ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2025, том 505, с. 46–49

МАТЕМАТИКА

УДК 517.54

РЕФЕРИРОВАНИЕ ХУДОЖЕСТВЕННОЙ ЛИТЕРАТУРЫ ПОСРЕДСТВОМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

© **2025 г.** **Д. А. Григорьев**[1,*], **Д. И. Чернышев**[1,**]

Представлено кем-то
Поступило 16.08.2025
После доработки 20.08.2025
Принято к публикации 31.08.2025

Работа исследует методы сжатия художественных текстов с помощью языковых моделей и предлагает улучшенные подходы для точного реферирования в условиях ограниченного контекста.

Ключевые слова и фразы: LLM, реферирование, литература, книги, краткий пересказ

ВВЕДЕНИЕ

Реферирование художественной литературы Автоматическое реферирование текста — одна из ключевых задач в области обработки естественного языка. Суть этой задачи заключается в создании информативной аннотации исходного текста с сохранением основного смысла содержания. В последние годы, с появлением больших языковых моделей, резко возрос интерес к автоматизации реферирования в самых разных жанрах текстов, включая художественные произведения. В отличие от научных, новостных или технических текстов, художественные произведения характеризуются высокой степенью стилистической и семантической сложности. Нелинейность повествования, образность, метафоричность и стилистические приёмы делают задачу написания краткого содержания особенно трудоёмкой. Ограниченное контекстное окно современных моделей дополнительно осложняет работу с длинными произведениями.

Теоретически автоматическое реферирование может выполняться двумя основными способами: извлекающим (выбор ключевых фрагментов текста) и абстрактивным (генерация нового текста на основе содержания оригинала). Для художественной литературы более уместен второй подход, поскольку он позволяет передать смысл и стиль произведения, не нарушая его целостности.

НАБОР ДАННЫХ

Рис. 1. Гистограмма с количеством слов в текстах

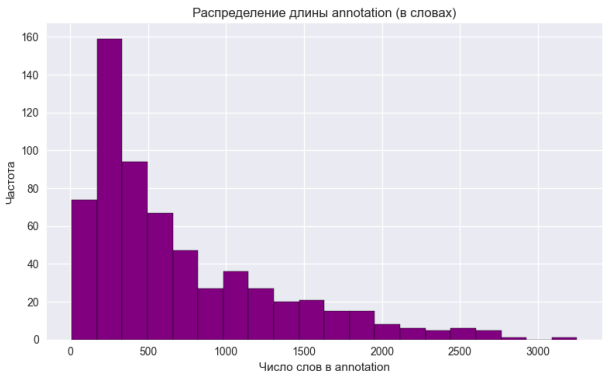

Рис. 2. Гистограмма с количеством слов в аннотациях

На момент начала исследования не существовало открытых и репрезентативных корпусов, предназначенных специально для задачи реферирования художественных текстов на русском языке. С целью проведения экспериментов и оценки различных подходов к генерации аннотаций был создан собственный корпус, состоящий из художественных текстов и соответствующих кратких пересказов. В качестве источника для аннотаций был выбран ресурс «Народный Брифли» [1] — платформа, где пользователи публикуют краткие пересказы литературных произведений. Несмотря на вариативность качества и

[1] *Московский государственный университет им. М. В. Ломоносова, Москва, Россия*
[*] *E-mail: dagrig14@yandex.ru*
[**] *E-mail: chdanorbis@yandex.ru*

стиля пользовательских аннотаций и наличие нерелевантной информации, такой как учебные вопросы или редакторские замечания, после тщательной предварительной обработки удалось получить достаточно надёжный и чистый набор данных. Художественные тексты были отобраны из электронной библиотеки LibRuSec — одного из крупнейших русскоязычных ресурсов художественной литературы. Отбор произведений осуществлялся на основании наличия аннотаций на выбранном ресурсе [1]. Каждый текст проходил автоматическую предварительную обработку: удалялась метаинформация (например, заголовки, описания глав и технические вставки), после чего текст форматировался в единый стандартизированный вид, подходящий для дальнейшего использования в моделях. Важно отметить, что при создании корпуса использовались только тексты, находящиеся в общественном достоянии или распространяемые свободно с разрешения правообладателей, что обеспечивает соблюдение требований авторского права.

Получившийся корпус включал в себя:

- более 600 пользовательских пересказов с ресурса «Народный Брифли»;
- исходные произведения из электронной библиотеки LibRuSec;

Тексты аннотаций проходили автоматическую очистку от HTML-тегов, комментариев и служебных пометок с помощью LLM Meta–Llama 3–70B–Instruct. Затем производился поиск по датасету LibRuSec и собиралась коллекция, состоящая из пар "текст книги - аннотация". На рисунке 1 показано распределение текстов в зависимотсти от количества слов в них. На рисунке 2 аналогичная информация об аннотациях.

МЕТОДОЛОГИЯ

Базовые и модифицированные стратегии

Иерархический метод. Суть этого метода [ссылка!] заключается в том, что текст разбивается на фрагменты (чанки), для каждого из которых отдельно генерируется локальная аннотация. Эти фрагменты затем объединяются в группы, и из полученных аннотаций снова формируется краткое содержание следующего уровня. Последний уровень представляет собой итоговую аннотацию всего произведения.

«Чертёжный» метод (Text-Blueprint). Данный метод [ссылка!] ориентирован на построение промежуточного плана аннотации перед генерацией текста. План формируется в виде набора вопросноответных пар, что повышает управляемость генерации и обеспечивает структурированность результата. Сначала модель формирует список вопросов, отражающих ключевые события, темы и персонажей текста. Далее к каждому вопросу автоматически подбирается краткий ответ. Эта структура служит планом, по которому генерируется итоговая аннотация.

Иерархический метод с фильтрацией узлов. Является модифицированным иерархическим методом. Направлен на ускорение генерации за счет удаления потенциально излишних частей информации, что повышает плотность полезной информации в итоговых рефератах. Для исключения «воды» и дублирующих фрагментов на каждом уровне иерархии мы теперь выполняем глобальную проверку семантической близости между всеми промежуточными аннотациями.

Алгоритм следующий:

1. Пусть на текущем уровне имеются аннотации $\{S_i\}_{i=1}^{n}$.
2. Вычисляем эмбеддинги $\mathbf{e}_i = \mathrm{Encoder}(S_i)$ и нормируем их.
3. Составляем матрицу косинусных сходств

$$M_{ij} = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \, \|\mathbf{e}_j\|}, \quad i, j = 1, \ldots, n.$$

4. Для каждой аннотации S_j находим

$$m_j = \max_{i<j} M_{ji},$$

   то есть максимальную степень схожести с любой предыдущей в списке.
5. Если $m_j < \theta$ (где $\theta = 0.85$), то сохраняем S_j, иначе отбрасываем.
6. Гарантируем, что S_1 всегда остаётся, чтобы не получилось пустого уровня.

Эмбеддинги получаются с помощью SentenceTransformer (модель USER-bge-m3) и при вычислении на GPU обеспечивается высокая скорость обработки.

«Чертёжный» метод с кластеризацией вопросов. Для снижения числа запросов к модели и повышения структурности:

1. Для каждого чанка C_i сгенерировать вопросы $Q_i = \{q_{i1}, \ldots, q_{im}\}$.
2. Вычислить эмбеддинги $E_i = \{\mathbf{e}_{i1}, \ldots, \mathbf{e}_{im}\}$.
3. Объединить все $\{\mathbf{e}_{ij}\}$ и применить алгоритм K-means.
4. Из каждого кластера c случайно выбрать несколько вопросов.
5. Для кластера c сформировать обобщённый вопрос Q_c^*:

$$Q_c^* = \mathrm{LLM}\big(\,\mathrm{concat}(q \in c)\big).$$

6. Использовать $\{Q_c^*\}$ как чертёж для генерации итоговой аннотации.

Такой подход позволяет уменьшить число обращений к LLM, что позволяет ускорить скорость генераций, как будет показано в таблице 2.

ОЦЕНИВАНИЕ МЕТОДОВ

Для объективного сравнения описанных подходов и моделей в задаче реферирования художественных текстов использовались четыре группы метрик.

ROUGE-L — основана на длине наибольшей общей подпоследовательности (LCS) между сгенерированной аннотацией S и эталонной R:

$$\mathrm{Precision} = \frac{\mathrm{LCS}(S,R)}{|S|}, \quad \mathrm{Recall} = \frac{\mathrm{LCS}(S,R)}{|R|},$$

$$\text{ROUGE-L} = \frac{2\,\mathrm{Precision}\,\cdot\,\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

BERTScore — семантическое качество на уровне токенов. Для каждой пары токенов предсказания и эталона вычисляется косинусное сходство их эмбеддингов в модели USER-bge-m3. Затем:

$$P = \frac{1}{|S|}\sum_{t\in S}\underset{u\in R}{\mathrm{maxsim}}(e_t, e_u), \quad R = \frac{1}{|R|}\sum_{u\in R}\underset{t\in S}{\mathrm{maxsim}}(e_u, e_t),$$

$$\mathrm{BERTScore} = \frac{2\,P\,R}{P+R}.$$

Полнота покрытия ключевых вопросов (Coverage) — доля заранее сгенерированных по эталонному тексту вопросов, на которые модель «отвечает» в аннотации:

$$\mathrm{Coverage} = \frac{\#\{q_i\colon P(\text{“да”} \mid q_i, S) > 0.75\}}{N},$$

где N — общее число вопросов, а $P(\text{“да”} \mid q_i, S)$ — вероятность наличия ответа на вопрос q_i в тексте S, оцененная LLM.

Совпадение ответов (AnswerSimilarity) — среднее семантическое сходство между сгенерированными ответами a_i^{pred} и эталонными a_i^{ref} на те же ключевые вопросы:

$$\mathrm{AnswerSimilarity} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{sim}\big(a_i^{\mathrm{pred}},\, a_i^{\mathrm{ref}}\big),$$

где sim — косинусное сходство эмбеддингов, полученных через USER-bge-m3.

Использование нескольких метрик, учитывающих как поверхностное совпадение текста (ROUGE-L), так и глубокое семантическое сходство (BERTScore, AnswerSimilarity), а также степень охвата заранее заданных вопросов (Coverage), обеспечивает всестороннюю и устойчивую оценку качества аннотаций.“‘

ПАРАМЕТРЫ ЭКСПЕРИМЕНТОВ

Все представленные в работе измерения выполнены на наборе из 100 художественных произведений, отобранных так, чтобы исходные тексты не превышали по длине 800 000 символов. Для всех методов генерируемые аннотации ограничивались максимумом в 500 слов.

Текст на вход разбивался на чанки фиксированного размера в 2000 токенов. Токенизация выполнялась с помощью `AutoTokenizer` модели `DeepPavlov/rubert-base-cased` в стандартном режиме. Для воспроизводимости всех случайных процедур использовался фиксированный seed ($random\_seed = 42$).

В **иерархическом методе с фильтрацией узлов** для оценки избыточности промежуточных аннотаций на каждом уровне вычислялась матрица косинусных сходств между их эмбеддингами. Порог схожести был установлен равным $\theta = 0.85$: если для аннотации S_j существует предыдущая S_i с косинусным сходством выше этого порога, то S_j отбрасывается как избыточная. Такой выбор порога обеспечивает компромисс между сохранением значимой информации и устранением дублирования, что эмпирически привело к заметному уменьшению объёма промежуточных представлений без существенной деградации качества.

В **чертёжном методе с кластеризацией вопросов** количество кластеров для K-means выбирается по эвристике, подобранной эмпирически:

$$n_{\text{clusters}} = \max\left(2,\ \left\lceil \sqrt{N_{\text{questions}}} \right\rceil\right),$$

где $N_{\text{questions}}$ — общее число первоначально сгенерированных вопросов по всем чанкам. Гарантируется минимум в два кластера, что позволяет даже при небольших наборах вопросов получать структурированное чертёжное представление.

Временные показатели измерялись как среднее значение (в секундах) времени генерации одной книги по каждому методу для 100 книг. В случае всех четырех методов учитывалось суммарное время всех этапов (включая генерацию промежуточных аннотаций / планов, фильтрацию и финальную агрегацию).

Проводились первоначальные замеры скорости работы методов на небольших текстах, полученные результаты в секундах (среднее по трем запускам) представлены в таблице 1. Результаты подтверждают, что модификации позволяют повысить скорость генерации.

Таблица 1. Время генерации аннотации (в секундах) для текста размером 81,049 символов (11 чанков). Усреднено по трём запускам.

Модель	Иерархический	Иерархический с фильтрацией	Чертежный	Чертежный с кластеризацией
RuadaptQwen2.5-7B-Lite-Beta	84.64	25.70	103.66	78.99
DeepSeek V3	237.83	72.42	292.80	268.75
Qwen3-235B-A22B	113.24	39.45	215.63	145.20
tpro	472.23	127.38	421.65	185.94
yagpt5lite	34.17	14.08	99.70	27.26

РЕЗУЛЬТАТЫ

В таблице 2 приведены сравнительные результаты работы описанных выше методов генерации кратких пересказов художественных текстов. Для каждой из исллдованных моделей измерялись метрики качеств и время выполнеия в зависимости от метода: базовые чертежный и иерархический методы, а также их усовершенствованные версии чертежный метод с кластеризацией вопросов и иерархический метод с фильтрацией узлов.

Таблица 2 показывает, что модифицированные варианты методов действительно существенно ускоряют обработку: среднее время генерации сокращается в 1.5-3 раза в зависимости от модели (например, DeepSeek V3: 315.67 $\Longrightarrow$ 132.60) Однако выигрыш по скорости сопровождается умеренным снижением качества: падение BERTScore и ROUGE-L чаще всего укладывается в 1-2 пункта и находится в пределах стандартных отклонений.

Таблица 2. Результаты по методам и моделям

Модель	Метрики	Чертежный	Чертежный с кластеризацией	Иерархический	Иерархический с фильтрацией
RuadaptQwen2.5-7B-Lite-Beta	bertscore	56.1 ± 4.9	54.0 ± 4.0	55.4 ± 2.9	55.8 ± 2.9
	rouge-l	10.1 ± 3.9	7.7 ± 2.8	8.6 ± 2.5	8.7 ± 2.5
	time	126.84	76.66	68.86	53.59
RuadaptQwen3-32B-Instruct-v2	bertscore	58.9 ± 3.6	55.3 ± 3.3	57.3 ± 2.9	57.7 ± 3.3
	rouge-l	10.6 ± 3.2	7.8 ± 2.1	11.0 ± 2.4	10.7 ± 2.4
	time	376.28	271.42	211.72	159.11
yagpt5lite	bertscore	61.1 ± 3.8	61.5 ± 3.3	62.5 ± 3.5	62.1 ± 3.2
	rouge-l	15.8 ± 5.1	14.3 ± 4.4	16.9 ± 5.1	16.4 ± 4.7
	time	113.34	42.15	31.02	27.39
Qwen3-235B-A22B	bertscore	61.6 ± 3.3	59.3 ± 3.4	61.2 ± 3.0	60.9 ± 2.7
	rouge-l	15.8 ± 4.5	12.2 ± 3.6	14.9 ± 4.0	14.8 ± 3.7
	time	200.30	149.11	103.49	83.06
DeepSeek V3	bertscore	58.0 ± 4.0	58.4 ± 3.6	60.0 ± 3.1	60.0 ± 2.9
	rouge-l	12.6 ± 4.6	11.2 ± 3.9	13.7 ± 3.9	13.5 ± 3.7
	time	315.67	132.60	196.77	147.21
tpro	bertscore	59.0 ± 4.9	58.2 ± 3.7	59.4 ± 3.0	59.5 ± 3.3
	rouge-l	14.7 ± 4.9	11.8 ± 3.9	13.8 ± 3.1	13.5 ± 3.0
	time	259.35	161.33	276.45	230.21

ЗАКЛЮЧЕНИЕ

В заключение, был создан первый открытый датасет, объединяющий в себе тексты книг и аннотации к ним с открытого сайта "Народный Брифли". В работе также предложены улучшенные методы реферирования художественных текстов с помощью LLM: иерархический с фильтрацией и чертёжный с кластеризацией. Оба метода позволяют значительно ускорить генерацию (до 3 раз) при минимальной потере качества. Подходы показали устойчивые результаты на созданном корпусе и пригодны для обработки длинных произведений в условиях ограниченного контекста.

СПИСОК ЛИТЕРАТУРЫ

[1] *Народный Брифли.* Электронная библиотека кратких пересказов литературных произведений. `https://wiki.briefly.ru/` (дата обращения: 16.07.2025).

REFERENCES

[1] *Народный Брифли.* Электронная библиотека кратких пересказов литературных произведений. `https://wiki.briefly.ru/` (дата обращения: 16.07.2025).