ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2026. Т. 30, № 2 (112). С. 64-73 http://journal.uust.ru/index.php/Vestnik

УДК 004.912 doi 10.54708/19926502_2026_30211264
Код ГРНТИ 16.31.21

Сравнительный анализ методов автоматического реферирования научных текстов на русском языке

А.В. Кан[1], А.А. Хорошилов[1], В.Н. Гокарев[2*],
А.Ф. Беззубов[2], А.А. Хорошилов[1]

[1]Всероссийский институт научной и технической информации Российской академии наук, Московский авиационный институт, г. Москва, Россия
[2]Центральный научно-исследовательский институт Министерства обороны России, г. Москва, Россия

Аннотация. Автоматическое реферирование научных текстов остается актуальной задачей обработки естественного языка, особенно для русскоязычных корпусов, где применение современных методов ограничено недостатком специализированных моделей и наборов данных. В данной работе проведен систематический сравнительный анализ экстрактивных и абстрактивных моделей суммаризации на материале русскоязычного мультимодального датасета научных статей. Для оценки качества применен комплексный набор метрик: классические статистические (ROUGE, TF-IDF) и современные нейросемантические (BERTScore, косинусная мера между эмбеддингами). Результаты показывают, что экстрактивные методы (TextRank, LexRank) обеспечивают высокую лексическую точность (ROUGE-1 F1 = 0,38; TF-IDF = 0,87), в то время как нейросетевые абстрактивные модели (rut5_base_sum_gazeta, mbart_ru_sum_gazeta) демонстрируют превосходство по семантическим метрикам (BERTScore F1 = 0,70; косинусное сходство ≈ 0,80), приближаясь к уровню человеческого реферирования. Исследование показывает, что выбор модели должен определяться конкретной задачей: экстрактивные методы предпочтительны для формализованных текстов, тогда как абстрактивные обеспечивают более естественное и гибкое изложение.

Ключевые слова: *автоматическое реферирование, суммаризация текстов, обработка естественного языка, нейросетевые модели, метрики качества, русский язык, научные тексты.*

***vgokarev@mail.ru**

Введение

Экспоненциальный рост объема научной литературы создает потребность в эффективных методах автоматической конденсации ключевой информации. Автоматическое реферирование направлено на создание компактных и семантически эквивалентных представлений исходных текстов, что особенно актуально для научно-технической документации.

Существующие подходы делятся на экстрактивные (извлечение предложений), абстрактивные (генерация нового текста) и гибридные [1, 2]. Критическим аспектом является оценка качества, где традиционные статистические метрики (напр., ROUGE [3]) измеряют лексическое сходство, а современные нейросетевые – семантическую близость.

Для русского языка существует дефицит специализированных моделей и датасетов для реферирования научных текстов, так как большинство корпусов [4–6] ориентированы на новостной домен. Это создает разрыв между потребностями научного сообщества и доступными инструментами.

Цель данной работы – провести систематический сравнительный анализ экстрактивных и абстрактивных методов автоматического реферирования научных текстов на русском языке с использованием комплексного набора классических и нейросемантических метрик. Исследование направлено на выявление сильных и слабых сторон каждого класса моделей, определение подходов, наиболее приближенных к человеческому реферированию, и формулирование рекомендаций по выбору методов для практических приложений в документных системах и научных библиотеках.

Классы моделей реферирования

В рамках экспериментального исследования использовались представители двух основных классов методов автоматического реферирования: экстрактивные и абстрактивные методы.

Экстрактивные методы предполагают выбор наиболее значимых предложений, фраз или фрагментов из исходного документа без их изменения. Сформированное краткое содержание является подмножеством оригинального текста, что обеспечивает точное воспроизведение исходной информации, но может ограничивать связность и естественность текста.

Метод TextRank [7] представляет собой графовый алгоритм, основанный на ранжировании предложений по принципу PageRank. Метод строит граф, где вершины соответствуют предложениям, а рёбра отражают их семантическую близость, после чего выделяет наиболее значимые узлы. Метод LexRank [8] аналогичен TextRank, но использует иной механизм вычисления весов рёбер, основанный на косинусной мере TF-IDF представлений предложений. Лингвистический суммаризатор основан на системе правил, учитывающих синтаксическую структуру, позицию предложения в тексте, наличие ключевых слов и терминов. Алгоритм реализует детерминированную логику отбора, что обеспечивает воспроизводимость результатов и высокую интерпретируемость принятых решений.

Абстрактивные методы предполагают генерацию принципиально новых формулировок, обобщающих содержание исходного текста:

– mT5_multilingual_XLSum: многоязычная T5 [9] модель, дообученная на датасете XL-Sum [6].

– mbart_ru_sum_gazeta: модель на базе BART [10], дообученная на русскоязычном новостном корпусе Gazeta [4].

– rut5_base_sum_gazeta: русскоязычная версия T5 [9, 11], также дообученная на корпусе Gazeta [4].

– t5_summary_en_ru_zh_base_2048: многоязычная T5-модель с контекстным окном до 2048 токенов.

Метрики оценки качества реферирования

Оценка качества автоматического реферирования требует применения разнородных метрик, способных охватить как поверхностные лексические характеристики, так и глубинную семантическую адекватность. Для комплексной оценки качества автоматического реферирования использованы две группы метрик: статистические, основанные на поверхностных лексических характеристиках, и нейросемантические, учитывающие глубокие контекстуальные связи.

Статистические метрики основаны на измерении лексического совпадения между автоматически сгенерированным рефератом и эталонным текстом. Они отличаются высокой скоростью вычисления и прозрачностью алгоритмов, но склонны недооценивать семантически эквивалентные перефразирования. Семейство метрик ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [3] оценивает совпадение n-грамм (ROUGE-1), биграмм (ROUGE-2) и самых длинных общих подпоследовательностей (ROUGE-L) между рефератами. Косинусная мера TF-IDF векторов измеряет лексическое сходство через косинус угла между TF-IDF векторами текстов.

Нейросетевые метрики базируются на предобученных языковых моделях, вычисляющих контекстуальные представления текстов. Они способны фиксировать семантическую близость независимо от буквальных совпадений, но требуют значительных вычислительных ресурсов. Метрика BERTScore [12] вычисляет семантическое сходство на основе контекстуализированных эмбеддингов токенов из BERT-подобных моделей. Косинусная мера между эмбеддингами оценивает семантическую близость полных текстов с помощью sentence-encoder моделей [13]. Для надежности оценка усреднялась по четырем моделям. В исследовании использовались четыре модели эмбеддингов, адаптированные для

русского языка: rubert-tiny2, LaBSE-ru-sts, multilingual_en_uk_pl_ru и rubert-base-cased. Использование нескольких моделей эмбеддингов позволяет повысить объективность оценки за счет усреднения результатов и компенсации специфических смещений отдельных моделей.

Комплексный набор метрик был выбран исходя из необходимости охватить весь спектр характеристик качества реферирования. Статистические метрики (ROUGE, TF-IDF) обеспечивают быстрые вычисления и прозрачную интерпретацию, фиксируя степень лексического совпадения и сохранения исходных формулировок. Нейросемантические метрики (BERTScore, эмбеддинги) дополняют анализ оценкой глубинной семантической близости, учитывая перефразирования и синонимию. Такой подход позволяет разграничить модели, склонные к буквальному копированию (высокие статистические метрики), и модели, ориентированные на семантически адекватную генерацию (высокие нейросемантические метрики при умеренных статистических). Комбинация статистических и нейросетевых подходов позволяет получить сбалансированную оценку [14], выявляя как сильные, так и слабые стороны различных моделей реферирования.

Описание набора данных и методология эксперимента

В качестве основы для экспериментов использовался русскоязычный мультимодальный датасет для автоматического реферирования научных статей [15], включающий 480 документов из восьми предметных областей: экономика, история, информационные технологии, журналистика, право, лингвистика и медицина. Каждая область представлена 60 статьями, что обеспечивает сбалансированное распределение по дисциплинам и позволяет оценить устойчивость моделей к доменной специфике.

Эксперимент проводился в два этапа, каждый из которых раскрывает различные аспекты качества моделей суммаризации. На первом этапе выполнялось сравнение автоматически сгенерированных рефератов (СР) с полными оригинальными текстами (ОТ). На втором этапе осуществлялось сопоставление СР с авторскими рефератами (АР), подготовленными человеком и принятыми за эталон.

Ключевой принцип методологии заключается в том, что качество модели оценивается не только по прямой близости СР к АР, но и по тому, насколько соотношение «СР → ОТ» приближается к эталонному «АР → ОТ». Если значения СР–ОТ по метрикам заметно превышают АР–ОТ, это указывает на чрезмерное копирование фрагментов из оригинала, нехарактерное для качественного человеческого резюмирования. Обратная ситуация, когда СР–ОТ значительно ниже АР–ОТ, сигнализирует о потере информативности или искажении содержания.

Результаты метрики ROUGE

Метрика ROUGE оценивает степень совпадения n-грамм между текстами, фиксируя лексическую близость на поверхностном уровне. Анализ метрик ROUGE при сравнении автоматически сгенерированных рефератов с оригинальными текстами выявляет четкую дифференциацию между лингвистическим суммаризатором и нейросетевыми моделями.

Для пары ОТ-АР (Табл. 1) высокие значения точности (особенно для ROUGE-1 и ROUGE-L) отражают то, что лексемы авторских рефератов в значительной степени встречаются в исходных текстах – ожидаемый эффект, поскольку реферат является конденсацией содержания. Низкие значения полноты и F1-меры объясняются большой разницей в объёме: реферат содержит лишь небольшую долю всех n-грамм оригинала.

Таблица 1. Средние значения метрики ROUGE между ОТ и АР.

Метрика	Точность	Полнота	F1-мера
ROUGE-1	0,4907	0,0301	0,0522
ROUGE-2	0,2877	0,0189	0,0323
ROUGE-L	0,4744	0,0280	0,0487

При сравнении с оригинальными текстами (ОТ-СР) (Табл. 2) лингвистический суммаризатор демонстрирует максимальную точность (0,9048) и наивысшую F1-меру (0,2260), что свидетельствует о его способности извлекать фрагменты, насыщенные ключевыми униграммами оригинала. Методы TextRank демонстрирует наилучшие F1-значения по всем трём измерениям (ROUGE-1 F1 = 0,3798), с очень высокой точностью и полнотой (0,9749, 0,2463).

Таблица 2. Средние значения метрики ROUGE между ОТ и СР.

Модель	ROUGE-1			ROUGE-2	ROUGE-L
	Точность	Полнота	F1-мера		
Лингвистический суммаризатор	0,9048	0,1398	0,2260	0,1754	0,1995
TextRank	**0,9749**	**0,2463**	**0,3798**	**0,2781**	**0,3798**
LexRank	0,4656	0,0179	0,0329	0,0205	0,0310
mT5_multilingual_XLSum	0,4138	0,0144	0,0268	0,0075	0,0262
mbart_ru_sum_gazeta	0,7209	0,0427	0,0775	0,0437	0,0746
rut5_base_sum_gazeta	0,6997	0,0377	0,0688	0,0371	0,0673
t5_en_ru_zh_base_2048	0,7266	0,0423	0,0766	0,0409	0,0752

Нейросетевые модели показывают контрастирующую картину: точность в диапазоне 0,413–0,727, но крайне низкую полноту (0,015–0,043). Это объясняется тем, что трансформеры генерируют лаконичные рефераты с перефразированием, что снижает лексическое совпадение с объемным оригиналом.

Динамика метрики ROUGE-2 усиливает выявленные тенденции: лингвистический метод сохраняет лидерство (F1: 0,1754), тогда как нейросетевые модели демонстрируют еще более выраженное снижение показателей из-за чувствительности ROUGE-2 к точным последовательностям из двух слов.

При сравнении с авторскими рефератами (сравнительная пара СР-АР) (Табл. 3) структура результатов изменяется. Данные в таблице демонстрируют значения ROUGE-1 для этой пары.

Таблица 3. Средние значения метрики ROUGE между АР и СР.

Модель	ROUGE-1			ROUGE-2	ROUGE-L
	Точность	Полнота	F1-мера		
Лингвистический суммаризатор	0,1042	0,2876	0,1247	0,0545	0,1111
TextRank	0,0742	**0,3639**	0,1025	0,0520	0,0894
LexRank	**0,1667**	0,1108	0,1151	0,0708	0,1049
mT5_multilingual_XLSum	0,1103	0,1154	0,0779	0,0254	0,0758
mbart_ru_sum_gazeta	0,1611	0,2210	**0,1439**	**0,0599**	**0,1366**
rut5_base_sum_gazeta	0,1580	0,1878	0,1326	0,0567	0,1252
t5_en_ru_zh_base_2048	0,1313	0,1644	0,1100	0,0401	0,1043

Абстрактивные нейросетевые модели (mbart и rut5) занимают ведущие позиции по F1 при сравнении с авторскими рефератами (F1 = 0,1439 и 0,1326 соответственно).

Экстрактивные методы имеют сопоставимые или несколько меньшие показатели. Во всех случаях абсолютные значения F1 невысоки (максимум 0,125), что указывает на существенное расхождение между автоматическими выводами и авторским рефератом в лексической форме. Лингвистический суммаризатор сохраняет высокую F1-меру (0,1247), более чем вдвое превосходящую эталонное значение оригинального текста (0,0522)

и сбалансированные показатели (ROUGE-1 точность – 0,104, полнота – 0,2876, F1-мера – 0,1247).

Низкие абсолютные F1 при сравнении с АР подчёркивают ограниченную способность метрики ROUGE адекватно фиксировать семантическое совпадение при сильном перефразировании.

Важно отметить существенное изменение соотношения точности и полноты – нейросетевые модели теперь показывают более сбалансированные показатели (точность 0,110–0,161, полнота 0,115–0,221), что указывает на их способность к перефразированию с сохранением лексического разнообразия, характерного для человеческих аннотаций.

Результаты расчета косинусной меры TF-IDF

Косинусное сходство TF-IDF векторов отражает лексико-тематическую близость текстов. Значение около 0,46 для пары ОТ-АР (Рис. 1, Табл. 4) отражает типичную степень лексико-тематической конденсации при человеческом реферировании.

Для пары ОТ-СР метод TextRank демонстрирует наивысшее сходство (0,8686), значительно превосходя эталонное значение. Это подтверждает, что экстрактивные методы сохраняют и повторно используют те же терминологические маркеры, которые присутствуют в оригинале. Нейросетевые модели показывают более низкие значения (0,39–0,59), что указывает на перефразирование и введение новых лексических единиц.

При сравнении с авторскими рефератами (АР-СР) все модели показывают заметно более низкие значения (0,23–0,42). TextRank вновь занимает лидирующую позицию (0,4154), хотя его значение ниже, чем при сравнении с оригиналом. Это указывает на то, что авторские рефераты содержат иной лексико-стилевой профиль, чем прямые фрагменты, извлекаемые экстрактивными методами.

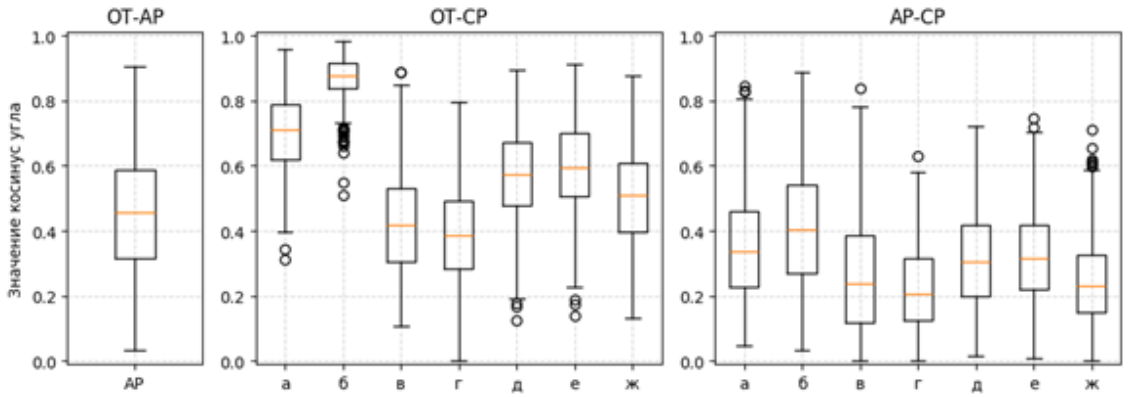

Рисунок 1. Среднее значение косинусной меры TF-IDF для различных пар сравнения (а – лингвистический суммаризатор, б – TextRank, в – LexRank, г – mT5_multilingual_XLSum, д – mbart_ru_sum_gazeta, е – rut5_base_sum_gazeta, ж – t5_summary_en_ru_zh_base_2048).

Таблица 4. Среднее значение косинусной меры TF-IDF для различных пар сравнения.

Модель	ОТ-АР	ОТ-СР	АР-СР
Авторский реферат	0,4582	–	–
Лингвистический суммаризатор	–	0,7008	0,3555
TextRank	–	**0,8684**	**0,4154**
LexRank	–	0,4223	0,2696
mT5_multilingual_XLSum	–	0,3867	0,2250
mbart_ru_sum_gazeta	–	0,5682	0,3119
rut5_base_sum_gazeta	–	0,5950	0,3286
t5_summary_en_ru_zh_base_2048	–	0,5051	0,2488

Как видно из данных в Табл. 4 на Рис. 2, экстрактивные методы (TextRank, лингвистический суммаризатор) демонстрируют максимальную близость к оригинальным текстам по TF-IDF, в то время как абстрактивные модели показывают более умеренные значения, отражая их способность к перефразированию.

Результаты BERTScore

Метрика BERTScore использует контекстуальные эмбеддинги для оценки семантической близости на уровне смысла (Рис. 2, Табл. 5). Уровень F1 ≈ 0,67 для пары ОТ-АР является эталонным ориентиром для оценки машинных моделей, отражая характерный для человеческого суммирования компромисс между точностью и полнотой.

Для пары ОТ-СР лингвистический суммаризатор и TextRank продемонстрировали наивысшие значения F1 (0,7468 и 0,7490), подтверждая, что экстрактивные модели сохраняют лексические структурные элементы, за счет чего растет оценка при сравнении с ОТ. Среди нейросетевых моделей лучшие результаты показала rut5_base_sum_gazeta (F1 = 0,6966), за которой следует mbart_ru_sum_gazeta (F1 = 0,6905).

При сравнении пары АР-СР, наоборот, – нейросетевые модели rut5_base_sum_gazeta и mbart_ru_sum_gazeta показали наивысшие результаты (0,6934 и 0,6890, соответственно), что указывает на значительное пересечение смысловых элементов с авторскими рефератами, остальные модели продемонстрировали устойчивые значения F1 в диапазоне 0,67–0,69, близкие к эталонному уровню (ОТ-АР: 0,6706).

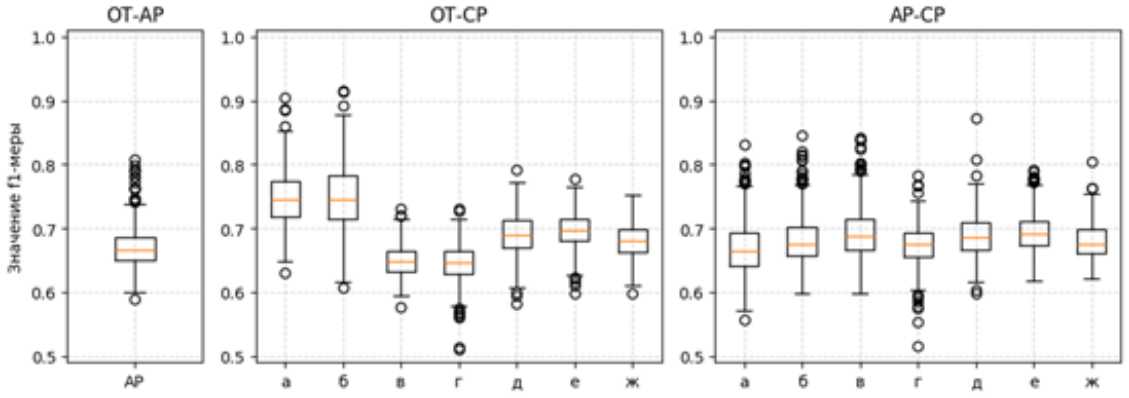

Рисунок 2. Средние значение косинусной меры BERTScore для различных пар сравнения (а – лингвистический суммаризатор, б – TextRank, в – LexRank, г – mT5_multilingual_XLSum, д – mbart_ru_sum_gazeta, е – rut5_base_sum_gazeta, ж – t5_summary_en_ru_zh_base_2048).

Таблица 5. Среднее значение F1-меры метрики BERTScore для различных пар сравнения

Модель	ОТ-АР	ОТ-СР	АР-СР
Авторский реферат	0,6705	–	–
Лингвистический суммаризатор	–	0,7468	0,6691
TextRank	–	**0,7490**	0,6830
LexRank	–	0,6493	0,6831
mT5_multilingual_XLSum	–	0,6454	0,6739
mbart_ru_sum_gazeta	–	0,6905	0,6890
rut5_base_sum_gazeta	–	0,6966	**0,6934**
t5_summary_en_ru_zh_base_2048	–	0,6796	0,6791

Анализ BERTScore (Табл. 5, Рис. 3) выявляет, что экстрактивные подходы обеспечивают более высокую семантическую точность относительно оригинальных текстов, в то время как абстрактивные модели демонстрируют более сбалансированное поведение, приближаясь к человеческому уровню при сравнении с авторскими рефератами.

Результаты расчета косинусной меры между эмбеддингами

Косинусное сходство между эмбеддингами полных текстов оценивает глубинную семантическую близость (Табл. 6). Среднее значение 0,89 для пары ОТ-АР служит эталоном семантической близости при человеческом реферировании. Наиболее высокое значение зафиксировано при использовании labse-ru-sts (0,9803).

Для пары АР-СР лингвистический суммаризатор продемонстрировал наивысшее среднее значение (0.9452), даже превышающее уровень человеческих аннотаций. Среди нейросетевых моделей лучшие результаты показали rut5_base_sum_gazeta (0.9236) и mbart_ru_sum_gazeta (0.9156).

Таблица 6. Среднее значение косинусной меры между эмбеддингами

Модель рефератора	Модель векторного представления текста				Среднее значение
	rubert-tiny2	labse-ru-sts	Multilingual_en_ru_uk	rubert-base-cased	
ОТ-АР					
Авторский реферат	0,8001	0,6987	0,7256	0,8213	0,7614
ОТ-СР					
Лингвистический рефератор	0,8697	0,7644	0,8592	0,9052	0,8496
TextRank	0,9362	0,7721	0,7657	0,9379	**0,8530**
LexRank	0,7754	0,6379	0,6390	0,7981	0,7126
mT5_multilingual_XLSum	0,7337	0,6426	0,6709	0,7747	0,7055
mbart_ru_sum_gazeta	0,6905	0,8234	0,7168	0,7245	0,7388
rut5_base_sum_gazeta	0,6966	0,8290	0,7359	0,7505	0,7530
t5_summary_en_ru_zh_base_2048	0,6796	0,8073	0,7178	0,7257	0,7326
АР-СР					
Лингвистический суммаризатор	0,8358	0,7127	0,7505	0,8208	0,7800
TextRank	0,8185	0,7152	0,7259	0,8186	0,7696
LexRank	0,8160	0,6759	0,7155	0,8314	0,7597
mT5_multilingual_XLSum	0,7672	0,6547	0,7075	0,7758	0,7263
mbart_ru_sum_gazeta	0,8337	0,7188	0,7465	0,8339	0,7832
rut5_base_sum_gazeta	0,8455	0,7303	0,7630	0,8398	**0,7947**
t5_summary_en_ru_zh_base_2048	0,8229	0,6833	0,7140	0,8106	0,7577

При сравнении с авторскими рефератами наилучшие результаты показали rut5_base_sum_gazeta (0,7947), mbart_ru_sum_gazeta (0,7832) и лингвистический суммаризатор (0,7800), что свидетельствует о высокой согласованности и стилистическом сходстве с авторскими рефератами.

Косинусная мера эмбеддингов демонстрирует существенно более высокие значения (0,75–0,85), чем статистические метрики, что связано с их способностью улавливать глубинные семантические соответствия. Все значения превышают 0,70, что говорит о высоком уровне семантической адекватности.

Обсуждение результатов

Комплексный анализ результатов выявляет фундаментальное различие между экстрактивными и абстрактивными подходами к автоматическому реферированию.

Экстрактивные методы, в частности TextRank, демонстрируют максимальную лексическую точность при сравнении с оригинальными текстами (ROUGE-1 F1 = 0,38, TF-IDF = 0,87). Это является ожидаемым артефактом их работы, а не показателем высокого

качества реферирования и объясняется полным сохранением формулировок исходного текста. Однако при сравнении с авторскими рефератами экстрактивные методы демонстрируют ограниченность: их значения по нейросемантическим метрикам не превосходят абстрактивные модели, а иногда уступают им. Это указывает на то, что дословное копирование предложений не обеспечивает естественности изложения, характерной для человеческого реферирования. Экстрактивные методы, несмотря на высокое формальное сходство с источником, генерируют результаты, структурно и лексически не похожие на рефераты, созданные человеком. Человек активно перефразирует и обобщает, что доказывается значительно более низким лексическим сходством его реферата с оригиналом.

Абстрактивные нейросетевые модели показали относительно низкие значения по статистическим метрикам (ROUGE-1 F1 = 0,08–0,14), что отражает их склонность к перефразированию и генерации новых формулировок. Однако по нейросемантическим метрикам они приближаются к человеческому уровню или даже превосходят его: BERTScore F1 = 0,69–0,70, косинусное сходство эмбеддингов ≈ 0,78–0,80. Особенно важно подчеркнуть, что ведущие абстрактивные модели (rut5_base_sum_gazeta, mbart_ru_sum_gazeta) дообучены преимущественно на новостных, а не на научных корпусах. Низкие показатели ROUGE у этих моделей могут быть вызваны не их «неумением» реферировать, а именно тем, что они избегают специфической научной терминологии, которой не было в обучающих данных, заменяя её более общими синонимами. Это приводит к семантически верному, но лексически отличному реферату.

Ограничения исследования

Необходимо отметить ряд ограничений проведённого исследования. Во-первых, датасет включает только научные статьи определённых предметных областей, что может ограничивать обобщаемость результатов на другие типы текстов. Во-вторых, оценка проводилась исключительно на автоматических метриках без привлечения экспертной оценки, что не позволяет полностью учесть субъективные аспекты качества рефератов. В-третьих, использованные нейросетевые модели дообучены преимущественно на новостных корпусах [6, 7], что может снижать их эффективность при работе с научно-технической терминологией. Дообучение на специализированных корпусах научных текстов потенциально способно повысить качество абстрактивного реферирования.

Заключение

В рамках настоящего исследования проведена комплексная оценка качества автоматической генерации рефератов научно-технических текстов на русском языке с использованием как традиционных лингвистических, так и современных нейросетевых моделей. Применение широкого спектра метрик – от классических статистических (ROUGE, TF-IDF) до современных нейросемантических (BERTScore, косинусная мера между эмбеддингами) – позволило получить всестороннюю картину производительности различных подходов.

Направления будущих исследований включают: дообучение абстрактивных моделей на специализированных корпусах научных текстов; интеграцию мультимодальных элементов (таблиц, рисунков) в процесс суммаризации; разработку метрик, учитывающих специфические характеристики научно-технических текстов; проведение экспертной оценки качества рефератов с привлечением специалистов предметных областей; исследование адаптивных методов, автоматически выбирающих оптимальную стратегию реферирования в зависимости от типа и содержания текста.

Литература:

1. Hasan K.S., Ng V. Automatic keyphrase extraction: A survey of the state of the art // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. Vol. 1. P. 1262–1273.

2. Fang F., Feng X., Peng X., Liao L., Lam W., Xu B. A survey on automatic text summarization: Progress, challenges and perspectives // Information Processing & Management. 2023. Vol. 60(3). Art. 103328.
3. Lin C.Y. ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004. P. 74–81.
4. Gusev I. Dataset for automatic summarization of Russian news // Proceedings of the International Conference “Dialogue”. 2019. Vol. 18. P. 122–127.
5. Scialom T., Dray P.A., Lamprier S., Piwowarski B., Staiano J. MLSUM: The multilingual summarization corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. P. 8051–8067.
6. 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. Brussels, Belgium: Association for Computational Linguistics, 2018. P. 1797–1807.
7. Mihalcea R., Tarau P. TextRank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004. P. 404–411.
8. Erkan G., Radev D.R. LexRank: Graph-based lexical centrality as salience in text summarization // Journal of Artificial Intelligence Research. 2004. Vol. 22. P. 457–479.
9. Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P.J. Exploring the limits of transfer learning with a unified text-to-text transformer // Journal of Machine Learning Research. 2020. Vol. 21(140). P. 1–67.
10. Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. P. 7871–7880.
11. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language / arXiv preprint. arXiv:1905.07213. arXiv, 2019.
12. Zhang T., Kishore V., Wu F., Weinberger K.Q., Artzi Y. BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR). ICLR, 2020.
13. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019. P. 3982–3992.
14. Smucker M.D., Allan J., Carterette B. A Comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. ACM Press, 2007. P. 623–632.
15. Tsanda A., Bruches E. Russian-Language Multimodal Dataset for Automatic Summarization of Scientific Papers / arXiv preprint.:2405.07886. arXiv, 2024.

Благодарности:
Статья выполнена в рамках государственного задания: FFFU-2025-0010.

Об авторах:
КАН Анна Владимировна, канд. техн. наук, доцент, заместитель директора Всероссийского института научно-технической информации Российской академии наук, kanav@viniti.ru.
ХОРОШИЛОВ Александр Алексеевич, д-р техн. наук, профессор, ведущий научный сотрудник Всероссийского института научно-технической информации Российской академии наук, khoroshilov@viniti.ru.
БЕЗЗУБОВ Александр Федорович, канд. техн. наук, доцент, ведущий научный сотрудник Центрального научно-исследовательского института Минобороны России, b17a52f@yandex.ru.

ГОКАРЕВ Вадим Николаевич, младший научный сотрудник Центрального научно-исследовательского института Минобороны России, vgokarev@mail.ru.
ХОРОШИЛОВ Алексей Алексеевич, кандидат технических наук, заведующий лабораторией Всероссийского института научно-технической информации Российской академии наук, alex_khor@viniti.ru.

Metadata:
Title: A comparative analysis of methods for automatic summarization of scientific texts in Russian.
Author 1: Anna Vladimirovna Kan, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, 20 Usievich st., 125190 Moscow, Russian Federation, kanav@viniti.ru, 0000-0001-9410-406X.
Author 2: Alexander Alexievich Khoroshilov, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, 20 Usievich st., 125190 Moscow, khoroshilov@viniti.ru.
Author 3: Vadim Nikolaevich Gokarev, Central Research Institute of the Ministry of Defense of the Russian Federation, 5 1st Khoroshevsky dr., 123007 Moscow, vgokarev@mail.ru, 0009-0002-4082-2925.
Author 4: Alexander Fedorovich Bezzubov, Central Research Institute of the Ministry of Defense of the Russian Federation, 5 1st Khoroshevsky dr., 123007 Moscow, b17a52f@yandex.ru.
Author 5: Alexey Alexievich Khoroshilov, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, 20 Usievich st., 125190 Moscow, alex_khor@viniti.ru.
Abstract: Automatic summarization of scientific texts remains a pressing problem in natural language processing, particularly for Russian-language corpora, where the application of modern methods is limited by the lack of specialized models and datasets. This paper conducts a systematic comparative analysis of extractive and abstractive summarization models using a Russian-language multimodal dataset of scientific articles. A comprehensive set of metrics was used to assess quality: classical statistical metrics (ROUGE, TF-IDF) and modern neural semantic metrics (BERTScore, cosine measure between embeddings). The results show that extractive methods (TextRank, LexRank) provide high lexical accuracy (ROUGE-1 F1 = 0.38; TF-IDF = 0.87), while neural network abstractive models (rut5_base_sum_gazeta, mbart_ru_sum_gazeta) demonstrate superiority in semantic metrics (BERTScore F1 = 0.70; cosine similarity = 0.79), approaching the level of human summarization. The study suggests that the choice of model should be determined by the specific task: extractive methods are preferable for formalized texts, while abstractive ones provide a more natural and flexible presentation.
Keywords: Automatic abstracting, text summarization, natural language processing, neural network models, quality metrics, Russian language, scientific texts.