Программные системы и вычислительные методы
Правильная ссылка на статью:
Дагаев А.Е., Попов Д.И. Сравнение автоматического обобщения текстов на русском языке // Программные системы и вычислительные методы. 2024. № 4. DOI: 10.7256/2454-0714.2024.4.69474 EDN: CSFMFC URL: https://nbpublish.com/library_read_article.php?id=69474

Сравнение автоматического обобщения текстов на русском языке

Дагаев Александр Евгеньевич

аспирант, кафедра «Информатика и информационные технологии», Московский политехнический университет

107023, Россия, г. Москва, ул. Большая Семёновская, 38

alejaandro@bk.ru

Попов Дмитрий Иванович

доктор технических наук

Сочинский государственный университет, профессор кафедры Информационных технологий и математики

354000, Россия, Краснодарский край, г. Сочи, ул. Пластунская, 94

damitry.popov@gmail.com

Статья из рубрики "Системный анализ, поиск, анализ и фильтрация информации"

DOI:

10.7256/2454-0714.2024.4.69474

EDN:

CSFMFC

Дата направления статьи в редакцию:

29-12-2023

Дата публикации:

07-11-2024

Аннотация: Предметом исследования в данной статье является обобщение текстов на русском языке с использованием моделей искусственного интеллекта. В частности, авторы сравнивают популярные модели GigaChat, YaGPT2, ChatGPT-3.5, ChatGPT-4, Bard, Bing AI и YouChat и проводят сравнительное исследование их работы на текстах

русского языка. В качестве исходных материалов для последующего обобщения в статье берутся наборы данных для русского языка, такие как Gazeta, XL-Sum и WikiLingua, а также для сравнения эффективности обобщения были взяты дополнительные наборы данных на английском языке CNN Dailymail и XSum. В статье применяются показатели: ROUGE, BLEU score, BERTScore, METEOR и BLEURT для оценки обобщения текстов. данной статье в качестве метода исследования используется сравнительный анализ данных, полученных в ходе автоматического обобщения с помощью моделей искусственного интеллекта. Научная новизна исследования заключается в проведении сравнительного анализа качества автоматического обобщения текстов на русском и английском языках с использованием различных нейросетевых моделей обработки естественного языка. Авторы исследования привлекли внимание к новым моделям GigaChat, YaGPT2, ChatGPT-3.5, ChatGPT-4, Bard, Bing AI и YouChat, рассматривая и анализируя их эффективность в задаче обобщения текста. Итоги обобщения на русском языке показывают, что YouChat демонстрирует самые высокие результаты по совокупности оценок, подчеркивая эффективность модели в обработке и генерации текста с более точным воспроизведением ключевых элементов содержания. В отличие от YouChat, модель Bard показала наихудшие результаты, представляя собой модель с наименьшей способностью к генерации связного и релевантного текста.

Ключевые слова:

обработка естественного языка, суммаризация текста, GigaChat, YaGPT2, ChatGPT-3, ChatGPT-4, Bard, Bing AI, YouChat, сжатие текста

Введение

Реферирование текста – это важное направление в области обработки естественного языка, которое имеет немалое значение для большого количества задач. Для обобщения текста модель искусственного интеллекта должна быть способна создавать связный и релевантный контент, одновременно сжимая главную информацию в более короткую форму вне зависимости от предметных областей. С помощью обобщения текстов можно определить и сопоставить качество работы нейросетей. В статье будет проведено сравнительное исследование популярных моделей искусственного интеллекта на текстах русского языка.

Связанные работы

Оценкой качества языковых моделей в последнее время занимались достаточно активно, но в основном исследования направлены на обработку английского языка, как международного. С русским языком, как основным, сравнения между GigaChat, YaGPT2, ChatGPT-3.5, ChatGPT-4, Bard, Bing AI и YouChat найдено не было.

В работе [6] было исследовано рекурсивное обобщение через GPT-3.5, а также методы отбора существенного контента для обобщения. В работе [7] указывают, что модели GPT с трудом идентифицирует важную информацию и более подвержены ошибкам при обобщении длинных текстовых форм. Исследование качества генерации с применением моделей GPT и дальнейший анализ показали, что качество для языков с высокой лингвистической частотностью выше, чем с низкой [8]. А в работе [9] отмечена слабая производительность GPT при работе с русским языком в мультиязычном наборе данных. В недавних исследованиях [1][2] показывается, что качество обобщения новостей с

помощью больших языковых моделей находится на сопоставимом с созданным человеком уровне.

Наборы данных

В качестве наборов данных для русского языка использованы:

Gazeta [3]. Набор данных содержит 63435 новостей, размещенных на сайте gazeta.ru.

XL-Sum [12]. В наборе представлены 1,35 миллионов аннотированных пар статей BBC на разных языках, в том числе на русском – 77803.

WikiLingua [16]. Многоязычный набор данных, созданный для оценки задачи обобщения. Материалы включают в себя статьи на 18 языках из WikiHow. На русском языке собраны 52928 статей.

Для сравнения эффективности обобщения были взяты дополнительные наборы данных на английском языке:

CNN Dailymail[11]. В набор включены новостные статьи CNN за период с апреля 2007 по апрель 2015 года и Daily Mail с июня 2010 по апрель 2015 года. Суммарное количество составляет 311672 статьи.

XSum [10]. Набор состоит из 226711 статей BBC за период с 2010 по 2017 год.

Из всех перечисленных наборов случайным образом выделены 100 оригинальных текстов, которые были унифицированы по длине в 1024 токена.

Показатели оценки

В данной работе использовались показатели оценки качества текстов ROUGE[4], BLEU score [5], BERTScore [13], METEOR[14] и BLEURT[15]. ROUGE используется для оценки качества текстов, созданных машинами. Он анализирует сходство между искусственно созданным текстом и эталонным. ROUGE определяет точность и полноту информации. Он может использоваться для оценки различных задач, включая обобщение текста и машинный перевод.

Среди разных вариантов ROUGE в работе были использованы:

ROUGE-1. Этот показатель вычисляет перекрытие униграмм (отдельных слов) между машинным и эталонным текстом. Это помогает оценить точность машинного перевода или обобщения текста.

ROUGE-2, который работает с биграммами, то есть с парами слов.

ROUGE-L, который анализирует длинные фразы в тексте и в результате измеряет сходство между сгенерированным текстом и ссылочным текстом с учетом последовательностей слов, то есть оценивает длину самой длинной общей подпоследовательности

Каждый из этих вариантов ROUGE позволяет по-разному оценить качество сгенерированного машиной текста. ROUGE-1 и ROUGE-2 фокусируются на перекрытии на уровне слов и биграмм, в то время как ROUGE-L рассматривает структуру и порядок слов в текстах.

BLEU score[5] представляет собой показатель, используемый для измерения качества машинного текста, в частности, в контексте машинного перевода и обобщения текста. Первоначально он был создан для оценки качества машинного перевода, однако сейчас применяется для многих других задач в области NLP. BLEU score оценивает сходство между машинно-созданным текстом с одним или более эталонными текстами, написанными людьми. Это достигается путем сравнения n-грамм в машинном тексте с n-граммами в эталонных текстах.

BERTScore метрика, которая оценивает схожесть между двумя текстами, используя векторные представления, полученные с использованием модели BERT Оценка BERTScore хорошо коррелируется с суждениями человека и обеспечивает более высокую эффективность выбора модели, чем существующие показатели, кроме этого она более устойчива к сложным примерам по сравнению с существующими метриками [13]. В статье используется F1 Score, которая рассчитывается как среднее гармоническое значение точности и запоминания. Это обеспечивает сбалансированный показатель, учитывающий как ложноположительные, так и ложноотрицательные результаты.

METEOR[14]. Метрика METEOR часто используется для оценки качества машинного перевода и предоставляет более подробную информацию, чем показатель BLEU. При этом учитывается не только точность и запоминаемость отдельных слов, но также рассматриваются основы слов, синонимы и порядок слов. Всё это сделано в целях обеспечения более целостной оценки качества.

BLEURT[15] представляет собой обученную метрику, основанную на BERT и RemBERT. В качестве входных данных используются пары текстов: кандидат-эталонный текст, и выводится оценка, показывающая, насколько хорошо кандидат владеет языком и передает основной смысл текста.

Результаты

Результаты по вышеуказанным показателям для набора данных Gazeta представлены в таблице 1.

Таблица 1 – Результаты на наборе Gazeta

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,17	0,09	0,17	6,71	0,71	0,16	0,14
YaGPT2	0,11	0,04	0,11	6,05	0,70	0,09	0,00
ChatGPT-3.5	0,29	0,12	0,28	5,88	0,72	0,19	0,14
ChatGPT-4	0,27	0,10	0,25	6,12	0,71	0,16	0,04
Bard	0,33	0,18	0,32	4,43	0,71	**0,26**	-0,06
Bing AI	0,33	0,15	0,31	5,09	0,72	0,23	0,03
YouChat	0,33	**0,19**	0,32	**9,60**	0,72	0,24	**0,22**

Результаты для набора XL-Sum приведены в таблице 2.

Таблица 2 – Результаты на наборе XL-Sum

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,15	0,05	0,14	5,69	0,68	0,09	-0,03
YaGPT2	0,10	0,02	0,10	5,98	0,66	0,04	-0,07
ChatGPT-	0,24	0,10	0,23	**6,60**	0,70	0,15	**0,20**

ChatGPT-3.5	0,24	0,10	0,23	**5,58**	0,70	0,19	**0,20**
ChatGPT-4	0,24	0,10	0,24	4,36	0,69	0,20	-0,09
Bard	0,24	0,10	0,23	4,48	0,69	0,20	-0,13
Bing AI	0,32	0,17	0,30	4,41	0,71	0,21	-0,09
YouChat	**0,38**	**0,23**	**0,36**	5,94	**0,73**	**0,26**	0,12

По набору WikiLingua результаты приведены в таблице 3.

Таблица 3 – Результаты на наборе WikiLingua

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,33	0,16	0,32	5,30	0,73	0,23	**0,16**
YaGPT2	0,20	0,05	0,19	5,07	0,72	0,09	0,12
ChatGPT-3.5	0,27	0,09	0,26	**5,11**	0,71	0,17	0,00
ChatGPT-4	0,23	0,05	0,21	4,54	0,70	0,14	0,04
Bard	0,34	0,17	0,33	5,04	0,75	0,24	0,08
Bing AI	0,41	0,24	0,39	4,51	0,75	0,29	0,09
YouChat	**0,56**	**0,36**	**0,54**	4,70	**0,83**	**0,47**	0,14

Для подсчета общей оценки каждому показателю был обозначен индивидуальный вес. Веса были распределены исходя из специфики задачи обобщения текстов, где больший вес выделен на семантическое сходство, комбинацию семантического и структурного сходства, а также на степень сжатия. Формула расчета:

$$S = 0{,}05 * M_1 + 0{,}05 * M_2 + 0{,}05 * M_3 + 0{,}1 * M_4 + 0{,}20 * M_5 + 0{,}20 * M_6 + 0{,}20 * M_7 + 0{,}15 * C \quad (1)$$

Где S – общая оценка,

M_1 – ROUGE-1;

M_2 – ROUGE-2;

M_3 – ROUGE-L;

M_4 – BLEU;

M_5 – BERTScore;

M_5 – METEOR;

M_6 – BLEURT;

C – Степень сжатие текста (%).

На рисунке 1 изображена диаграмма общих оценок для всех использованных наборов данных русского языка.

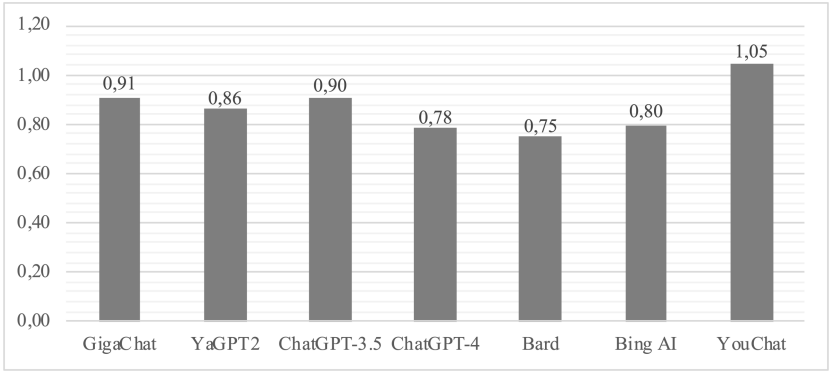

Рисунок 1 – Диаграмма общих оценок для наборов данных русского языка

YouChat показал самые высокие результаты на всех наборах данных по совокупности показателей. Тем самым подчеркивается эффективность модели в обработке и генерации текста, а также ее способность к точному воспроизведению ключевых элементов содержания.

Bard, в отличие от других моделей, выполняет генерацию связного и контекстуально релевантного текста худшего качества, что приводит к неудовлетворительным результатам при оценке сходства, обобщения и других задач обработки естественного языка. Также более низкие показатели могут указывать на трудности в восприятии тонкостей языка, что приводит к таким проблемам, как нерелевантная информация, отсутствие смысловой последовательности и неточности в воспроизведении основного контекста.

GigaChat лучше подошел к задаче обобщения по сравнению с ChatGPT-3.5, но в целом результаты находятся на сопоставимом по качеству уровне.

GigaChat показал воспроизведение контекста точнее YaGPT2 и продемонстрировал более осмысленную генерацию текста с общей повышенной способностью к обобщению.

Bard показал наименьшую итоговую оценку на наборах русского языка.

Результаты сжатия между входными и выходными данными отражены на рисунке 2.

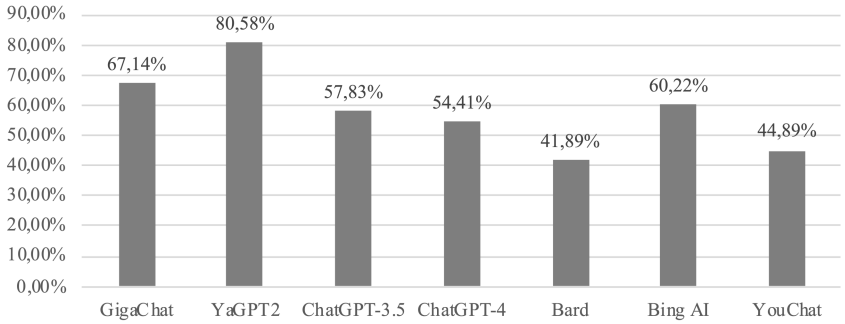

Рисунок 2 – Обобщенный график сжатия на наборах данных русского языка

Максимальное сжатие показал YaGPT2 – 80,58%, далее GigaChat – 67,14%, Bing AI – 60,22%, ChatGPT-3.5 – 57,83%, ChatGPT-4 – 54,41%, YouChat – 44,89%, а наиболее низкие результаты по сжатию получились у Bard – 41,89%.

Проведя обобщение на английском языке для набора данных CNN Dailymail были зафиксированы показатели, которые указаны в таблице 4. Следует отметить, что YaGPT2 не работает с текстами на английском языке, по этой причине он был исключен из

списка моделей для последующего анализа.

Таблица 4 – Результаты на наборе CNN Dailymail

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,32	0,16	0,30	6,27	0,72	0,16	-0,29
ChatGPT-3.5	0,28	0,11	0,25	5,35	0,71	0,13	-0,34
ChatGPT-4	0,30	0,09	0,27	5,02	0,71	0,16	-0,37
Bard	0,32	0,17	0,30	**7,04**	0,72	0,16	-0,55
Bing AI	0,33	0,15	0,31	5,63	0,72	0,16	-0,38
YouChat	**0,39**	**0,19**	**0,36**	5,48	**0,73**	**0,22**	**-0,28**

Результаты по набору XSum отражены в таблице 5

Таблица 5 – Результаты на наборе XSum

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,38	0,22	0,36	**6,01**	0,74	0,19	-0,27
ChatGPT-3.5	0,34	0,14	0,31	5,09	0,72	0,18	-0,28
ChatGPT-4	0,34	0,11	0,31	4,58	0,71	0,18	-0,36
Bard	0,36	0,20	0,35	4,31	0,73	0,15	-0,36
Bing AI	0,42	0,22	**0,40**	4,45	0,74	0,25	-0,36
YouChat	0,42	0,21	0,39	5,47	0,73	0,25	**-0,17**

Общая оценка на данных английского языка приведена на рисунке 3. Результаты показывают, что качество выполненного обобщения имеет зависимость от исходного языка.

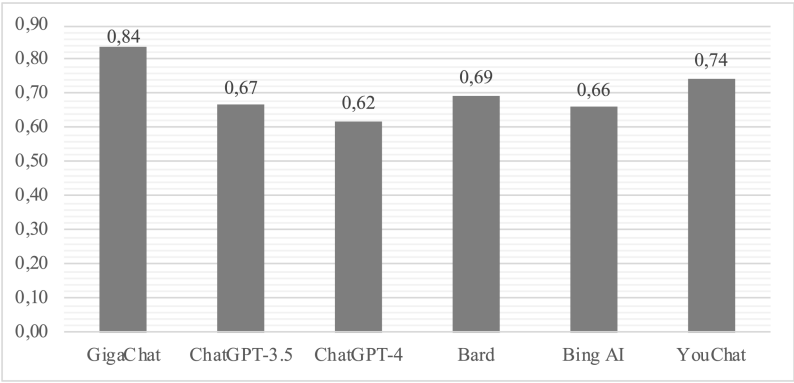

Рисунок 3 – Диаграмма общих оценок для наборов данных английского языка

GigaChat показал самый высокий балл среди других рассмотренных моделей.

ChatGPT-4 в наборах на английском языке демонстрируют самое низкое качество, а также сжимает текст меньше остальных моделей.

Результаты сжатия между входными и выходными данными представлены на рисунке 4.

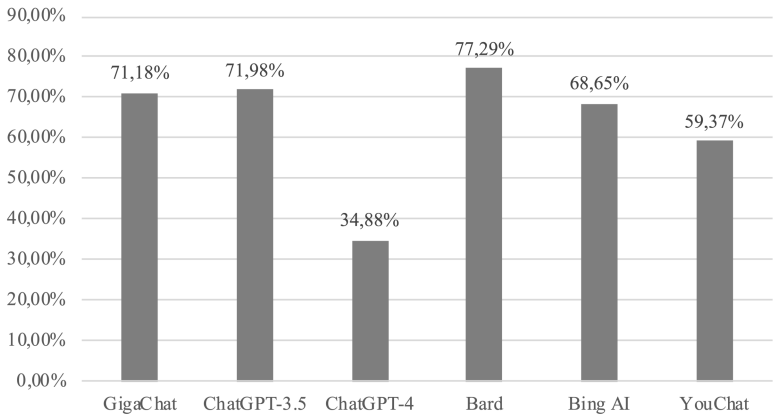

Рисунок 4 – Обобщенный график сжатия на наборах данных английского языка

Максимальное сжатие на англоязычных наборах данных показал Bard – 77,29%, ChatGPT-3.5 – 71,98%, GigaChat – 71,18%, Bing AI – 68,65%, YouChat – 59,37%, самые низкие показатели у ChatGPT-4 – 34,88%

В целом, все представленные модели показали приемлемые результаты по выбранным метрикам. Это свидетельствует о том, что их можно применять для решения задач обобщения текста. Однако стоит отметить, что исходя из разных возможностей моделей могут быть отличия в определенных сценариях использования, связанных с размером модели, типом задачи или языковыми особенностями, что требует дополнительного исследования.

Заключение

В данной работе было проведено сравнение качества автоматического обобщения текстов с помощью различных нейросетей в части обработки естественного языка, таких как GigaChat, YaGPT2, ChatGPT-3.5, ChatGPT-4, Bard, Bing AI и YouChat. Для этого был взят и предобработан набор данных, содержащий тексты на русском и для сравнения на английском языках. Затем выполнено обобщение одинакового списка текстов на каждой модели. После этого получены результаты по показателям ROUGE[4], BLEU score[5], BERTScore[13], METEOR[14] и BLEURT[15], которые сравнивали оригинальные и сгенерированные в ходе автоматического реферирования тексты. Были также получены результаты общей оценки между всеми показателями, где каждому показателю был выделен вес, исходя из важности для задачи обобщения текста.

Данные, полученные в ходе сравнения, будут способствовать более глубокому пониманию рассматриваемых моделей, помогая делать выбор при применении искусственного интеллекта для задач обобщения текстов в качестве основы для будущих разработок.

В дальнейшем планируется исследовать работу по обработке текстов между моделями с различными параметрами настроек.

Библиография

1. Goyal T., Li J. J., Durrett G. News summarization and evaluation in the era of gpt-3 //arXiv preprint arXiv:2209.12356. – 2022.
2. Zhang T. et al. Benchmarking large language models for news summarization //arXiv preprint arXiv:2301.13848. – 2023.
3. Gusev I. Dataset for automatic summarization of Russian news //Artificial Intelligence

and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9. – Springer International Publishing, 2020. – С. 122-134.
4. Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. – 2004. – С. 74-81.
5. Post M. A call for clarity in reporting BLEU scores //arXiv preprint arXiv:1804.08771. – 2018.
6. Bhaskar A., Fabbri A., Durrett G. Prompted opinion summarization with GPT-3.5 //Findings of the Association for Computational Linguistics: ACL 2023. – 2023. – С. 9282-9300.
7. Tang L. et al. Evaluating large language models on medical evidence summarization //npj Digital Medicine. – 2023. – Т. 6. – №. 1. – С. 158.
8. Hendy A. et al. How good are gpt models at machine translation? a comprehensive evaluation //arXiv preprint arXiv:2302.09210. – 2023.
9. Jiao W. et al. Is ChatGPT a good translator? Yes with GPT-4 as the engine //arXiv preprint arXiv:2301.08745. – 2023.
10. Narayan S., Cohen S. B., Lapata M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization //arXiv preprint arXiv:1808.08745. – 2018.
11. Nallapati R. et al. Abstractive text summarization using sequence-to-sequence rnns and beyond //arXiv preprint arXiv:1602.06023. – 2016.
12. Hasan T. et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages //arXiv preprint arXiv:2106.13822. – 2021.
13. Zhang T. et al. Bertscore: Evaluating text generation with bert //arXiv preprint arXiv:1904.09675. – 2019.
14. Banerjee S., Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments //Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. – 2005. – С. 65-72.
15. Sellam T., Das D., Parikh A. P. BLEURT: Learning robust metrics for text generation //arXiv preprint arXiv:2004.04696. – 2020.
16. Ladhak F. et al. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization //arXiv preprint arXiv:2010.03093. – 2020.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Предметом исследования в данной статье является сравнение качества автоматического обобщения (реферирования) текстов на русском языке с использованием различных моделей искусственного интеллекта. Методология исследования включает отбор и предобработку наборов текстовых данных на русском и английском языках, генерацию рефератов этих текстов различными моделями ИИ, а также оценку качества полученных рефератов с помощью стандартных метрик ROUGE, BLEU, BERTscore, METEOR и BLEURT Использованные в работе метрики (ROUGE, BLEU, BERTscore, METEOR и BLEURT формируют комплексный подход к оценке качества автоматического реферирования текстов, учитывающий различные аспекты: точность передачи отдельных слов и словосочетаний (ROUGE, BLEU), семантическое сходство и порядок слов в предложениях (BERTscore, METEOR), общую передачу смысла исходного текста (BLEURT). Каждая метрика имеет свои преимущества и недостатки. В целом их совместное использование позволяет получить наиболее объективную оценку и сравнить

эффективность разных моделей автоматического реферирования текстов. При этом результаты отдельных метрик могут несколько расходиться, что связано с учетом ими различных лингвистических факторов. Тема является актуальной, поскольку задача автоматического реферирования текстов активно исследуется в области обработки естественного языка и имеет множество практических применений. Сравнение эффективности различных подходов для русского языка ранее не проводилось. Научная новизна работы заключается в том, что впервые проведено сравнительное исследование качества обобщения текстов на русском языке с использованием ряда популярных моделей искусственного интеллекта. Стиль изложения научный, текст структурирован, основные разделы соответствуют логике исследования. Содержание достаточно полно раскрывает заявленную тему. Библиография актуальна, охватывает последние работы в данной предметной области. Результаты исследования представляют интерес для специалистов в области компьютерной лингвистики и обработки естественного языка. Могут найти применение при выборе оптимальных моделей ИИ для решения задач автоматического реферирования текстов. Таким образом, статья актуальна, обладает научной новизной и может быть рекомендована к публикации. Рекомендации для дальнейших исследований:

1. Расширение списка сравниваемых моделей автоматического реферирования за счет наиболее передовых и популярных архитектур.
2. Более подробный анализ влияния конфигурации моделей (размер, объем обучающих данных и т.д.) на качество реферирования.
3. Исследование особенностей применения моделей для текстов из разных предметных областей и на разных языках.
4. Разработка комбинированных подходов с использованием нескольких моделей на разных этапах процесса реферирования.
5. Сравнение с рефератами, составленными экспертами, для выявления недостатков существующих алгоритмов.

Проведение таких исследований позволит лучше понять возможности и ограничения современных моделей автоматического реферирования текстов.