ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В БИБЛИОТЕЧНОЙ ДЕЯТЕЛЬНОСТИ

УДК 004.93:02
https://doi.org/10.33186/1027-3689-2025-11-203-214

Особенности использования больших языковых моделей при составлении текстовых рефератов

М. В. Гончаров[1], К. А. Колосов[2]

[1, 2]ГПНТБ России, Москва, Российская Федерация

[1]goncharov@gpntb.ru
[2]kolosov@gpntb.ru

Аннотация. В условиях стремительного увеличения объёма издаваемой научной литературы автоматическое реферирование текстов с помощью технологий искусственного интеллекта становится актуальной задачей. Современные модели реферирования основаны на использовании предварительно обученных больших языковых моделей, развёртывание которых часто требует значительных аппаратных ресурсов. В то же время применение для реферирования текстов специализированных моделей, основанных на той же архитектуре трансформеров, не требует больших аппаратных ресурсов, что позволяет использовать их как на локальных серверах, так и в облачной среде со значительно меньшими затратами.

В статье приводятся результаты оценки на основе метрики ROUGE для рефератов, сформированных на больших языковых моделях MBart (специализированная модель) и T-lite (универсальная модель). Исходные текстовые массивы для анализа формировались из статей, опубликованных в журнале «Научные и технические библиотеки» в 2025 г. Проведённый анализ показал, что лучшие значения метрики ROUGE даёт использование модели MBart. Однако полученные данные не могут свидетельствовать о качестве содержания рефератов, формируемых сравниваемыми моделями, поскольку метрика ROUGE показывает лишь степень совпадения слов и фраз в реферате и эталонном тексте. Вывод авторов заключается в том, что достаточно «лёгкие» модели, такие как MBart, в библиотеках могут быть развёрнуты локально и без использования графического процессора, а это предпочтительнее для их широкого использования на практике.

Публикация подготовлена в рамках Государственного задания ГПНТБ России № 075-00548-25-02 от 05.11.2025 по выполнению работы

№ 720000Ф.99.1.БН60АА03000 по теме № 1024031200035-5-1.2.1;5.8.2 (FNEG-2025-0004).

Ключевые слова: автореферирование, большие языковые модели, LLM, трансформеры, MBart, T-lite, ROUGE

Для цитирования: Гончаров М. В., Колосов К. А. Особенности использования больших языковых моделей при составлении текстовых рефератов // Научные и технические библиотеки. 2025. № 11. С. 203–214. https://doi.org/10.33186/1027-3689-2025-11-203-214

ARTIFICIAL INTELLIGENCE IN LIBRARIES

UDC 004.93:02
https://doi.org/10.33186/1027-3689-2025-11-203-214

Specific aspects of using large language models for text abstracting

Mikhail V. Goncharov[1] and Kirill A. Kolosov[2]

[1, 2]Russian National Public Library for Science and Technology, Moscow, Russian Federation

[1]goncharov@gpntb.ru
[2]kolosov@gpntb.ru

Abstract. In the context of spike of science publications, the automatic abstracting based on AI technologies has become a relevant task. The existing abstracting models use the trained large language models which deployment requires significant hardware resources. Meanwhile, specialized models based on the same transformer architecture do not require such big resources and therefore, can be used both on local servers and in the cloud environment at a much lower cost. The authors discuss the results of the ROUGE assessment of the abstracts generated in the LLM MBart (specialized model) and T-lite (universal model). The original large scale prompt was formed of the articles published in “Scientific and

technical libraries" journal in 2025. The analysis findings evidences that MBart model demonstrates the better ROUGE metric value. However, the obtained data do not evidence on the quality of abstracts generated by the compared models, as the ROUGE metric shows just the match value for the words and phrases in the abstract and the reference text. The authors conclude that the "lightish" models, like MBart, may be deployed just locally in the libraries and without graphic processor, which would be more preferable for their practical common use.
The paper is prepared within the framework of the Government Order to RNPLS&T No. 075-00548-25-02 of November 5, 2025, Project No. 720000F.99.1.BN60AA03000 theme No. 1024031200035-5-1.2.1;5.8.2 (FNEG-2025-0004).

Keywords: automated abstracting, large language models, transformer, MBart, T-lite, ROUGE

Cite: Goncharov M. V., Kolosov K. A. Specific aspects of using large language models for text abstracting // Scientific and technical libraries. 2025. No. 11, pp. 203–214. https://doi.org/10.33186/1027-3689-2025-11-203-214

Стремительный рост объёма научной литературы, в особенности публикуемых статей, затрудняет понимание и обобщение информации исследователями. Особенно сложно анализировать информацию в такой быстро развивающейся области, как искусственный интеллект, где исследователям часто требуется синтезировать знания из многих источников. Резюмирование исследований – это не просто чтение статей, но и выявление наиболее важной информации, объединение идей из разных источников и их представление в ясной и лаконичной форме. Одной из ключевых задач является автоматическое реферирование текстов для создания кратких и информативных резюме больших текстов. Это особенно важно для научных исследований, где точность передаваемой информации имеет решающее значение, а резюме могут значительно сократить время, необходимое для изучения материала [**1**].

В последние годы в области реферирования текстов был достигнут значительный прогресс благодаря использованию больших языко-

вых моделей (LLM), основанных на архитектурах трансформеров. Трансформер (Transformer) изучает взаимосвязи между каждым словом и всеми другими словами в тексте с помощью механизма, известного как «самовнимание» [**2**].

Как отмечает Я. Л. Шрайберг: «Совершенствование нейронных сетей, в особенности так называемых “больших языковых моделей” (Large Language Models, LLM), то есть вычислительных моделей, построенных для обработки и генерации естественного языка, начиная с 2020 г. способствовали расцвету систем генеративного ИИ. Генеративные системы ИИ способны обучаться и выявлять шаблоны и структуры во входных данных на основе статистических алгоритмов, полученные выводы позволяют им генерировать новые данные…» [**3**].

Такой подход позволяет модели интерпретировать слова осмысленно и контекстно. Архитектура трансформеров достигла значительных успехов в различных областях искусственного интеллекта, включая обработку естественного языка, компьютерное зрение и обработку звука [**4**]. Примерами моделей на основе трансформеров являются GPT-3, GPT-4, BERT, T5, BART.

Наиболее доступным способом использования возможностей больших языковых моделей для составления рефератов являются чат-боты типа ChatGPT. Взаимодействие с такими чат-ботами может осуществляться через API с площадкой разработчика (поставщика ресурса) либо с локально скачанной моделью. Второй вариант предпочтительнее, но требует значительных вычислительных ресурсов, в особенности – наличия специализированной видеокарты или карты ускорителя искусственного интеллекта. В то же время использование диалогового режима общения с чат-ботом не технологично при работе с массивами текстовой информации. Для пакетной обработки предпочтительнее использовать программные решения на Python, использующие загрузку предварительно обученных больших языковых моделей на локальный сервер. В некоторых случаях, например, для модели MBart, развёртывание не требует значительных вычислительных ресурсов.

В настоящее время существует множество предварительно обученных больших языковых моделей и готовых фреймворков, которые позволяют пользователям сравнительно легко развернуть модель без необходимости обучать её с нуля. Однако качество рефератов, получа-

емых при использовании типовых моделей, не всегда соответствует ожиданиям пользователей. Прежде всего это связано с тематикой массивов данных, на которых проводилось обучение типовой модели. В большинстве случаев это информация новостных сайтов [**5**]. Для повышения качества формируемых рефератов актуальной задачей является дообучение большой языковой модели на массиве русскоязычных текстов научно-технической тематики.

Практическое использование отдельных больших языковых моделей для целей автоматического реферирования, как отмечается в ряде источников, показывает неудовлетворительные результаты. Причины низкой производительности этих моделей:

1) неадекватное понимание контекста;
2) повторы текста и отсутствие семантической связности;
3) возможность пропуска важной информации;
4) отсутствие структурных знаний;
5) невозможность обработки больших документов.

Последняя из перечисленных выше проблем (невозможность обработки больших документов) вызывает огромный интерес у исследователей. Как отмечается в [**2**], методы реферирования, основанные на больших языковых моделях, сталкиваются со значительными трудностями при обработке длинных текстов из-за высоких вычислительных затрат, ограничений памяти и потенциальной потери информации из-за ограничений длины токенов. Эти проблемы приводят к неэффективности и снижению плотности информации в рефератах, особенно в системах с низкими ресурсами. Для решения проблемы обработки больших документов в ряде публикаций рассматриваются варианты разбиения текста на фрагменты по тем или иным критериям с целью последующей обработки этих фрагментов с использованием больших языковых моделей и получения итогового реферата.

В публикации [**6**] был представлен наиболее простой метод разбиения исходного документа на фрагменты с использованием выборки последовательных частей текста по элементам его логической структуры (раздел, рубрика, страница). В приведённом исследовании разбивка текста и его обработка на модели проводилась вручную методом копирования и вставки, что, разумеется, не эффективно для работы с большим количеством обрабатываемых документов. Кроме того, при

таком подходе не учитывается информативность отобранных фрагментов, что влияет на качество итогового реферата.

Авторы публикации [**2**] рассматривают методику предобработки документов, подлежащих реферированию, с формированием текстового графа. В полученном графе связи между узлами, представляющими предложения, определяются на основе количества содержащихся в них общих слов. После создания текстового графа производится отбор наиболее ценных предложений с использованием энтропийного метода Карчи [**7**].

Энтропия Карчи – это метод измерения информации, который использует дробные производные, обеспечивая более гибкие и чувствительные результаты, чем классическая энтропия Шеннона. Процесс начинается с преобразования текста в графовую структуру, где каждое предложение является узлом, а связи между узлами взвешиваются по количеству общих слов. Информационное содержание каждого узла затем рассчитывается с помощью энтропии Карчи, основанной на правиле Лопиталя и дробных производных.

Энтропия Карчи предпочтительна из-за гибкости в измерении плотности информации благодаря дробной степенной структуре и совместимости с большими языковыми моделями. В отличие от энтропии Шеннона, которая ограничена целыми значениями, энтропия Карчи обеспечивает более детальный анализ с дробными значениями от 0 до 1. Это гарантирует, что в больших языковых моделях при реферировании будут представлены только наиболее содержательные предложения, что значительно сокращает время обработки и повышает точность. Кроме того, её графовый подход позволяет более эффективно оценивать плотность информации между предложениями. В то время как энтропия Шеннона лучше подходит для независимых переменных, энтропия Карчи даёт более релевантные результаты для сетевых структур данных.

В системах автоматического реферирования текста выбор предложений из исходного документа имеет решающее значение для определения качества создаваемого реферата. Поэтому крайне важно использовать структурированный процесс отбора предложений. Авторы публикации [**1**] предлагают модель, состоящую из пяти отдельных этапов:

1. Предварительная обработка: начальный этап включает в себя различные этапы предварительной обработки предложений исходного документа, подлежащих реферированию. Это подготавливает текст к последующей обработке, улучшая его качество и связность.

2. Построение текстовых графов: на этом этапе на основе обработанных текстов создаются текстовые графы. Эти графы отображают отношения и связи между различными предложениями, облегчая анализ их значимости.

3. Первичная оценка предложений: Третий этап охватывает первую фазу двухуровневого процесса оценки и отбора предложений. Здесь рассчитываются значения энтропии всех узлов графа. Это измерение служит для выявления и отбора наиболее ценных предложений на основе их информативности.

4. Интеграция с большими языковыми моделями: выбранные предложения затем вводятся в большие языковые модели для второго этапа оценки. На этом этапе LLM присваивают баллы кандидатам, уточняя оценку их релевантности и значимости.

5. Окончательный отбор и составление резюме: на последнем этапе выбираются наиболее ценные предложения от отобранных кандидатов, что завершается созданием резюме.

В публикации [**1**] предложен ещё один вариант предварительной обработки текстов документов большого объёма. Полные тексты документов делятся на взаимосвязанные фрагменты по 150 токенов каждый с перекрытием в 20 токенов. Такое перекрытие сохраняет контекстную непрерывность между последовательными фрагментами, а соблюдение границ предложений гарантирует, что разделение не нарушает семантический поток текста. Как утверждают авторы исследования, данную конфигурацию выбрали экспериментальным путём, опробовав различные варианты и обнаружив, что она обеспечивает наилучший баланс между сохранением контекста и вычислительной эффективностью.

В дальнейшем мы планируем провести анализ обоих описанных выше методов (на основе вычисления энтропии и на основе деления текстов на фрагменты по 150 токенов с перекрытием в 20 токенов) с точки зрения эффективности составления рефератов с использованием метрик.

Наиболее широко используемой метрикой оценки в системах реферирования является метрика ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Эта методология оценивает рефераты на основе N-грамм и последовательности слов [**8**]. ROUGE служит сравнительным критерием, количественно оценивающим степень совпадения между машинными и референтными резюме, созданными человеком. Диапазон оценок ROUGE составляет от 0 до 1, где более высокие значения указывают на большую степень совпадения между сгенерированным и референтным резюме. Следовательно, более высокий балл свидетельствует о более информативном и точном автоматическом резюме.

ROUGE-1 измеряет отношение количества перекрывающихся N-грамм в сгенерированных системой и эталонных (созданных специалистами) резюме к общему количеству N-грамм в эталонных резюме.

ROUGE-2 измеряет перекрытие биграмм (пар последовательных слов) между сгенерированными системой и эталонными резюме.

ROUGE-L фокусируется на самой длинной общей подпоследовательности между сгенерированными и эталонными сводками, независимо от их порядка.

ROUGE-1 оценивает перекрытие N-грамм, уделяя особое внимание сохранению ключевых терминов, в то время как ROUGE-2 анализирует перекрытие биграмм для измерения беглости и связности. ROUGE-L исследует наидлиннейшую общую подпоследовательность (LCS) для оценки структурного сходства на уровне предложений без строгой зависимости от порядка слов.

При этом стоит отметить, что каждая из вышеперечисленных метрик сама по себе является набором из трёх чисел, выражающих значения точности (англ. precision), полноты (англ. recall) и F1-меры (англ. F1-score) – среднее между значениями precision и recall.

Метрики ROUGE могут быть использованы для оценки качества составления рефератов отдельными большими языковыми моделями. В качестве примера приведём значения метрик ROUGE, полученных с использованием программного пакета rouge (https://github.com/pltrdy/rouge). Рефераты создавались на двух больших языковых моделях:

Модель MBart, дообученная на массиве публикаций gazeta.ru (https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta).

Модель Т-Банка (далее – T-lite) (https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1).

Обе модели существенно отличаются как по функциональным возможностям, так и по требованиям к аппаратным ресурсам, как показано в табл. 1.

Таблица 1

Модели и ресурсы, используемые при проведении анализа

Модель	Назначение	Аппаратные ресурсы
MBart	Автореферирование	Intel Xeon Gen2, виртуальные CPU 4 шт., объём оперативной памяти 16 384 МБ
T-lite	Нейросетевой чат-бот	Работа в среде Google Colab, графический процессор Nvidea A100

Тестовые текстовые массивы для анализа формировались на основе выборки десять статей из журнала «Научные и технические библиотеки» за 2025 г. Первый массив использует полные тексты статей в качестве источника для составления авторефератов, а аннотации, составленные авторами статьей, используются в качестве эталонных рефератов при вычислении метрик ROUGE. Второй массив использует первые три абзаца статей в качестве источника для составления авторефератов и их же в качестве эталонных текстов при вычислении метрик ROUGE. В табл. 2 и 3 приведены полученные результаты.

Таблица 2

Показатели ROUGE при обработке полных текстов статей

Модель	ROUGE-1 (F)	ROUGE-2 (F)	ROUGE-L (F)
MBart	0.185567	0.060120	0.170103
T-lite	0.125654	0.017326	0.111692

Таблица 3

Показатели ROUGE при обработке фрагментов текстов

Модель	ROUGE-1 (F)	ROUGE-2 (F)	ROUGE-L (F)
MBart	0.578431	0.530120	0.578431
T-lite	0.147887	0.028011	0.133802

Полученные данные не могут свидетельствовать о качестве содержания рефератов, формируемых изучаемыми моделями, так как анализируют лишь степень совпадения слов и фраз в реферате и эталонном тексте. Кроме того, аннотации, составленные авторами, нельзя считать полноценными рефератами статей. С визуальной точки зрения рефераты, составленные с использованием модели Т-Банка, получаются более интересными, содержат более развёрнутые предложения и обобщения. В то же время рефераты, составленные моделью MBart, более лаконично передают положения исходного текста, что, в общем, и подтверждается вычисленными метриками.

В любом случае при выборе модели генеративного ИИ для использования в системах автоматизации библиотек и открытых архивах придётся учитывать требуемые вычислительные мощности. Мощные модели, такие как модель Т-Банка, могут быть доступны в режиме удалённого сервиса, развёрнутого на сервере крупной библиотеки, имеющей возможности для развёртывания и поддержки дорогостоящего оборудования. В то же время достаточно «лёгкие» модели, такие как рассмотренная MBart, могут быть развёрнуты как на физическом сервере, так и обычном облачном сервере без использования графического процессора. В нашем дальнейшем исследовании мы планируем проанализировать варианты и разработать решения, использующие трансформеры, для автоматизированной пакетной обработки массивов статьей с целью составления рефератов, не требующие значительных вычислительных ресурсов и работающие локально, без необходимости выхода в интернет.

Список источников

1. **Achkar P., Gollub T., Potthast M.** Ask, Retrieve, Summarize: A Modular Pipeline for Scientific Literature Summarization // arXiv preprint arXiv:2505.16349. 2025.

2. **Uckan T.** A hybrid model for extractive summarization: Leveraging graph entropy to improve large language model performance // Ain Shams Engineering Journal. 2025. Vol. 16. № 5. (103348).

3. **Шрайберг Я. Л., Волкова К. Ю.** Вопросы авторского права в отношении произведений, созданных при помощи генеративного искусственного интеллекта // Научные и технические библиотеки. 2025. № 2. С. 115–130.

4. **Lin T. et al.** A survey of transformers // AI open. 2022. Vol. 3. P. 111–132.

5. **Gusev I.** Dataset for automatic summarization of Russian news // Conference on Artificial Intelligence and Natural Language. Cham : Springer International Publishing, 2020. P. 122–134.

6. **Бычкова Е. Ф., Колосов К. А.** Анализ возможностей автоматического реферирования статей на примере источников базы данных «Экология: наука и технологии» ГПНТБ России // Научные и технические библиотеки. 2023. № 10. С. 99–120.

7. **Karci A.** Fractional order entropy: New perspectives // Optik. 2016. Vol. 127. № 20. P. 9172–9177.

8. **Lin C. Y.** Rouge: A package for automatic evaluation of summaries // Text summarization branches out. 2004. P. 74–81.

References

1. **Achkar P., Gollub T., Potthast M.** Ask, Retrieve, Summarize: A Modular Pipeline for Scientific Literature Summarization // arXiv preprint arXiv:2505.16349. 2025.

2. **Uckan T.** A hybrid model for extractive summarization: Leveraging graph entropy to improve large language model performance // Ain Shams Engineering Journal. 2025. Vol. 16. № 5. (103348).

3. **Shrai`berg Ia. L., Volkova K. Iu.** Voprosy` avtorskogo prava v otnoshenii proizvedenii`, sozdanny`kh pri pomoshchi generativnogo iskusstvennogo intellekta // Nauchny`e i tekhnicheskie biblioteki. 2025. № 2. S. 115–130.

4. **Lin T. et al.** A survey of transformers // AI open. 2022. Vol. 3. P. 111–132.

5. **Gusev I.** Dataset for automatic summarization of Russian news // Conference on Artificial Intelligence and Natural Language. Cham : Springer International Publishing, 2020. P. 122–134.

6. **By`chkova E. F., Kolosov K. A.** Analiz vozmozhnostei` avtomaticheskogo referirovaniia statei` na primere istochnikov bazy` danny`kh «E`kologiia: nauka i tekhnologii» GPNTB Rossii // Nauchny`e i tekhnicheskie biblioteki. 2023. № 10. S. 99–120.

7. **Karci A.** Fractional order entropy: New perspectives // Optik. 2016. Vol. 127. № 20. P. 9172–9177.

8. **Lin C. Y.** Rouge: A package for automatic evaluation of summaries // Text summarization branches out. 2004. P. 74–81.

Информация об авторах / Authors

Гончаров Михаил Владимирович – канд. техн. наук, доцент, ведущий научный сотрудник, руководитель группы перспективных исследований и аналитического прогнозирования ГПНТБ России, Москва, Российская Федерация

goncharov@gpntb.ru

Колосов Кирилл Анатольевич – канд. техн. наук, ведущий научный сотрудник ГПНТБ России, Москва, Российская Федерация

kolosov@gpntb.ru

Mikhail V. Goncharov – Cand. Sc. (Engineering), Associate Professor, Leading Researcher, Head, Group for Perspective Research and Analytic Forecasting, Russian National Public Library for Science and Technology, Moscow, Russian Federation

goncharov@gpntb.ru

Kirill A. Kolosov – Cand. Sc. (Engineering), Leading Researcher, Russian National Public Library for Science and Technology, Moscow, Russian Federation

kolosov@gpntb.ru