Инженерный вестник Дона, №9 (2025)
ivdon.ru/ru/magazine/archive/n9y2025/10342

Комбинированный метод реферирования русскоязычных текстов

В.И. Шиян[1], *В.Н. Марков*[2]

[1]*Кубанский государственный университет, Краснодар*
[2]*Кубанский государственный технологический университет, Краснодар*

Аннотация: Статья посвящена разработке комбинированного метода реферирования русскоязычных текстов, объединяющего экстрактивные и абстрактивные подходы для преодоления ограничений существующих методов. Предлагаемому методу предшествуют этапы: предобработка текста, комплексный лингвистический анализ с использованием RuBERT, кластеризация на основе семантической близости. Метод включает экстрактивное реферирование через алгоритм TextRank и абстрактивную доработку с помощью нейросетевой модели RuT5. Эксперименты на новостном корпусе Газета.Ру подтвердили преимущество метода по точности, полноте, F-мере и метрикам ROUGE. Результаты показали превосходство комбинированного подхода над чисто экстрактивными методами, такими как TF-IDF и статистический, и абстрактивными методами, такими как RuT5 и mBART.
Ключевые слова: комбинированный метод, реферирование, русскоязычные тексты, TextRank, RuT5.

Введение в задачу реферирования

Реферирование – это процесс создания краткого изложения исходного текста, существенно сокращённого по объёму [1, 2]. Основная цель реферата – кратко передать ключевые идеи и главную информацию оригинального текста [3]. Такой подход особенно важен при работе с научными, юридическими и новостными текстами, где требуется точность и ясность изложения [4]. Русский язык с его богатой морфологией, сложной синтаксической структурой и семантическими нюансами, представляет дополнительные вызовы для реферирования, требуя учёта контекста и языковых особенностей [5, 6]. Современные методы реферирования делятся на три основные категории: экстрактивные, абстрактивные и комбинированные [7]. Экстрактивные методы выбирают важные предложения из исходного текста [8, 9], абстрактивные генерируют новые формулировки [10], а комбинированные сочетают оба подхода для достижения оптимального результата [11]. В данной статье представлен авторский метод комбинированного реферирования, демонстрирующий

новый подход к повышению качества рефератов посредством интеграции экстрактивных и абстрактивных технологий.

Предложенный комбинированный метод

Стоит отметить, что хотя данная статья посвящена комбинированному методу реферирования русскоязычных текстов, представляется важным кратко описать этапы, которые привели к необходимости его разработки и применения, чтобы пояснить место предложенного подхода.

На рис. 1 приведены этапы обработки текстов, показана последовательность шагов и потоки данных между ними.

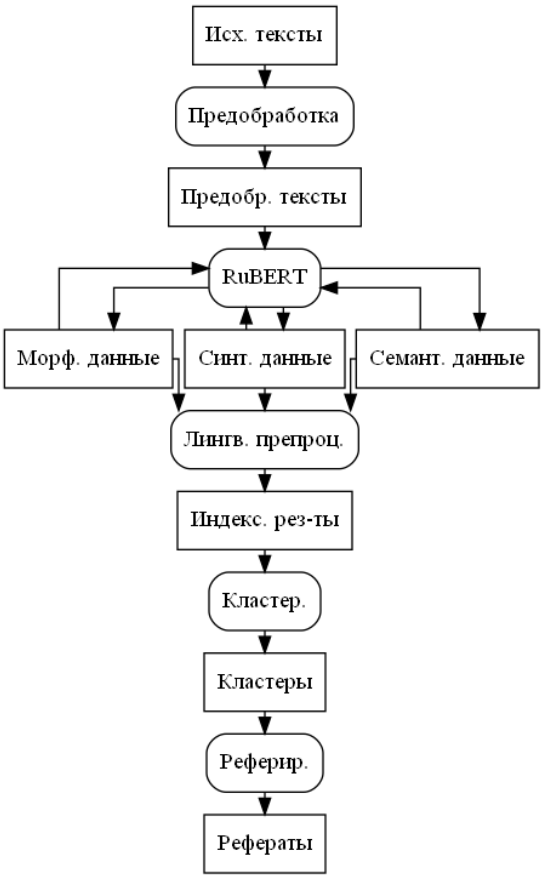

Рис. 1. – Этапы обработки текстов

Предобработка текстов. Исходные тексты разбиваются на отдельные предложения и токены, что облегчает их дальнейший анализ, при этом учитываются морфологические и другие особенности русского языка [12, 13].

Комплексный анализ с использованием BERT. Проводится комплексный анализ текстов с использованием модели RuBERT, адаптированной для русскоязычных текстов [14]. RuBERT – это предтренированная нейросетевая модель на основе архитектуры двунаправленных кодировщиков текстовых представлений на основе трансформеров (Bidirectional Encoder Representations from Transformers – BERT), обученная на больших корпусах русскоязычных текстов, таких как Википедия и новостные статьи. Модель RuBERT реализует 3 основных вида лингвистического анализа:

– морфологический анализ, позволяющий определить морфологические характеристики слов: части речи, падежи, числа, времена глаголов и другие признаки, существенные для понимания внутренней структуры текста;

– синтаксический анализ, выявляющий взаимосвязи между словами в предложении, идентифицирующий подлежащие, сказуемые, дополнения и прочие синтаксические элементы [15];

– семантический анализ, раскрывающий смысловые связи между словами с учётом языковых явлений: синонимии, антонимии, омонимии, многозначности и контекстуальных значений [16].

Индексирование результатов. Индексирование результатов анализа осуществляется для оптимизации последующей обработки текстовых данных [17]. Такой подход обеспечивает возможность эффективного сопоставления и группировки отдельных текстовых фрагментов. На основе полученных индексов производится вычисление сходства между текстами с использованием меры семантической близости для всех пар предложений. Данная мера позволяет количественно оценить, насколько тексты близки по смыслу, и основана на сравнении структур, полученных с использованием RuBERT.

Кластеризация текстов. Тексты объединяются в кластеры с использованием той же меры семантической близости, что позволяет сгруппировать семантически схожие фрагменты для дальнейшего анализа.

Комбинированный метод реферирования. Существующие подходы к автоматическому реферированию имеют ряд ограничений. Например, экстрактивные методы, основанные на отборе наиболее значимых предложений исходного текста, часто приводят к рефератам с избыточностью и нарушенной логической связностью. В свою очередь, абстрактивные модели могут допускать фактологические ошибки или упускать важные детали исходного текста. Для устранения перечисленных недостатков был предложен комбинированный метод реферирования, объединяющий преимущества обоих подходов.

На рис. 2 приведён комбинированный метод реферирования, основанный на алгоритме TextRank и нейросетевой модели RuT5.

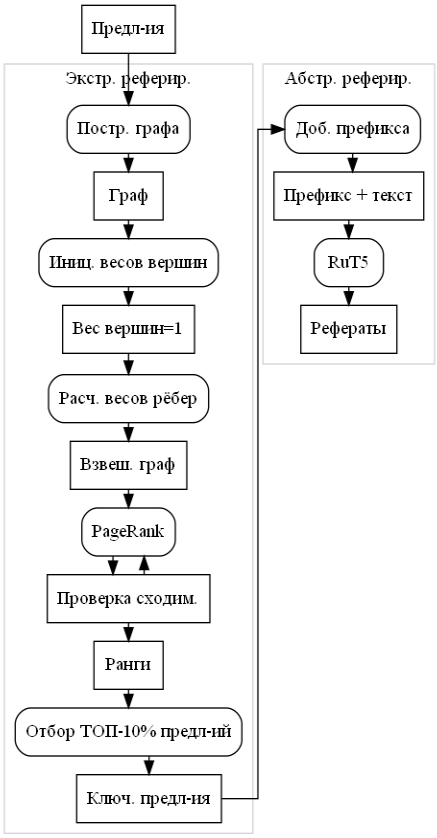

Рис. 2. – Комбинированный метод реферирования

Предложенный комбинированный метод состоит из двух последовательных этапов: экстрактивного реферирования на основе алгоритма TextRank [18] и последующей абстрактивной доработки полученных результатов с использованием нейросетевой модели RuT5 [19].

Экстрактивное реферирование. На первом этапе для каждого кластера формируется краткая выжимка путём отбора наиболее значимых предложений посредством метода TextRank.

Алгоритм TextRank – это графовый метод ранжирования, адаптированный для обработки текстов, и позволяющий выделить наиболее значимые предложения на основе их взаимосвязей. Алгоритм включает следующие шаги:

Построение графов. Для каждого кластера текстов формируется отдельный граф. Все предложения текстов биективно отображаются в вершины графа.

Инициализация весов вершин. Всем вершинам задаётся одинаковый начальный вес, равный 1.

Вычисление весов рёбер. В отличие от классического TextRank, который традиционно использует меру близости на основе TF-IDF весов слов, в предлагаемом методе используется мера семантической близости.

Для каждого кластера отдельно рассчитывается мера семантической близости между всеми парами предложений из всех текстов данного кластера. Каждая пара предложений (s_1, s_2) формируется таким образом, что предложение s_1 берётся из одного текста кластера, а предложение s_2 – из другого текста того же кластера. Мера семантической близости рассчитывается по формуле:

$$Prox(s_1, s_2) = \sum_{i=1}^{5} p_i \cdot \tilde{w}_i(s_1, s_2), \quad (1)$$

$$\sum_{i=1}^{5} p_i = 1, \tag{2}$$

$$0 \le p_i \le 1, \tag{3}$$

где $\tilde{w}_1(s_1, s_2)$ – мера близости IDF-весов слов; $\tilde{w}_2(s_1, s_2)$ – мера близости TF-IDF весов слов; $\tilde{w}_3(s_1, s_2)$ – мера близости синтаксических отношений слов; $\tilde{w}_4(s_1, s_2)$ – мера близости семантических связей слов; $\tilde{w}_5(s_1, s_2)$ – мера близости семантических ролей слов; p_i, $i = \overline{1, 5}$ – весовые коэффициенты, определяющие относительный вклад каждой из мер близости. Экспериментально установленные оптимальные значения коэффициентов: $p_1 = 0,294$, $p_2 = 0,06$, $p_3 = 0,198$, $p_4 = 0,15$, $p_5 = 0,298$.

Если полученное значение меры близости превышает заранее заданный порог, равный 0,07, то между соответствующими вершинами графа создаётся ребро, вес которого равен вычисленному значению меры семантической близости.

Итеративное вычисление весов вершин по формуле PageRank. Веса предложений рассчитываются итеративно по формуле PageRank:

$$PR(s_i) = (1-d) + d \cdot \sum_{s_j \in In(s_i)} \frac{PR(s_j) \cdot w_{ji}}{\sum\limits_{s_k \in Out(s_j)} w_{jk}}, \tag{4}$$

где d – коэффициент затухания, отражающий вероятность перехода по ребру, $0 \le d \le 1$, оптимальное значение которого, как показали эксперименты, составляет 0,85; $In(s_i)$ – множество вершин, из которых исходят рёбра в вершину s_i; $Out(s_j)$ – множество вершин, в которые направлены рёбра из вершины s_j; w_{ji} – вес ребра от s_j к s_i, $0 \le w_{ji} \le 1$.

Итерации алгоритма выполняются до тех пор, пока не будет достигнуто априори заданное количество шагов. В ходе проведённых

экспериментов использовалось 20 итераций. Однако дополнительно можно использовать условие ранней остановки для проверки сходимости. Например, алгоритм можно остановить раньше, если изменения весов вершин между двумя соседними итерациями становятся меньше заранее заданного порога ε, $\varepsilon > 0$, например, $\varepsilon = 10^{-3}$.

Отбор предложений. После вычисления итоговых весов предложений из графа кластера выбираются предложения с наибольшими значениями PageRank-весов. Эти предложения составляют примерно 10% от общего числа предложений всех текстов данного кластера и формируют экстрактивный реферат кластера.

Абстрактивная доработка. На втором этапе полученные экстрактивные рефераты сглаживаются и дорабатываются при помощи нейросетевой модели RuT5. Эта модель основана на архитектуре трансформера преобразования текста в текст (Text-to-Text Transfer Transformer – T5), которая рассматривает любые задачи обработки естественного языка как задачи преобразования одного текста в другой. Для реферирования входной текст дополняется специальным префиксом «Сгенерировать реферат:», после чего модель формирует реферат.

Модель обучена на корпусе Газета.Ру, состоящем из русскоязычных текстов, что обеспечивает её адаптацию к особенностям данного типа текстов. Эта модель улучшает связность рефератов и устраняет типичные недостатки экстрактивного реферирования, такие как избыточность или отсутствие логической последовательности. Она преобразует экстрактивные выжимки в более последовательные и легко читаемые тексты, максимально приближенные к рефератам, написанным человеком.

Показатели качества предложенного комбинированного метода

Для оценки качества предложенного комбинированного подхода реферирования использовались такие показатели, как точность, полнота и F-мера, а также метрики ROUGE, *chrF* , *METEOR* и *BLEU* . Перечисленные показатели позволяют комплексно оценить качество полученных рефератов по различным критериям: содержания, структуры и лаконичности текста.

Результаты эксперимента

В рамках настоящей работы был проведен эксперимент по оценке эффективности различных методов автоматического реферирования текстов. В качестве корпуса текстов для проведения эксперимента использовались статьи новостного ресурса Газета.Ру. Полученные результаты представлены далее.

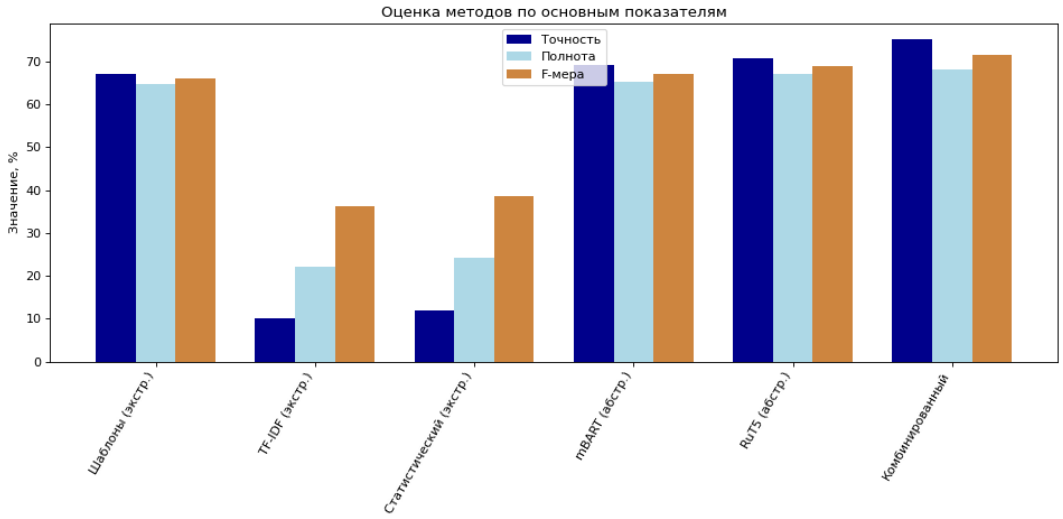

Рис. 3. – Оценка методов по основным показателям

Из рис. 3 видно, что наилучшие результаты по точности, полноте и F-мере продемонстрировал комбинированный метод, который сочетает в себе преимущества экстрактивного и абстрактивного подходов. Среди чисто абстрактивных методов наибольшую эффективность показала модель RuT5, немного опередив модель mBART по всем трём показателям. Экстрактивные

методы значительно уступают абстрактивным и комбинированному подходу по качеству реферирования. Из них наиболее успешным оказался шаблонный метод, в то время как методы, основанные на TF-IDF и статистическом подходе, показали низкие результаты по всем трём показателям.

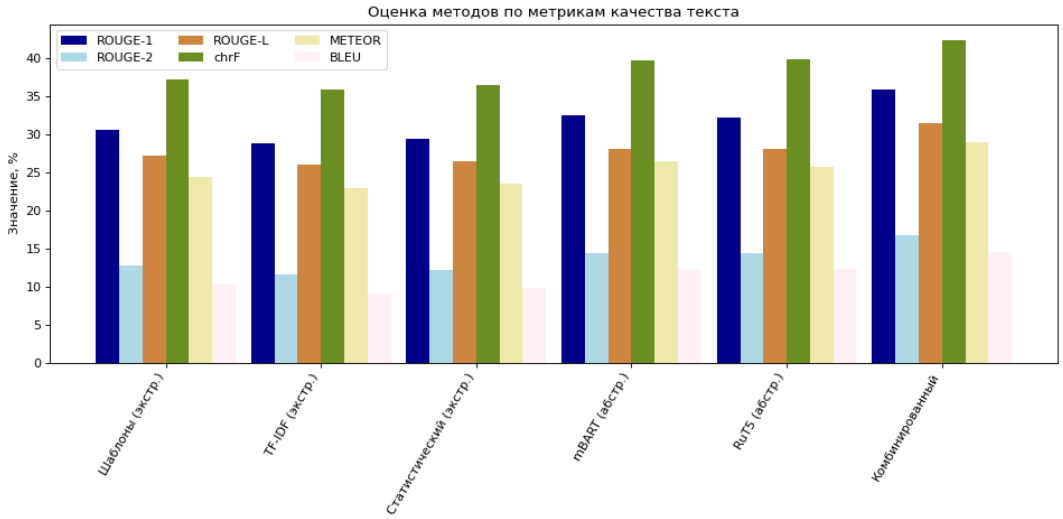

Рис. 4. – Оценка методов по метрикам качества текста

Из рис. 4 видно, что по всем рассмотренным метрикам качества текста лидирует предложенный комбинированный метод. Абстрактивные модели mBART и RuT5 демонстрируют близкие друг к другу результаты, которые существенно превосходят результаты экстрактивных методов. Среди экстрактивных подходов наиболее эффективным вновь оказывается шаблонный метод, однако разница между ним и другими экстрактивными методами в данном случае менее выражена. Методы TF-IDF и статистический показывают самые низкие значения по всем метрикам.

Комбинированный метод обладает явным преимуществом по сравнению с другими подходами. Абстрактивные модели mBART и RuT5 также показываются высокие результаты, значительно опережая экстрактивные методы. Шаблонный подход, хотя и уступает абстрактивным моделям, всё же превосходит остальные экстрактивные методы. Наименее

эффективными оказываются методы TF-IDF и статистический, показывающие низкие результаты по всем рассмотренным показателям.

Заключение

Предложенный комбинированный метод реферирования демонстрирует значительное улучшение качества рефератов за счёт синергии экстрактивного и абстрактивного этапов. Анализ на корпусе Газета.Ру подтвердил его превосходство над альтернативными подходами: по показателям точности, полноты, F-меры и метрик ROUGE он существенно опережает экстрактивные методы и абстрактивные модели. Ключевые преимущества – устранение избыточности, обеспечение связности текста и высокая семантическая адекватность. Этот метод эффективно справляется со специфическими сложностями русского языка, открывая тем самым возможности для его применения в обработке научных, юридических и новостных текстов.

Литература

1. Осминин П.Г. Современные подходы к автоматическому реферированию и аннотированию // Вестник Южно-Уральского государственного университета. 2012. Вып. 25. С. 134–135.

2. Луканин А.В. Автоматическая обработка естественного языка. Челябинск: Изд. центр ЮУрГУ, 2011. 70 с.

3. Ступин В.С. Система автоматического реферирования методом симметричного реферирования // Труды Междунар. конф. «Диалог-2004». Компьютерная лингвистика и интеллектуальные технологии. М.: Наука, 2004. С. 579–591.

4. Михаилян А. Некоторые методы автоматического анализа естественного языка, используемые в промышленных продуктах. 2000. URL: citforum.ru/programming/digest/avtestlang.shtml

5. Харламов А.А. Автоматический структурный анализ текстов // Открытые системы. 2002. №10. С. 16–22.

6. Кутукова Е.С. Технология Text mining // Перспективные инновации в науке, образовании, производстве и транспорте: сб. науч. тр. SWorld. Одесса, 2013. С. 136–138.

7. Бурмистров А.С., Свиридова О.В. Экспертная оценка программных продуктов для аннотирования документов // Постулат. 2017. №5. URL: e-postulat.ru/index.php/Postulat/article/viewFile/567/588

8. Яцко В.А. Симметричное реферирование: теоретические основы и методика // НТИ. Сер. 2. Информационные процессы и системы. 2002. №5. С. 18–28.

9. Вичева О.Н. Подходы к автоматическому обзорному реферированию группы текстов одной тематики // Проблемы современной прикладной лингвистики: сб. науч. ст. Минск: МГЛУ, 2014. С. 246–252.

10. Осминин П.Г. Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод: дис. … канд. филол. наук (10.02.21). Челябинск: ЮУрГУ, 2016. 239 с.

11. Челышев Э.А., Раскатова М.В., Мишин А.А., Щёголев П. Автоматическое реферирование текстов: обзор алгоритмов и подходов к оценке качества // Инженерный вестник Дона. 2023. №12. URL: ivdon.ru/ru/magazine/archive/n12y2023/8871

12. Шиян В.И., Марков В.Н. Комплексный анализ русскоязычных текстов на основе нейросетевых моделей трансформерного типа // Инженерный вестник Дона. 2025. №5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10038

13. Bharti S.K., Babu K.S., Jena S.K. Automatic Keyword Extraction for Text Summarization: A Survey. 2017. URL: arxiv.org/ftp/arxiv/papers/1704/1704.03242.pdf

14. Kupiec J., Pederson J., Chen F. A trainable document summarizer // Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval. Seattle, 1995. P. 68–73.

15. Luhn H.P. The Automatic Creation of Literature Abstracts // IBM Journal of Research and Development. 1958. Vol. 2. №2. pp. 159–165. doi: 10.1147/RD.22.0159

16. Salton G. Dynamic Information and Library Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975. 523 p.

17. Edmundson H.P. New methods in automatic extracting // Journal of the ACM (JACM). 1969. Vol. 16. №2. pp. 264–285.

18. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2004. pp. 404–411. URL: aclanthology.org/W04-3252

19. Zmitrovich D. et al. A Family of Pretrained Transformer Language Models for Russian // arXiv preprint arXiv:2309.10931. 2023. URL: arxiv.org/abs/2309.10931

References

1. Osminin P.G. Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. 2012. Vyp. 25. pp. 134–135.

2. Lukanin A.V. Avtomaticheskaya obrabotka estestvennogo yazyka [Automatic natural language processing]. Chelyabinsk: Izd. tsentr YuUrGU, 2011. 70 p.

3. Stupin V.S. Trudy Mezhdunar. konf. «Dialog-2004». Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Moscow, 2004. pp. 579–591.

4. Mikhailyan A. Nekotorye metody avtomaticheskogo analiza estestvennogo yazyka, ispol'zuemye v promyshlennykh produktakh [Some methods of automatic natural language analysis used in industrial products]. 2000. URL: citforum.ru/programming/digest/avtestlang.shtml

5. Kharlamov A.A. Otkrytye sistemy. 2002. №10. pp. 16–22.

6. Kutukova E.S. Perspektivnye innovatsii v nauke, obrazovanii, proizvodstve i transporte: sb. nauch. tr. SWorld. Odessa, 2013. pp. 136–138.

7. Burmistrov A.S., Sviridova O.V. Postulat. 2017. №5. URL: http://e-postulat.ru/index.php/Postulat/article/viewFile/567/588

8. Yatsko V.A. NTI. Ser. 2. Informatsionnye protsessy i sistemy. 2002. №5. pp. 18–28.

9. Vicheva O.N. Problemy sovremennoy prikladnoy lingvistiki: sb. nauch. st. Minsk, 2014. pp. 246–252.

10. Osminin P.G. Postroenie modeli referirovaniya i annotirovaniya nauchno-tekhnicheskikh tekstov, orientirovannoy na avtomaticheskiy perevod: dis. … kand. filol. nauk (10.02.21) [Building a model for summarization and annotation of scientific and technical texts aimed at machine translation: Cand. philol. sci. diss.]. Chelyabinsk: YuUrGU, 2016. 239 p.

11. Chelyshev E.A., Raskatova M.V., Mishin A.A., Shchegolev P. Inzhenernyj vestnik Dona. 2023. №12. URL: ivdon.ru/ru/magazine/archive/n12y2023/8871

12. Shiyan V.I., Markov V.N. Inzhenernyj vestnik Dona. 2025. №5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10038

13. Bharti S.K., Babu K.S., Jena S.K. Automatic Keyword Extraction for Text Summarization: A Survey. 2017. URL: arxiv.org/ftp/arxiv/papers/1704/1704.03242.pdf

14. Kupiec J., Pederson J., Chen F. Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval. Seattle, 1995. pp. 68–73.

15. Luhn H.P. IBM Journal of Research and Development. 1958. Vol. 2. №2. pp. 159–165. doi: 10.1147/RD.22.0159

16. Salton G. Dynamic Information and Library Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975. 523 p.

17. Edmundson H.P. Journal of the ACM (JACM). 1969. Vol. 16. №2. P. 264–285.

18. Mihalcea R., Tarau P. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2004. P. 404–411. URL: aclanthology.org/W04-3252

19. Zmitrovich D. et al. A Family of Pretrained Transformer Language Models for Russian. preprint arXiv: 2309.10931. 2023. URL: arxiv.org/abs/2309.10931.

Дата поступления: 30.06.2025

Дата публикации: 25.08.2025