ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 308–317

УДК 004.89

ПОИСК ИСКУССТВЕННО СГЕНЕРИРОВАННЫХ ТЕКСТОВЫХ ФРАГМЕНТОВ В НАУЧНЫХ ДОКУМЕНТАХ

© 2023 г. Г. М. Грицай[1,2,*], А. В. Грабовой[1,2,3,], А. С. Кильдяков[1,***], Ю. В. Чехович[1,3,****]**

Представлено академиком РАН А.Л. Семеновым
Поступило 02.09.2023 г.
После доработки 15.09.2023 г.
Принято к публикации 18.10.2023 г.

Недавние достижения в области текстовых генеративных моделей позволяют получать искусственные тексты, едва отличимые от написанных человеком при беглом прочтении. Прогресс подобных моделей ставит новые задачи перед научным сообществом, ведь их развитие влечет за собой появление и распространение ложной информации, спама, способствует распространению неэтичных практик. В области обработки естественного языка уже разработано большое количество методов для детектирования текстов, полученных при помощи моделей машинного обучения, включая большие языковые модели. Однако улучшению методов выявления искусственных текстов происходит одновременно с улучшением методов генерации текстов, поэтому требуется изучение появляющихся моделей, искусственных текстов – результатов их работы и модернизации существующих подходов к детекции. В настоящей работе представлен детальный анализ ранее созданных методов детекции, а также исследование лексических, синтаксических и стилистических особенностей генерируемых фрагментов. В вычислительном эксперименте сравниваются различные методы детектирования машинной генерации в документах с точки зрения их дальнейшего применения для научных и учебных текстов. Эксперименты проводились для русского и английского языков на собранных авторами наборах данных. Разработанные методы позволили довести качество детектирования до значения 0.968 по метрике F1-score для русского и до 0.825 для английского языков соответственно. Созданные методы используются в практических системах для выявления сгенерированных фрагментов в научных, исследовательских и выпускных работах.

Ключевые слова: машинно-сгенерированный текст, обработка естественного языка, множественная проверка гипотез, перефразирование, детекция сгенерированных текстов

DOI: 10.31857/S2686954323601677, **EDN:** GQRWLF

1. ВВЕДЕНИЕ

Развитие методов машинного обучения и глубоких нейросетевых моделей в применении к задачам естественного языка позволяет получать искусственно созданные тексты, степень сходства которых с человеческими с годами становится лишь выше. Механизм самовнимания (англ. self-attention) [1], огромные объемы данных, собранных из всех доступных источников сети Интернет и новейшие методики дообучения моделей с архитектурой трансформер в основе, используемые в подходах генерации текстовых фрагментов, позволяют добиться высокого уровня “осмысленности” у подобных текстов. Алгоритмы генерации текстовых последовательностей успешно встраиваются в диалоговые системы, которые обеспечивают взаимодействие с пользователями в формате чата. Такие системы являются помощником для человека при решении задач, связанных с естественным языком. Современные чат-боты ChatGPT [2], Jasper [3], Google Bard [4], GigaChat [5], YaGPT [6], способны предоставить ответ на любой корректно поставленный вопрос, а также обладают большим количеством дополнительных возможностей: перевод, суммаризация текста, перефразировка, написание программного кода и др. Текстовые генеративные модели имеют возможность отвечать на вопросы, ответы на которые встречались в материале обучения при решении оптимизационной задачи выбора параметров модели. Из-за

[1]*Компания “Антиплагиат”, Москва, Россия*
[2]*Московский физико-технический институт (национальный исследовательский университет), Москва, Россия*
[3]*Федеральный исследовательский центр “Информатика и управление” Российской академии наук, Москва, Россия*
**E-mail: gritsai@ap-team.ru*
***E-mail: grabovoy@ap-team.ru*
****E-mail: kildyakov@ap-team.ru*
*****E-mail: chehovich@ap-team.ru*

огромного объема данных для обучения модель генерирует ответ, который далеко не всегда соответствуют действительности. Ключевым принципом, используемым при создании текстов или генерации ответов, является минимизации уровня перплексии (англ. perplexity) модели, но такой принцип не обеспечивает наличие в ответе только правдивой информации. В последнее время генеративные сервисы все чаще используются для составления текстовых документов, при этом пользователи далеко не всегда проверяют информацию, полученную в результате генерации. Поэтому важно иметь возможность детектировать искусственно созданные текстовые последовательности, полученные различными генеративными моделями в автоматическом режиме, для их дальнейшего более тщательного анализа.

Генеративные модели активно используются при подготовке научных и учебных работ [7]. При этом зачастую авторы не обеспечивали должную проверку сгенерированным утверждениям. Чатботы использовались для того, чтобы быстро и без дополнительных усилий создать текст необходимого объема. Еще одним применением генеративных моделей стало перефразирование текста для того, чтобы затруднить системам обнаружения заимствований выявление того, что текст был ранее опубликован [8]. Таким образом, чат-боты стали рассматриваться недобросовестными авторами как инструмент быстрой генерации научного или учебного текста, гарантированно проходящего “проверку на плагиат”. Это обеспечивает актуальность решения задачи создания надежного средства детекции машинно-сгенерированного текста в научных или учебных документах.

Следует отметить, что в современных языковых генеративных моделях существует возможность получать текст варьируемой длины. Во многих алгоритмах есть ограничение на максимальное количество символов для генерации за один шаг, однако существуют способы, которые позволяют сгенерировать связную последовательность символов произвольной длины, используя итеративный подход. Текст таких работ чаще всего не генерируется моделями целиком: как правило, это умелая комбинация машинно-сгенерированных и написанных человеком [9, 10] фрагментов. Тексты научных и учебных документов имеют несколько особенностей: значительная длина в несколько десятков, а иногда и несколько сотен страниц, наличие четкой структуры, в рамках которой текст разбивается на введение, содержание, основная часть, заключение и библиография. Разработанные ранее методы детекции не решают проблемы детекции документов большого объема. Длинные текстовые последовательности содержат последовательность фрагментов, которые ранее разработанные методы анализируют независимо. Данный подход ведет к большому числу ложно-положительных срабатываний, а также возможно к неполному анализу исследуемого документа.

В данной работе предлагается метод, который повышает полноту детектирования искусственно созданных фрагментов в научных работах. Анализ использования языковых генеративных моделей показывает, что сгенерированные фрагменты чаще всего встречаются в введении, основной части и заключении, реже в библиографии и содержании. В работе предлагается алгоритм, который выявляет сгенерированные фрагменты в наиболее подходящих разделах документа. После исключения из анализа наименее используемых компонентов структуры документа, длина работы все также имеет большое число символов и страниц. Полная обработка текстовой последовательности современными моделями детекции не представляется возможным, поэтому необходимо фрагментировать поданный документ.

Стоит отметить, что длины фрагментов, на которые разбивается анализируемый документ, оказывает влияние на результат детекции [12]. Оптимальная длина зависит от языка документа и анализируемого раздела. К примеру, при базовом решении задачи детекции при помощи модели, основанной на архитектуре трансформер, XLM-RoBERTa [11], максимально возможная длина входной последовательности – 512 токенов. В среднем, для русского языка это порядка 4000 символов. Таким образом, имеет место подход с делением текста отобранных частей поданного документа на фрагменты, где их средняя длина близка к указанному значению.

В настоящей работе:

- проанализирована эволюция методов детектирования искусственных текстов, позволяющая проследить зависимость признаков, необходимых для успешной классификации фрагментов, полученных от самых простых моделей до последних изданных;
- собрана выборка документов, включающая в себя сгенерированные и настоящие тексты из разных научных областей;
- предложен подход, включающий фрагментацию и идею множественного тестирования гипотез, для детектирования наличия сгенерированных отрывков в научных работах;
- протестированы методы детекции машинно-сгенерированных фрагментов при помощи моделей, основанных на архитектуре трансформер, проведены эксперименты с конкатенацией ручных признаков к выходу нейросетевого кодировщика, мультиязычным обучением и подменой в сгенерированных частях некоторых документов на переводные и перефразированные.

2. ОБЗОР ЛИТЕРАТУРЫ

Качество текста, сгенерированного нейросетевыми моделями, приближается к “человекоподобности” пару десятков лет. В данном разделе представленны основные работы по способам генерации текстовых последовательностей и методов их детекции, включая базовые и современные методы генерации.

В работе [13] искусственные тексты получаются при помощи комбинации различных видов алгоритмов на n-граммах, а сам метод детекции использовал эмпирические, синтаксические и семантические признаки, объединение которых затем использовалось в качестве входного признакового описания в классификаторе AdaBoost [14]. Начиная с данных алгоритмов генерации, величина перплексии, характеризующая качество предсказания выборки, настоящих и искусственных текстов стала отражать кардинальное различие. В данной работе отмечена важность исследования признаков, коррелирующих с частями речи, частота использования которых указывает на принадлежность к тому или иному представленному классу. Подход с ручным порождением признаков оставался популярен довольно длительный период времени. В работе [15], посвященной исследованию по борьбе с фейковыми новостями, расширили используемое признаковое пространство, добавив туда подсчет видов пунктуации и индекс удобочитаемости (англ. Flesch Reading Ease). Данная величина отражает меру определения сложности восприятия текста читателем. В качестве классификатора использовался метод опорных векторов (англ. SVM). В работе отмечалось, что существенный минус описанных подходов – плохая переносимость на иные домены при наличии в обучающей выборке лишь одного.

Появление архитектуры трансформер и механизама самовнимания (англ. Self Attention) привело к значительному росту качества генерации искусственных текстов. Языковые авторегрессионные модели на каждом шаге предсказывают следующий токен. В связи с этим появляется разнообразие стратегий для декодирования токенов, такие как greedy, beam search и другие, а также способы сэмплирования, такие как top-k, top-p, pure. С другой стороны, в качестве детекторов наравне с классическими методами машинного обучения добавляются BERT-подобные модели [16], используя токен фрагмента текста [CLS], а также дообучения энкодеров с несколькими миллионами параметров. В работе [17] проанализировали влияние способов сэмплирования и выявили, что наибольшую переносимость и устойчивость при классификации обеспечивает обучение на примерах, полученных при помощи top-p подхода, в то время как использование top-k сильно понижало уровень оценки при подаче на вход детектора документов из области, которая не содержалась в тренировочной части. При большом обзоре области искусственных текстов авторы [18] вновь продемонстрировали возможность достижения высокого качества при наличии top-p сгенерированных текстов в выборке, при этом были отмечены сильные стороны модели RoBERTa [19], которая существенно улучшила качество в семействе BERT-подобных моделей. Спустя небольшой промежуток времени подход с дообучением модели RoBERTa на текстах, полученных при помощи способа сэмплирования top-p, закрепил за собой статус state-of-the-art (SOTA) решения. Действительно, данная комбинация позволяет достигать высокого качества детектирования для текстов, полученных как с более ранних моделей генерации, так и с недавно представленных. Параллельно с этим авторами [20] были предложены алгоритмы с использованием фактической структуры фрагментов и комбинацией с графовыми нейросетями, однако данные подходы не позволяли существенно повышать качество детектирования и оказывались вычислительно неэффективными.

В моделях для классификации с архитектурой трансформер существует ограничение на длину подаваемого фрагмента. В более новых работах [21] отмечалась зависимость качества детекции от длины входной последовательности. Для разных языков оптимальная длина варьируется, но в каждом существует плато, позволяющее зафиксировать качество классификации. Все те же классификаторы, основанные на трансформерах, часто оказываются особенно восприимчивыми к состязательным атакам (англ. adversarial attacks). Иными словами, злонамеренное манипулирование входными данными модели машинного обучения с целью заставить ее выдать неправильные предсказания. Используя эту идею, авторы [22] предложили разбавлять сгенерированную часть выборки примерами с исправлениями. Под исправлениями подразумевается использование текстов, в которых часть предложений заменена, к примеру, перефразированными. Для реализации такого подхода было предложено использовать SOTA энкодер-декодер модель – T5 [23, 24], которая по результатам экспериментов отлично с этим справляется. На этом использование T5 в области детекции машинной генерации не заканчивается, в недавно представленной работе [23] данная архитектура всецело решает задачу классификации подозрительных фрагментов, ограничив словарь выходов у декодера до нужного количества классов.

К настоящему времени разработано большое количество методов борьбы с искусственными текстами, однако состязание генерации и детекции очевидно продолжится до тех пор, пока обе области будут развиваться. Мы полагаем, что

универсальный качественный детектор, подходящий для любых типов текстов доменов и любых языков, создать скорее всего не получится. Тем не менее учет значительной специфики, присущей научным и учебным работам, позволит создать детектор высокого качества.

3. ПОСТАНОВКА ЗАДАЧИ

Пусть $\mathbf{W}$ – это алфавит, содержащий минимальные элементы текстовых последовательностей, – символы. Вводится множество документов, где каждый документ представляется конечной комбинацией символов алфавита:

$$\mathbb{D} = \{[t_j]_{j=1}^{n} | t_j \in \mathbf{W}, n \in \mathbb{N}\}.$$

Зададим выборку из N документов:

$$\mathbf{D} = \bigcup_{i=1}^{N} D^i, \quad D^i \in \mathbb{D}.$$

Требуется найти в каждом документе искусственные последовательности символов. Для детектирования сгенерированного текста будет использоваться подход классификации. Иными словами, необходимо разделить текст исходного документа на фрагменты и провести бинарную классификацию каждого из них.

Зададим модель ϕ в виде суперпозиции двух преобразований $\mathbf{f}$ и $\mathbf{g}$. Преобразование $\mathbf{f}$ представляет собой разделение текста на фрагменты, а преобразование $\mathbf{g}$ – классификатор текстовых последовательностей на наличие искусственных частей:

$$\phi = \mathbf{f} \circ \mathbf{g},$$

$$\phi : \mathbb{D} \to \mathbb{T}, \quad \mathbb{T} = \{[t_{s_j}, t_{f_j}, c_j]_{j=1}^{J} | t_{s_j} = t_{f_{j-1}},$$
$$s_j \in \mathbb{N}_0, f_j \in \mathbb{N}, c_j \in \mathbf{C}\},$$

где J – количество фрагментов после этапа фрагментации документа, t_{s_j} – стартовый индекс j-го фрагмента, t_{f_j} – завершающий индекс j-го фрагмента, c_j – класс j-го фрагмента.

В текущей постановке $\mathbf{C} = \{0,1\}$, где метка $c_j = 1$ соответствует тексту, который машинно-сгенерирован, $c_j = 0$ соответствует отрывку, который был написан человеком.

Первый элемент суперпозиции – преобразование, позволяющее разделить текст на непересекающиеся фрагменты:

$$\mathbf{f} : \mathbb{D} \to \mathbf{T}^*,$$

$$\mathbf{T}^* = \{[t_{s_j}, t_{f_j}]_{j=1}^{J} | t_{s_j} = t_{f_{j-1}}, s_j \in \mathbb{N}_0, f_j \in \mathbb{N}\},$$

где s_j обозначает стартовый индекс j-го фрагмента и f_j обозначает завершающий индекс j-го фрагмента. Таким образом, $\mathbf{T}^*$ представляет собой множество всех возможных непересекающихся фрагментов, которые полностью покрывают документ.

Вторым преобразованием проводится бинарная классификация каждого текстового фрагмента.

$$\mathbf{g} : \mathbf{T}^* \to \mathbf{C}.$$

Для подбора оптимального преобразования каждому документу из заданной выборки $\mathbf{D}$ поставлена в соответствие метка c_j, которая распространяется также на все фрагменты данного документа. Далее задача состоит в том, чтобы найти бинарный классификатор, который минимизирует эмпирический риск в наборе данных $\mathbf{D}$:

$$\hat{g} = \arg\min_{g \in \mathfrak{F}} \sum_{D^i \in \mathbf{D}} \sum_{x_j, c_j \in D^i} [g(x_j) \neq c_j], \qquad (1)$$

где x_j – фрагмент документа D^i, а $\mathfrak{F}$ – набор всех рассмотренных алгоритмов для классификации.

4. ПРОБЛЕМЫ МНОЖЕСТВЕННЫХ СРАВНЕНИЙ

При описанном подходе детекции машинно-сгенерированного текста чем больше оказывается длина документа, тем больше возникает фрагментов, которые приходят на вход для модели классификации. Сформулируем задачу проверки фрагмента на языке статистических гипотез:

$$H_0 : \hat{g}(\mathit{fragment}) = 0,$$

$$H_1 : \hat{g}(\mathit{fragment}) = 1.$$

Зафиксируем уровень значимости $\alpha = 0.05$. При проверке поставленной гипотезы возникает ситуация, когда во время множественного тестирования, а именно классификации m фрагментов исходного документа с уровнем значимости α, будут накапливаться ошибки первого рода (ложные отклонения гипотезы). Верхняя оценка вероятности того, что хотя бы один из них будет неверным:

$$P(\mathit{false\ positive}) = 1 - (1 - \alpha)^m,$$

величина которой достаточно велика уже при небольших значениях m. Требуется контролировать ошибку первого рода, для этого существует несколько способов корректировки уровня значимости, различаются они лишь мощностью тестов. Для соблюдения баланса в описанных сущностях вводятся две меры, контролирующие ошибки первого рода: family-wise error rate (FWER) – групповая вероятность ошибки и false discovery rate (FDR) – ожидаемая доля ложных отклонений.

$$FWER = P(V > 0), \quad FDR = \mathbb{E}\left(\frac{V}{V + S}\right),$$

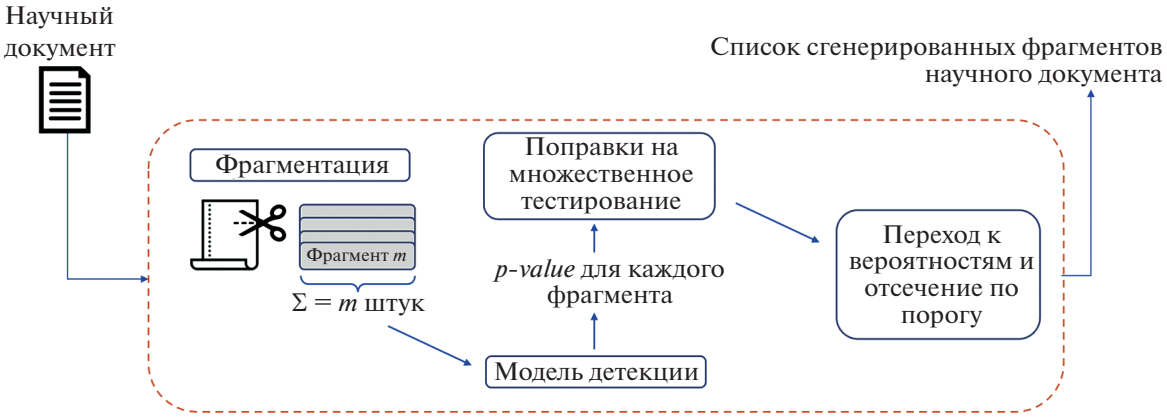

Рис. 1. Полный цикл работы алгоритма детекции сгенерированных фрагментов в научных документах.

где V – число ложно положительных результатов (ошибок первого рода), а S – число истинно положительных результатов.

Одним из распространенным методов контроля на FWER на уровне α является метод Холма [26]. Первый шаг – сортировка по возрастанию уровней *p-value*. Далее, уровень значимости при этом меняется в таком порядке:

$$\alpha_1 = \frac{\alpha}{m}, \quad \alpha_2 = \frac{\alpha}{m-1}, \ldots,$$
$$\alpha_i = \frac{\alpha}{m-i+1}, \ldots, \quad \alpha_m = \alpha.$$

Следом процесс выстраивается таким образом: если $p_1 >= \alpha_1$, то все нулевые гипотезы не отвергаются, иначе отвергается первая, и продолжаем.

Другой метод, позволяющий добиться высокого уровня мощности, – метод Бенджамини-Хохберга [27], где контролируется FDR. После сортировки по возрастанию *p-value* уровень значимости в данном подходе меняется по следующему закону:

$$\alpha_1 = \frac{\alpha}{m}, \quad \alpha_2 = \frac{2\alpha}{m}, \ldots,$$
$$\alpha_i = \frac{i\alpha}{m}, \ldots, \quad \alpha_m = \alpha.$$

Здесь процесс выстроен с иной стороны, в отличие от метода Холма: если $p_m < \alpha_m$, то необходимо отвергнуть все гипотезы, иначе – не отвергать m-ю, и продолжить.

Замечание 1. *Для любой процедуры множественного тестирования гипотез FDR <= FWER.*

Из замечания следует, что в большинстве случаев метод Бенджамини-Хохберга оказывается мощнее метода Холма, он отвергает не меньше гипотез с теми же α_i. В связи с этим в данной работе выбор был сделан в сторону метода Бенджамини-Хохберга. На рис. 1 изображен полный цикл работы алгоритма детекции машинно-сгенерированных фрагментов для домена научных документов. Данная комбинация идеи фрагментации и статистических поправок позволяет провести проверку научной работы всецело, не теряя при этом части текста из-за ограничения моделей детекции и не накапливая ошибку отклонения верной гипотезы.

5. НАБОР ДАННЫХ

Высокий уровень полноты детекции и большая обобщающая способность модели достигаются при обучении на мультидоменных текстах. Чем больше разнородных по тематикам документов будет на стадии обучения, тем выше будут метрики качества классификации при использовании детектора в реальных задачах поиска искусственных отрывков [28]. В текущей работе для проведения экспериментов были выбраны два языка: русский и английский. На английском языке существует большое число наборов данных в открытых источниках, на русском – выбор ограничен. В рамках данной работы проведены анализ имеющихся датасетов и объединение некоторых из них для каждого представленного языка. Стоит отметить, что наборов данных для детекции машинно-сгенерированных текстов, примеры из которых сравнимы с публикуемыми статьями и исследовательскими работами по размеру, в открытом доступе не найти. В рамках вычислительного эксперимента предлагается работа с текстами, средняя длина которых приблизительно равна длине потенциального фрагмента из описанного ранее алгоритма, наборов с такими данными преобладающее количество среди опубликованных. Помимо этого, были выбраны тексты, относящиеся к различным тематикам. Подробное описание структуры данных представлено в табл. 1.

Таблица 1. Структура и описание собранного набор данных для русского и английского языков

Часть датасета	Источник	Количество текстов	Средняя длина	Медианная длина
Ru-human	MGTDR	31.500	3.810	2.260
	Yandex Q	11.700	1.950	1.515
	Rus. Essays	1.000	3.366	3.176
Ru-machine	MGTDR	31.500	3.665	3.929
	Alpaca	7.600	1.361	1.272
	Saiga	4.000	1.190	1.145
	Rus. Essays	1.000	3.414	3.287
En-human	DeepFake	31.500	2.315	1.682
	HC3	12.600	1.150	1.156
En-machine	DeepFake	31.500	2.398	1.877
	GPT4	12.600	1.560	1.502

• В получившемся наборе для русского языка часть настоящих текстов представлена датасетами: Open access dataset for machine-generated text detection in Russian (**MGTDR**) [29] – статьи с ресурса Википедии, **Yandex-Q** [30] – сборник ответов на вопросы на популярном интернет-ресурсе, **Russian Essays** – сборник школьник сочинений по гуммунитарным тематикам. Искусственная часть: **MGTDR** – сгенерированные тексты по первым строкам настоящих статей с ресурса Википедии, **Alpaca** [31] – задания и ответы, сгенерированные ChatGPT, **Saiga** [32] – диалоги, сгенерированные ChatGPT на заданные темы, **Russian Essays** – сгенерированные тексты ChatGPT по темам сочинений из настоящего сборника.

• В английской части датасета настоящие тексты были представлены выборками **DeepFake** [33] – письменные работы, новости, истории, научные работы и **GPT4 on Instructions** [34], а сгенерированные противоположной частью набора – **DeepFake** – коллекция текстов от 27 языковых генеративных моделей, в том числе LLaMa, OpenAI и др. и **HC3** [35].

Для более детального анализа собранных текстов были подсчитаны статистики по представленным частям речи в используемых текстах. Для сбора информации о частях речи в русском языке использовалась библиотека Pymorphy [36], для английского – NLTK [37]. На рис. 2 показано их соотношение для русского и английского языков. Видно, что статистические значения у искусственных и настоящих текстов близкие, но отличаются.

6. ОПИСАНИЕ ЭКСПЕРИМЕНТОВ

Для получения качественной модели классификации текстовых фрагментов в данной работе проведено несколько экспериментов с разными подходами дообучения моделей, основанных на архитектуре трансформер. В качестве базового решения используется мультиязычная модель XLM-RoBERTa. После проведения анализа ранее опубликованных работ в данной работе были поставлены пять экспериментов помимо базового решения:

• **Ручные признаки:** кодировщики, с архитектурой трансформер, обучаясь на огромном количестве текстов, извлекают информацию из подаваемого фрагмента в виде векторного представления. При исследовании статистик в датасетах были выделены 26 числовых статистик, на базе которых построено признаковое пространство. BERT-подобные модели описывают свойства поданного текста в пространстве 768 признаков на базе [CLS] токена. В данном подходе предлагается конкатенация признаков, собранных вручную, с признаками, полученными на выходе кодировщика. После операции конкатенации вектора подаются на вход классификатору. На рис. 3 сравниваются статистики для русской и английской выборок для части ручных признаков.

• **Мультиязычное дообучение:** в данном эксперименте строится зависимость качества модели в зависимости при использовании смешанных данных из двух различных языков. Для экспериментов используется мультиязычная модель-энкодер XLM-RoBERTa.

• **Перевод текстов:** в данном эксперименте проверяется гипотеза о том, что использование переводных текстов улучшает предсказания классификатора на тестовом наборе данных. Для перевода с английского на русский и в обратную сторону использован сервис автоматического перевода Google Translate [38]. Были созданы дополнительные три тренировочные выборки для каждого языка, в которых 10, 25 и 50% сгенерированных данных заменены переводами соответственно.

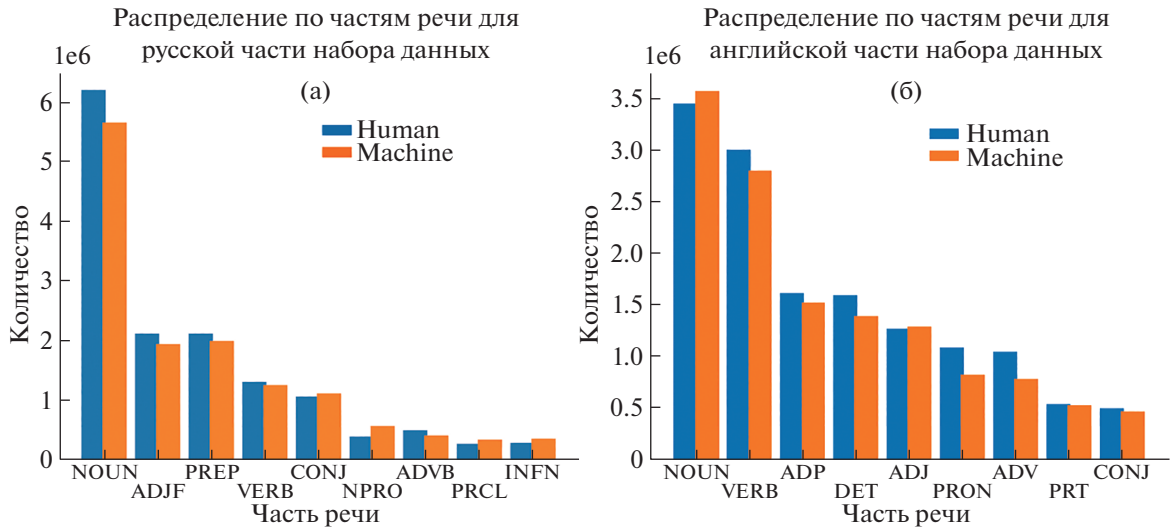

Рис. 2. На рисунке (а) представлено количественное разбиение текстов русской части данных по частям речи, а на рисунке (б) для английского. Здесь NOUN – существительное, ADJ(F) – прилагательное (полное), PREP – предлог, VERB – глагол, CONJ – союз, NPRO – местоимение, ADV(B) – наречие, PRCL – частица, INFN – инфинитив, ADP – дополнения (предлоги), DET – артикли, числительные, PRON – местоимения, PRT – частицы.

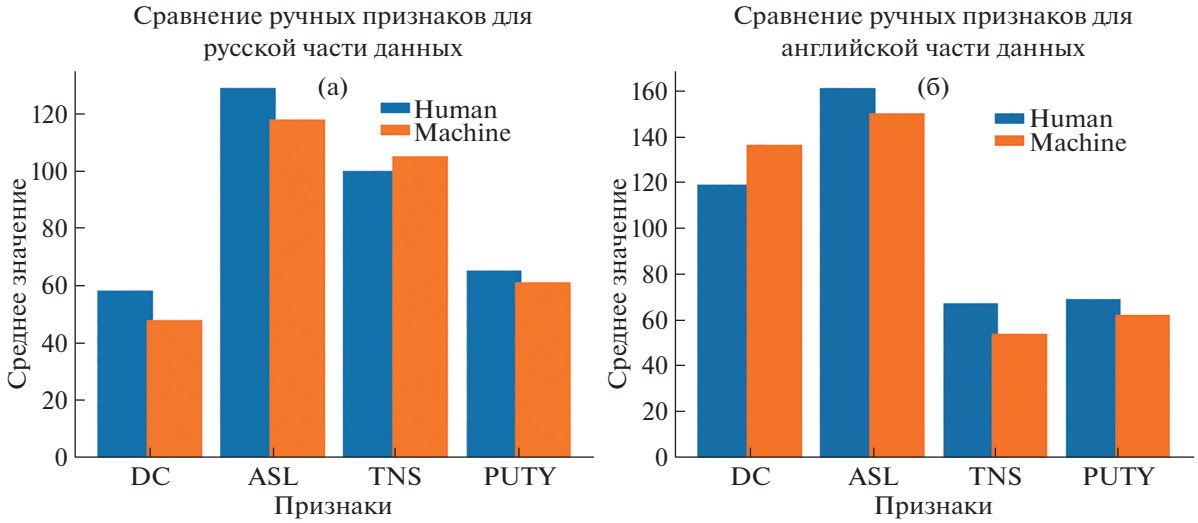

Рис. 3. На рисунке представлены статистики на основе ручных признаков для (а) русской части данных и (б) английской. Здесь DC – количество цифр во фрагментах, ASL – средняя длина предложения, TNS – общее количество стоп-слов, PUTY – уникальные пунктуационные знаки, FRS – значение индекса удобочитаемости.

• **Перефразирование текстов:** классификаторы на основе архитектуры трансформер восприимчивы к состязательным атакам. В частности, к атакам с переписыванием разделов сгенерированной статьи [39, 40] и атакам с перефразированием. Например, когда модель с архитектурой трансформер используется для перефразирования результатов другой генеративной модели. Для повышения устойчивости модели проводится эксперимент с подмешиванием в сгенерированную часть выборок перефразированных фрагментов. Были созданы дополнительные две тренировочные выборки для обоих языков. В первом таком наборе в каждом сгенерированном тексте перефразируется от 0 до 100% предложений. Во втором – в половине сгенерированных текстов перефразируется 50 до 100% предложений. Для получения перефразирования текстов для русского языка использовалась дообученнная модель на основе T5: rut5-base-paraphraser [41], а для английского – t5 sentence paraphraser [42].

• **T5:** в данном эксперименте исследуются модели T5 для решения задачи классификации. В

Таблица 2. Сводная таблица результатов вычислительного эксперимента

Язык	Эксперимент	F1-score	Precision	Recall
ru	базовое решение	0.955	0.958	0.955
	ручные признаки	0.960	0.962	0.959
	мультиязычное обучение	0.964	0.964	0.966
	перевод текстов 10%	0.955	0.958	0.955
	перевод текстов 25%	0.958	0.961	0.958
	перевод текстов 50%	0.966	0.968	0.966
	парафраз предложений 100%	**0.968**	**0.970**	**0.968**
	парафраз предложений 50%	0.964	0.965	0.963
	T5 энкодер	0.953	0.956	0.952
	T5 полная модель	0.952	0.954	0.954
en	базовое решение	0.796	0.855	0.802
	ручные признаки	0.801	0.856	0.807
	мультиязычное обучение	0.823	0.867	0.828
	перевод текстов 10%	0.812	0.859	0.815
	перевод текстов 25%	0.821	0.865	0.826
	перевод текстов 50%	**0.825**	**0.868**	**0.830**
	парафраз предложений 100%	0.822	0.866	0.827
	парафраз предложений 50%	0.816	0.862	0.817
	T5 энкодер	0.795	0.854	0.801
	T5 полная модель	0.787	0.850	0.794

первом подходе используется полная энкодер-декодер модель, а задача ставится в предсказании одной из двух меток “human” либо “machine”. Во втором – от полной модели T5 взят энкодер и поверх него добавлен нейросетевой классификатор, данный подход схож с классическим дообучением BERT-подобных моделей, однако выбран кодировщик от генеративной модели.

Выборки для тестирования остаются зафиксированными во всех описанных выше экспериментах и не отличаются от изначально построенных. Во всех экспериментах дообучение было произведено на 2 эпохах, с размером батча, равным 16, и с темпом обучения (англ. learning rate) – $4e-5$. При токенизации применялась обрезка фрагментов по максимальной длине входной последовательности 512 токенов, в силу ограничений используемой модели.

7. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

В табл. 2 приведены результаты поставленных экспериментов с различными техниками дообучения модели с архитектурой трансформер. Результаты эксперимента схожи для обоих заявленных языков. Для языков различается рост метрик качества, при максимально достигнутых метриках качества. В случае русского языка лучшее качество достигается в рамках подхода с перефразированием каждого текста в сгенерированной части собранного набора данных с диапазоном покрытия 0–100% предложений, для английского – замена половины сгенерированной части собранного набора данных на переводы текстов с русского языка. Заметим, что оба подхода решают задачу с потенциальными состязательными атаками, проводимыми над моделью классификации. Результатом исследования показано, что переведенные и перефразированные тексты позволяют повысить качество и обобщающую способность модели детекции машинно-сгенерированных фрагментов.

8. ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены варианты решения задачи поиска искусственно сгенерированных фрагментов в текстовых последовательностях. Проведены анализ области и исследование существующих методов для поиска машинной генерации. Предложен подход детекции в домене научных работ, а именно: разбиение исходного документа по разделам, фрагментация наиболее значимых и последующая их подача на вход дообученному классификатору, основанному на архитектуре трансформер. На выходе классификатора фрагменты одного текста проходят процедуру поправок на множественное тестирование, что

позволяет уменьшить систематическое накопление ложноположительных срабатываний алгоритма. Для проведения экспериментов с дообучением нейросетевой модели собран набор данных для русского и английского языков, включающий в себя различные по тематикам документы. Проведены эксперименты с конкатенацией вручную собранных признаков с выходом нейросетевого кодировщика, дообучением энкодер-декодер модели, мультиязычным обучением и интеграцией в сгенерированные части наборов данных фрагментов с переводами и перефразированием. Последний эксперимент показал наибольший прирост качества относительно базового решения, по метрике F1-score для русского языка удалось добиться отметки 0.968, а для английского – 0.825. Это объясняется способностью нейросетевой модели становиться более устойчивой к состязательным атакам, так как большое количество текстов порождаются именно подобными способами – перефразированием или переводом.

СПИСОК ЛИТЕРАТУРЫ

1. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin*. Attention Is All You Need, CoRR. 2017.
2. *ChatGPT by OpenAI*. Available at: https://chat.openai.com.
3. *Jasper*. Available at: https://www.jasper.ai
4. *Google Bard*. Available at: https://bard.google.com/?hl=ru
5. *GigaChat by SberDevices*. Available at: https://developers.sber.ru/portal/products/gigachat
6. *YaGPT by Yandex*. Available at: https://yandex.ru/project/alice/yagpt
7. *Lenta.ru* Москвич защитил написанный нейросетью диплом. Доступ по ссылке: https://lenta.ru/news/2023/02/01/neiroset/
8. *Николаев В.В., Рахконен М.Е.* Применение различных инструментов и использование чат-бота “ChatGpt” при написании научных работ, проверяемых в программе “Антиплагиат”, Профессиональное юридическое образование и наука. 2023. Т. 1 (9). С. 78–81.
9. *Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, Hai Hu* ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models, arXiv, 2023.
10. *Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, Xiaozhong Liu*. AI vs. Human – Differentiation Analysis of Scientific Content Generation, arXiv, 2023.
11. *Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov.* Unsupervised Cross-lingual Representation Learning at Scale, CoRR. 2019.
12. *Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, Xuanjing Huang.* A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models, arXiv, 2023.
13. *Badaskar Sameer, Agarwal Sachin, Arora Shilpa*. Identifying Real or Fake Articles: Towards better Language Modeling, Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II, 2008.
14. *Yoav Freund, Robert E. Schapire*. A Short Introduction to Boosting, 1999.
15. *Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea*. Automatic Detection of Fake News, CoRR. 2017.
16. *Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, Lichao Sun.* A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, 2023.
17. *Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, Douglas Eck*. Automatic Detection of Generated Text is Easiest when Humans are Fooled, CoRR. 2019.
18. *Ganesh Jawahar, Muhammad Abdul-Mageed, Laks V.S. Lakshmanan*. Automatic Detection of Machine Generated Text: A Critical Survey, CoRR. 2020.
19. *Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov.* RoBERTa: A Robustly Optimized BERT Pretraining Approach, CoRR. 2019.
20. *Zhong Wanjun, Tang Duyu, Xu Zenan, Wang Ruize, Duan Nan, Zhou Ming, Wang Jiahai, Yin Jian*. Neural Deepfake Detection with Factual Structure of Text, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.
21. *Gritsay German, Grabovoy Andrey, Chekhovich Yury.* Automatic Detection of Machine Generated Texts: Need More Tokens, 2022 Ivannikov Memorial Workshop (IVMEM). 2022.
22. *Hans W.A. Hanley, Zakir Durumeric.* Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites, CoRR. 2023.
23. *Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, CoRR. 2020.
24. *Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel.* mT5: A massively multilingual pre-trained text-to-text transformer, CoRR. 2020.
25. *Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, Bhiksha Raj*. GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content, CoRR. 2023.
26. *Sture Holm*. A Simple Sequentially Rejective Multiple Test Procedure, Board of the Foundation of the Scandinavian Journal of Statistics, Wiley, Volume 6. 1979.
27. *Yoav Benjamini, Yosef Hochberg.* Controlling the False Discovery Rate: A Practical and Powerful Approach to

Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological), Volume 57. 1995.

28. *Rodriguez Juan Diego, Hay Todd, Gros David, Shamsi Zain, Srinivasan Ravi.* Cross-Domain Detection of GPT-2-Generated Technical Text, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.

29. *Open access dataset for machine-generated text detection in Russian* Available at: https://data.mendeley.com/datasets/4ynxfp3w53/1.

30. *Answers scraped from Yandex Q.* Available at: https://huggingface.co/datasets/its5Q/yandex-q

31. *Dataset of ChatGPT-generated instructions in Russian*. Available at: https://huggingface.co/datasets/IlyaGusev/ru_turbo_alpaca

32. *Dataset of ChatGPT-generated chats in Russian*. Available at: https://huggingface.co/datasets/IlyaGusev/ru_turbo_saiga

33. *Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, Yue Zhang.* Deepfake Text Detection in the Wild, arXiv. 2023.

34. *Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, Jianfeng Gao.* Instruction Tuning with GPT-4, arXiv. 2023.

35. *Guo Biyang, Zhang Xin, Wang Ziyuan, Jiang Minqi, Nie Jinran, Ding Yuxuan, Yue Jianwei, Wu Yupeng.* How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection, arXiv. 2023.

36. *Khachay M.Yu., Konstantinova N., Panchenko Al., Ignatov, Dmitry I., Labunets V.G.* Morphological Analyzer and Generator for Russian and Ukrainian Languages, Springer International Publishing. 2015.

37. *Edward Loper, Steven Bird* NLTK: The Natural Language Toolkit, CoRR. 2002.

38. *Google Translate* Available at: https://translate.google.com/?hl=ru

39. *Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, Chelsea Finn.* DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, arXiv. 2023.

40. *Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, Bimal Viswanath.* Deepfake Text Detection: Limitations and Opportunities, arXiv. 2022.

41. *Paraphraser for Russian sentences* Available at: https://huggingface.co/cointegrated/rut5-base-paraphraser

42. *Paraphraser for English sentences* Available at: https://huggingface.co/ramsrigouthamg/t5_sentence_paraphraser

ARTIFICIALLY GENERATED TEXT FRAGMENTS SEARCH IN ACADEMIC DOCUMENTS

G. M. Gritsay[a,b], A. V. Grabovoy[a,b,c], A. S. Kildyakov[a], and Yu. V. Chekhovich[a,c]

[a] *Antiplagiat Company, Moscow, Russia*

[b] *Moscow Institute of Physics and Technology (National Research University), Moscow, Russia*

[c] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia*

Presented by Academician of the RAS A.L. Semenov

Recent advances in text generative models make it possible to create artificial texts that look like human-written texts. A large number of methods for detecting texts obtained using large language models have already been developed. But the improvement of detection methods occurs simultaneously with the improvement of generation methods. Therefore, it is necessary to explore new generative models and modernize existing approaches to their detection. In this paper, we present a large analysis of existing detection methods, as well as a study of lexical, syntactic and stylistic features of the generated fragments. Taking into account the developments, we have tested the most qualitative, in our opinion, methods of detecting machine-generated documents for their further application in the scientific domain. Experiments were conducted for Russian and English languages on the collected datasets. The developed methods improved the detection quality to a value of 0.968 on the F1-score metric for Russian and 0.825 for English, respectively. The described techniques can be applied to detect generated fragments in scientific, research and graduate papers.

Keywords: machine-generated text, natural language processing, multiple hypothesis testing, paraphrasing, detection of generated texts