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

МАТЕМАТИКА

УДК 517.54

ОЦЕНКА ОБЩИХ И СПЕЦИАЛЬНЫХ ЗНАНИЙ В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ ДЛЯ РУССКОГО ЯЗЫКА ПОСРЕДСТВОМ ВОСПРОИЗВЕДЕНИЯ ЭНЦИКЛОПЕДИЧЕСКИХ СТАТЕЙ

© **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

На момент начала исследования не существовало открытых и репрезентативных корпусов, предназначенных специально для задачи реферирования художественных текстов на русском языке. Был сформирован собственный набор данных, включающий:

- более 600 пользовательских пересказов с ресурса «Народный Брифли»;
- исходные произведения из электронной библиотеки LibRuSec (публичное достояние или тексты с разрешения правообладателей);

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".

ОЦЕНИВАНИЕ МЕТОДОВ

На момент начала исследования в русскоязычном сегменте отсутствовали открытые и репрезентативные наборы данных, специально предназначенные для задачи автоматического реферирования

[1] *Московский государственный университет им. М. В. Ломоносова, Москва, Россия*

[*] *E-mail: dagrig14@yandex.ru*

[**] *E-mail: chdanorbis@yandex.ru*

художественных текстов. В отличие от аналогичных англоязычных проектов, которые уже включали крупные корпуса литературных произведений с высококачественными аннотациями, для русскоязычных текстов подобные ресурсы не были представлены. Это существенно затрудняло объективное тестирование и сравнение эффективности методов автоматического реферирования в условиях художественных произведений 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 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, содержащего свыше 400 тысяч текстов. Отбор произведений осуществлялся на основании наличия аннотаций на выбранном ресурсе. 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.

Для иллюстрации структуры сформированного корпуса ниже представлен пример пары «фрагмент художественного текста — соответствующая аннотация».

ПРИМЕНЕНИЕ МЕТОДОВ

Экспериментальная часть включает проверку базовых стратегий (иерархическая, итеративная, «чертёжная») и двух усовершенствований.

Влияние предварительной очистки Очистка исходного текста от технических артефактов положительно сказалась на ROUGE–L и BERTScore для моделей MetaLlama 370B и RuadaptQwen2.5-32BProBeta, повысив также полноту покрытия ключевых вопросов.

Сравнение базовых методов Иерархический и итеративный подходы продемонстрировали сопоставимое качество (ROUGE–L ≈ 0.48, BERTScore ≈ 0.70), заметно превосходя псевдо-генерацию без доступа к исходному тексту.

Иерархический метод с фильтрацией узлов Добавление фильтра по семантической близости (порог 0.85) позволило убрать избыточные фрагменты и ускорить генерацию на длинных текстах в среднем на 15 %.

«Чертёжный» метод с кластеризацией вопросов Кластеризация эмбеддингов вопросов алгоритмом k-means и последующее сэмплирование 10–30 вопросов из каждого кластера уменьшили число обращений к модели и ускорили работу метода в 1.22 раза без потери качества.

ОЦЕНИВАНИЕ МЕТОДОВ

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.

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.

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

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).