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

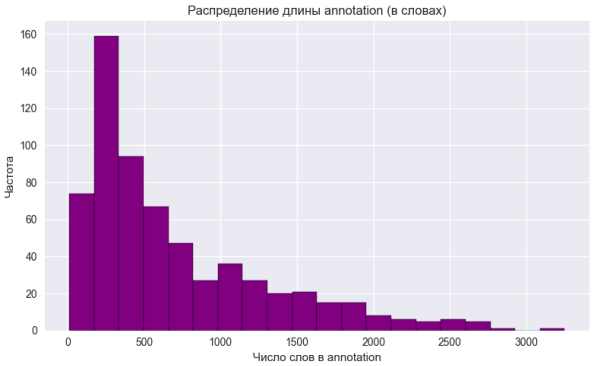

Рис. 2. Гистограмма с количеством слов в аннотациях

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 Пусть входной текст D имеет длину $L \gg W$, где W — размер контекстного окна LLM, а $C < W$ — длина одного чанка. Обозначим уровень иерархии $l = 0, 1, \ldots, L$, и гиперпараметр контроля длины на уровне l как G_l.

1. Разбиваем D на $n_0 = \lceil L/C \rceil$ чанков
$$C_i, \quad i = 1, \ldots, n_0.$$
2. На уровне $l = 0$ для каждого чанка генерируем локальную аннотацию
$$S_i^{(0)} = \mathrm{Summarize}(C_i), \quad i = 1, \ldots, n_0.$$
3. Для $l = 1, 2, \ldots$ до тех пор, пока не останется единственной аннотации:
   - Задаём порог длины
   $$T_l = W \; - \; G_l.$$
   - Инициализируем $i \leftarrow 1$, $j \leftarrow 1$.
   - Пока $i \leqslant n_{l-1}$:
     - Найдём наибольшее $m \geqslant 1$, такое что
     $$\sum_{t=i}^{i+m-1} \left| S_t^{(l-1)} \right| \; \leqslant \; T_l.$$
     - Объединяем эти аннотации и генерируем
     $$S_j^{(l)} \; = \; \mathrm{Summarize}\big(S_i^{(l-1)} \oplus \cdots \oplus S_{i+m-1}^{(l-1)}\big).$$
     - Обновляем $i \leftarrow i + m$, $j \leftarrow j + 1$.
   - Получаем $n_l = j - 1$ аннотаций уровня l.
4. Итоговая аннотация — единственный элемент $S^{(L)}$.

«Чертёжный» метод (Text-Blueprint) Метод строит промежуточный план в виде вопросов и ответов перед генерацией текста. Для всего текста T или каждого чанка модель последовательно:

1. Генерирует список вопросов $\{q_i\}$, охватывающих ключевые элементы сюжета.
2. Для каждого q_i формирует краткий ответ a_i.
3. Собирает последовательность $(q_1, a_1), \ldots, (q_m, a_m)$ как «чертёж».
4. По этому «чертежу» генерирует итоговое резюме:
$$S = \mathrm{LLM}\big((q_1, a_1) \oplus \cdots \oplus (q_m, a_m)\big).$$

Иерархический метод с фильтрацией узлов 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 вычисляются на GPU, что обеспечивает высокую 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 на r кластеров.
4. Из каждого кластера c случайно выбрать 10–30 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 1.

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:

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

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

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:

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

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.“‘

РЕЗУЛЬТАТЫ

В таблице 1 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 1. Результаты по методам и моделям

Модель	Метрики	Чертежный	Чертежный с кластеризацией	Иерархический	Иерархический с фильтрацией
RuadaptQwen2.5-7B-Lite-Beta	bertscore	58.7 ± 3.8	57.7 ± 3.7	59.6 ± 3.5	59.3 ± 3.4
	rouge-l	14.1 ± 4.8	12.2 ± 4.5	14.4 ± 4.3	13.8 ± 4.3
	time	159.13	77.30	106.18	79.58
RuadaptQwen3-32B-Instruct-v2	bertscore	56.8 ± 5.8	53.7 ± 5.2	55.6 ± 3.4	55.7 ± 3.6
	rouge-l	10.4 ± 4.5	7.6 ± 3.6	10.2 ± 2.9	9.9 ± 2.5
	time	201.88	140.41	182.41	147.32
yagpt5lite	bertscore	61.0 ± 3.6	61.4 ± 3.2	62.3 ± 3.2	62.1 ± 3.3
	rouge-l	15.8 ± 5.2	14.0 ± 4.4	16.7 ± 5.0	16.5 ± 4.7
	time	98.45	27.06	24.97	24.34
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

Как видно из таблицы 1, 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.

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).