УДК 004.032.26 **DOI** 10.15622/ia.24.1.10

А.Н. Голубинский, А.А. Толстых, М.Ю. Толстых

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ АННОТАЦИЙ НАУЧНЫХ СТАТЕЙ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Голубинский А.Н., Толстых А.А., Толстых М.Ю. **Автоматическая генерация аннотаций научных статей на основе больших языковых моделей.**

Аннотация. Предложена концепция автоматизации процесса аннотирования научных материалов (русскоязычных научных статей) и выполнена ее практическая реализация посредством технологий машинного обучения, дообучения больших языковых моделей. Обозначена актуальность корректного и рационального составления аннотаций, выделена проблематика, касающаяся установления баланса между затратами времени на аннотирование и обеспечением соблюдения ключевых требований к аннотации. Проанализированы основы аннотирования, представленные в семействе стандартов по информации, библиотечному и издательскому делу, приведены классификация аннотаций и требования к их наполнению и функционалу. Схемографически представлено существо и содержание процесса аннотирования, типовая структура объекта исследования. Проанализирован вопрос интеграции в процесс аннотирования цифровых технологий, особое внимание уделено преимуществам внедрения машинного обучения и технологий искусственного интеллекта. Кратко описан цифровой инструментарий, применяемый для генерации текста в приложениях обработки естественного языка. Отмечены его недостатки для решения поставленной в данной научной статье задачи. В исследовательской части обоснован выбор модели машинного обучения, применяемый для решения задачи условной генерации текста. Проанализированы существующие предобученные большие языковые модели и с учетом постановки задачи и имеющихся ограничений вычислительных ресурсов выбрана модель ruT5-base. Приведено описание датасета, включающего научные статьи из журналов, включенных в перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней кандидата и доктора наук. Охарактеризована методика разметки данных, основанная на работе токинезатора предобученной большой языковой модели, графически и таблично приведены численные характеристики распределений датасета и параметры конвейера обучения. Для оценки модели использована метрика качества ROUGE, для оценки результатов – метод экспертных оценок, включающий грамматику и логику в качестве базовых критериев. Качество автоматической генерации аннотаций сопоставимо с реальными текстами, отвечает требованиям информативности, структурированности и компактности. Статья может представлять интерес для аудитории ученых и исследователей, стремящихся оптимизировать свою научную деятельность в части интеграции в процесс написания статей инструментов цифровизации, а также специалистам, занимающимся обучением больших языковых моделей.

Ключевые слова: аннотация, генерация, большие языковые модели, цифровизация, машинное обучение.

1. Введение.

Научные публикации являются важным источником сведений и знаний в области академических и прикладных исследований и разработок. Когда научные материалы опубликованы, первая часть, с которой начинается ознакомление читателей, после

самого названия и сведений об авторах, – это, как правило, аннотация. Она представляет собой краткое изложение статьи, которое должно передавать емкое и лаконичное сообщение, являть сжатый обзор всей статьи и излагать ее суть.

Зачастую аннотация научной работы составляется в конце ее подготовки и оформления, когда у автора сформировалось четкое представление о существе, ходе и итогах исследования, уверенность в его завершенности [1, 2]. При этом автор готов обозначить корректную характеристику темы научной работы, ее проблемы, выделить объект и предмет, цели и задачи, а также указать результаты решения обозначенной проблемы в выбранной предметной области.

Корректно написанная аннотация может служить одновременно нескольким целям: позволить читателям оперативно понять суть научного материала, чтобы решить, ознакомиться ли с ним целиком; настроить внимание респондентов к тому, чтобы следить за ходом представления сведений, анализом и аргументацией в тексте научного исследования; помочь читателям запомнить ключевые аспекты научного материала.

Процесс аннотирования является важной задачей как для авторов-исследователей, так и других потребителей научного контента. При этом можно отметить сопутствующую проблематику, заключающуюся в отсутствии универсальных методов аннотирования, субъективности автора-составителя к реализации требований релевантности и точности аннотации, трудоемкости и времязатратности самого процесса, семантической неоднозначности результатов.

Наиболее разумным в данном контексте видится способ решения указанных затруднений, заключающийся в применении гибридных методов составления аннотации: комбинирование ручного аннотирования и автоматизированных методов, базирующихся на использовании современных цифровых технологий. Цифровой инструментарий, например, в виде технологий искусственного интеллекта, может выполнять первоначальное аннотирование, которое затем будет проверяться и корректироваться автором научной публикации, что позволит снизить временные затраты на составление текста и повысить качество результата.

2. Стандартизация и основы аннотирования. В России действует ряд стандартов, устанавливающих требования к содержанию, построению и оформлению текста аннотации. Ввиду тематической ориентированности данного исследования на аннотирование именно научных статей, рассмотрим стандарты из

семейства Системы стандартов по информации, библиотечному и издательскому делу [3 – 6].

В целом, положения [3] и [4] нормативных технических документов идентичны, за исключением некоторых особенностей. Ниже приведены результаты сравнительного анализа нормативных документов (рисунок 1), а также расширенная классификация аннотаций (рисунок 2).

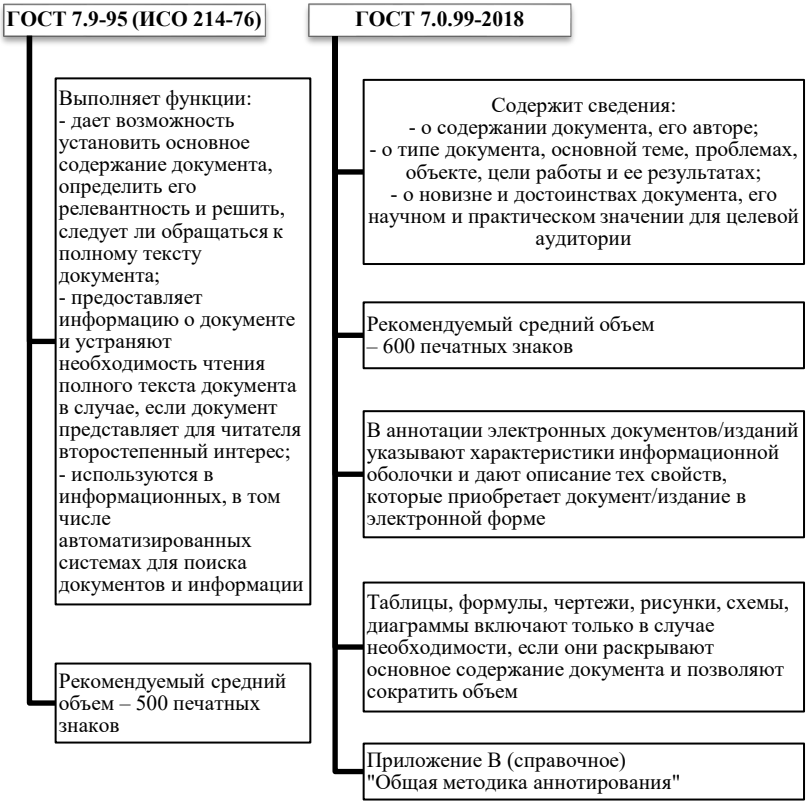

Рис. 1. Сравнительный анализ нормативных технических документов, регламентирующих построение и оформление текста аннотации (индикативного реферата) на документ (ключевые различия)

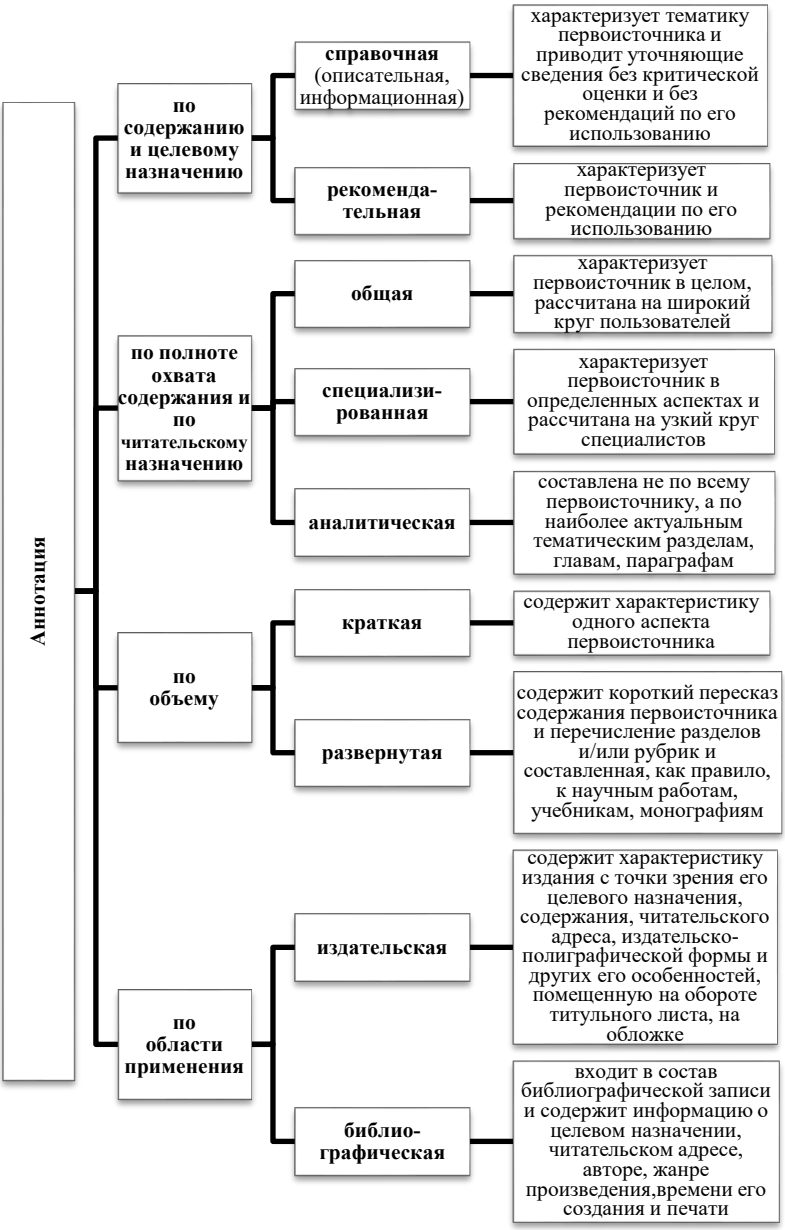

Рис. 2. Классификация и виды аннотаций по ГОСТ Р 7.0.99-2018

Под аннотацией на рисунке 2 понимается краткая характеристика первичного документа с точки зрения его назначения, содержания, вида, формы и других особенностей.

Таким образом, редакция [4] шире и детальней, кроме того, она содержит отдельное приложение, в котором раскрывается методика аннотирования, ключевые моменты которой для удобства восприятия приведены графически (рисунок 3).

Рис. 3. Основы методики аннотирования согласно требованиям нормативных технических документов

Можно также отметить, что аннотация должна представлять собой краткое отдельное резюме статьи, состоящее из нескольких предложений по каждому из следующих ключевых моментов в вопросных формах (рисунок 4).

Рис. 4. Типовая структура аннотации

Эмпирически установлены и логически обоснованы некоторые запреты в содержании аннотации. Так, например [7], не следует повторять текст самой статьи (исключить перенос предложений из основного текста научного материала), а также ее название. В тексте аннотации не должны приводиться таблицы, внутритекстовые сноски, обилие цифр. Следует избегать синтаксических конструкций, несвойственных языку научных и технических документов, нецелесообразно применять сложные грамматические конструкций, вводные слова, общие формулировки.

Стоит также отметить, что многие поисковые системы и библиографические базы данных используют аннотации вместе с заголовками научных статей для определения ключевых терминов в процессе индексации опубликованных научных трудов.

3. Интеграция цифровых технологий в процесс аннотирования. Цифровизация аннотирования представляет собой процесс применения цифровых технологий [8 – 10] для улучшения, автоматизации и ускорения процесса создания аннотаций и последующей работы с ними (рисунок 5).

Рис. 5. Основные аспекты цифровизации аннотирования

С технической стороны процесс аннотирования включает выделение ключевых элементов исходного материала (меток) и формирование фактически метаданных к тексту. Указанные процедуры могут быть автоматизированы и оптимизированы посредством использования современных передовых цифровых технологий – машинного обучения, в частности в приложениях обработки естественного языка (Natural language processing, NLP) и компьютерного зрения [11].

В динамично трансформирующемся ландшафте академических исследований инструменты на базе искусственного интеллекта совершают революцию в мастерстве написания текста. В России лидируют ChatGPT, YandexGPT2 [12], justGPT, GigaChat, которые предлагают авторам-исследователям эффективные способы свести обширные тексты научных материалов в краткие изложения, сэкономить время, улучшить качество контента и избежать плагиата. Также доступно использование приложений и расширений браузеров (например, Hypothesis, Kami), платформ для коллективной работы (например, Google Docs, Overleaf) и менеджеров/приложений для оркестрации процесса аннотирования в ручном формате (например, Zotero, Mendeley, Evernote).

В целом, функционал указанных сервисов заключается в автоматическом извлечении метаданных из добавленных текстов или файлов, настройке фильтрации для улучшения семантического поиска,

работе с библиографическими данными. Однако в своем большинстве они являются платными, не адаптированы под нюансы отечественного научного знания: высока вероятность некорректного извлечения метаданных из русских источников в отсутствии унифицированных международных идентификаторов; неполноценный перевод частей интерфейса и технической документации к сервисам усложняет их использование для русскоязычных пользователей. Более того, указанные сервисы являются закрытыми с точки зрения информации о моделях, датасетах и методике обучения, что не позволяет провести корректное сравнение с открытыми решениями.

Кроме того, применение указанного цифрового инструментария затрагивает вопросы этики научных исследований и академического мошенничества [13]. Последнее включает в себя преднамеренные попытки обмана и плагиат, фабрикацию данных, искажение исторических источников, подделку доказательств, заказ работ, выдачу чужих работ за свои, так называемую «двойную» сдачу материалов (например, одной и той же статьи в несколько редакций различных журналов), выборочное сокрытие нежелательных или неприемлемых результатов и кражу идей. Например, плагиат может появляться при генерации текстов с помощью больших языковых моделей, выдаваемых за оригинальные результаты. Выполнение работ на заказ упрощается за счёт автоматизации и сокращения времени на написание текста работы и обзора литературы. «Двойная» сдача материалов может реализовываться за счёт быстрого автоматического перефразирования словесных конструкций исходного материала без изменения семантической составляющей (гипотез, результатов экспериментов, выводов и т.д.). Научным и академическим сообществом отмечаются риски недобросовестного применения цифровых технологий в отношении развития научного знания, вместе с тем принимаются соответствующие меры реагирования в виде так называемых «карательных» (применение строгих санкций к нарушителям) и ценностных (разработка и внедрение этических кодексов, пропаганда этичного научного поведения и формирования честной академической среды). Очевидно соблюдение баланса между использованием аппарата цифровой трансформации и интеллектуальной авторской деятельностью.

Таким образом, применение передовых цифровых решений может способствовать обеспечению ясности, грамматической точности и актуальности текста, в том числе аннотации, удовлетворяя широкий спектр академических потребностей.

4. Исследовательская часть. Выбор модели обучения. Задача аннотирования научных статей является задачей условной генерации текста, т.е. создания последовательности слов (символов) на основе заданного контекста, тематики или условий [14 – 21].

Как правило, условная генерация текста может быть реализована с помощью двух базовых подходов. Первый заключается в использовании предопределенного шаблона для генерации текста на основе различных входных данных. Например, используя предопределенный шаблон, искусственный интеллект может сгенерировать определенное описание продукта на основе типа продукта, его характеристик и преимуществ. Второй способ использует метод неконтролируемого обучения, называемый глубоким обучением, который изучает нюансы языковых структур и функционирует с условием наличия больших объемов входных данных. Данный алгоритм более гибкий, может генерировать более естественный язык по сравнению с подходом на основе шаблонов.

Условная генерация текста имеет широкие практические применения в различных отраслях. К основным областям использования относятся: создание тематического контента (статьи, описания), поддержка клиентов с использованием чат-ботов; перевод (предоставление оперативных и точных интерпретаций посредством анализа входного языка и применения соответствующей языковой структуры и правил использования) и др.

В настоящее время существует несколько предобученных больших языковых моделей (large language model, LLM), предназначенных для условной генерации, в связи с чем решение поставленной в работе задачи сводится к дообучению (finetuning) одной из предобученных LLM. В работе [14] было представлено семейство LLM, предобученных на корпусе текстов, б**о**льшая часть которого являлась текстами на русском языке.

В наборе LLM [14] описаны следующие модели для условной генерации из семейства LLM T5: ruT5-base и ruT5-large. Выбор данного семейства моделей обусловлен ограниченностью вычислительных ресурсов: из открытых источников известно, что модели с большим числом параметров, например Llama 2 [15], GPT-3 [16] и подобные, показывают более высокие результаты, однако требуют гораздо больше вычислительных мощностей для обработки запросов (например, дообучение Llama 2 требует около 112 GB GPU в режиме fp32, GPT-3 – около 80 GB GPU, в то время как ruT5-base – около 18 GB GPU). Кроме того, рассмотренные модели имеют тенденцию создавать текст, который чрезмерно повторяется или не

отражает нюансы человеческого языка, так как обучены на корпусе, преимущественно, содержащего тексты на английском языке. В рамках эксперимента были доступны 16 GB GPU. В связи с этим была выбрана модель ruT5-base, содержащая 222×10^6 параметров (весов), является моделью трансформера [20] для русского языка, состоит из энкодера и декодера, решает задачу генерацию текстов, а также может быть обучена на широком списке NLP-задач.

Описание датасета и методики разметки данных. Для решения задачи аннотирования научных статей необходимо подготовить соответствующий датасет: пары «текст статьи» – «аннотация». Для уменьшения размера датасета принято решение ограничения предметной области научных статей: экономка и юриспруденция, педагогика, а также технические науки в контексте правоохранительной деятельности.

В качестве репозитория научных статей выбраны выпуски за последние 5 лет следующих журналов: Вестник Московского университета МВД России имени В.Я. Кикотя, Вестник Краснодарского университета МВД России, Вестник Воронежского института МВД России. Издания являются научно-практическими журналами, освещающими актуальные проблемы образовательного процесса, общественных, технических (информационных) и гуманитарных наук. Их корреспондентами являются как именитые ученые, так и молодые авторы: ученые деятели, преподаватели, студенты (курсанты и слушатели) высших учебных заведений, научно-педагогические кадры, практические работники и служащие правоохранительных органов, интересующиеся актуальными проблемами научного знания, участвующие в процессе обмена информацией и ведения конструктивного научного диалога.

Первоначально размечено 825 научных статей из 22 томов. После первичной обработки датасета, заключающейся в объединении статей, аннотаций и метаданных (название журнала, название статьи), исключено 15 статей (2%), аннотации к которым отсутствуют. Таким образом, исходный датасет для обучения составляет 810 пар «текст статьи» – «аннотация».

Для дальнейшего контроля хода обучения исходный датасет разбит на обучающую и валидационную выборки в соотношении 80/20 (обучающая часть – 648 пар; валидационная – 162 пары).

На рисунке 6 приведено распределение длин статей (в символах, включая пунктуационные знаки) в датасете, на рисунке 7 – распределение длин аннотаций.

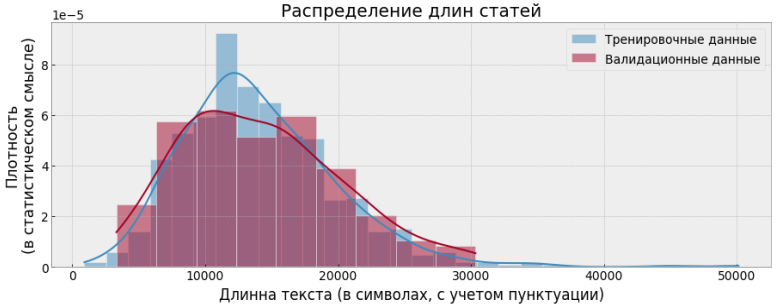

Рис. 6. Распределение длин статей в датасете

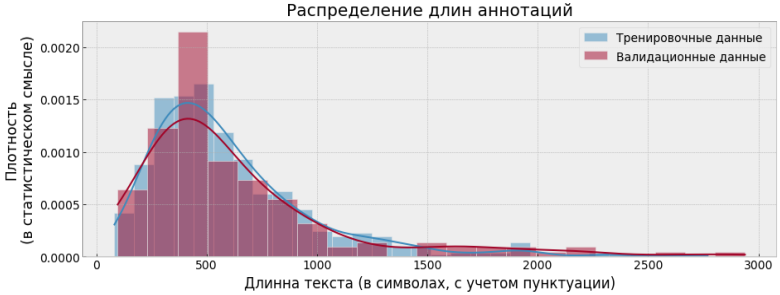

Рис. 7. Распределение длин аннотаций в датасете

Из рисунков 6 и 7 видно, что основная часть статей содержит не боле 25 000 символов, а аннотаций – 1 500 (оценка по 95% квантилю). Численные характеристики распределений (для всего датасета) приведены в таблице 1.

Таблица 1. Численные характеристики распределений датасета

Тип	**25% квантиль**	**Медиана**	**75% квантиль**	**95% квантиль**	**Математическое ожидание**
Текст	10197,5	13196,5	17356,5	24120,7	14043,6
Аннотация	349,7	510,5	741,2	1325,9	602,0

Подобное распределение обусловлено стандартизированным требованиями к размеру аннотации, о которых говорилось ранее, применяемыми научными журналами, в которых были размещены статьи. На рисунке 8 приведены наиболее часто встречающиеся слова в тексте статей, на рисунке 9 – в аннотациях.

Рис. 8. Наиболее частые слова в текстах статей

Рис. 9. Наиболее частые слова в аннотациях статей

Анализ наиболее частых слов в статьях и аннотациях к ним из датасета (рисунок 8 и 9) позволяет говорить об общей направленности текстов и совпадает с тематическими рубрикаторами исследуемых журналов.

Следующим этапом подготовки данных для обучения LLM является выбор максимального размера входных данных в токенах. Под токенами понимается результат алгоритма представления языковой сущности (слово, часть слова или отдельный символ) в виде целого числа (с добавлением нуля, $\mathbb{N} \cap \{0\}$), сам алгоритм в свою очередь называется «токенизатором» [17]. Для дообучения LLM необходимо использовать тот же токинизатор (алгоритм преобразования текста в численное представление) [17]. Токены вступают фундаментальными единицами информации, которые модели обрабатывают и производят. Эффективность модели часто

можно проследить по тому, насколько хорошо происходит преобразование токенов.

Благодаря использованию токинезатора предобученной LLM, получены распределения длин текстов и аннотаций, численные распределения которых приведены в таблице 2.

Таблица 2. Распределения длин текстов и аннотаций в токенах

Тип	Текст	Аннотация
40% квантиль	2596	103
50% квантиль	2805	103
60% квантиль	2807	103
70% квантиль	3343	103
80% квантиль	3334	118

Исходя из того, что на располагаемых вычислительных мощностях (16 GB GPU) не представляется возможным дообучить LLM со входом (текст) более 3000 токенов (требует более 16 GB GPU в режиме fp16), принято решение использовать методику ограничения длины входа (отбрасывания всех токенов после 3000). Верхняя граница длины входа LLM определяется с одной стороны ограниченностью вычислительных ресурсов, доступных для обучения, а с другой – статистическим распределением длин текстов в исходном датасете.

Для выхода модели применяется 128 токенов (недостающие токены заменяются специальным токеном <pad> [17]). Такое количество обеспечивает удовлетворительный размер аннотации научной статьи (3-4 предложения) и позволяет наиболее эффективно использовать имеющиеся данные для обучения. Под эффективностью в данном случае понимается то, что длинны аннотаций в датасете достаточно близко укладываются в 128 токенов (нет больших последовательностей <pad>), кратность степени двойки обусловлена архитектурой (следующий размер выхода – 256 токенов).

Конвейер обучения. Время обучения модели составляет 53 часа. Ее численные параметры приведены ниже:

- темп обучения (learning rate): 2×10^{-5};
- затухание весов (weights decay): 0.01;
- максимальная длинна входа (в токенах): 3000;
- максимальная длинна аннотации (в токенах): 128;
- параметры архитектуры T5 взяты без изменений из [14].

Мониторинг обучения происходит с помощью метрики ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Впервые данная метрика была предложена в [18], она является специализированной метрикой для задачи автоматической аннотации текстов. ROUGE основана на измерении пересечения между выходом модели (результатом условной генерации) и целевыми аннотациями, написанными человеком. Иными словами, производится подсчет совпадений слов и словосочетаний в сгенерированном тексте и в целевом, кроме того, метрика не чувствительна к регистру. Более высокие баллы (близкие к 1) указывают на эффективность с точки зрения сохранения ключевой информации из исходного текста при создании аннотации. На рисунке 10 приведены график изменения ROUGE при обучении модели.

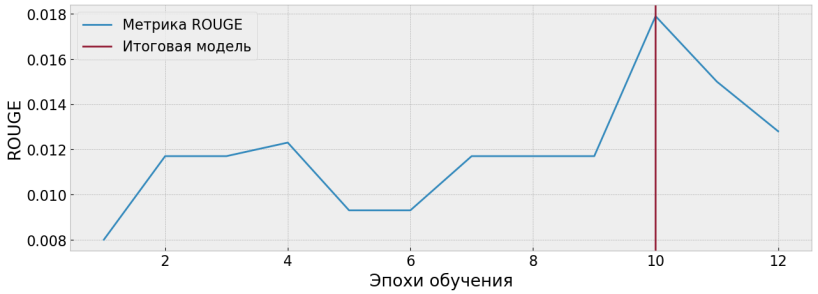

Рис. 10. Динамика дообучения модели

Алгоритм работы выбранной метрики заключается следующем. На этапе предварительной обработки сгенерированные аннотации анализируются для устранения любого шума или нерелевантной информации (например, знаков препинания, стоп-слов), которые могут помешать процессу оценки. Далее выполняется извлечение признаков, таких как n-граммы и прочие показатели сходства, которые получают как из сгенерированного системой текста, так и из исходных аннотаций, что обеспечивает основу для сравнения двух текстов. После этого с использованием различных методов, таких как статистика совместной встречаемости n-грамм, вычисление коэффициентов перекрытия слов, выполняется расчет оценок сходства путем сравнения признаков, извлеченных из сгенерированного моделью текста с признаками из исходных аннотаций.

Заключительными этапами алгоритма являются агрегация оценок сходства, нормализация и интерпретация. При агрегации отдельные оценки сходства, полученные для каждого типа признаков,

объединяются для получения единой оценки ROUGE, представляющей общую эффективность сгенерированного моделью текста аннотации. Окончательный балл ROUGE часто нормируется на отрезок [0, 1], при этом более высокие баллы указывают на более высокую эффективность модели с точки зрения сохранения ключевой информации из исходного текста.

Помимо метрики ROUGE, каждую эпоху проводится субъективная оценка на основе 5 сгенерированных пар «текст – аннотация», что позволяет дополнительно верифицировать результаты с точки зрения экспертной оценки. В таблице 3 приведена динамика данной оценки с 10 эпохи обучения, значимым являлся балл за логику (непротиворечивость и соответствие аннотации тексту исходной статьи), баллы усреднены по парам и экспертам.

Таблица 3. Динамика изменения оценки качества генерации в процессе обучения сети с 10 эпохи обучения

Эпохи	**Баллы**
10 эпоха	8,0 баллов
11 эпоха	7,8 баллов
12 эпоха	7,5 баллов

Методика оценки результатов. Оценка обучения модели – важный процесс, необходимый для подтверждения ее результативности и эффективности, качества и производительности. Профессиональное сообщество в сфере технологий искусственного интеллекта и машинного обучения не согласовало унифицированные требования к типовой методологической базе оценки моделей. В основном это связано с проблемами формализации измерений качества семантической информации.

Формальная оценка качества облегчается с помощью структурированного инструментария на основе эмпирических данных, уточненных экспертным консенсусом.

Для верификации результатов обучения модели, разработанной в данном исследовании, используется экспертная оценка [22]. В качестве субъектов экспертирования выступают обучающиеся старших курсов ведомственного вуза, у которых имеется опыт выполнения научных исследований, участия в научно-представительских мероприятиях, написания научных статей по тематикам из собранного датасета, что свидетельствует о приемлемом уровне экспертной компетентности и сопоставляется с

идентифицируемыми задачами оценки и измеримостью результатов. Оценка проводилась по валидационной выборке.

Отметим, что в данном контексте толкование дефиниции «эксперт» относится к пониманию лица как деятельного субъекта, включенного в механизмы принятия решений. Обучающиеся, обладающие правами и возможностями принятия заключений относительно вопросов экспертирования, могут не являться специалистами и профессионалами в оцениваемой области, но будут реализовывать ролевые экспертные роли согласно условиями и критериям процесса оценивания результатов обучения модели. Такой подход также реализует концепцию студентоориентированности [23], актуальную для отечественной системы образования, при которой понимание студента сводится не просто к его идентификации как штатного участника образовательного процесса, а индивидуального субъекта, продуцирующего систему рефлексии в рамках единого общественно значимого процесса воспитания и обучения в интересах человека, семьи, общества и государства.

Предлагается методика оценки эффективности модели на основе двух критериев: оценка грамматики и оценка логики. Каждой аннотации эксперты выставляют оценку по 10 бальной шкале. Под грамматикой в данном контексте понимаются любые синтаксические, грамматические и иные ошибки, позволяющие идентифицировать аннотацию как сгенерированную. Например, грамматической будет являться ошибка изменения алфавита посередине слова (кириллица / латиница), некорректное написание слов, отсутствие пробелов и т. д.

В качестве примера, в приведенной ниже аннотации (автоматически сгенерированной) полужирным шрифтом выделены грамматические ошибки:

Рассматривается личность преступника как базовый элемент криминалистической характеристики преступлений, **совершаемых террористической направленности**, с точки зрения надлежащего субъекта преступления, а также его мотиваци**ю**, целеполагани**е**.

В приведенной аннотации отсутствует согласованность в спряжении слов, ошибки в окончаниях. Данная аннотация была оценена 18 экспертами в среднем в 6,1 балл по показателю грамматика.

Второй параметр оценки – логика, показывающая семантическую корректность аннотации, а также соответствие

аннотации тексту исходной статьи. Для оценивания данного параметра в распоряжении экспертов имеются исходные тексты статей.

В оценке принимает участие 51 эксперт, каждый из которых оценивает 40 пар «статья-аннотация». Эксперты осведомлены, что каждый из наборов содержит реальные аннотации. Для каждого эксперта составлен набор из 20 реальных и 20 сгенерированных аннотаций, перемешанных в случайном порядке. Информация об источнике конкретной аннотации (реальная / сгенерированная) экспертам недоступна. Вместе с тем общий набор данных для оценки содержит всего 200 пар «статья-аннотация».

Таким образом, каждую пару «статья-аннотация» оценивают, в среднем, 20 экспертов. Подобный подход позволяет минимизировать единичные ошибки экспертов, дает более точную оценку качества сгенерированных текстов.

Ниже приведены сгенерированные аннотации (сохранена орфография и пунктуация) с высокой средней оценкой экспертов (выше 8,5 баллов).

В статье рассматриваются предпосылки возникновения и развития азартных игр, а также особенности их организации и проведения. Анализируются законодательные акты Российской Федерации, регулирующие деятельность по организации и проведению игорных заведений. Формулируются предложения по совершенствованию законодательства в данной сфере.

Рассматриваются вопросы защиты прав и свобод человека и гражданина. Формулируются предложения по совершенствованию механизма реализации конституционных обязанностей граждан РФ. Предлагается классификация способов защиты гражданских прав по материально-правовым основаниям: репрессивные, пресекательные, восстановительные и компенсационные.

В статье рассмотрен вопрос создания математической модели поддержки процесса временного перераспределения трудовых ресурсов в проектно-ориентированных организационных системах. Предложен алгоритм, позволяющий автоматизировать предварительный отбор кандидатов на роли в проектах, выполняемых в проектных организациях. Предложены алгоритмы распределения участников по проектам при достаточно большом количестве проектов и составляющих их операций.

Исследуются проблемы, возникающие при реализации административного надзора за лицами, судимыми за насильственные преступления против половой неприкосновенности несовершеннолетних. Выявлены основные направления деятельности участкового уполномоченного полиции в сфере профилактики преступлений и других правонарушений.

В статье рассмотрены некоторые области деятельности органов внутренних дел по использованию технологии больших данных (Big Data). Рассмотрены некоторые проблемы и вызовы, которые могут быть затруднины при использовании технологий больших данных.

Как видно из сгенерированных текстов в примерах, присутствуют незначительные орфографические ошибки, которые могут быть исправлены в автоматическом режиме с помощью компьютерных программ, осуществляющих проверку заданного текста на предмет наличия в нем орфографических, пунктуационных, а также стилевых ошибок.

Следует отметить, что несмотря на орфографические ошибки, семантическое содержание приведенных примеров аннотаций является корректным.

Оценка результатов. Результаты оценки модели до дообучения неудовлетворительные: по критерию «грамматика» – 7,8±1,33 балла, логика – 1,52±0,79 балла. Для более удобного восприятия информации распределение оценок экспертов представлено в виде диаграммы «ящик с усами» (boxplot) на рисунке 11 [24].

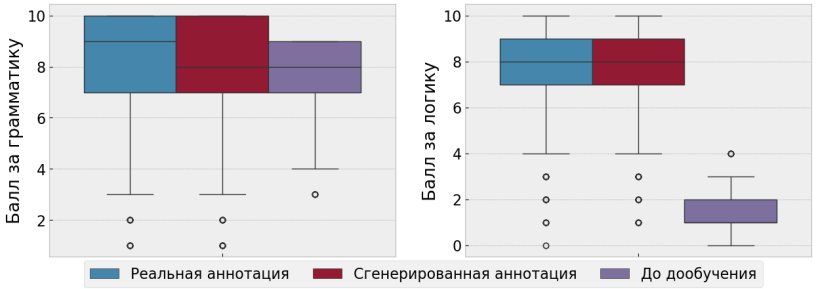

Рис. 11. Диаграмма «ящик с усами» для оценок экспертов

Из анализа рисунка 10 можно сделать следующий вывод: распределения оценок для реальных и сгенерированных аннотаций практически неотличимы.

Для подтверждения выдвинутого тезиса о неразличимости распределений проведён статистический тест Колмогорова-Смирнова для гипотезы о том, что выборки взяты из одного распределения вероятностей [25]. Для оценок грамматики p-value составляет 0,842; для оценок логики – 0,941. Таким образом, статистический тест подтверждает факт статистической неразличимости оценок качества реальных и сгенерированных аннотаций.

На рисунке 12 приведена альтернативная визуализация оценок экспертов.

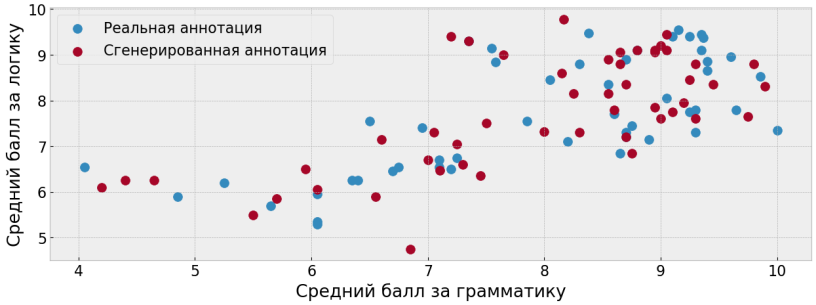

Рис. 12. Скаттерограмма оценок экспертов, усреднённых для каждой пары «статья-аннотация»

На рисунке 12 цветом обозначен источник аннотации – реальная или сгенерированная. Точки разных цветов сильно перемешаны, что не позволяет провести четкую классификацию в данных координатах. Данный факт позволяет говорить о сопоставимости качества сгенерированных и реальных аннотаций.

Рисунок 13 визуализирует результаты в разрезе по областям науки. В валидационных данных содержались 6 областей науки: юриспруденция, педагогика, информационная безопасность, психология, социология, история.

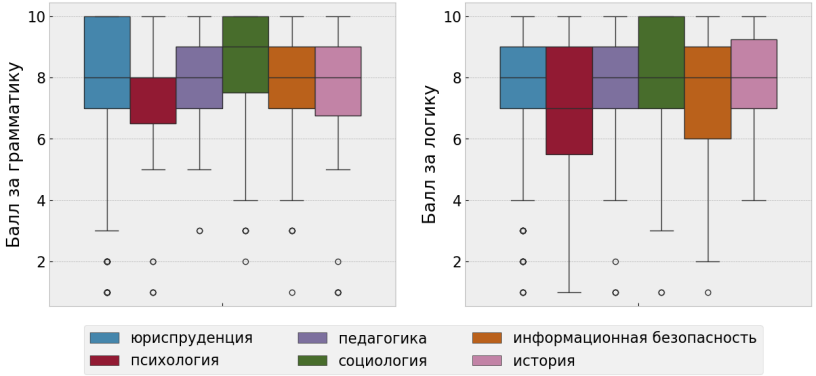

Рис. 13. Диаграмма «ящик с усами» для оценок экспертов в разрезе по областям науки

Из рисунка видно, что модель генерирует аннотации к статьям из различных областей науки с одинаковым качеством. Таким образом, для расширения области применения модели целесообразно обогащать датасет парами «текст-аннотация» из различных областей науки. Однако, вопрос поведения модели при постепенном расширении датасета необходимо исследовать отдельно – существует ли граница, после которой разнообразие предметных областей начнёт ухудшать качество генерации? Данный вопрос является темой дальнейших исследований.

На рисунке 14 приведены результаты оценки модели на дополнительных тестовых данных. Тестовые данные были собраны после обучения, они не учувствовали ни в обучении, ни в валидации. Объем дополнительных тестовых данных – 112 пар «аннотация – текст», источники, методика сбора и оценки аналогичны источникам и методикам для тренировочных и валидационных данных.

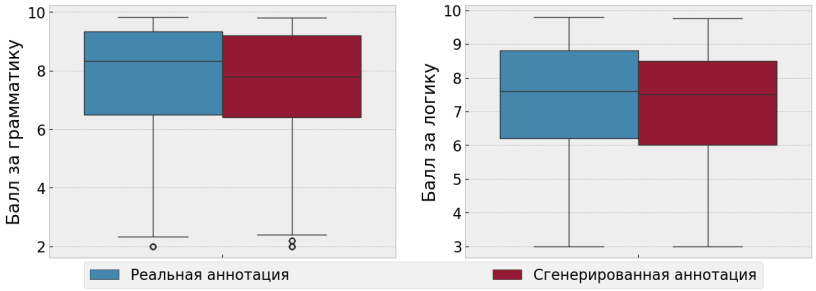

Рис. 14. Диаграмма «ящик с усами» для оценок экспертов на тестовых данных

Оценка на новых данных практически не отличается от оценки на валидации. Таким образом, можно говорить о достижении сетью возможности обобщения, а не только заучивания обучающих данных.

5. Заключение и выводы. В данной статье проработан вопрос совершенствования процесса аннотирования научных статей, имеющий высокую актуальность ввиду очевидной необходимости оптимизации способов составления кратких характеристик первичных научных документов с учетом их особенностей (назначения, содержания, формы). Необходимость также подтверждается увеличением объемом научных сведений, при которых значение аннотации определяется одной из ее функций – помощь исследователям в оперативном нахождении релевантного научного материала и извлечение ключевой информации. Аннотирование научных статей также способствует структурированию знаний,

выделению ключевых идей и результатов, а в условиях общего возрастания объемов научной информации позволяет проводить систематизацию и управление научными знаниями посредством навигационных репозиториев и баз данных.

Автоматизация процесса составления аннотации в современных реалиях должна производиться на базе цифровой трансформации. Так, технология блокчейн может быть интегрирована в процесс составления аннотации для обеспечения прозрачности, защиты интеллектуальной собственности и гарантии подлинности и целостности представленных в научных трудах сведений. Инструментарий предобученных больших языковых моделей позволяет значительно сократить время, необходимое для анализа больших массивов научных текстов, повысить точность и корректность извлечения информации.

В данном научном исследовании решена задача автоматической генерации аннотаций к научным статьям. Качество генерации сопоставимо с реальными аннотациями, а также отвечает требованиям информативности, структурированности и компактности, которые отмечены в действующих стандартах по издательскому оформлению статей в печатных и электронных научных, периодических и продолжающихся сборниках. Сгенерированные тексты согласуются с типовой структурой аннотации: содержат справочную информацию, цель, описание подходов, результатов и краткие выводы. Они реализуют задачу по отображению существенных признаков содержания научных трудов, позволяющих выявить их научное, теоретическое или практическое значение для целевой аудитории, новизну, отличить конкретный научный материал от других, аналогичных по тематике и целевому назначению, представляют информацию о достоинствах статей. Также полученные тексты аннотаций выполняют установленные стандартами функции: дают возможность установить основное содержание документа, определить его релевантность; предоставляют базовые сведения о научной статье, устраняют необходимость чтения полного текста документа; могут быть использованы в системах поиска документов и информации.

Основным элементом решения является собранный и размеченный датасет, который позволяет провести дообучение базовой языковой модели. Датасет состоял из 825 научных материалов, подготовленных по тематике решения актуальных проблем образовательного процесса, общественных, технических (информационных), гуманитарных, экономических и юридических наук.

Эффективность генерации модели верифицируется с помощью предложенной методики экспертной оценки на основе балльной системы от 1 до 10 с двумя параметрами: логика и грамматика. Под грамматикой понимаются любые синтаксические, грамматические и иные ошибки, позволяющие идентифицировать аннотацию как сгенерированную; под логикой – смысловая корректность аннотации. Обработка результатов экспертной оценки показала, что распределение оценок сгенерированных и реальных аннотаций статистически неразличимы, что свидетельствует о высоком качестве генерации языковой модели.

Разработка внедрена в учебный процесс государственного вуза в виде программного продукта (веб-приложения), используемого в научном обеспечении и сопровождении образовательного процесса, оказывающего помощь в подготовке квалифицированных научных специалистов. Веб-приложение позволяет сформировать краткую характеристику научного материала с точки зрения его тематики, содержания, новизны и других особенностей. Работа основана на функционировании большой языковой модели архитектуры T5, дообученной на корпусе из тысячи размеченных научных публикаций, содержащих результаты научных и прикладных исследований в области экономики, юриспруденции, педагогики, а также технических наук в контексте правоохранительной деятельности

Дальнейшее исследование предполагает дообучение моделей и оценку сгенерированных аннотаций с точки зрения требований нормативных документов, а также рассмотрение дообучения мультиязычных больших языковых моделей для задачи генерации аннотаций к научным статьям на разных языках. Помимо этого, остаётся открытым вопрос исследования качества генерации модели при постепенном расширении датасета текстами из различных предметных областей.

В заключении целесообразно отметить, что концепция предлагаемой разработки позиционирует свое применение в качестве системы поддержки принятия решений. Крайне важно использовать результаты генерации в сочетании с собственными авторскими знаниями предметной области на базе персонального критического мышления, анализа и интерпретации данных.

Литература

1. Жмудь В.А. Методы научных исследований: учебное пособие. Москва: Ай Пи Ар Медиа. 2024. 344 с.
2. Мейлихов Е.З. Искусство писать научные статьи: научно-практическое руководство. Долгопрудный: Издательский Дом «Интеллект». 2020. 335 с.

3. ГОСТ 7.9-95 (ИСО 214-76). Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования // М.: Госстандарт России. 1995.
4. ГОСТ Р 7.0.99-2018 (ИСО 214:1976). Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования // М.: Госстандарт России. 2018.
5. ГОСТ 7.86-2003. Система стандартов по информации, библиотечному и издательскому делу. Издания. Общие требования к издательской аннотации // М.: Госстандарт России. 2003.
6. ГОСТ Р 7.0.7-2021. Система стандартов по информации, библиотечному и издательскому делу. Статьи в журналах и сборниках. Издательское оформление // М.: Госстандарт России. 2021.
7. Курицкая Е.В. Технология написания аннотации к техническому тексту // Актуальные вопросы современного языкознания и тенденции преподавания иностранных языков: теория и практика: Материалы III Всероссийской научно-практической конференции (Кострома, 20 октября 2022 г.). Кострома: Военная академия радиационной, химической и биологической защиты имени Маршала Советского Союза С.К. Тимошенко (г. Кострома) Министерства обороны Российской Федерации. 2023. С. 93–99.
8. Schmarzo B. The Economics of Data, Analytics, and Digital Transformation: The theorems, laws, and empowerments to guide your organization's digital transformation // Packt Publishing. 2020. 260 p.
9. Reinsel D., Gantz J., Rydning J. The Digitization of the World From Edge to Core // An IDC White Paper. 2018. 28 p.
10. Толстых М.Ю. К вопросу обеспечения процессов цифровой трансформации в системе обучения // Цифровая трансформация образования: современное состояние и перспективы: Сборник научных трудов по материалам II Международной научно-практической конференции (Курск, 17–18 ноября 2023 г.). Курск: Курский государственный медицинский университет, 2024. С. 439–442.
11. Хлыбова М.А. Цифровые технологии в обучении написанию аннотаций в магистратуре неязыкового вуза // Филологический аспект. 2023. № 05(22). С. 55–58.
12. Солдатенкова Ю.А. YandexGPT и ChatGPT: характеристика, сравнение и основные отличия нейросетей // Моя профессиональная карьера. 2023. Т. 3. № 55. С. 277–284.
13. Lal K., Sharma B. Research Integrity & Ethics Scientific Misconduct // National Seminar on Academic Integrity and Research Ethics. At: DIT University, Dehradun. 2023. pp. 129–143.
14. Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Kadulin V., Markov S., Shavrina T., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian: arXiv:2309.10931. arXiv. 2023.
15. Touvron H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models: arXiv:2307.09288. arXiv. 2023.
16. Brown T.B. et al. Language Models are Few-Shot Learners: arXiv:2005.14165. arXiv. 2020.
17. Tunstall L., Werra L. von, Wolf T. Natural Language Processing with Transformers, Revised Edition. 1st edition. Sebastopol: O'Reilly Media, Inc. 2022. 406 p.
18. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. Barcelona. 2004. pp. 74–81.

19. Ravenscroft J., Oellrich A., Saha S., Liakata M. Multi-label Annotation in Scientific Articles – The Multi-label Cancer Risk Assessment Corpus // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 4115–4123.
20. Sun J., Wang Y., Li Z. An Improved Template Representation-based Transformer for Abstractive Text Summarization // IEEE International Joint Conference on Neural Network. 2020. pp. 1–8.
21. Amusat O., Hegde H., Mungall C.J., Giannakou A., Byers N.P., Gunter D., Fagnan K., Ramakrishnan L. Automated Annotation of Scientific Texts for ML-based Keyphrase Extraction and Validation. arXiv.2311.05042. arXiv, 2023.
22. Гуцыкова С.В. Метод экспертных оценок: теория и практика. Москва: Издательство «Институт психологии РАН». 2011. 144 с.
23. Щеглов И.А. Роль студентоориентированного подхода в социализации экспертизы // Гуманитарный вестник. 2021. № 4(90). С. 1–15.
24. Уилке К. Основы визуализации данных. Пособие по эффективной и убедительной подаче информации. Москва: Бомбора, 2024. 352 с.
25. Иванов Б.Н. Теория вероятностей и математическая статистика: учебное пособие для вузов. Издание третье. Санкт-Петербург: Лань. 2024. 224 с.

Голубинский Андрей Николаевич — д-р техн. наук, доцент, и.о. заместителя директора по научной работе, Институт проблем передачи информации им. А.А. Харкевича Российской академии наук. Область научных интересов: машинное обучение, нейросетевое моделирование, автоматизированные системы управления с элементами искусственного интеллекта, обработка речевых сигналов. Число научных публикаций — 242. annikgol@mail.ru; Большой Каретный переулок, 19/1, 127051, Москва, Россия; р.т.: +7(495)650-2235.

Толстых Андрей Андреевич — канд. техн. наук, инженер-программист, ООО «РТК». Область научных интересов: искусственные нейронные сети, машинное обучение, обучение с подкреплением. Число научных публикаций — 63. tolstykh.aa@yandex.ru; проспект Высоковольтный, 1/49, 127566, Москва, Россия; р.т.: +7(910)242-7955.

Толстых Марина Юрьевна — канд. техн. наук, доцент, кафедра международной информационной безопасности, Московский государственный лингвистический университет; доцент кафедры, кафедры специальных информационных технологий учебно-научного комплекса информационных технологий, Московского университета МВД России им. В.Я. Кикотя. Область научных интересов: информационная безопасность, цифровая трансформация, машинное обучение. Число научных публикаций — 109. marina_lion@mail.ru; улица Коптевская, 63, 127299, Москва, Россия; р.т.: +7(920)440-3845.

DOI 10.15622/ia.24.1.10

A. GOLUBINSKIY, A. TOLSTYKH, M. TOLSTYKH

AUTOMATIC GENERATION OF SCIENTIFIC ARTICLES ABSTRACTS BASED ON LARGE LANGUAGE MODELS

Golubinskiy A., Tolstykh A., Tolstykh M. **Automatic Generation of Scientific Articles Abstracts Based on Large Language Models.**

Abstract. The concept of automation of the process of annotation of scientific materials (Russian-language scientific articles) is proposed and its practical implementation is carried out by means of machine learning technologies, and additional training of large language models. The relevance of correct and rational compilation of annotations is indicated, and the problems related to establishing a balance between the time-consuming process of annotation and ensuring compliance with key requirements for annotation are highlighted. The basics of annotation presented in the family of standards on information, librarianship, and publishing are analyzed, and the classification of annotations and requirements for their content and functionality is given. The essence and content of the annotation process, and the typical structure of the research object are presented schematically. The issue of integration of digital technologies into the annotation process is analyzed, and special attention is paid to the advantages of introducing machine learning and artificial intelligence technology. The digital toolkit used to generate text in natural language processing applications is briefly described. Its shortcomings for solving the problem posed in this scientific article are noted. The research part substantiates the choice of the machine learning model used to solve the problem of conditional text generation. The existing pre-trained large language models are analyzed and, considering the problem statement and existing limitations of computing resources, the ruT5-base model is selected. A description of the dataset is given, including scientific articles from journals included in the list of peer-reviewed scientific publications in which the main scientific results of dissertations for the degrees of candidate and doctor of science should be published. The data labeling technique based on the operation of the tokenizer of the pre-trained large language model is characterized, and the numerical characteristics of the dataset distributions and the parameters of the training pipeline are presented graphically and in tables. The ROUGE quality metric is used to evaluate the model, and the expert assessment method, including grammar and logic as basic criteria, is used to evaluate the results. The quality of automatic annotation generation is comparable to real texts and meets the requirements of information content, structure and compactness. The article may be of interest to an audience of scientists and researchers seeking to optimize their scientific activities in terms of integrating digitalization tools into the process of writing articles, as well as to specialists involved in training large language models.

Keywords: annotation, generation, large language models, digitalization, machine learning.

References

1. Zhmud' V.A. Metody nauchnyh issledovanij: uchebnoe posobie [Scientific research methods: textbook]. Moscow: Aj Pi Ar Media. 2024. 344 p. (In Russ.).
2. Mejlihov E.Z. Iskusstvo pisat' nauchnye stat'i: nauchno-prakticheskoe rukovodstvo [The art of writing scientific articles: a scientific and practical guide]. Dolgoprudnyj: Izdatel'skij Dom «Intellekt». 2020. 335 p. (In Russ.).
3. GOST R 7.9-95 (ISO 214-76). Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Referat i annotacija. Obshhie trebovanija [System of standards on

information, librarianship and publishing. Informative abstract and indicative abstract. General requirements]. M.: Gosstandart Rossii. 1995. (In Russ.).
4. GOST R 7.0.99-2018 (ISO 214:1976). Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Referat i annotacija. Obshhie trebovanija [System of standards on information, librarianship and publishing. Abstract and annotation. General requirements]. M.: Gosstandart Rossii. 2018. (In Russ.).
5. GOST 7.86-2003. Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Izdanija. Obshhie trebovanija k izdatel'skoj annotacii [System of standards on information, librarianship and publishing. Editions. General requirements for publishing annotations]. M.: Gosstandart Rossii. 2003. (In Russ.).
6. GOST R 7.0.7-2021. Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Stat'i v zhurnalah i sbornikah. Izdatel'skoe oformlenie [System of standards on information, librarianship and publishing. Articles in magazines and collections. Publishing design]. M.: Gosstandart Rossii. 2021. (In Russ.).
7. Kurickaja E.V. Tehnologija napisanija annotacii k tehnicheskomu tekstu [Technology for writing annotations for technical texts] Aktual'nye voprosy sovremennogo jazykoznanija i tendencii prepodavanija inostrannyh jazykov: teorija i praktika : Materialy III Vserossijskoj nauchno-prakticheskoj konferencii [Current issues of modern linguistics and trends in teaching foreign languages: theory and practice: Materials of the III All-Russian Scientific and Practical Conference]. Kostroma: Voennaja akademija radiacionnoj, himicheskoj i biologicheskoj zashhity imeni Marshala Sovetskogo Sojuza S.K. Timoshenko (g. Kostroma) Ministerstva oborony Rossijskoj Federacii. 2023. pp. 93–99. (In Russ.).
8. Schmarzo B. The Economics of Data, Analytics, and Digital Transformation: The theorems, laws, and empowerments to guide your organization's digital transformation. Packt Publishing. 2020. 260 p.
9. Reinsel D., Gantz J., Rydning J. The Digitization of the World from Edge to Core. An IDC White Paper. 2018. 28 p.
10. Tolstyh M.J. K voprosu obespechenija processov cifrovoj transformacii v sisteme obuchenija [On the issue of ensuring digital transformation processes in the education system]. Cifrovaja transformacija obrazovanija: sovremennoe sostojanie i perspektivy: Sbornik nauchnyh trudov po materialam II Mezhdunarodnoj nauchno-prakticheskoj konferencii [Digital transformation of education: current state and prospects: Collection of scientific papers based on the materials of the II International Scientific and Practical Conference]. Kursk: Kurskij gosudarstvennyj medicinskij universitet. 2024. pp. 439–442. (In Russ.).
11. Hlybova M.A. [Digital technologies in teaching annotation writing in a master's program at a non-linguistic university]. Filologicheskij aspekt – The philological aspect. 2023. no. 05(22). pp. 55–58. (In Russ.).
12. Soldatenkova J.A. [YandexGPT and ChatGPT: characteristics, comparison and main differences between neural networks]. Moja professional'naja kar'era – My professional career. 2023. vol. 3. no. 55. pp. 277–284. (In Russ.).
13. Lal K., Sharma B. Research Integrity & Ethics Scientific Misconduct. National Seminar on Academic Integrity and Research Ethics. At: DIT University, Dehradun. 2023. pp. 129–143.
14. Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Kadulin V., Markov S., Shavrina T., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian: arXiv:2309.10931. arXiv. 2023.
15. Touvron H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models: arXiv:2307.09288. arXiv. 2023.

16. Brown T.B. et al. Language Models are Few-Shot Learners: arXiv:2005.14165. arXiv. 2020.
17. Tunstall L., Werra L. von, Wolf T. Natural Language Processing with Transformers, Revised Edition. 1st edition. Sebastopol: O'Reilly Media, Inc. 2022. 406 p.
18. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. Barcelona. 2004. pp. 74–81.
19. Ravenscroft J., Oellrich A., Saha S., Liakata M. Multi-label Annotation in Scientific Articles – The Multi-label Cancer Risk Assessment Corpus. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 4115–4123.
20. Sun J., Wang Y., Li Z. An Improved Template Representation-based Transformer for Abstractive Text Summarization. IEEE International Joint Conference on Neural Network. 2020. pp. 1–8.
21. Amusat O., Hegde H., Mungall C.J., Giannakou A., Byers N.P., Gunter D., Fagnan K., Ramakrishnan L. Automated Annotation of Scientific Texts for ML-based Keyphrase Extraction and Validation. arXiv.2311.05042. arXiv, 2023.
22. Gucykova S.V. [Expert assessment method: theory and practice] Metod jekspertnyh ocenok: teorija i praktika. Moscow: Institut psihologii RAN. 2019. 144 p. (In Russ.).
23. Shheglov I.A. [The role of the student-centered approach in the socialization of expertise]. Gumanitarnyj vestnik – Humanitarian Bulletin. 2021. no. 4(90). pp. 1–15. (In Russ.).
24. Uilke K. Osnovy vizualizacii dannyh. Posobie po jeffektivnoj i ubeditel'noj podache informacii [Basics of data visualization. A Guide to Effectively and Persuasively Presenting Information]. Moscow: Bombora, 2024. 352 p. (In Russ.).
25. Ivanov B.N. Teorija verojatnostej i matematicheskaja statistika: uchebnoe posobie dlja vuzov[Probability theory and mathematical statistics: a textbook for universities]. The third edition. Sankt-Peterburg: Lan'. 2024. 224 p. (In Russ.).

Golubinskiy Andrey — Ph.D., Dr.Sci., Associate Professor, Acting deputy director for research, Institute for Information Transmission Problems (Kharkevich Institute) Russian Academy of Sciences. Research interests: machine learning, neural network modeling, automated control systems with artificial intelligence elements, speech signal processing. The number of publications — 242. annikgol@mail.ru; 19/1, Bolshoy Karetny Lane, 127051, Moscow, Russia; office phone: +7(495)650-2235.

Tolstykh Andrey — Ph.D., Software engineer, OOO "RTK". Research interests: artificial neural networks, machine learning, reinforcement learning. The number of publications — 63. tolstykh.aa@yandex.ru; 1/49, Vysokovoltny Av., 127566, Moscow, Russia; office phone: +7(910)242-7955.

Tolstykh Marina — Ph.D., Associate professor, Department of international information security, Moscow State Linguistic University; Associate professor of the department, Department of special information technologies of the educational and scientific complex of information technologies, Moscow University of the Ministry of Internal Affairs of Russia. V.Ya. Kikotya. Research interests: information security, digital transformation, machine learning. The number of publications — 109. marina_lion@mail.ru; 63, Koptevskaya St., 127299, Moscow, Russia; office phone: +7(920)440-3845.