как 68 % из второй — классифицированы верно. Эксперты отмечали, что сгенерированные тексты носят поверхностный характер, а также отличаются неясностью формулировок. Анализу русскоязычного материала посвящены работы [15, 16].

В [12] описан поэтапный процесс написания научного текста при помощи ChatGPT, предполагающий формулировку научной проблемы, введение, обзор литературы, методологию и представление результатов с последующим обсуждением. Авторы оценивают потенциальные возможности и ограничения ИИ, а также дают рекомендации по применению системы при написании разных разделов статьи.

При этом выявлены недостатки при формировании обзоров литературы при помощи систем ИИ. Например, ошибки при перифразировании или плагиат [17].

Результаты показывают, что сгенерированные ИИ тексты содержат формальные повторы словосочетаний или конструкций (например, из промптов), смысловые повторы и уже упомянутые искажения.

Одной из проблем, связанных с использованием систем ИИ, является недостоверная информация. Для определения автоматически сгенерированного текста можно проверить отдельные позиции (например, список литературы), а также использовать специальные приложения. Несмотря на то, что существуют системы-детекторы, которые используются для идентификации автоматически созданных текстов (например, GPTZero[1], Writer[2], ZeroGPT[3]), все они с разной степенью успешности справляются с поставленной задачей. В работе [18] наряду с хорошо известными признаками (такими как удобочитаемость или сложность текстов) вводится признак «обратной связи от искусственного интеллекта» (AI feedback feature), связанный с ответами («да», «нет», «не знаю»), которые выдает сама система на вопрос о том, был ли текст создан автоматически.

Методы и материал исследования

В настоящей работе были отобраны по 15 аннотаций научных статей на русском языке, опубликованных в научных журналах по компьютерной лингвистике[4] и по германистике[5] за 2021 год[6]. Выбор текстов из разных областей языкознания продиктован тем, что статьи по компьютерной лингвистике содержат большое количество терминов, заимствованных из английского языка без перевода или калькирования, в том числе с сохранением иноязычного написания (например, *эмбеддинги* vs *embeddings*), что делает такие тексты схожими по своей структуре со статьями, посвященными информационным технологиям или компьютерным наукам, и отличным от второй группы статей по германистике, имеющим более описательно-нарративный характер. Таким образом, обеспечивается гетерогенность исследуемых наборов текстов. Далее заголовки статей были использованы для формирования набора данных сгенерированных аннотаций при помощи ChatGPT (модель GPT-4o mini). Дополнительно было задано ограничение на длину выдаваемых текстов: она могла варьироваться от 160 до 270 слов (диапазон длин в наборе оригинальных текстов). Поскольку ChatGPT запоминает диалог в чате, каждый промпт задавался в отдельной новой сессии. Общий объем данных составил 60 аннотаций, разделенных по двум типам (30 оригинальных и 30 сгенерированных текстов) и репрезентирующих две темы — по компьютерной лингвистике и германистике.

В работе были решены следующие задачи. Во-первых, выполнен анализ с целью выявления особенностей автоматически сгенерированных текстов путем сравнения их с текстами, написанными авторами-людьми. Тексты проверялись по следующим критериям: использование общепринятых научных терминов; соответствие правилам русского языка; соблюдение формальных правил при составлении аннотаций; логическая связанность. Во-вторых, двумя способами оценивалась возможность разделения текстов на два класса (созданные автоматически с использованием ИИ или нет). Для этого были привлечены эксперты (эксперт-германист и эксперт в области компьютерной лингвистики), которые соотносили аннотации, соответствующие их профессиональным областям, с двумя классами. Аннотации предъявлялись в случайном порядке, эксперты не знали о том, какой процент составляют оригинальные и сгенерированные тексты. Также были использованы системы Smodin[7], ZeroGPT и GPTZero (наиболее популярные среди сервисов для выявления участия ИИ при написании текста) для определения того, что текст был написан ИИ. При оценке текста программы указывали вероятность того, что он был создан автоматически или человеком, далее результаты были приведены к бинарной шкале. В-третьих, был применен критерий обратной связи, полученный от системы ChatGPT. Эксперты и ChatGPT оценивали тексты из двух наборов данных по бинарной шкале (1 — написанный ИИ, 0 — написанный человеком).

Результаты

Анализ различий в аннотациях. Несмотря на ограничение размера аннотаций, длина сгенерированных ИИ аннотаций в ряде случаев не соответствовала требованию и была короче. Тексты аннотаций отличаются своей формальной структурой: выделяются абзацы, которые отсутствуют в аннотациях, написанных авторами-людьми. В целом их отличает более четкая организация, соответствующая правилам IMRAD (Introduction — Materials and Methods — Results — Discussion).

Первые предложения всех сгенерированных аннотаций по германистике включали перифразированные названия статей, которые были использованы в качестве промптов. В то время как в текстах, посвященных компьютерной лингвистике, содержится только две трети подобных примеров, при этом в ряде случаев предложения дополняются уточнениями. Примеры автоматически созданных аннотаций, посвященных двум темам, приведены в табл. 1.

В сгенерированных ИИ аннотациях присутствуют общие формулировки, которые носят поверхност-

[1] [Электронный ресурс]. Режим доступа: https://gptzero.me/, свободный. Яз. англ. (дата обращения: 24.08.2024).

[2] [Электронный ресурс]. Режим доступа: https://writer.com/, свободный. Яз. англ. (дата обращения: 24.08.2024).

[3] [Электронный ресурс]. Режим доступа: https://www.zerogpt.com/, свободный. Яз. англ. (дата обращения: 24.08.2024).

[4] [Электронный ресурс]. Режим доступа: https://www.dialog-21.ru/media/5617/_-dialog2021scopus.pdf, свободный. Яз. рус. (дата обращения: 24.08.2024).

[5] [Электронный ресурс]. Режим доступа: https://germanphilology.spbu.ru/issue/view/647, свободный. Яз. рус. (дата обращения: 24.08.2024).

[6] Для обучения модели ChatGPT 3.5 были использованы данные, ограниченные по времени сентябрем 2021, что послужило причиной для выбора статей именно этого временного диапазона.

[7] [Электронный ресурс]. Режим доступа: https://smodin.io/, свободный. Яз. англ. (дата обращения: 24.08.2024).

Таблица 1. Примеры первых предложений в сгенерированных аннотациях
Table 1. Examples of the first sentences in generated abstracts

Название статьи	Первое предложение в сгенерированной аннотации
германистика	
Фастнахтшпиль «Странствующий школяр в раю» Ганса Сакса и его переводы на русский и английский языки	*Статья посвящена анализу фастнахтшпиля «Странствующий школяр в раю» Ганса Сакса, одного из ярких представителей немецкой театральной традиции XVI века, и его переведенных версий на русский и английский языки*
Лингвистическое отражение ценностных ориентаций в текстах песен немецких и русских рок-музыкантов (на примере Рио Райзера и Александра Башлачёва)	*Статья посвящена исследованию лингвистического отражения ценностных ориентаций в текстах песен немецких и русских рок-музыкантов на примере Рио Райзера и Александра Башлачёва*
Этнолингвистическая специфика устойчивых сравнений с компонентом-онимом в немецком языке (на фоне русского языка)	*В статье исследуется этнолингвистическая специфика устойчивых сравнений с компонентом-онимом в немецком языке на фоне русского языка*
Сочетаемость деонтического показателя с коммуникативным глаголом в перволичных конструкциях: семантика и прагматика (на материале русского и немецкого языков)	*В статье рассматривается сочетаемость деонтического показателя с коммуникативным глаголом в перволичных конструкциях на материалах русского и немецкого языков*
Проблемы языковой критики в немецком и российском медиадискурсах	*Статья «Проблемы языковой критики в немецком и российском медиадискурсах» посвящена анализу проблематики языковой критики в медиадискурсах двух культур: немецкого и российского*
компьютерная лингвистика	
Окуломоторное повседневное общение: как выбрать хорошую метрику	*В последние годы изучение окуломоторных паттернов стало важным инструментом для исследования различных аспектов человеческого общения*
Russian SuperGLUE 1.1: пересматривая невыученные уроки русскоязычных NLP-моделей	*В статье «Russian SuperGLUE 1.1: пересматривая невыученные уроки русскоязычных NLP-моделей» рассматривается обновлённый набор данных для оценки производительности моделей обработки естественного языка (NLP) на русском языке*
Эффективность правиловых эвристик в решении тестовых сетов Russian SuperGLUE	*В данной статье рассматривается проблема применения правиловых эвристик для решения задач, входящих в тестовые сеты Russian SuperGLUE – крупномасштабного набора заданий для оценки производительности моделей обработки естественного языка на русском языке*
Межъязыковой перенос без дообучения толковой языковой модели для обнаружения семантических сдвигов	*В статье «Межъязыковой перенос без дообучения толковой языковой модели для обнаружения семантических сдвигов» рассматривается подход к выявлению семантических изменений в текстах с использованием межъязыкового переноса без необходимости дополнительного обучения языковой модели*
Переключение к рабочей деятельности в инклюзивной мастерской: мультимодальный анализ взаимодействия	*В статье представлена результативная работа, посвящённая мультимодальному анализу взаимодействия в контексте инклюзивной мастерской, с акцентом на процессы переключения к рабочей деятельности*

ный характер («*В заключение рассматриваются перспективные направления для будущих исследований*», «*Сделан вывод о важности создания подобных ресурсов для изучения аргументации как в теоретическом, так и в практическом аспекте*»), а также упоминаются квазитермины («*Результаты исследования открывают пути для дальнейших исследований и практических приложений в области коммуникационных наук*»). Также аннотации из статей по компьютерной лингвистике, предложенные ИИ, содержат аббревиатуры терминов на английском, которые приводятся в дополнение к терминам на русском языке. Например, «*моделей обработки естественного языка (NLP)*» или «*без использования оптического распознавания символов (OCR)*». Также цитируются названия статей, что несвойственно авторским текстам. В ряде примеров содержатся ошибки или искажения в лексической сочетаемости («*В статье представлена результативная работа*», «*набор данных является расширением предыдущих версий и включает дополнительные задачи*», «*с более сложными и нюансированными задачами*», «*дают новые перспективы*») и не соблюдается научный стиль («*качество сканов варьируется*», «*авторы также обсуждают уроки, извлеченные из предыдущих экспериментов*», «*автор сосредоточен*», «*предоставляет ценные инсайты для лингвистов*», «*предоставляя новые инсайты*»). В тек-

стах также встречается нарушение синтаксической сочетаемости: «*В статье представлен сравнительный лингвокультурологический анализ слов и антислов года в немецком и русском языках, исследующих изменения в языковом и культурном контексте двух разных социокультурных сред*». В данном случае причастие «исследующих» формально должно относиться к существительному «анализ» (однако не указывается субъект действия) и быть использовано в форме единственного числа, однако в приведенном примере оно может быть связано с формами «слов» и «антислов» или «языках», что приводит к противоречию. Вместе с тем ИИ приводит неуместные уточнения, которые не относятся к теме статьи «Динамика лексического состава русской художественной прозы (на материале частотных словарей корпуса русских рассказов 1900–1930)» («*подчеркивая роль писателей-реалистов и авангардистов в формировании нового лексического облика*»), дополняет текст деталями и примерами, которые не упоминались в промпте. Например, в сгенерированной аннотации к статье «Контрастивный анализ урбанонимов исторических городов Германии и России» перечислены такие города как Берлин, Дрезден, Санкт-Петербург и Казань, которые не рассматривались в исследовании. В двух случаях автоматически созданные аннотации сопровождались перечислением ключевых слов, хотя это не было сформулировано в промпте. В половине аннотаций использовались одинаковые конструкции («*особое внимание уделяется/ уделено*»).

Оригинальные аннотации (в отличие от сгенерированных) отличаются большей детализированностью и более широким охватом содержания статей, а значит, как следствие, большей информативностью: они содержат указание на использованные методы, в них также приводятся полученные результаты и количественные характеристики (последнее в случае аннотаций к статьям по компьютерной лингвистике). В трети текстов встречается авторское «мы», которое не характерно для аннотаций, написанных системой ИИ.

Определение участия ИИ в создании аннотаций. Результаты оценивались при помощи метрик recall, precision, accuracy и F1-score. В табл. 2 представлена матрица ошибок классификации, в которой показаны результаты приписывания принадлежности объекта к некоторому классу со следующей расшифровкой: True Positive (TP) — истинно положительные метки; False Positive (FP) — ложно положительные метки; False Negative (FN) — ложно отрицательные метки; True Negative (TN) — истинно отрицательные метки.

Recall определяется как количество правильных ответов (меток) относительно всех правильных:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Precision — количество правильных ответов (меток) относительно всех выданных:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

F1-score представляет собой гармоническое среднее между полнотой и точностью:

$$\text{F1} - \text{score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}.$$

Accuracy понимается как доля правильных ответов (меток) алгоритма:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

В табл. 3 приводятся данные для всех классификаторов в целом при оценке аннотаций статей на обе темы. Наиболее низкие результаты при оценке аннотаций при сравнении с другими инструментами были показаны ChatGPT. Согласно мере accuracy, менее половины текстов по компьютерной лингвистике (43 %) были верно определены упомянутой системой: в большин-

Таблица 2. Результаты классификации
Table 2. Classification results

Предсказанный класс	Истинный класс	
	сгенерированные аннотации	оригинальные аннотации
Аннотации, определенные как сгенерированные	TP	FP
Аннотации, не определенные как сгенерированные	FN	TN

Таблица 3. Оценка аннотаций
Table 3. Abstracts evaluation

Метрика	Компьютерная лингвистика					Германистика				
	Эксперт	ChatGPT	Smodin	GPTZero	ZeroGPT	Эксперт	ChatGPT	Smodin	GPTZero	ZeroGPT
Precision	0,63	0,42	0,76	0,80	0,60	0,88	0,33	1,00	0,67	0,54
Recall	0,47	0,33	0,87	0,27	1,00	1,00	0,07	0,67	0,13	1,00
Accuracy	0,60	0,43	0,80	0,60	0,50	0,93	0,47	0,83	0,53	0,57
F1-score	0,54	0,37	0,81	0,40	0,75	0,94	0,12	0,80	0,22	0,70

стве случаев она ошибочно определяла тексты, как созданные автоматически или, наоборот, как написанные человеком. В то время как для второй темы процент верно определенных аннотаций в обоих классах был выше и составил 47 %. Экспертная оценка оказалась более точной: 60 % верно определенных аннотаций по компьютерной лингвистике и 93 % по германистике.

Системы-детекторы, определяющие участие ИИ при создании текстов, демонстрируют разные результаты. Наиболее высокие показатели по метрикам precision, accuracy и F1-score достигаются системой Smodin, в то время как по критерию recall — системой ZeroGPT, которая верно определяет все тексты, написанные ИИ.

Предположение о том, что тематика текста может оказывать влияние на определение авторства с использованием автоматических методов, подтвердилось только в случае системы ChatGPT: для аннотаций по компьютерной лингвистике показатели в целом выше (т. е. их проще дифференцировать по двум типам), чем для текстов, посвященных второй теме. Описательно-нарративный характер изложения в текстах по германистике обнаруживает схожесть с автоматически сгенерированным содержанием и, следовательно, вызывает сложности для системы при их дифференциации. Можно заключить, что специальные детекторы справляются с задачей лучше, а в случае упомянутых текстов системой Smodin также достигается стопроцентная точность. Однако экспертная оценка достигла наибольших значений по трем из четырех критериев в случае аннотаций по германистике.

Заключение

Использованная модель ChatGPT-4o mini позволяет создавать научные тексты, которые отличаются четкой формальной структурой и обнаруживают сходство с оригинальными аннотациями в том, что касается формальной корректности использования средств для реализации речевого общения. Тем не менее сгенерированные тексты содержательно носят поверхностный характер, в них не всегда соблюдается научный стиль, присутствуют повторы конструкций и перифразирование названий статей, что не встречается в рассмотренных аннотациях, написанных авторами-людьми. Они нуждаются не только в дальнейшей редакторской правке (поскольку в ряде случаев нарушено согласование и присутствует неоднозначность), но и в проверке упоминаемых фактов и терминов.

Автоматические сгенерированные аннотации дают возможность читателям познакомиться с общим содержанием статей, тогда как оригинальные аннотации дают образ еще и личности исследователя, стоящего за текстом статьи (благодаря наличию четко расставленных смысловых акцентов, субъективно важных для самого автора, и не обязательных для знакомства с тематикой работы в первом приближении).

Наиболее высокие результаты при определении того, написан ли текст при помощи искусственного интеллекта, был продемонстрирован системой Smodin и экспертом (в случае текстов по германистике), в то время как оценка, произведенная другими программами, имеет низкие показатели recall, precision, accuracy и F1-score.

Дальнейшая работа может быть связана с дообучением моделей, увеличением объема оригинальных текстов, а также с формированием более сложных промптов, касающихся содержания и накладывающих ограничения на генерируемые аннотации.

Литература

1. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // Proc. of the 1st International Conference on Learning Representations, ICLR 2013 — Workshop Track Proceedings. 2013. P. 1–12.
2. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems. 2013. V. 26. P. 3111–3119.
3. Orăsan C. ChatGPT for translators: a survey // Proc. of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications. 2023. P. 61–63.
4. Castilho S., Mallon C.Q., Meister R., Yue S. 2023. Do online machine translation systems care for context? What about a GPT model? // Proc. of the 24th Annual Conference of the European Association for Machine Translation. 2023. P. 393–417.
5. Rajpurkar P., Jia R., Liang P. Know what you don't know: Unanswerable questions for SQuAD // Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018. P. 784–789. https://doi.org/10.18653/v1/p18-2124
6. Bahak H., Taheri F., Zojaji Z., Kazemi A. Evaluating ChatGPT as a question answering system: A comprehensive analysis and comparison with existing models // arXiv. 2023. arXiv:2312.07592. https://doi.org/10.48550/arXiv.2312.07592
7. Alkaissi H., McFarlane S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing // Cureus. 2023. V. 15. N 2. P. e35179. https://doi.org/10.7759/cureus.35179

References

1. Mikolov T., Corrado G., Chen K., Dean J. Efficient estimation of word representations in vector space. *Proc. of the 1st International Conference on Learning Representations, ICLR 2013 — Workshop Track Proceedings*, 2013, pp. 1–12.
2. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013, vol. 26, pp. 3111–3119.
3. Orăsan C. ChatGPT for translators: a survey. *Proc. of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, 2023, pp. 61–63.
4. Castilho S., Mallon C.Q., Meister R., Yue S. 2023. Do online machine translation systems care for context? What about a GPT model? *Proc. of the 24th Annual Conference of the European Association for Machine Translation*, 2023, pp. 393–417.
5. Rajpurkar P., Jia R., Liang P. Know what you don't know: Unanswerable questions for SQuAD. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789. https://doi.org/10.18653/v1/p18-2124
6. Bahak H., Taheri F., Zojaji Z., Kazemi A. Evaluating ChatGPT as a question answering system: A comprehensive analysis and comparison with existing models. *arXiv*, 2023, arXiv:2312.07592. https://doi.org/10.48550/arXiv.2312.07592
7. Alkaissi H., McFarlane S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 2023, vol. 15, no. 2, pp. e35179. https://doi.org/10.7759/cureus.35179

8. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove // Nature. 2023. V. 613(7945). P. 620–621. https://doi.org/10.1038/d41586-023-00107-z
9. Ide K., Hawke P., Nakayama T. Can ChatGPT be considered an author of a medical article? // Journal of Epidemiology. 2023. V. 33. N 7. P. 381–382. https://doi.org/10.2188/jea.JE20230030
10. Dwivedi Y.K., Kshetri N., Hughes L., Slade E.L., Jeyaraj A., Kar A.K., Wright R. et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy // International Journal of Information Management. 2023. V. 7. P. 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
11. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use // Nature. 2023. V. 613(7945). https://doi.org/10.1038/d41586-023-00191-1
12. Rahman M., Terano H.J.R., Rahman N., Salamzadeh A., Rahaman S. ChatGPT and academic research: A review and recommendations based on practical examples // Journal of Education, Management and Development Studies. 2023. V. 3. N 1. P. 1–12. https://doi.org/10.52631/jemds.v3i1.175
13. Herbold S., Hautli-Janisz A., Heuer U., Kikteva Z., Trautsch A. A large-scale comparison of human-written versus ChatGPT-generated essays // Scientific Reports. 2023. V. 13. N 1. P. 18617. https://doi.org/10.1038/s41598-023-45644-9
14. Gao Y., Wang R., Hou F. How to design translation prompts for ChatGPT: An empirical study // arXiv. 2023. arXiv:2304.02182. https://doi.org/10.48550/arXiv.2304.02182
15. Kolmogorova A.V., Margolina A.V. Written vs generated text: "naturalness" as a textual and psycholinguistic category // Научный результат. Вопросы теоретической и прикладной лингвистики. 2024. Т. 10. № 2. С. 71–99. https://doi.org/10.18413/2313-8912-2024-10-2-0-4
16. Черкасова М.Н., Тактарова А.В. Признаки сгенерированного текста в академическом дискурсе: проблема идентификации // Филологические науки. Вопросы теории и практики. 2024. Т. 17. № 7. С. 2226–2232. https://doi.org/10.30853/phil20240307
17. Aydın Ö., Karaarslan E. OpenAI ChatGPT generated literature review: Digital twin in healthcare // Emerging Computer Technologies 2. İzmir Akademi Dernegi, 2022. P. 22–31. https://doi.org/10.2139/ssrn.4308687
18. Mindner L., Schlippe T., Schaaff K. Classification of human- and AI-generated texts: Investigating features for ChatGPT // Lecture Notes on Data Engineering and Communications Technologies. 2023. V. 190. P. 152–170. https://doi.org/10.1007/978-981-99-7947-9_12

8. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, 2023, vol. 613(7945), pp. 620–621. https://doi.org/10.1038/d41586-023-00107-z
9. Ide K., Hawke P., Nakayama T. Can ChatGPT be considered an author of a medical article? *Journal of Epidemiology*, 2023, vol. 33, no. 7, pp. 381–382. https://doi.org/10.2188/jea.JE20230030
10. Dwivedi Y.K., Kshetri N., Hughes L., Slade E.L., Jeyaraj A., Kar A.K., Wright R. et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 2023, vol. 7, pp. 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
11. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, 2023, vol. 613(7945). https://doi.org/10.1038/d41586-023-00191-1
12. Rahman M., Terano H.J.R., Rahman N., Salamzadeh A., Rahaman S. ChatGPT and academic research: A review and recommendations based on practical examples. *Journal of Education, Management and Development Studies*, 2023, vol. 3, no. 1, pp. 1–12. https://doi.org/10.52631/jemds.v3i1.175
13. Herbold S., Hautli-Janisz A., Heuer U., Kikteva Z., Trautsch A. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 2023, vol. 13, no. 1, pp. 18617. https://doi.org/10.1038/s41598-023-45644-9
14. Gao Y., Wang R., Hou F. How to design translation prompts for ChatGPT: An empirical study. *arXiv*, 2023, arXiv:2304.02182. https://doi.org/10.48550/arXiv.2304.02182
15. Kolmogorova A.V., Margolina A.V. Written vs generated text: "naturalness" as a textual and psycholinguistic category. *Research Result. Theoretical and Applied Linguistics*, 2024, vol. 10, no. 2, pp. 71–99. https://doi.org/10.18413/2313-8912-2024-10-2-0-4
16. Cherkasova M.N., Taktarova A.V. Attributes of generated text in academic discourse: the problem of identification. Philology. *Theory & Practice*, 2024, vol. 17, no. 7, pp. 2226–2232. (in Russian). https://doi.org/10.30853/phil20240307
17. Aydın Ö., Karaarslan E. OpenAI ChatGPT generated literature review: Digital twin in healthcare. *Emerging Computer Technologies 2*. İzmir Akademi Dernegi, 2022, pp. 22–31. https://doi.org/10.2139/ssrn.4308687
18. Mindner L., Schlippe T., Schaaff K. Classification of human- and AI-generated texts: Investigating features for ChatGPT. *Lecture Notes on Data Engineering and Communications Technologies*, 2023, vol. 190, pp. 152–170. https://doi.org/10.1007/978-981-99-7947-9_12

Авторы

Хохлова Мария Владимировна — кандидат филологических наук, доцент, доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, sc 56088078800, https://orcid.org/0000-0001-9085-0284, m.khokhlova@spbu.ru
Корышев Михаил Витальевич — кандидат филологических наук, доцент, доцент, декан факультета, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, sc 57844284200, https://orcid.org/0000-0001-8946-4431, m.koryshev@spbu.ru

Authors

Maria V. Khokhlova — PhD (Philology), Associate Professor, Associate Professor, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation, sc 56088078800, https://orcid.org/0000-0001-9085-0284, m.khokhlova@spbu.ru
Mikhail V. Koryshev — PhD (Philology), Associate Professor, Associate Professor, Dean, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation, sc 57844284200, https://orcid.org/0000-0001-8946-4431, m.koryshev@spbu.ru

Статья поступила в редакцию 25.08.2024
Одобрена после рецензирования 09.10.2024
Принята к печати 24.11.2024

Received 25.08.2024
Approved after reviewing 09.10.2024
Accepted 24.11.2024