оригинальная статья

https://elibrary.ru/hfslfk

Автоматическая генерация новостных заголовков при помощи нейронной сети RuGPT-3 (влияние обучающего датасета на результативность модели)

Шамигов Федор Федорович
Национальный исследовательский Томский государственный университет, Россия, Томск
https://orcid.org/0009-0005-4933-2764
fshamigov@mail.ru
Резанова Зоя Ивановна
Национальный исследовательский Томский государственный университет, Россия, Томск
eLibrary Author SPIN: 4102-0940
https://orcid.org/0000-0002-0550-991X
Scopus Author ID: 56896043500

Аннотация: В статье представлены результаты проверки гипотезы о влиянии типа датасета (отдельные рубрики новостей vs их совокупность) на качество автоматической генерации заголовков новостных статей. Актуальность работы обусловлена высокой конкурентоспособностью СМИ в цифровом пространстве, где успех новостного агентства часто зависит от скорости публикации. На это, в свою очередь, во многом влияет использование технологий автоматической генерации текста новости в целом и его заголовка. Цель – проверить гипотезу: обучение модели RuGPT-3 на статьях отдельных рубрик и на их совокупности даст разные результаты в качестве генерируемого заголовка. Новизна работы: мы сравнили качество генерации на отдельных рубриках и их совокупности, в то время как большинство исследователей в этой сфере, как правило, обучают модели на всей совокупности сразу. Поставлена следующая задача: изучить влияние типов датасетов на качество генерируемых заголовков. Задача решалась с использованием модели RuGPT-3 на материале новостных статей корпуса Lenta.ru. Данные были организованы в три датасета: рубрики *наука* и *спорт* (по 6900 статей каждая), а также совокупность этих рубрик (6900 статей). В результате исследования гипотеза подтвердилась. Модель, обученная на совокупности рубрик, генерирует более качественные с точки зрения формальной метрики ROUGE заголовки, она достигла среднего значения F-мер 0,22 (против 0,174 на *науке* и 0,196 на *спорте*). Более того, в процедурах качественного анализа генерируемых заголовков было выявлено, что они обладают естественностью, соответствуют большинству правил эффективного заголовка (длина до 10 слов, предикативность, прошедшее время предиката, в основном действительный залог, отсутствие предлога или числа в начале, отсутствие относительных указателей времени), однако не всегда соотносятся с правилом о соответствии материалу. Статья имеет поисковый характер, перспективы исследования видятся в проведении дополнительных экспериментов с другими типами датасетов.

Ключевые слова: новость, новостной заголовок, автоматическая генерация, машинное обучение, модель RuGPT-3, нейронные сети, метрика ROUGE

Цитирование: Шамигов Ф. Ф., Резанова З. И. Автоматическая генерация новостных заголовков при помощи нейронной сети RuGPT-3 (влияние обучающего датасета на результативность модели). *Виртуальная коммуникация и социальные сети.* 2025. Т. 4. № 1. С. 62–70. https://doi.org/10.21603/2782-4799-2025-4-1-62-70

Поступила в редакцию 06.11.2024. Принята после рецензирования 20.01.2025. Принята в печать 21.01.2025.

full article

Automatic Generation of News Headlines Using RuGPT-3 Neural Network: Effect of Training Dataset on Model Performance

Fedor F. Shamigov
National Research Tomsk State University, Russia, Tomsk
https://orcid.org/0009-0005-4933-2764
fshamigov@mail.ru
Zoya I. Rezanova
National Research Tomsk State University, Russia, Tomsk
eLibrary Author SPIN: 4102-0940
https://orcid.org/0000-0002-0550-991X
Scopus Author ID: 56896043500

Abstract: News agencies compete in the digital space, where the success often depends on the promptness of publication, which can be provided by automatic headline generation technologies. This study examined the effect of dataset types on the quality of headline generation, i.e., the impact of dataset type (individual news categories vs. their combination) on the quality of automatic news headlines. The initial hypothesis was that training the RuGPT-3 model on thematic sets of articles and on their totality would give different generated headlines. The authors used the RuGPT-3 model and news articles published by Lenta.ru. The research included three datasets: the categories of *science* and *sports* (6,900 articles each) and their combination (6,900 articles). The results confirmed the hypothesis: the model trained on the combined dataset generated higher-quality headlines as measured by the formal ROUGE metric, achieving an average F-score of 0.22 (compared to 0.17 for *science* and 0.2 for *sports*). The generated headlines looked authentic and conformed to the good headline practice, i.e., length (⩽10 words), predicativity, past tense, active voice, no opening prepositions or figures, no relative time indicators, etc. However, the headlines were not always consistent with the content.

Keywords: news, news headline, automatic generation, machine learning, RuGPT-3 model, neural networks, ROUGE metric

Citation: Shamigov F. F., Rezanova Z. I. Automatic Generation of News Headlines Using RuGPT-3 Neural Network: Effect of Training Dataset on Model Performance. *Virtual Communication and Social Networks*, 2025, 4(1): 62–70. (In Russ.) https://doi.org/10.21603/2782-4799-2025-4-1-62-70

Received 6 Nov 2024. Accepted after review 20 Jan 2025. Accepted for publication 21 Jan 2025.

Введение

Мы живем в информационную эпоху – время компьютеров и информационных технологий, и с каждым годом количество людей, использующих Интернет, и объемы информации, которая производится в мире, значительно увеличиваются. По оценкам экспертов, ожидается, что за 2025 г. будет произведено примерно 463 млрд гигабайт информации[1]. Огромные объемы информации ставят задачи по упрощению и ускорению работы с ней, которые решаются в настоящее время с использованием искусственного интеллекта (ИИ), машинного обучения (ML) и обработки естественного языка (NLP).

Искусственный интеллект охватывает широкий спектр технологий, среди которых NLP направлена на выполнение различных задач. Одной из таких задач является автоматическое реферирование (суммаризация) неструктурированных текстовых данных, т.е. извлечение ключевой информации из текста. Суммаризация делится на два основных вида: экстрактивная (квазиреферирование) и абстрактивная (реферирование). Экстрактивная суммаризация заключается в выделении ключевых фраз или предложений из исходного текста при сохранении их в неизменном виде.

[1] How Much Data Is Created Every Day in 2024? *Techjury*. 12 Mar 2024. URL: https://techjury.net/blog/how-much-data-is-created-every-day/ (accessed 19 Oct 2024).

В результате абстрактивной суммаризации новый текст передает основные идеи исходного материала с использованием уникальных формулировок. Оба вида суммаризации широко применяются исследователями в разных сферах: юридической, научной, финансовой и др. [Головизнина 2022; Горбачев, Синицын 2023; Дорош и др. 2022; Жигалов и др. 2023; Коротких, Носенко 2021; Сорокина 2024a; 2024b; Abualigah et al. 2020; Alami et al. 2021; Chen et al. 2019; Jalil et al. 2021; Yadav et al. 2024; Zhou et al. 2021].

Суммаризация востребована и в СМИ. Как было отмечено, с появлением и распространением Интернета традиционный уклад СМИ претерпел значительные изменения. В этой быстро меняющейся среде особенно важным становится эффективное представление информации, где структура и цель новостных материалов играют ключевую роль. Следовательно, понимание того, как новости конструируются, является неотъемлемой частью их успешного распространения.

Новость рассматривается с разных точек зрения, что находит отражение в вариантах определений в работах ряда авторов [Ахмадулин 2020; Иванова 2022; Колесниченко 2020: 15–22; Макушин 2014; Троицкий 2017]. Однако можно проследить общие черты, которые выделяются в большинстве определений: новость – это всегда некоторое событие, которое «выбивается» из обыденности, и информация о котором имеет значение для целевой аудитории. Базовая функция новости – информирующая, выполнение этой функции обеспечивается ее соответствием прежде всего требованиям оперативности и значимости (влиятельности), достоверности (объективности, точности) и лаконичности. Вследствие этого к новостям предъявляется требование наличия специфического стиля и жесткой структуры. Основные элементы структуры новостной статьи: заголовок, лид и текст [Амзин 2011: 13–20; Дьякова 2011; Колесниченко 2020: 6–22].

Заголовок – это анонс события, о котором идет речь в статье. Он содержит в себе основную информацию, максимальную «выжимку» статьи. Заголовок многие ученые считают важнейшей частью новости. Заголовки имеют вероятность быть прочитанными примерно в 90 % случаев, т.к. взгляд человека «скользит» по ним в поисках интересующей информации. Если же заголовок не завлек человека, то новостная статья с большей вероятностью не будет им прочитана. В связи с этим выделяются две главные функции заголовка: информативная и контактная. Заголовок призван сообщить о чем статья и подвигнуть на ее прочтение, чему способствуют качества привлекательности, краткости, удобочитаемости, информативности. Эффективный заголовок, т.е. такой, который выполняет функции, указанные выше, должен обладать следующими основными свойствами [Амзин: 2011; Колесниченко 2020]:

1) иметь длину около 65–75 знаков, максимум – чуть больше 100, при этом слов – не более 10;
2) соответствовать материалу;
3) ясно давать понять суть новости;
4) иметь ссылки на источник при выражении предположений, допущений;
5) предпочтительно иметь действительный залог.

И, напротив, заголовок не должен [Амзин: 2011; Колесниченко 2020]:

1) иметь малоизвестные, непонятные аудитории слова и сокращения;
2) иметь крылатые выражения, пословицы, поговорки в неизменном виде;
3) иметь относительные указатели времени (типа *сегодня*, *завтра*, *вчера* и т.д.);
4) объединять несколько тем в заголовках;
5) содержать излишнюю детализацию;
6) иметь в начале предлог или число, а в конце точку.

Перед учеными, работающими в сфере применения технологий ИИ с целью повышения скорости и качества распространения новостного потока, в качестве одной из важных задач стоит повышение качества автоматического генерирования заголовков. Зарубежные и российские исследователи проводили ряд экспериментов с автоматической суммаризацией частей новостных статей с помощью нейронных сетей, используя как квазиреферирование, так и абстрактное реферирование.

В целом алгоритмы квазиреферирования или сочетания квазиреферирования с абстрактным реферированием показывают значительно более высокий результат, чем алгоритмы абстрактного реферирования. Например, P. Muniraj и соавторы использовали для суммаризации новостных статей гибридную нейросеть архитектуры Seq2Seq, обученную на 50 тыс. новостных статей с применением как квазиреферирования, так и абстрактного реферирования, и получили среднее по F-мерам на валидационной выборке примерно 0,39 (чем ближе это значение к 1, тем более похож сгенерированный заголовок на оригинальный), т. е. совпадение примерно на 39 % [Muniraj et al. 2023]. Такой же гибридный подход использовали в работе [Kumari et al. 2023].

 https://doi.org/10.21603/2782-4799-2025-4-1-62-70

N. Kumari и коллеги получили максимальный результат по качеству заголовка, около 70 %.

При применении абстрактного реферирования результаты были несколько ниже. В работе [Gupta et al. 2022] с помощью модели T5, обученной на 2225 новостях, представлен максимальный результат при генерировании заголовка по среднему значению F-мер метрики ROUGE, примерно 0,4. В другом исследовании с нейронной сетью на той же архитектуре, обученной на 2755 новостных статьях, учеными был достигнут результат по метрике ROUGE около 0,5 [Hayatin et al. 2021]. Эксперименты с рекуррентной нейронной сетью показали значение метрики примерно 0,41 [Yao et al. 2020], с моделью LSTM – 0,34 [Jiang et al. 2021], с моделью BERT – 0,44 [Ma et al. 2022] и 0,45 [Bao, Zhang 2023].

В работах российских исследователей при использовании метода LexRank достигнуты следующие значения: ROUGE-1 = 0,25, ROUGE-2 = 0,09, и ROUGE-L = 0,18 [Белякова, Беляков 2020]. На основе использования нейронной сети с архитектурой encoder-decoder, обученной на корпусе из примерно 660 тыс. статей, А. А. Шевчук получил значение метрики BLEU около 0,11 [Шевчук 2020]. Д. А. Аишева обучила модель с архитектурой *Transformer* тремя корпусами новостных статей (в сумме более 2 млн) и получила среднее по F-мерам ROUGE около 0,16[2]. И. Гусевым на основе обучения трех моделей для генерации заголовков MBART, ruT5 и RuGPT-3 на корпусе из 61 тыс. новостных статей получено среднее по F-мерам ROUGE 0,25, 0,25, и 0,185[3] соответственно.

Особенность абсолютного большинства работ по генерации частей новостных статей заключается в том, что исследователи используют для обучения нейронных сетей корпусы, состоящие из новостей различных рубрик (политика, наука, спорт, экономика и т. д.).

Актуальность работы обусловлена высокой конкурентоспособностью СМИ в цифровом пространстве, где успех новостного агентства часто зависит от скорости публикации. Новизна работы: мы сравнили качество генерации на отдельных рубриках и их совокупности, в то время как большинство исследователей в этой сфере, как правило, обучают модели на всей совокупности сразу. Цель – проверить гипотезу: обучение модели RuGPT-3 на статьях отдельных рубрик и на их совокупности даст разные результаты в качестве генерируемого заголовка.

Методы и материалы

Поставленная в работе цель решалась на основе применения русскоязычной версии модели GPT-3 – RuGPT-3 Medium с 350 млн параметров, относящейся к семейству нейронных сетей с архитектурой Transformer. Нейронные сети с архитектурой Transformer выгодно отличаются от других, т. к. обладают позиционным кодированием, механизмом внимания и механизмом самовнимания [Vaswani et al. 2017].

В качестве материала для обучения модели мы использовали корпус новостных статей Lenta.ru (корпус Lenta.ru v1.0) библиотеки Corus. Обработка корпуса производилась на языке программирования Python в сервисе Google Colab. После загрузки корпуса в Google Colab из него было сделано три отдельных датасета: в первый извлечены 6900 новостей из рубрики *наука*, во второй – 6900 новостей из рубрики *спорт*, в третий – смесь из 6900 статей обеих рубрик. После создания датасетов они были «очищены» от лишних символов, которые могли возникнуть при кодировке (например, *\xa0*, *<...&gt*, *&amp* и др.).

После обработки датасеты были разделены на обучающую (80 %) и валидационную (20 %) выборки: 5520 статей для обучения и 1380 для валидации. Из всей информации мы оставили только заголовки и тексты статей. Датасеты переведены в формат JSON для обучения. Обучение длилось 6 эпох и заняло около 2,5 часов для каждой модели (примерно 7,5 часов в сумме) в сервисе Kaggle с использованием процессора Intel Xeon 2.30GHz и графического ускорителя Tesla P100.

Результаты

Для оценки качества генерации заголовков использовалась метрика ROUGE. При оценке модели, обученной на рубрике *наука*, были получены значения, представленные в таблице 1. Среднее F-мер для модели, обученной на рубрике *наука*, составило 0,174, т. е. заголовок, сгенерированный этой моделью, идентичен оригинальному в среднем примерно на 17,4 %.

[2] Результаты представлены в магистерской работе Д. А. Аишевой «Модификация нейронной сети Transformer для генерации новостных заголовков на русском языке». Томск, 2021.

[3] RuGPT3MediumSumGazeta. *HuggingFace*. URL: https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta (accessed 19 Oct 2024).

В таблице 2 приведены примеры оригинальных и сгенерированных заголовков для идентичных новостей рубрики *наука*.

При оценке по метрике ROUGE модели, обученной на рубрике *спорт*, получены следующие значения (табл. 3). Среднее F-мер для модели, обученной на рубрике *спорт*, составило 0,196.

Заголовки, сгенерированные моделью, обученной на новостях рубрики *спорт*, обозначены в таблице 4.

В таблице 5 отражены значения, полученные при оценке по метрике ROUGE модели, которая обучена на совокупности новостей рубрик *наука* и *спорт*. Среднее F-мер для модели, обученной на совокупности новостей рубрик *наука* и *спорт*, составило 0,22.

Заголовки, сгенерированные моделью, обученной на совокупности новостей рубрик *наука* и *спорт*, представлены в таблице 6.

Обученные модели, согласно метрике ROUGE, показывают качество заголовков: 17,4 % на рубрике *наука*; 19,6 % – *спорт*; 22 % – *наука* и *спорт*.

Табл. 1. Значения метрики ROUGE модели, обученной на рубрике *наука*
Tab. 1. ROUGE metric values for the model trained on the science category

Метрика	F-мера	P (точность)	R (полнота)
ROUGE-1	0,218	0,234	0,212
ROUGE-2	0,091	0,097	0,089
ROUGE-L	0,212	0,228	0,206

Табл. 2. Оригинальные и сгенерированные заголовки (модель, обученная на рубрике *наука*)
ab. 2. Original and generated headlines (model trained on the science category)

Оригинальный заголовок	Сгенерированный заголовок
На Земле нашли следы смертоносной космической катастрофы	Раскрыта причина вымирания динозавров
Предсказаны масштабные разрушения в северных странах	Раскрыта главная опасность глобального потепления
Созданы вечно текущие капли первородной материи	Физики впервые получили капли кварк-глюонной плазмы
Объяснена смертельная опасность соли	Ученые раскрыли механизм гипертонии
Найдена неизвестная причина гибели клеток у людей	Названа главная причина смерти клеток

Гипотеза подтвердилась: в нашем эксперименте совокупность рубрик дает более качественный результат (на 4,6 % качественнее в сравнении с моделью *наука* и на 2,4 % качественнее в сравнении с моделью *спорт*).

При количественной оценке заголовков было выявлено, что среднее количество слов в оригинальных заголовках в рубрике *наука* составило около 7 (6,56), в сгенерированных – около 6 (5,86); в рубрике *спорт* – около 8 (7,52), в сгенерированных – около 7 (7,095); в совокупности рубрик *наука* и *спорт* – около 7 (6,59), в сгенерированных – около 6 (5,68). Таким образом, первое правило эффективного заголовка соблюдается в текстах, сгенерированных всеми тремя моделями. Помимо этого, они имеют в своем составе глагол; обладают предикативностью; глагол представлен в прошедшем времени (*раскрыли*..., *опровергли*..., *победил*... и др.),

Табл. 3. Значения метрики ROUGE модели, обученной на рубрике *спорт*
Tab. 3. ROUGE metric values for the model trained on the sports category

Метрика	F-мера	P (точность)	R (полнота)
ROUGE-1	0,244	0,256	0,248
ROUGE-2	0,107	0,113	0,11
ROUGE-L	0,237	0,249	0,241

Табл. 4. Оригинальные и сгенерированные заголовки (модель, обученная на рубрике *спорт*)
Tab. 4. Original and generated headlines (model trained on the sports category)

Оригинальный заголовок	Сгенерированный заголовок
МОК потребовал вернуть медали трех олимпийских чемпионов из России	Российские бобслеисты потребовали вернуть медали Олимпиады в Сочи
Моуринью сравнил футболистов с мебелью	Моуринью раскритиковал «Манчестер Юнайтед»
Россиянин не глядя отбил летевшую в пустые ворота шайбу пяткой	Вратарь «Тампа-Бэй Лайтнинг» совершил сэйв пяткой в матче регулярного чемпионата НХЛ
В деле российских биатлонистов нашли след Родченкова	Глава Союза биатлонистов России прокомментировал расследование австрийской полиции
Российский биатлон отреагировал на обвинения австрийской полиции	Сборную России обвинили в нарушении антидопинговых правил

 https://doi.org/10.21603/2782-4799-2025-4-1-62-70

Табл. 5. Значения метрики ROUGE модели, обученной на рубриках *наука* и *спорт*
Tab. 5. ROUGE metric values for the model trained on the science and sports categories

Метрика	F-мера	P (точность)	R (полнота)
ROUGE-1	0,267	0,293	0,254
ROUGE-2	0,13	0,144	0,124
ROUGE-L	0,262	0,288	0,250

Табл. 6. Оригинальные и сгенерированные заголовки (модель, обученная на рубриках *наука* и *спорт*)
Tab. 6. Original and generated headlines (model trained on the science and sports category)

Оригинальный заголовок	Сгенерированный заголовок
Роналду предложили вечно считать лучшим футболистом	Тренер «Наполи» высказался о Роналду
Рогозин нашел мистический след в отмене пуска с Восточного	Российского вице-премьера обвинили в мистификации
Российский биатлон обрел нового президента	Драчев победил на выборах президента СБР
Названы гонорары победителей турнира UFC в России	Названы гонорары бойцов на UFC 136
Опровергнута гипотеза об эволюции человека	Ученые опровергли популярную теорию эволюции

а в страдательной конструкции – в совершенном виде (*названа*…, *раскрыта*… и др.); отсутствуют предлоги и числа в начальной позиции и относительные указатели времени.

При анализе соответствия сгенерированных заголовков материалу было замечено, что во всех моделях во многих случаях сгенерированные заголовки хоть и не повторяют оригинальные, но очень точно передают суть новости. Например, (1) оригинальный и (2) сгенерированный заголовки (табл. 2): (1) *Созданы вечно текущие капли первородной материи* и (2) *Физики впервые получили капли кварк-глюонной плазмы*. В новости сообщается, что физики впервые смогли получить капли кварк-глюонной плазмы в коллайдере. Сгенерированный заголовок отражает информацию более конкретно и даже, возможно, более точно передает суть события. Другой пример: (1) *Моуринью сравнил футболистов с мебелью* и (2) *Моуринью раскритиковал «Манчестер Юнайтед»* (табл. 4). Объединяет заголовки лишь фамилия. В самой новости Моуринью, тренер футбольной команды «Манчестер Юнайтед», высказывает мнение, что *футбольная команда – это больше, чем бездумная покупка игроков*, и, приводя сравнение, отмечает: *Вы должны много трудиться, тратить силы на поиск лучшей мебели, чтобы жить в обустроенном доме*[4]. Обозначим, что в целом сгенерированный заголовок довольно точно передает суть.

Но некоторые заголовки соответствуют материалу не в полной мере. Например, статья, посвященная влиянию излучения от вспышки сверхновой на Землю и на морскую мегафауну (табл. 2), была озаглавлена: (1) *На Земле нашли следы смертоносной космической катастрофы*. Динозавры вообще не упоминаются, однако нейросеть сгенерировала заголовок: (2) *Раскрыта причина вымирания динозавров*.

Относительно седьмого правила эффективного заголовка из нашего перечня: сгенерированные заголовки соответствуют материалу примерно в половине случаев, а в другой половине либо передают суть не совсем точно, либо совсем неточно, и эта характеристика проявляется при обучении и на совокупности, и на одиночных рубриках.

Также при наблюдении сгенерированных заголовков было установлено, что, безотносительно к их соответствию или несоответствию материалу статьи, в большинстве случаев они выглядят настолько естественно, что их невозможно отличить от заголовков, придуманных человеком.

Заключение

При обучении модели на новостных статьях совокупности рубрик (*наука* и *спорт*) качество генерации заголовка, согласно метрике ROUGE, получается не только выше в сравнении с моделями, обученными на отдельных рубриках, но и в большинстве случаев превосходит результаты, полученные ранее на материале русского языка: А. Ю. Беляковой и Ю. Д. Беляковым – на 17,3 % [Белякова, Беляков 2020]; А. А. Шевчуком – на 0,11 %; [Шевчук 2020], Д. А. Аишевой – на 0,16 %[5]; превосходит один из результатов в исследовании И. Гусева на 0,185, но уступают двум другим – 0,25 и 0,25[6].

[4] Моуринью сравнил футболистов с мебелью. *Lenta.ru*. 15.12.2018. URL: https://lenta.ru/news/2018/12/15/mebel/ (дата обращения: 19.10.2024).
[5] Результаты представлены в магистерской работе Д. А. Аишевой…
[6] RuGPT3MediumSumGazeta. HuggingFace…

Результаты можно учитывать для проведения дальнейших экспериментов при разработке систем автоматической суммаризации новостных статей. Отметим, что публикуемые данные имеют поисковый характер и необходим дальнейший тщательный и объемный анализ генерируемых заголовков с привлечением большего числа рубрик, новостных статей, других моделей и мощного оборудования.

Конфликт интересов: Авторы заявили об отсутствии потенциальных конфликтов интересов в отношении исследования, авторства и / или публикации данной статьи.

Conflict of interests: The authors declared no potential conflict of interests regarding the research, authorship, and / or publication of this article.

Критерии авторства: Авторы в равной степени участвовали в подготовке и написании статьи.

Contribution: All the authors contributed equally to the study and bear equal responsibility for information published in this article.

Литература / References

Амзин А. А. Новостная интернет-журналистика. М.: Аспект Пресс, 2011. 142 с. [Amzin A. A. *Journalism of online news*. Moscow: Aspekt Press, 2011, 142. (In Russ.)] https://elibrary.ru/sbfkaz

Ахмадулин Е. В. «Новость» как основа журнализма. *Гуманитарный вектор*. 2020. Т. 15. № 5. С. 149–154. [Akhmadulin E. V. The "News" as the basis of journalism. *Humanitarian Vector*, 2020, 15(5): 149–154. (In Russ.)] https://doi.org/10.21209/1996-7853-2020-15-5-149-154

Белякова А. Ю., Беляков Ю. Д. Обзор задачи автоматической суммаризации текста. *Инженерный вестник Дона*. 2020. № 10. С. 142–159. [Belyakova A. Yu., Belyakov Yu. D. Overview of text summarization methods. *Inzhenernyi vestnik Dona*, 2020, (10): 142–159. (In Russ.)] https://elibrary.ru/ayyyfq

Головизнина В. С. Автоматическое реферирование текстов. *Информационные технологии и нанотехнологии (ИТНТ-2022)*: VIII Междунар. конф. (Самара, 23–27 мая 2022 г.) Самара: Самарский ун-т, 2022. [Goloviznina V. S. Automatic abstracting of texts. *Information technologies and nanotechnology (ITNT-2022)*: Proc. VIII Intern. Conf., Samara, 23–27 May 2022. Samara: Samara University, 2022. (In Russ.)] https://elibrary.ru/evsbxc

Горбачев А. Д., Синицын А. В. Сравнительный анализ алгоритмов суммаризации текста для проектирования и разработки программного комплекса. *Развитие современной науки и технологий в условиях трансформационных процессов*: XI Междунар. науч.-практ. конф. (Москва, 12 мая 2023 г.) СПб.: Печатный цех, 2023. С. 43–52. [Gorbachev A. D., Sinitsyn A. V. Comparative analysis of text summarization algorithms for the design and development of a software package. *The development of modern science and technology in the context of transformational processes*: Proc. XI Intern. Sci.-Prac. Conf., Moscow, 12 May 2023. St. Petersburg: Pechatnyj ceh, 2023, 43–52. (In Russ.)] https://elibrary.ru/nonvjs

Дорош М., Райковский Д. И., Пугин К. В. Задача суммаризации текста. *Инновации. Наука. Образование*. 2022. № 49. С. 2036–2044. [Dorosh M., Rajkovsky D. I., Pugin K. V. Text summary problem. *Innovatsii. Nauka. Obrazovanie*, 2022, (49): 2036–2044. (In Russ.)] https://elibrary.ru/znzfhc

Дьякова Т. В. Основные принципы и структура новостных сообщений. *Lingua mobilis*. 2011. № 2. С. 102–105. [Dyakova T. V. Basic principles and structure of news mes-sages. *Lingua mobilis*, 2011, (2): 102–105. (In Russ.)] https://elibrary.ru/rodaws

Жигалов А. Ю., Гришина Л. С., Болодурина И. П. Исследование моделей искусственного интеллекта для автоматического аннотирования и реферирования текстов. *Цифровые технологии в образовании, науке, обществе*: XVII Всерос. науч.-практ. конф. (Петрозаводск, 22–24 ноября 2023 г.) Петрозаводск: ПетрГУ, 2023. С. 36–38. [Zhigalov A. Yu., Grishina L. S., Bolodurina I. P. Research of artificial intelligence models for automatic and abstracting of texts. *Digital technologies in education, science, and society*: Proc. XVII All-Russian Sci.-Prac. Conf., Petrozavodsk, 22–24 Nov 2023. Petrozavodsk: PetrSU, 2023, 36–38. (In Russ.)] https://elibrary.ru/tugzpu

Иванова С. В. Новость как дискурсивный жанр: не отсутствующая структура. *Terra Linguistica*. 2022. Т. 13. № 3. С. 7–14. [Ivanova S. V. News as a genre of discourse: A non-missing structure. *Terra Linguistica*, 2022, 13(3): 7–14. (In Russ.)] https://doi.org/10.18721/JHSS.13301

Колесниченко А. В. Практическая журналистика. 3-е изд. М.: Московский ун-т, 2020. 191 с. [Kolesnichenko A. V. *Practical journalism.* 3rd ed. Moscow: Moscow University, 2020, 191. (In Russ.)]

Коротких Е. Г., Носенко Н. В. Семантико-прагматическая компрессия текста в обучении английскому языку для специальных целей. *Современные проблемы науки и образования*. 2021. № 2. [Korotkikh E. G., Nosenko N. V. Semantic and pragmatic text compression in teaching English for special purposes. *Modern problems of science and education*, 2021, (2). (In Russ.)] https://doi.org/10.17513/spno.30665

Макушин А. Б. Современная трактовка понятия новости в условиях медиаконвергенции. *Вестник Кемеровского государственного университета*. 2014. № 2-2. С. 187–189. [Makushin A. B. Modern treatment of the concept of news in the conditions of media convergence. *Vestnik Kemerovskogo gosudarstvennogo universitet*a, 2014, (2-2): 187–189. (In Russ.)] https://elibrary.ru/smmxjz

Сорокина С. Г. Интеллектуальная обработка текстовой информации: обзор автоматизированных методов суммаризации. *Виртуальная коммуникация и социальные сети*. 2024a. Т. 3. № 3. С. 203–222. [Sorokina S. G. Intelligent text processing: A review of automated summarization methods. *Virtual Communication and Social Networks*, 2024a, 3(3): 203–222. (In Russ.)] https://doi.org/10.21603/2782-4799-2024-3-3-203-222

Сорокина С. Г. Особенности применения технологии автоматической суммаризации к научным публикациям. *Три «Л» в парадигме современного гуманитарного знания: лингвистика, литературоведение, лингводидактика*: Всерос. науч.-практ. конф. (Москва, 23 ноября 2023 г.) М.: Языки Народов Мира, 2024b. С. 132–138. [Sorokina S. G. Applying automatic summarization technology to academic publications. *Three L's of modern humanities: Linguistics, literary studies, and linguadidactics*: Proc. All-Russian Sci.-Prac. Conf., Moscow, 23 Nov 2023. Moscow: Yazyki Narodov Mira, 2024b, 132–138. (In Russ.)] https://elibrary.ru/duydpi

Троицкий Ю. Л. Новости как литература: об одной экспериментальной практике. *Новый филологический вестник*. 2017. № 3. С. 52–59. [Troitskiy Yu. L. News as literature: Experimental practice. *The New Philological Bulletin*, 2017, (3): 52–59. (In Russ.)] https://elibrary.ru/yllsqd

Шевчук А. А. Автоматическая генерация новостных заголовков с применением нейронной сети encoder-decoder. *Актуальные проблемы лингвистики и литературоведения*: VI (XX) Междунар. конф. (Томск, 18–19 апреля 2019 г.) Томск: ООО СТТ, 2020. С. 100–101. [Shevchuk A. A. Encoder-decoder neural network for automatic news headline generation. *Relevant issues of linguistics and literary studies*: Proc. VI (XX) Intern. Conf., Tomsk, 18–19 Apr 2019. Tomsk: STT LLC, 2020, 100–101. (In Russ.)] https://elibrary.ru/oqgvly

Abualigah L., Bashabsheh M. Q., Alabool H., Shehab M. Text summarization: A brief review. *Recent advances in NLP: The case of arabic language*, eds. Abd Elaziz M., Al-qaness M. A. A., Ewees A. A., Dahou A. Cham: Springer, 2020, 1–15. https://doi.org/10.1007/978-3-030-34614-0_1

Alami N., Mallahi M. E., Amakdouf H., Qjidaa H. Hybrid method for text summarization based on statistical and semantic treatment. *Multimedia Tools and Applications*, 2021, 80(13): 19567–19600. https://doi.org/10.1007/s11042-021-10613-9

Bao G., Zhang Y. A general contextualized rewriting framework for text summarization. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2023, 31: 1624–1635. https://doi.org/10.1109/TASLP.2023.3268569

Chen D., Ma S., Harimoto K., Bao R., Su Q., Sun X. Group, extract and aggregate: Summarizing a large amount of finance news for forexmovement prediction. *Proceedings of the Second Workshop on Economics and Natural Language Processing*, eds. Hahn U., Hoste V., Zhang Z. Hong Kong: Association for Computational Linguistics, 2019, 41–50. https://doi.org/10.18653/v1/D19-5106

Gupta A., Chugh D., Anjum, Katarya R. Automated news summarization using transformers. *Sustainable advanced computing*, eds. Aurelia S., Hiremath S. S., Subramanian K., Biswas S. Kr. Springer, 2022, 249–259. https://doi.org/10.1007/978-981-16-9012-9_21

Hayatin N., Ghufron K. M., Wicaksono G. W. Summarization of COVID-19 news documents deep learning-based using transformer architecture. *TELKOMNIKA. Telecommunication Computing Electronics and Control*, 2021, 19(3): 754–761. https://doi.org/10.12928/TELKOMNIKA.v19i3.18356

Jalil Z., Nasir J. A., Nasir M. Extractive multi-document summarization: A review of progress in the last decade. *IEEE Access*, 2021, 9: 130928–130946. https://doi.org/10.1109/ACCESS.2021.3112496

Jiang J., Zhang H., Dai C., Zhao Q., Feng H., Ji Z., Ganchev I. Enhancements of attention-based bidirectional LSTM for hybrid automatic text summarization. *IEEE Access*, 2021, 9: 123660–123671. https://doi.org/10.1109/ACCESS.2021.3110143

Kumari N., Sharma N., Singh P. Performance of optimizers in text summarization for news articles. *Procedia Computer Science*, 2023, 218: 2430–2437. https://doi.org/10.1016/j.procs.2023.01.218

Ma T., Pan Q., Rong H., Qian Y., Tian Y., Al-Nabhan N. T-BERTSum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, 2022, 9(3): 879–890. https://doi.org/10.1109/TCSS.2021.3088506

Muniraj P., Sabarmathi K. R., Leelavathi R., Balaji S. HNTSumm: Hybrid text summarization of transliterated news articles. *International Journal of Intelligent Networks*, 2023, 4: 53–61. https://doi.org/10.1016/j.ijin.2023.03.001

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need. *31st International Conference on Neural Information Processing Systems (NIPS'17)*: Proc. Conf., Long Beach, 4–9 Dec 2017. NY: Curran Associates, 2017, 6000–6010. https://doi.org/10.48550/arXiv.1706.03762

Yadav A. K., Ranvijay, Yadav R. S., Maurya A. K. Graph-based extractive text summarization based on single document. *Multimedia Tools and Applications*, 2024, 83(7): 18987–19013. https://doi.org/10.1007/s11042-023-16199-8

Yao K., Zhang L., Du D., Luo T., Tao L., Wu Y. Dual encoding for abstractive text summarization. *IEEE Transactions on Cybernetics*, 2020, 50(3): 985–996. https://doi.org/10.1109/TCYB.2018.2876317

Zhou H., Ren W., Liu G., Su B., Lu W. Entity-aware abstractive multi-document summarization. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, eds. Zong Ch., Xia F., Li W., Navigli R. Stroudsburg: Association for Computational Linguistics, 2021, 351–362. https://doi.org/10.18653/v1/2021.findings-acl.30