УДК 004.89
DOI: 10.26102/2310-6018/2024.47.4.038

Оценка качества интеллектуального перефразирования текстов на русском языке

А.Е. Дагаев[1]✉, Д.И. Попов[2]

[1]Московский политехнический университет, Москва, Российская Федерация
[2]Сочинский государственный университет работы, Сочи, Российская Федерация

Резюме. Данное исследование посвящено разработке интегральной метрики для оценки качества моделей перефразирования текстов, что отвечает актуальной задаче создания комплексных и объективных методов оценки. В отличие от предыдущих исследований, преимущественно фокусирующихся на англоязычных наборах данных, настоящее исследование акцентирует внимание на наборах данных русского языка, которые до настоящего времени оставались недостаточно изученными. Использование таких датасетов, как Gazeta, XL-Sum и WikiLingua (для русского языка), а также CNN Dailymail и XSum (для английского языка), обеспечивает многоязычную применимость предложенного подхода. Предлагаемая метрика сочетает лексические (ROUGE, BLEU), структурные (ROUGE-L) и семантические (BERTScore, METEOR, BLEURT) критерии оценки с распределением весов, исходя из важности каждой метрики. Результаты демонстрируют превосходство моделей ChatGPT-4 на русскоязычных наборах и GigaChat на англоязычных наборах, тогда как модели Gemini и YouChat показывают ограниченные возможности в достижении семантической точности вне зависимости от языка датасета. Оригинальность исследования заключается в объединении метрик в единую систему, что делает возможным более объективное и комплексное сравнение языковых моделей. Исследование вносит вклад в область обработки естественного языка, предлагая инструмент для оценки качества языковых моделей.

Ключевые слова: обработка естественного языка, перефразирование текста, GigaChat, YandexGPT 2, ChatGPT-3.5, ChatGPT-4, Gemini, Bing AI, YouChat, Mistral Large.

Для цитирования: Дагаев А.Е., Попов Д.И. Оценка качества интеллектуального перефразирования текстов на русском языке. *Моделирование, оптимизация и информационные технологии.* 2024;12(4). URL: https://moitvivt.ru/ru/journal/pdf?id=1763 DOI: 10.26102/2310-6018/2024.47.4.038

Evaluation of the quality of intelligent text paraphrasing in Russian

A.E. Dagaev[1]✉, D.I. Popov[2]

[1]Moscow Polytechnic University, Moscow, the Russian Federation
[2]Sochi State University, Sochi, the Russian Federation

Abstract. The study focuses on the development of an integral metric for evaluating the quality of text paraphrasing models, addressing the pressing need for comprehensive and objective evaluation methods. Unlike previous research, which predominantly focuses on English-language datasets, this study emphasizes Russian-language datasets, which have remained underexplored until now. The inclusion of datasets such as Gazeta, XL-Sum, and WikiLingua (for Russian) as well as CNN Dailymail and XSum (for English) ensures the multilingual applicability of the proposed approach. The proposed metric combines lexical (ROUGE, BLEU), structural (ROUGE-L), and semantic (BERTScore, METEOR, BLEURT) evaluation criteria, with weights assigned based on the importance of each metric. The results highlight the superiority of ChatGPT-4 on Russian datasets and GigaChat on English datasets, whereas models such as Gemini and YouChat exhibit limited capabilities in achieving semantic accuracy

regardless of the dataset language. The originality of this research lies in the integration of multiple metrics into a unified system, enabling more objective and comprehensive comparisons of language models. The study contributes to the field of natural language processing by providing a tool for assessing the quality of language models.

Keywords: natural language processing, text paraphrasing, GigaChat, YandexGPT 2, ChatGPT-3.5, ChatGPT-4, Gemini, Bing AI, YouChat, Mistral Large.

For citation: Dagaev A.E., Popov D.I. Evaluation of the quality of intelligent text paraphrasing in Russian. *Modeling, Optimization and Information Technology.* 2024;12(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1763 DOI: 10.26102/2310-6018/2024.47.4.038

Введение

Перефразирование текста является важным направлением, оказывающим влияние на большое количество задач в области обработки естественного языка. Актуальность данной темы подтверждается тем, что такие тексты (парафразы) используются в разнообразных сферах, включая задачи информационного поиска, где они способствуют улучшению точности и полноты результатов, а также в задачах повышения автоматизированной оценки качества систем машинного перевода, где ручное редактирование эталонных текстов может оказаться дорогим. Помимо того, существующие исследования сосредоточены в основном на англоязычных моделях, тогда как разработка и оценка моделей для русского языка остаются недостаточно изученными.

Для эффективного перефразирования модель искусственного интеллекта должна создавать связное и релевантное содержание, при этом сохраняя исходный смысл [1, 2]. Однако, несмотря на значительный прогресс в развитии моделей искусственного интеллекта, большая часть исследований сосредоточена на английском языке, тогда как модели для русского языка остаются менее изученными. Существующие подходы к оценке качества перефразирования зачастую опираются на отдельные метрики, такие как ROUGE или BLEU, которые оценивают лишь частные аспекты текста и не всегда дают полное представление о его качестве.

В данном исследовании проводится сравнительный анализ популярных моделей искусственного интеллекта для перефразирования текстов на русском языке. Особое внимание уделяется тому, насколько эффективно эти модели генерируют перефразирования, которые не только точны и последовательны, но, что важно, соответствуют контексту и сохраняют оригинальную семантику текста [3, 4].

Оригинальность использованного подхода состоит в объединении лексических, структурных и семантических метрик в рамках интегрального показателя, что делает возможным более объективное и комплексное сравнение языковых моделей. Включение метрик BERTScore и BLEURT обеспечивает более глубокий анализ, выходящий за рамки традиционных методов.

Целью данного исследования является разработка интегрального показателя, объединяющего разные метрики для более объективной оценки качества моделей перефразирования. Этот показатель позволит учитывать лексические, семантические и структурные аспекты текста.

Для решения поставленной цели были выполнены следующие задачи:

1. Проведен выбор ключевых метрик для оценки качества перефразирования.
2. Определены весовые коэффициенты для каждой метрики на основе их значимости.
3. Подобраны репрезентативные наборы данных для русского и английского языков.

4. Реализована методика вычисления интегрального показателя.

5. Проведен сравнительный анализ производительности популярных языковых моделей.

Несмотря на значительный прогресс в разработке и оценке языковых моделей большинство исследований сосредоточены на англоязычном контенте. Между тем, сравнительные исследования моделей, таких как GigaChat, YandexGPT 2, ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Gemini, Bing AI, YouChat и Mistral Large, для русского языка остаются недостаточно изученными. Важно отметить, что современные подходы к оценке качества перефразированных текстов часто используют методы, основанные на сравнении с эталонами (reference-based), однако исследование [5] показывает, что метрики без использования эталонных текстов (reference-free) демонстрируют лучшую производительность по сравнению с метриками, использующими эталонные тексты (reference-based). В работе [6] подчеркивается важность использования многоязычных наборов данных, что особенно актуально для оценки качества перефразирования текстов на разных языках, включая русский. Значимость содержания исходных текстов при оценке эффективности языковых моделей отражена в работе [7].

Оценка качества перефразирования текста с использованием искусственного интеллекта включает в себя различные методологии и применения. Han и др. (2022) предложили оценочные рамки с использованием NLP, которые включают лексические и синтаксические особенности, что позволяет более структурировано оценивать задачи перефразирования и перевода [8]. В другом исследовании [9] была представлена автоматическая система для оценки текстовых обобщений, ориентированная на проверку фактической согласованности, полноту информации и коэффициент сжатия. Эта система также адаптируема для анализа эффективности перефразирования текстов. Nicula и др. (2021) представили автоматическую оценку качества перефразирования с использованием рекуррентных нейронных сетей и языковых моделей, подчеркивая важность оценки лексического, синтаксического и семантического сходства для своевременной обратной связи в образовательных целях [10]. Таким образом, для сравнительного анализа эффективности моделей искусственного интеллекта необходимо использовать разные языковые наборы данных и использовать комплекс метрик для оценки ключевых аспектов полученных текстов.

Материалы и методы

В исследовании для оценки качества интеллектуального перефразирования используются наборы данных на русском языке, а также контрольные наборы на английском языке. Их выбор имеет важное значение для обеспечения всесторонней оценки возможностей моделей в разных языках и типах текста. В работе используются следующие наборы данных на русском языке:

Gazeta [11]. Набор данных Gazeta состоит из 63,435 новостных статей с российского сайта новостей «gazeta.ru». Этот набор данных предоставляет крупный источник современных новостных текстов, охватывающих широкий спектр тем и содержащий разные стили написания. Он особенно полезен для оценки производительности моделей на журналистских текстах, требующих точного перефразирования при сохранении целостности исходной информации.

XL-Sum [12]. Многоязычный набор данных, включающий 1,35 миллионов аннотированных пар статей на разных языках, из которых 77,803 пары на русском языке. Этот набор данных хорошо подходит для задач перефразирования, предоставляя сложную текстовую среду для моделей. Он охватывает разнообразные темы и жанры, повышая надежность оценки.

WikiLingua [13]. Многоязычный набор данных, содержащий 52,928 статей на русском языке из WikiHow. Этот набор данных включает учебные тексты, которые имеют иную структуру по сравнению с новостными статьями.

В качестве наборов данных на английском языке были взяты:

CNN Dailymail [14]. Набор данных CNN Dailymail включает 311,672 новостные статьи с сайтов CNN (с апреля 2007 по апрель 2015 года) и Daily Mail (с июня 2010 по апрель 2015 года). Этот набор данных широко используется в исследованиях по обобщению и перефразированию текста, предлагая большой корпус высококачественных, хорошо структурированных текстов.

XSum [15]. Набор данных, содержащий 226,711 статей BBC, опубликованных в период с 2010 по 2017 годы. Содержащиеся статьи также охватывают широкий спектр тем, включая новости, аналитические блоки, репортажи и другие журналистские материалы.

WikiLingua [13]. Английская часть набора данных WikiLingua, подобная его русскому аналогу и содержащая инструктивные статьи из WikiHow в количестве 141,457. Она предлагает схожий контекст, но на английском языке, что позволяет проводить кросс-лингвистическое сравнение способностей моделей по перефразированию текстов.

В исследовании применена случайная выборка из 100 оригинальных текстов, стандартизированных до 1024 токенов.

Для оценки перефразирования важно использовать комбинацию нескольких метрик, чтобы получить всесторонние и точные выводы эффективности моделей [16]. Комплексный подход уменьшает риск того, что итоговая оценка будет подвержена влиянию сильных или слабых сторон случайной метрики. Расчет интегрального показателя произведен по формуле 1.

$$K_{\text{интегр}} = \sum_{i=1}^{|SRC|} K_i \cdot \sum_{j=1}^{|Metr|} g_j \cdot m_{ij}, \quad (1)$$

где K_i – вес набора данных; g_j – вес метрики; m_{ij} – значение метрики j для набора данных i.

В работе были использованы следующие метрики:

ROUGE-1. Измеряет пересечение униграмм (на уровне слов), позволяя оценить сохранение основного содержания и лексическое сходство между сгенерированным и эталонным текстами [17].

ROUGE-2. Оценивает пересечение биграмм (пар слов) [17], предоставляя информацию о связности сгенерированного текста.

ROUGE-L. Анализирует самую длинную общую подпоследовательность, что помогает понять структурное сходство и сохранение длинных фраз в сгенерированном тексте [17].

BLEU. Измеряет точность *n*-грамм (последовательностей слов) в сгенерированном тексте по сравнению с эталонными текстами [18]. Данный показатель особенно полезен для оценки грамматической и синтаксической правильности сгенерированного текста. Однако он не всегда демонстрирует высокую корреляцию с человеческими оценками семантического содержания.

BERTScore. Использует векторное представление слов BERT для вычисления сходства между сгенерированными и эталонными текстами на более глубоком семантическом уровне. Он учитывает особенности языка и контекстуальную релевантность, что важно для оценки семантической точности перефразированных

текстов. BERTScore демонстрирует высокую корреляцию с оценками человека и обеспечивает надежное измерение семантического содержания [19].

METEOR. Эта метрика учитывает точность, полноту *n*-грамм, а также включает синонимы, основы и порядок слов [20]. Она предоставляет всестороннюю оценку качества текста, поскольку учитывает разнообразные аспекты лексики и структуры. METEOR стремится устранить ограничения BLEU, уделяя особое внимание адекватности текста, что позволяет более точно сопоставить его с оригиналом по качеству перефразирования.

BLEURT. Показатель сочетает преимущества векторного представления слов BERT и обученной модели оценки для вычисления высокоточного параметра качества сгенерированных текстов. BLEURT оценивает как лингвистическое качество, так и сохранение смысла, делая его всеобъемлющей метрикой для задач генерации текста [21].

Используя эти метрики в комбинации, можно достичь более сбалансированной и всесторонней оценки перефразирования. Каждая метрика охватывает разные аспекты анализа текста, от лексического и структурного сходства до семантической точности и контекстуальной релевантности. В Таблице 1 отражено распределение весов использованных метрик.

Таблица 1 – Распределение весов метрик
Table 1 – Distribution of metric weights

Метрика	Вес	Обоснование
ROUGE-1	0,05	Ограниченная значимость для оценки семантики текста
ROUGE-2	0,05	Важно для связности, но фокусируется только на лексике
ROUGE-L	0,10	Отражает структурное сходство и сохранение порядка слов
BLEU	0,10	Полезно для грамматической оценки, но ограничено в учете семантики
BERTScore	0,30	Высокая корреляция с человеческими оценками
METEOR	0,20	Учитывает синонимы и порядок слов, компенсирует ограничения BLEU
BLEURT	0,20	Сочетает семантический и лингвистический анализ, высокая точность

Результаты

Для русскоязычных наборов данных результаты по всем указанным ранее показателям представлены в Таблице 2. Итоговые оценки моделей были получены путем усреднения значений метрик для каждой модели на всех наборах данных.

Таблица 2 – Показатели метрик на всех русскоязычных наборах данных
Table 2 – Metrics performance on all Russian-language datasets

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,71	0,57	0,69	0,61	0,82	0,68	0,75
YandexGPT2	0,72	0,57	0,69	0,61	0,82	0,67	0,74
ChatGPT-3.5	0,72	0,59	0,70	0,61	0,83	0,69	0,75
ChatGPT-4	0,75	0,61	0,72	0,63	0,86	0,71	0,78
ChatGPT-4o	0,74	0,60	0,71	0,64	0,85	0,70	0,77
Gemini	0,69	0,56	0,66	0,59	0,78	0,64	0,71
Bing AI	0,72	0,57	0,69	0,61	0,82	0,68	0,76
YouChat	0,69	0,55	0,66	0,58	0,79	0,65	0,72
Mistral Large	0,72	0,58	0,69	0,62	0,82	0,68	0,77

В таблице 3 представлены расчеты интегрального показателя.

Таблица 3 – Показатели интегрального расчета на русскоязычных наборах данных
Table 3 – Integral calculation indicators on Russian-language datasets

	$K_{\text{интегр}}$
ChatGPT-4	0,76
ChatGPT-4o	0,75
ChatGPT-3.5	0,73
Mistral Large	0,73
Bing AI	0,73
GigaChat	0,73
YandexGPT 2	0,73
YouChat	0,70
Gemini	0,69

На Рисунке 1 изображена диаграмма полученных оценок среди всех наборов данных на русском языке.

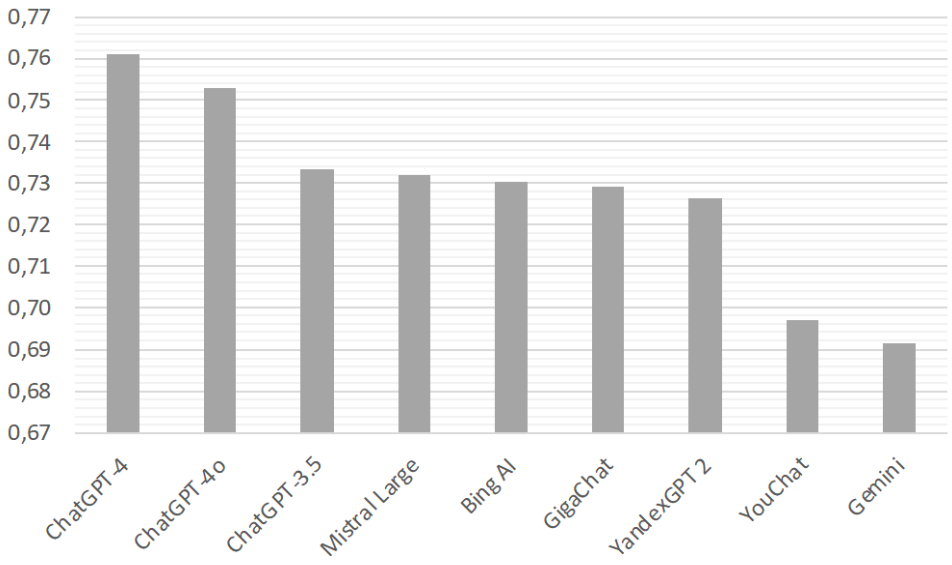

Рисунок 1 – Диаграмма оценок для наборов данных на русском языке
Figure 1 – Rating chart for datasets in Russian

На англоязычных наборах данных был исключен YandexGPT 2 из списка моделей для дальнейшего сравнения по причине отсутствия возможности работы с англоязычными текстами. В Таблице 4 представлены результаты по всем наборам на английском языке.

Таблица 4 – Показатели метрик на всех англоязычных наборах данных
Table 4 – Metrics performance on all English-language datasets

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	METEOR	BLEURT
GigaChat	0,73	0,59	0,72	0,63	0,79	0,72	0,80
ChatGPT-3.5	0,70	0,55	0,65	0,61	0,75	0,69	0,74
ChatGPT-4	0,74	0,59	0,68	0,64	0,77	0,69	0,76
ChatGPT-4o	0,74	0,60	0,71	0,64	0,78	0,69	0,77
Gemini	0,70	0,56	0,67	0,60	0,74	0,66	0,73
Bing AI	0,72	0,57	0,69	0,61	0,75	0,68	0,75
YouChat	0,69	0,55	0,66	0,59	0,75	0,65	0,72
Mistral Large	0,72	0,57	0,69	0,63	0,75	0,68	0,75

В таблице 5 представлены расчеты интегрального показателя.

Таблица 5 – Показатели интегрального расчета на англоязычных наборах данных
Table 5 – Integral calculation indicators on English-language datasets

	$K_{\text{интегр}}$
GigaChat	0,74
ChatGPT-4o	0,73
ChatGPT-4	0,72
Mistral Large	0,71
Bing AI	0,71
ChatGPT-3.5	0,70
Gemini	0,69
YouChat	0,69
GigaChat	0,74

На Рисунке 2 изображена диаграмма оценок среди всех наборов данных английского языка.

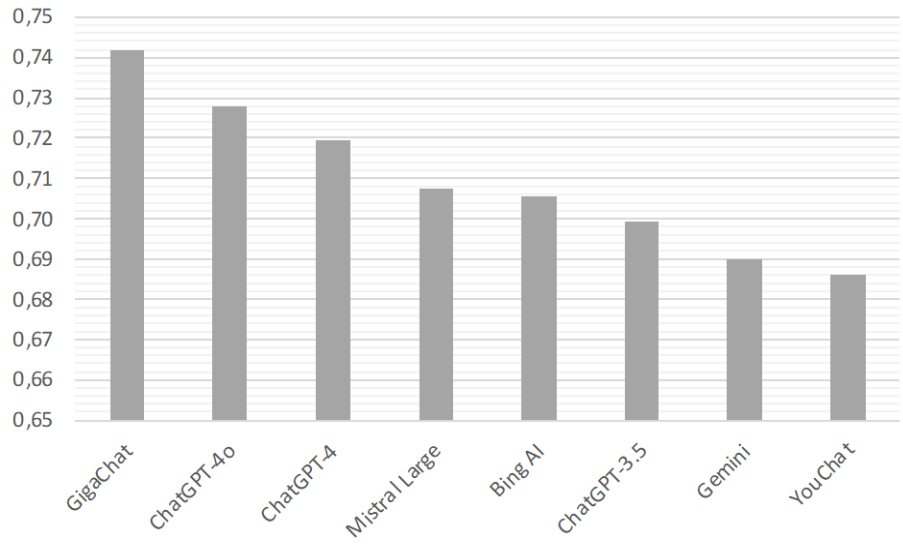

Рисунок 2 – Диаграмма оценок для наборов данных на английском языке
Figure 2 – Rating chart for datasets in English

Обсуждение

Исследование подтверждает высокую эффективность предложенной интегральной метрики для оценки качества перефразирования, применимой к современным языковым моделям. Оригинальность подхода состоит в комплексном использовании нескольких метрик, каждая из которых обладает собственным весом, что позволяет учитывать лексические, семантические и структурные особенности текста.

Наибольшую эффективность на русскоязычных наборах данных продемонстрировали модели ChatGPT-4 и ChatGPT-4o, показав высокую точность и связность генерируемых текстов. Это подчеркивает их эффективность в обработке и генерации текста, а также его способность точно воспроизводить ключевые элементы содержания. Высокие значения метрик свидетельствуют о его превосходстве в сохранении структуры и смысла исходного текста, обеспечивая при этом точное и естественное перефразирование.

Gemini продемонстрировал более низкие результаты как по подавляющему количеству метрик, так и по интегральной оценке. Низкие показатели этой модели делают ее менее предпочтительным выбором для задач, требующих высокого качества перефразирования.

В то же время на англоязычных наборах данных наилучшие результаты показал GigaChat. Полученная оценка свидетельствует о лучшей способности генерировать связные, точные и контекстуально релевантные тексты. Следующими в рейтинге стали ChatGPT-4o и ChatGPT-4, которые также демонстрирует высокую производительность, подтверждая свою способность эффективно справляться с такими задачами, как и на русскоязычных наборах. ChatGPT-3.5 отстает по качеству в сравнении с новыми моделями, что подчеркивает прогресс, достигнутый в последующих версиях.

Модели Gemini и YouChat показали низкие результаты на англоязычных наборах данных. Учитывая отставание и на русскоязычных наборах, можно отметить, что данные модели имеют сложности в поддержке контекстуальной и семантической точности. Это также выразилось в низких значениях соответствующих метрик, что говорит об ограниченной способности этих моделей воспроизводить сложные языковые конструкции и семантические связи.

Научная новизна исследования заключается в использовании интегральной метрики, способной объективно сравнивать модели перефразирования на разных языках, а также в выявлении сильных и слабых сторон современных языковых моделей. Включение метрик, обеспечивающих высокую корреляцию с человеческими оценками, существенно усиливает достоверность результатов оценки.

Практическая значимость работы заключается в предоставлении исследователям инструмента для объективного анализа и оценки качества моделей, что может быть полезным как в задачах автоматизированного перефразирования, так и в задачах автоматического обобщения текстов и машинного перевода.

Заключение

В данном исследовании была представлена и подтверждена эффективность интегральной метрики для оценки качества перефразирования, генерируемого современными языковыми моделями. Метрика обеспечивает сбалансированную и комплексную оценку лингвистических, структурных и семантических характеристик перефразированных текстов. Интегральная метрика вносит вклад в оценку языковых моделей, предлагая инструмент для оценки качества на разных языках и задачах. Ее внедрение может способствовать дальнейшему развитию более точных, согласованных и семантически насыщенных языковых моделей.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Xie J., Agrawal A. Emotion and Sentiment Guided Paraphrasing. In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, *13 July 2023, Toronto, Canada*. Association for Computational Linguistics; 2023. pp. 58–70. https://doi.org/10.18653/v1/2023.wassa-1.7
2. Krishna K., Song Y., Karpinska M., Wieting J., Iyyer M. Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense. In: *Advances in Neural Information Processing Systems: 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, *10–16 December 2023, New Orleans, USA*. Curran Associates; 2024. https://doi.org/10.48550/arXiv.2303.13408

3. Sadasivan V.S., Kumar A., Balasubramanian S., Wang W., Feizi S. Can AI-Generated Text be Reliably Detected? arXiv. URL: https://doi.org/10.48550/arXiv.2303.11156 [Accessed 14th November 2024].
4. Verma D., Lal Y.K., Sinha S., Van Durme B., Poliak A. Evaluating Paraphrastic Robustness in Textual Entailment Models. arXiv. URL: https://doi.org/10.48550/arXiv.2306.16722 [Accessed 14th November 2024].
5. Shen L., Liu L., Jiang H., Shi S. On the Evaluation Metrics for Paraphrase Generation. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 07–11 December 2022, Abu Dhabi, United Arab Emirates.* Association for Computational Linguistics; 2022. pp. 3178–3190.
6. Weston J., Lenain R., Meepegama U., Fristed E. Generative Pretraining for Paraphrase Evaluation [Preprint]. arXiv. URL: https://doi.org/10.48550/arXiv.2107.08251 [Accessed 14th November 2024].
7. Sharma S., Joshi A., Mukhija N., Zhao Y., Bhathena H., Singh P., Santhanam S., Biswas P. Systematic review of effect of data augmentation using paraphrasing on Named entity recognition. In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 28 November – 09 December 2022, New Orleans, USA.*
8. Han T., Li D., Ma X., Hu N. Comparing product quality between translation and paraphrasing: Using NLP-assisted evaluation frameworks. *Frontiers in Psychology*. 2022;13. https://doi.org/10.3389/fpsyg.2022.1048132
9. Ahn J., Khosmood F. Evaluation of Automatic Text Summarization using Synthetic Facts. arXiv. URL: https://doi.org/10.48550/arXiv.2204.04869 [Accessed 14th November 2024].
10. Nicula B., Dascalu M., Newton N., Orcutt E., McNamara D.S. Automated Paraphrase Quality Assessment Using Recurrent Neural Networks and Language Models. In: *Intelligent Tutoring Systems: 17th International Conference, ITS 2021: Proceedings, 07–11 June 2021, Online*. Cham: Springer; 2021. pp. 333–340. https://doi.org/10.1007/978-3-030-80421-3_36
11. Gusev I. Dataset for Automatic Summarization of Russian News. In: *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020: Proceedings*, *07–09 October 2020, Helsinki, Finland*. Cham: Springer; 2020. pp. 122–134. https://doi.org/10.1007/978-3-030-59082-6_9
12. Hasan T., Bhattacharjee A., Islam M.S., Mubasshir K., Li Y.-F., Kang Y.-B., Rahman M.S., Shahriyar R. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, *01–06 August 2021, Online*. Association for Computational Linguistics; 2021. pp. 4693–4703. https://doi.org/10.18653/v1/2021.findings-acl.413
13. Ladhak F., Durmus E., Cardie C., McKeown K. WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, *16–20 November 2020, Online*. Association for Computational Linguistics; 2020. pp. 4034–4048. https://doi.org/10.18653/v1/2020.findings-emnlp.360
14. Nallapati R., Zhou B., Dos Santos C., Gülçehre Ç., Xiang B. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, *11–12 August 2016, Berlin, Germany*. Berlin: Association for Computational Linguistics; 2016. pp. 280–290. https://doi.org/10.18653/v1/K16-1028

15. Narayan S., Cohen S.B., Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, *31 October – 04 November 2018, Brussels, Belgium*. Association for Computational Linguistics; 2018. pp. 1797–1807. https://doi.org/10.18653/v1/D18-1206
16. Patil O., Singh R., Joshi T. Understanding Metrics for Paraphrasing. arXiv. URL: https://doi.org/10.48550/arXiv.2205.13119 [Accessed 14th November 2024].
17. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, *25–26 July 2004, Barcelona, Spain*. Association for Computational Linguistics; 2004. pp. 74–81.
18. Zhang T., Kishore V., Wu F., Weinberger K.Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT. In: *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, 26–30 April 2020, Addis Ababa, Ethiopia*. Addis Ababa: International Conference on Learning Representations; 2020. pp. 1–43. https://doi.org/10.48550/arXiv.1904.09675
19. Banerjee S., Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, *29 June 2005, Ann Arbor, USA*. Association for Computational Linguistics; 2005. pp. 65–72.
20. Post M. A Call for Clarity in Reporting BLEU Scores. In: *Proceedings of the Third Conference on Machine Translation: Research Papers*, *31 October – 01 November 2018, Brussels, Belgium*. Association for Computational Linguistics; 2018. pp. 186–191. https://doi.org/10.18653/v1/W18-6319
21. Sellam T., Das D., Parikh A. BLEURT: Learning Robust Metrics for Text Generation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *05–10 July 2020, Online*. Association for Computational Linguistics; 2020. pp. 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Дагаев Александр Евгеньевич, аспирант кафедры информатики и информационных технологий, Московский политехнический университет, Москва, Российская Федерация.
e-mail: a.e.dagaev@staff.mospolytech.ru

Alexander E. Dagaev, postgraduate student of the Department of Informatics and Information Technologies, Moscow Polytechnic University, Moscow, the Russian Federation.

Попов Дмитрий Иванович, доктор технических наук, профессор, Сочинский государственный университет, Сочи, Российская Федерация.
e-mail: damitry.popov@gmail.com

Dmitry I. Popov, Doctor of Technical Science, Professor, Sochi State University, Sochi, the Russian Federation.

Статья поступила в редакцию 05.12.2024; одобрена после рецензирования 23.12.2024; принята к публикации 25.12.2024.

The article was submitted 05.12.2024; approved after reviewing 23.12.2024; accepted for publication 25.12.2024.