обзорная статья
https://elibrary.ru/zdvtuu

Интеллектуальная обработка текстовой информации: обзор автоматизированных методов суммаризации

Сорокина Светлана Геннадьевна
Первый Московский государственный медицинский университет им. И. М. Сеченова Минздрава России (Сеченовский университет), Россия, Москва
eLibrary Author SPIN: 3957-8400
https://orcid.org/0000-0002-8667-6743
lana40ina@mail.ru

Аннотация: Интерес к инновационным технологическим стратегиям и современным цифровым инструментам обработки информации значительно возрос в связи с необходимостью управления большими массивами неструктурированных данных. Автоматизированная суммаризация – важный инструмент в различных областях, требующих эффективного анализа и обработки больших объемов текстовой информации. В статье представлен обзор актуальных парадигм и сервисов автоматизированной суммаризации на основе междисциплинарных исследований в области лингвистики, компьютерных технологий и искусственного интеллекта. Особое внимание уделено синтаксическим и лексическим приемам, используемым нейросетевыми моделями для сжатия текста. В качестве примера рассмотрены сервисы *QuillBot, Summate.it, WordTune, SciSummary, Scholarcy* и *OpenAI ChatGPT*. Выявлено, что современные модели автоматизированной суммаризации успешно применяют экстрактивные и абстрактивные методы для создания резюме разного качества и объема. Экстрактивный подход основан на выделении наиболее значимых предложений в исходном тексте. Абстрактивные алгоритмы создают новые формулировки, сохраняя основную мысль оригинального текста. Автоматизированные суммаризаторы эффективно используют приемы сжатия текста (устранение избыточной информации, упрощение сложных конструкций и обобщение данных), присущие человеку в процессе обработки текстовой информации. Эти технологии обеспечивают высокую точность и связность генерируемых резюме, хотя каждая модель имеет свои ограничения. Для достижения оптимальных результатов важно учитывать специфику задачи и выбирать подходящую модель суммаризации: экстрактивную – для краткости и точности; абстрактивную – для более глубокой смысловой обработки текстовых данных.

Ключевые слова: автоматизированная суммаризация, авторезюмирование, экстрактивная суммаризация, абстрактивная суммаризация, нейронные сети, искусственный интеллект, междисциплинарные исследования

Цитирование: Сорокина С. Г. Интеллектуальная обработка текстовой информации: обзор автоматизированных методов суммаризации. *Виртуальная коммуникация и социальные сети.* 2024. Т. 3. № 3. С. 203–222. https://doi.org/10.21603/2782-4799-2024-3-3-203-222

Поступила в редакцию 03.06.2024. Принята после рецензирования 04.09.2024. Принята в печать 09.09.2024.

VIRTUAL COMMUNICATION
& SOCIAL NETWORKS 2024, 3(3)

review article

Intelligent Text Processing: A Review of Automated Summarization Methods

Svetlana G. Sorokina
Sechenov First Moscow State Medical University, Russia, Moscow
eLibrary Author SPIN: 3957-8400
https://orcid.org/0000-0002-8667-6743
lana40ina@mail.ru

Abstract: Interest in innovative technological strategies and modern digital tools has increased significantly due to the need to manage large amounts of unstructured data. This paper reviews current paradigms and services for automated summarization, developed based on interdisciplinary research in linguistics, computer technologies, and artificial intelligence. It focuses on syntactic and lexical techniques employed by neural network models for text compression. The paper presents performance examples of such AI-powered services as *QuillBot, Summate.it, WordTune, SciSummary, Scholarcy*, and *OpenAI ChatGPT*. The contemporary automated models proved effective in using extractive and abstractive methods to generate summaries of varying quality and length. The extractive approach relies on identifying the most significant sentences from the original text, while abstractive algorithms create new sentence structures that preserve the main idea of the original content. Automated summarizers effectively utilize text compression techniques that are inherent to human approach to text processing, e.g., they exclude redundant information, simplify complex structures, and generalize data. These technologies provide high accuracy and coherence in the generated summaries, though each summarization model has its limitations. Optimal results depend on the specifics of the task at hand: extractive models provide brevity and precision while abstractive ones allow for deeper semantic processing. Automated summarization is becoming an important tool in various fields that require effective analysis and processing of large text data.

Keywords: automated summarization, auto summary, extractive summarization, abstractive summarization, neural networks, artificial intelligence, interdisciplinary research

Citation: Sorokina S. G. Intelligent Text Processing: A Review of Automated Summarization Methods. *Virtual Communication and Social Networks*, 2024, 3(3): 203–222. (In Russ.) https://doi.org/10.21603/2782-4799-2024-3-3-203-222

Received 2 Jun 2024. Accepted after review 4 Sep 2024. Accepted for publication 9 Sep 2024.

Введение

С ростом доступности интернет-ресурсов и цифровых технологий многократно увеличился и объем доступной информации, что вызывает некоторые сложности у исследователей в обновлении своих знаний. Информационный шум затрудняет обработку получаемых данных и выявление значимых исследований и создают проблему информационной перегрузки и принятия решения [Вертинова и др. 2022]. При этом важная и качественная информация часто теряется среди менее релевантных данных, что, как бы это ни было парадоксально, может приводить к изоляции от ключевых научных тенденций. Эпоха экспонентного развития технологий требует нового технологического подхода и новых цифровых инструментов при обработке огромного количества неструктурированной информации, которую представляют собой тексты и документы [Иванюкович и др. 2023; Малышева, Лычагина 2022].

Одним из инструментов обработки естественного языка (*Natural Language Processing* – NLP) является автоматизированное резюмирование текста, или автоматизированная суммаризация [Белов и др. 2020; Мусаев, Григорьев 2021; Перелетов 2021; Yadav et al. 2022]. Она базируется на лингвистических знаниях, достижениях в области компьютерных технологий и искусственного интеллекта (ИИ) и, следовательно, является объектом междисциплинарных исследований [Сорокина 2023]:

- специалисты в области машинного обучения и ИИ, во-первых, занимаются разработкой и оптимизацией алгоритмов, способных

анализировать, понимать и генерировать текст [Белякова, Беляков 2020], во-вторых, стремятся создать коды для оптимизации обзора больших данных и технологий, позволяющие автоматически сжимать тексты без потери смысла;

- специалисты в области компьютерных наук разрабатывают программное обеспечение и архитектуру систем, которые поддерживают алгоритмы машинного обучения и NLP [Горбачев, Синицын 2023; Abualigah et al. 2020];
- когнитивная наука и психолингвистика исследуют, как человеческий мозг обрабатывает язык, что лежит в основе создания алгоритмов, имитирующих человеческое понимание текста [Безлепкин, Зайкова 2021];
- лингвистика же, владея знаниями о структуре языка, помогает в создании эффективных алгоритмов для обработки и понимания естественного языка в процессе суммаризации с целью адекватной интерпретации семантики и синтаксиса текста.

Резюме, создаваемые моделями автоматизированной суммаризации, нацелены на уменьшение времени, затрачиваемого на просмотр литературных источников, и предоставление ученым возможности быстро оценить соответствие определенной публикации исследовательским задачам. Этот метод облегчает процесс самостоятельного обучения студентов и начинающих ученых [Соколова, Чалова 2020] и вносит вклад в оптимизацию образовательного процесса в целом [Толстых 2017]. Кроме того, автоматизированная суммаризация способствует улучшению доступности научных данных и их пониманию широкой аудиторией, в том числе лицами без специализированных технических знаний.

Текстовая суммаризация – процесс создания краткого и содержательного изложения ключевых идей исходного текста [Головизнина 2022], т. е. выделение существенных элементов текста и отсеивание второстепенных деталей, повторений и элементов, не несущих значимой информационной нагрузки, но создающих информационный шум. Другими словами, кроме способности технического сжатия исходного текста, главным становится и сохранение ключевого контента, смысла [Arana-Catania et al. 2021], что крайне важно для научных статей, в которых каждая деталь может иметь особую значимость [Арефьева 2018]. Реализация технологии суммаризации обеспечивает быстрый доступ к сути текста и устраняет необходимость в его детальном чтении. Это способствует эффективному усвоению обширного объема знаний, содержащихся в научных публикациях, и играет важную роль в облегчении работы исследователей в контексте управления информацией в научной среде.

Цель – обзор различных видов автоматизированной суммаризации, включая существующие интеллектуальные суммаризаторы, приемы сжатия текста, анализ синтаксических и лексических особенностей суммарных текстов, представляющих новый контент и являющихся ценным объектом для лингвистического исследования.

Результаты

Основные парадигмы автоматизированной суммаризации

Обзор современных исследований интеллектуальных моделей и технологий суммаризации [Dehru et al. 2021; Huang et al. 2020; Khurana et al. 2023; Pramita Widyassari et al. 2022] позволяет выделить определенные подходы к процессу автоматизированного резюмирования текста.

1. На основе количества исходных текстов, предназначенных для резюмирования, различают одно-текстовую (монотекстовую) суммаризацию (*single-document summarization*) [Lamsiyah et al. 2020], которая применяется к одному документу или тексту, и многотекстовую (мультитекстовую) суммаризацию *(multi-document summarizatio*n) [Khan et al. 2015; Thaiprayoon et al. 2021]. Мультитекстовая суммаризация нацелена на создание сжатого и содержательного изложения основной информации, представленной в группе документов или текстов, объединенных общей темой. Эта задача имеет более сложный характер реализации по сравнению с монотекстовой суммаризацией из-за наличия текстовой гетерогенности, избыточности информации и необходимости учета различных точек зрения [Wolhandler et al. 2022].

2. В зависимости от языка исходного материала инструменты автоматизированной суммаризации делятся на три типа:

- моноязычные (*monolingual*) суммаризаторы работают с исходными и итоговыми текстами на одном языке [Kutlu et al. 2010];
- многоязычные (multilingual) суммаризаторы обрабатывают входные данные на нескольких языках и создают резюме на них же [Hovy, Lin 1998];
- межъязыковые *(cross-lingual*) суммаризаторы принимают текст на одном языке (например,

английском), а резюме генерируется на другом (например, русском) [Linhares Pontes et al. 2020].

3. По характеру содержания исследователи выделяют три основные категории автоматизированных резюме [Чернышкова и др. 2023]:

3.1. Индикативные резюме предоставляют пользователю общую идею о содержании исходного документа, позволяя оценить, стоит ли изучать оригинальный текст подробнее. Как правило, такие резюме составляют около 5 % от объема исходного текста и служат ориентиром для понимания тематики документа [Bhat et al. 2018].

3.2. Информативные резюме, напротив, стремятся охватить все основные темы оригинального документа в краткой форме. Данные резюме занимают около 20 % от общего объема текста и предоставляют более детальный обзор материала [Ghodratnama et al. 2021; Saggion, Lapalme 2002].

3.3. Оценочные резюме включают критическую оценку или анализ автора по заданной теме и являются, скорее, критическим обзором. Однако создание таких резюме – сложная задача для современных автоматизированных суммаризаторов из-за необходимости учитывать субъективные мнения и оценки [Ježek, Steinberger 2008].

Процесс автоматизированной суммаризации предполагает два различных подхода: извлекающий, или экстрактивный (*extractive summarization*) [Bhargava, Sharma 2020; Sharma, Sharma 2022], который заключается в извлечении ключевых предложений или фраз непосредственно из исходного текста без его изменения; обобщающий, или абстрактивный (*abstractive summarization*) [Gupta, Gupta 2019], основанный на перефразировании исходного текста, создании новых предложений, которые сохраняют основные идеи и содержание исходного текста, но выражаются другими словами. То есть в результате суммаризации могут быть получены два вида резюме: экстрактивное и абстрактивное.

Технологическая основа автоматизированной суммаризации

Основу составляют различные алгоритмы, суть работы которых описывается в данной статье нематематическим языком с целью получения общего представления о принципах отбора элементов исходного текста для суммарного текста. Алгоритмы, используемые в процессе автоматизированной суммаризации, функционируют на:

1. Основе анализа центроидов (*centroid-based methods*) [Thaiprayoon et al. 2021]. Такие методы определяют значимость элементов текста, соотнося их с центроидом, представляющим собой некий ключевой элемент в наборе данных [Puduppully et al. 2023]. Исходя из сходства предложений с центроидом, выбираются наиболее релевантные или представительные предложения для включения в суммарный текст.

2. Построении графовых структур (*graph-based methods*) [Полякова, Зайцев 2022; Mihalcea 2004; Yadav et al. 2024]. Методы на данной основе применяются для визуализации и анализа текстовых элементов, а также для изучения их взаимосвязей [Belwal et al. 2022]. Текст преобразуется в граф, узлы которого представляют элементы текста (например предложения), а ребра отражают взаимосвязи между этими элементами. Взаимосвязи могут основываться на различных критериях (семантическое сходство, последовательность текста или лексическая связь). Используя алгоритмы или другие методы ранжирования, система оценивает важность каждого узла в графе. Узлы, имеющие много связей или связанные с другими важными узлами, получают более высокий рейтинг, на основе которого самые важные из них включаются в суммарный текст.

3. Терминологическом анализе (*term-based methods*) [Orasan et al. 2004]. Методы, фокусирующиеся на анализе терминов [Гринев-Гриневич и др. 2022], отдельных слов или фраз в исходном тексте, определяют их контекстуальную важность и необходимость выбора предложений, содержащих данные термины, для включения в суммарный текст. В этих методах часто используются статистические меры, такие как частота слов (*Term Frequency* – TF) и обратная частота документа (*Inverse Document Frequency* – IDF), для вычисления веса каждого термина в тексте [Mishra et al. 2023]. TF измеряет, как часто термин появляется в документе, а IDF оценивает уникальность термина по всей коллекции документов, снижая вес слов, которые встречаются повсеместно и не несут значительной информационной ценности.

Формула TF-IDF объединяет концепции TF и IDF для определения важности каждого слова в каждом документе [Jalilifard et al. 2021; Lubis et al. 2021]. В результате предложения, содержащие термины с высоким весом, считаются более значимыми и, следовательно, чаще выбираются для включения в суммарный текст.

4. Принципе восходящего внимания (*bottom-up attention*) [Gehrmann et al. 2018]. Этот алгоритм анализирует конкретные элементы текста, детерминанты,

 https://doi.org/10.21603/2782-4799-2024-3-3-203-222

т. е. ключевые слова, фразы, имена собственные, которые затем используются алгоритмом для выявления общей сути исходного документа.

5. Онтологиях (*ontology-based methods*) [Mohan et al. 2016], представляющих собой иерархические описания концептов и их взаимосвязей в определенной области и обеспечивающих более глубокое понимание таких семантических аспектов текста, как синонимия и полисемия. На основе более глубокого анализа сути исходного текста создаются более точные и содержательные суммарные тексты[1].

Методы на основе центроидов, графов и терминов направлены на извлечение и использование существующих предложений из исходного текста без создания новых формулировок или предложений, что является характерной чертой экстрактивной суммаризации. С задачами абстрактивной суммаризации (понимания и интерпретации текста) справляются алгоритмы, использующие онтологии, или структурированное представление знаний.

Задачи разработки и оптимизации алгоритмов суммаризации решаются специалистами в области информационных технологий. Лингвистический же интерес направлен на анализ результатов функционирования этих алгоритмов, т. е. суммарного текста и конечного резюме экстрактивного или абстрактивного вида.

Подготовка текста к автоматизированной суммаризации

Эффективность автоматизированной суммаризации обеспечивается тщательной подготовкой исходного текста [Полонский, Федосова 2021]:

1) удаление несущественных элементов текста (избыточные пробелы, заголовки, нумерация строк, нерелевантные изображения и графики).
2) применение NLP-техник, таких как токенизация / разбиение текста на отдельные слова или предложения, лемматизация, стемминг / приведение слов к их базовой форме, удаление стоп-слов, не несущих значимой информативной нагрузки (предлоги, союзы), а также устранение орфографических и грамматических ошибок, которые могут повлиять на анализ текста.
3) деление текста в неструктурированном виде на абзацы с заголовками и подзаголовками, а в некоторых случаях добавление метаданных или аннотации, которые помогают алгоритму лучше понять структуру текста, его контекст, специфику.

Все эти операции выполняются разработчиками и инженерами данных на этапе создания и обучения моделей ИИ, за счет чего конечный пользователь может использовать инструменты автоматизированной суммаризации без необходимости заботиться о технических деталях предварительной обработки текста. В зависимости от цели суммаризации (создание обзора, извлечение ключевых фактов, сокращение текста для обзора) исследователь может выбрать соответствующую модель ИИ, способную создавать разные виды суммарных текстов, т. к. во многие модели интегрированы необходимые алгоритмы [Жигалов и др. 2023].

Подходы к суммаризации текстов

Виды автоматизированной суммаризации:

1. **Экстрактивная суммаризация** – процесс создания сокращенной версии текста, при котором ключевые предложения или фразы извлекаются из исходного текста без существенных изменений, обеспечивая сохранение первоначального стиля и смысла [Chen et al. 2019; Jalil et al. 2021; Mutlu et al. 2019; Yadav et al. 2021]. Данный процесс осуществляется на основе алгоритмической оценки важности и длины предложений, частоты встречаемости слов и других статистических параметров.

В процессе суммаризации особое внимание уделяется предложениям, содержащим ключевые слова, встречающиеся в заголовке или главной теме исходного документа. Это обусловлено тем, что они (ключевые слова) несут важную информацию об основной идее текста. Позиция предложения также имеет значение: предложения, находящиеся в начале / конце абзацев или документа, часто рассматриваются как более важные, т. к. устанавливают тему всего абзаца / обобщают его главную мысль и подчеркивают ее значимость.

Значимыми (важными) детерминантами являются также имена собственные, именованные объекты, т. е. имена людей, названия организаций, географические названия, отражающие основные понятия действительности [Пенцова 2015] и указывающие на уникальные концепции или сущности, необходимые для понимания исходного контента. В качестве маркеров важности выступают слова-связки

[1] Корешкова Т. Семантический анализ для автоматической обработки естественного языка. *Научно-технический центр ФГУП «ГРЧЦ»*. 08.09.2021. URL: https://rdc.grfc.ru/2021/09/semantic_analysis/#post-1707-_Toc69397630 (дата обращения 03.05.2024).

consequently (следовательно), *in conclusion* (в заключении), *as a result* (в результате), *thus* (таким образом), *briefly* (вкратце), которые детерминируют предложения, содержащие ключевые выводы или резюмирующие части текста.

Формальные признаки текста, такие как выделение слов курсивом, жирным шрифтом или подчеркивание, тоже могут быть использованы для оценки важности [Сорокина 2024]. Эти признаки часто указывают на то, что автор текста считает выделенные элементы особенно значимыми. Кроме того, в процессе оценки важности учитывается и длина предложений: более длинные предложения могут содержать больше информации и, следовательно, считаться более важными для понимания общего контента. Однако это не всегда справедливо, поскольку короткие предложения, например парцеллированные, иногда эмоционально окрашены, концентрированно информативны и ясно выражают ключевые идеи [Гурьева 2020].

Кроме длины предложений существуют и другие элементы текста, которые алгоритмы суммаризации могут относить к детерминантам, хотя они не добавляют значимости к общему содержанию или смыслу исходного документа, а лишь создают информационный шум. К ним относятся:

- неинформативные повторяющиеся фразы, которые не вносят в текст новую информацию и повторяют уже сказанное;
- нерелевантные детали, не имеющие прямого отношения к основной теме или цели текста;
- техническая терминология, чаще всего несущественная для общего понимания текста, но сложная или вовсе незнакомая суммаризатору;
- отступления от основной темы, обычно вводящие в заблуждение или отвлекающие от основного содержания;
- конкретные примеры, которые, хотя и могут быть иллюстративными, не всегда необходимы для понимания основных идей текста;
- стилистические приемы, используемые для украшения или эмфазы, но не несущие существенной информативной нагрузки в научном тексте.

Однако современные суммаризаторы обучены распознавать эти отвлекающие элементы текста.

2. Особое место в обработке естественного языка занимает **абстрактивная суммаризация**, в основе которой лежит NLP. В отличие от экстрактивной суммаризации, которая выбирает и комбинирует предложения непосредственно из оригинального текста, абстрактивная создает новые предложения, передающие суть содержания [Zhou et al. 2021]. Результат такой суммаризации – не просто набор отдельных экстрагированных из текста фраз, а совершенно новый контент, семантически последовательный и синтаксически связный [Полякова, Зайцев 2022]. При этом нейросеть способна сохранять регистр, ключевые термины, эмоциональную окраску исходного текста, а также генерировать новые синтаксические и лексические конструкции, не используемые в оригинальном тексте. Таким образом демонстрируется способность ИИ эффективно использовать разнообразные лексические средства для выражения мыслей.

3. **Гибридная суммаризация** объединяет экстрактивные и абстрактивные подходы, сочетая их сильные стороны для создания более точных и осмысленных резюме. Так, экстрактивные методы выделяют ключевые части исходного текста, сохраняя оригинальные формулировки и данные, что обеспечивает высокую точность. Однако эти методы часто сталкиваются с проблемой отсутствия связности и плавности, т.к. просто копируют фрагменты текста. Абстрактивные же методы позволяют перефразировать и преобразовывать исходную информацию, создавая связные и более естественные резюме, но могут вводить неточные или избыточные данные, не присутствующие в оригинале. Гибридные модели стремятся устранить эти недостатки, интегрируя точность экстрактивных методов с креативностью абстрактивных, формируя резюме, которые не только сохраняют важную информацию, но и имеют логичную и удобочитаемую форму [Дорош и др. 2022; Alami et al. 2021; Ghadimi, Beigy 2022; Xiao et al. 2020].

Инструменты автоматизированной суммаризации

В рамках текущего обзора были протестированы современные инструменты автоматизированной суммаризации текста, использующие ИИ и технологии NLP. Выбор таких моделей, как *WordTune, QuillBot, SciSummary, Scholarcy* и *Summate.it*, обусловлен тем, что они являются предобученными моделями и не требуют от пользователя наличия специальных навыков программирования, что делает их доступными и простыми в применении.

Характеризуя NLP-сервисы *WordTune* и *QuillBot*, надо отметить, что это многофункциональные модели, включающие в себя целые наборы инструментов для обработки больших объемов данных. Они функционируют как ИИ-компаньоны для работы

 https://doi.org/10.21603/2782-4799-2024-3-3-203-222

с текстом. Кроме осуществления функции суммаризации, эти инструменты помогают пользователям улучшать, перефразировать и сокращать текст.

Специально для научного сообщества были разработаны ИИ-сервисы *SciSummary* и *Scholarcy*. Их разработчики стремились создать инструмент, который бы помог исследователям и студентам быстро обрабатывать научные статьи и извлекать из них ключевую информацию. В основе работы данных сервисов лежит модель GPT (*Generative Pre-trained Transformers*), что позволяет им обучаться в процессе работы.

Так как в настоящее время, по мнению исследователей, лучшим функционалом обработки неструктурированных языковых данных обладают генеративные предобученные трансформационные модели GPT, то данные технологии также были протестированы в качестве инструмента суммаризации на примере трансформационной модели *OpenAI ChatGPT-4*[2] [Aydın, Karaarslan 2023; Azaria 2022]. Для обработки текста с помощью этого инструмента требуется гиперссылка на исходный текст, который нейросеть способна самостоятельно найти в соответствующей онлайн-базе данных.

Необходимо отметить, что, хотя эти инструменты оперируют текстовым материалом на нескольких языках, они обучены на базе англоязычных текстов. Принимая этот факт и авторский интерес во внимание, используемые в ходе исследования исходные тексты и, следовательно, созданные резюме представлены на английском языке.

ИИ-инструменты оценки эффективности автоматизированной суммаризации

Среди методов оценки эффективности автоматизированной суммаризации исследователи выделяют такие метрики, как ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), BLEU (*Bilingual Evaluation Understudy*) и METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) [Saadany, Orasan 2021].

Объединяя различные подметрики, ROUGE предоставляет исследователям и практикам систематический и количественный способы объективной оценки схожести между сгенерированными резюме и эталонными документами [Shinde et al. 2021; Sri, Dutta 2021]. Например, ROUGE-N оценивает совпадение последовательностей слов в созданном резюме и эталонном документе, анализируя семантическую информацию, содержащуюся в *n*-граммах, т. е. устойчивых словосочетаниях, или коллокациях [Ganesh et al. 2022]. ROUGE-L измеряет самую длинную общую подпоследовательность, что подчеркивает значимость общих слов в процессе оценки.

Метрика BLEU, изначально созданная для машинного перевода, также была применена к тестам суммаризации текста, что расширило использование количественной оценки. BLEU концентрируется на точности n-грамм, оценивая, насколько сгенерированное резюме соответствует эталонным документам. Совмещение метрик ROUGE и BLEU предлагает комплексную оценку таких аспектов, как полнота, точность и лексическое совпадение [Alam et al. 2003; Supriyono et al. 2024].

Еще одним значимым инструментом для оценки семантики в суммаризации текста является метрика METEOR. Она выходит за рамки простого сопоставления n-грамм, включая в анализ синонимы и стемминг [Cao, Zhuge 2020; Gao et al. 2020; Guadalupe Ramos et al. 2019; Gupta et al. 2016]. Их учет улучшает способность метрики выявлять выражения с одинаковым значением, обеспечивая более детализированное представление о семантическом качестве сгенерированных резюме [Goldstein et al. 2000; Gupta et al. 2023].

Дополнительное измерение различных подходов к суммаризации текста может быть представлено экспертными оценками лингвистов, помогающими восполнить разрыв между автоматическими метриками и субъективной природой понимания языка человеком [Gupta, Patel 2020]. Хотя автоматические метрики, безусловно, привносят объективность в процесс оценивания и повышают его эффективность, человеческие оценки предоставляют качественный слой для оценки за счет включения человеческого суждения. В таких оценках эксперты могут комплексно оценить резюме, исходя из связности, информативности и общего понимания [Fabbri et al. 2021; Mohammed Badry et al. 2013; Supriyono et al. 2024]. Следовательно, цель исследования – анализ лингвистических приемов сжатия текста, используемых нейросетью для резюмирования текстового материала.

Приемы сжатия текста

Сжатие текста происходит по двум взаимосвязанным направлениям: содержательная компрессия и языковая [Ивановская и др. 2021; Моисеенко и др. 2020; Степанюк 2021]. Одно без другого невозможно,

[2] На базе технологий OpenAI также работает сервис Summate.it.

поскольку языковые приемы могут быть применены только в том случае, если в тексте выявлено главное, а также чтобы передать основное содержание, необходимо владеть языковыми приемами сжатия текста [Коротких, Носенко 2021]. В предыдущем разделе были описаны ИИ-метрики, используемые для анализа содержательного соответствия резюме исходному тексту, т.е. для оценки эффективности содержательной компрессии.

Первым ожидаемым результатом применения инструментов автоматизированной суммаризации является сжатие исходного текста, уменьшение количества страниц, слов, знаков и других количественных показателей [Uçkan, Karci 2020]. Тестовое использование обозначенных ранее сервисов автоматизированной сумммаризации показывает, что с задачей количественной компрессии наилучшим образом справился суммаризатор *OpenAI Summate.it*. Объем текста значительно уменьшился: полученные резюме по количеству слов составили всего 1–3 % от исходных текстов. Подобным образом проявили свой функционал сжатия модели *SciSummary* и *OpenAI ChatGPT-4*, резюме которых не превысили 6 % от исходных текстов. Однако в случае применения моделей *WordTune* и *Scholarcy* к тому же текстовому материалу объем резюме составил 21–30 % от исходного текста.

Рассмотрим наиболее распространенные способы языкового сжатия текста.

1. **Исключение** (*exclusion*) подразумевает удаление из текста несущественной, малозначимой информации, например деталей, которые не влияют на общий смысл текста или его основную идею. Это могут быть описательные элементы, избыточные примеры, повторения одной и той же мысли и лишние пояснения, не добавляющие новой информации. Благодаря этому приему резюме становится более четким и сфокусированным, позволяя читателю быстрее и легче усваивать суть текста, не отвлекаясь на детали, которые не имеют ключевого значения.

Современные сервисы автоматизированной суммаризации способны значительно сокращать объем исходного текста, исключая крупные фрагменты без потери ключевого смысла. В приведенном далее отрывке резюме, созданного ИИ-моделью *WordTune*, наглядно продемонстрирован прием исключения примеров исходного текста (выделенных в оригинале) в сочетании с объединением предложений для создания более сжатого и связного изложения:

- (T2) *In the current status of precision agriculture,* ***there are*** *several issues,* ***such as unsustainable resource utilization, long-term monoculture, intensive animal farming, environmental compromises, uneven distribution of digitization, food safety issues, inefficient agrifood supply chain, and lack of awareness of and inertia toward novel changes. These issues*** *prevent achieving efficiency, productivity, and sustainability from agricultural production and escalate unintended impacts on ecosystems* (исходный текст, 61 слово) – *In the current status of precision agriculture, several issues prevent achieving efficiency, productivity, and sustainability, and escalate unintended impacts on ecosystems* (резюме, 21 слово).

В следующем примере исключается дополнительное пояснение (выделенное в оригинале), но фокус внимания не смещается и удерживается на необходимости альтернативных метрик оценки креативности:

- (T2) *Alternative metrics such as a potential creativity indicator are needed,* ***but they may be harder to quantify as they will necessarily be less tangible*** (исходный текст) – *Alternative metrics such as a potential creativity indicator are needed* (резюме).

2. **Упрощение** (*simplification*) подразумевает процесс преобразования сложных предложений и терминов в более понятные формы при сохранении основного смысла и ключевой информации [Al-Thanyyan, Azmi 2021; Maddela et al. 2021]. Данный метод особенно полезен в образовательных и научно-популярных текстах, где важно и необходимо сделать материал доступным для широкой аудитории. Процесс упрощения текста включает различные техники, такие как изменение синтаксической структуры предложений через замену и синтаксическую синонимию, объединение нескольких предложений в одно, замена сложных конструкций простыми, использование более ясных и общепринятых слов вместо специализированных терминов, а также синонимическая замена сложных выражений на их более понятные аналоги. Эти подходы помогают адаптировать информацию таким образом, чтобы она была легко воспринимаемой и доступной и при этом сохраняла содержательную ценность.

 https://doi.org/10.21603/2782-4799-2024-3-3-203-222

В результате синтаксической трансформации упрощаются сложноподчиненные предложения с определительными придаточными:

- (T1) *Problems* ***that require a convergence approach*** *are those that involve nonlinearity* (исходный текст) – *Convergence approaches are used to solve problems involving nonlinearity* (резюме);
- (T3) *Solar cells are devices* ***that convert sunlight directly into electricity****; typical semiconductor materials are utilized to form a PV solar cell device* (исходный текст) – *Solar cells use semiconductor materials to convert sunlight directly into electricity* (резюме).

Во втором примере также исключается повторяющийся термин *solar cell device*.

Значительного упрощения текста можно достичь путем сокращения длины предложений и объединения нескольких предложений в одно (в оригинале выделены фразы, сохраненные в полученном резюме):

- (T2) ***In the current status of precision agriculture, there are several issues****, such as unsustainable resource utilization, long-term monoculture, intensive animal farming, environmental compromises, uneven distribution of digitization, food safety issues, inefficient agrifood supply chain, and lack of awareness of and inertia toward novel changes. These issues* ***prevent achieving efficiency, productivity, and sustainability from agricultural production and escalate unintended impacts on ecosystems*** (исходный текст) – *In the current status of precision agriculture, several issues prevent achieving efficiency, productivity, and sustainability, and escalate unintended impacts on ecosystems* (резюме).

В предыдущих примерах упрощение осуществлялось путем последовательного извлечения информации. В следующем же демонстрируется подход, основанный на трансформации (сохраненные фразы выделены):

- (T1) *To the greatest extent possible, funding for* ***convergence processes*** *should allow for* ***problem identification*** *to occur after funding has been granted, and for desired products and outcomes to be* ***flexible*** *and* ***moving targets*** *as a reflection of the learning and transformation that should occur in a convergence process* (исходный текст) – *The problem identification process in a convergence process should be flexible and allow for moving targets* (резюме).

Нейросети обучены приемам упрощения текста на основе синонимических замен, что подтверждается трансформациями текста далее. В них коннотативно окрашенные лексемы *reveal*, *unearth*, *center* заменены стилистически нейтральными лексемами *find*, *identify*, *focus*, широко используемыми в научных, технических и повседневных контекстах:

- (T4) *The analysis of metaphor-related research studies published between 2015 and 2020* ***revealed****…* (исходный текст) – *A systematic review of metaphor-related research studies published between 2015 and 2020* ***found****…* (резюме);
- (T4) <…> *the thematic analysis* ***unearthed*** *potential gaps and under-researched areas* (исходный текст) – *The thematic analysis* ***identified*** *gaps and under-researched areas* (резюме);
- (T4) <…> *metaphor studies in this review* ***centered more on*** *written discourse than spoken data* (исходный текст) – *The reviewed studies* ***focused more on*** *written discourse than spoken data* (резюме).

В следующих резюме, созданных сервисом *WordTune*, упрощение осуществляется через замену синонимами целых фраз, позволяющих избегать излишней повторяемости, заменяя повторы или сложные фразы на их эквиваленты, что делает текст разнообразнее и понятнее для читателя [Wilber et al. 2021]:

- (T1) *Convergence research* ***may provide opportunities*** *to confront and navigate Arctic change* (исходный текст) – *Convergence research* ***can help*** *confront and navigate Arctic change* (резюме);
- (T3) *Green technology sources* ***play an important role*** *in sustainably providing energy supplies* (исходный текст) – *Green technology sources* ***are important*** *in sustainably providing energy supplies* (резюме)*;*
- (T3) *The global community* ***is starting to shift towards utilizing*** *sustainable energy sources and reducing dependence on traditional fossil fuels as a source of energy* (исходный текст) – *Decision-makers* ***are switching to*** *renewable energy sources and reducing dependence on traditional fossil fuels* (резюме).

3. **Обобщение** (*generalization*) как прием сжатия текста предполагает сведение частных случаев к более общим понятиям или выводам. Этот прием помогает сократить текст за счет объединения схожих идей и выведения общих закономерностей [Dönicke et al. 2021; Hupkes et al. 2020]. Например, создавая резюме, нейросеть может обобщать информацию из нескольких предложений:

- (T2) ***Smart agriculture*** *is an evolving field that leverages technological innovations to transform traditional farming practices.* ***The integration of digital technologies*** *into agriculture has opened up new opportunities and possibilities, revolutionizing the way farmers manage their crops, resources, and operations.* <…> *Leveraging machine vision technology,* ***the Internet of Things (IoT), and artificial intelligence (AI)*** *can lead to enhanced precision and efficiency in agricultural processes, benefiting both farmers and the environment* (исходный текст) – *The integration of digital technologies, such as big data analytics, machine vision technology, the Internet of Things (IoT), and artificial intelligence (AI), is revolutionizing precision agriculture and paving the way for smart farming* (резюме).

Особого внимания заслуживают примеры предложений, обобщающих и суммирующих информацию, изложенную в нескольких разделах исходного текста, в виде параллельных синтаксических конструкций:

- (T2) *New trends in precision agriculture include the use of* ***big data analytics for*** *decision making,* ***machine vision technology for*** *accurate data collection,* ***the IoT for*** *real-time monitoring and control,* ***AI and machine learning for*** *data analysis and prediction,* ***guidance systems for*** *optimized field operations, and* ***blockchain technology for*** *secure data sharing* (резюме).

Лексическая креативность нейросетей

Абстрактивные резюме могут содержать лексические конструкции, отсутствующие в исходном тексте, т. к. нейронная сеть использует свои собственные датасеты для формулирования идей. Так, в оригинальной статье (T3), анализирующей будущие перспективы применения солнечной энергии, лексема *future* используется в сочетаниях, таких как *the foreseeable future* (обозримое будущее), *the near future* (ближайшее будущее), *a bright future* (яркое будущее), однако прилагательное *promising*, описывающее будущее многообещающим, не встречается. В суммарном же тексте читаем: *The future of solar energy looks* ***promising***.

В настоящее время и пользователи, и разработчики признают, что генеративные предобученные трансформационные модели наилучшим образом обрабатывают неструктурированные языковые данные. При анализе функции суммаризации, выполняемой генеративной моделью *OpenAI ChatGPT-4*, было отмечено, что она, действительно, генерирует сочетания слов, которые не используются в оригинальном тексте.

Например, в (T1) существительное *concept* встречается в словосочетаниях *the concept of ecological resilience, the concept of a solution*, а в суммарном тексте читаем *the concept of convergence research*. Исходный текст (T2) описывает различные виды исследований (*statistical analysis, image analysis, soil analysis, on-site analysis, real-time analysis*), резюме же дает качественную характеристику *an in-depth analysis* (глубокий анализ). Аналогично в (T2) встречаются сочетания *a comprehensive understanding, comprehensive frameworks*, в суммарном – *a comprehensive examination* (комплексное, обстоятельное исследование).

ChatGPT демонстрирует способности к обобщению и выявлению основных идей исходного текста. Так, в исходной статье анализируются возможности конвергентного подхода к решению комплексных задач, и суммарный текст указывает, что в оригинале описывается необходимость механизмов широкой поддержки исследований: *broad-based support mechanisms* (T1), при этом прилагательное *broad-based* в исходном тексте не упоминается.

В другом резюме одной из проблем использования солнечной энергии является ее нестабильность, «прерывистость»: *the challenges of solar energy, such as intermittency* (T3), однако в исходном тексте существительное *intermittency* отсутствует.

Согласно Кембриджскому словарю, лексема *intermittency* означает *the fact of stopping and starting repeatedly or with periods of time in between*. Вероятно, нейросеть сделала такой вывод из следующего предложения: *It is important to mention here the operational challenges of solar energy in that it does not work at night*,

has less output in cloudy weather and does not work in sandstorm conditions[3] (Здесь важно упомянуть эксплуатационные проблемы солнечной энергетики, поскольку она не работает ночью, имеет меньшую мощность в пасмурную погоду и не работает в условиях песчаных бурь[4]).

Помимо этого, при создании резюме искусственный интеллект использует соответствующие клише и глаголы косвенной речи: *provides an* ***in-depth analysis of***, ***focuses on***, ***it provides a comprehensive examination***, ***the paper discusses***, ***highlights***, ***examines***, ***emphasizes***. В результате чего полученное резюме, кроме значительного сжатия объема, демонстрирует признаки довольно успешного резюме, написанного в академическом стиле, обобщающего содержание всего текста, имеющего адекватную структуру и использующего средства логической связи.

В качестве примеров приведем отрывки из некоторых резюме, полученных при тестированиии сервиса SciSummary:

- (T1) *The article* <…> ***discusses*** *the increasing importance of collaboration in modern science to address complex societal problems. It* ***highlights*** *the U.S. National Science Foundation's prioritization of convergence research as a means to solve such challenging issues.* ***The authors provide*** *their understanding of the objectives of convergence research and* ***outline*** *the conditions and processes essential for successful convergence research*.
- (T2) ***The paper discusses*** *the application of advanced digital technologies in precision agriculture,* <…>. ***It emphasizes*** *the importance of site-specific management decisions in agriculture,* ***considering*** *factors such as soil and climate properties,* <…> ***The paper discusses*** *the role of big data analytics, machine vision, the Internet of Things (IoT), artificial intelligence (AI), machine learning (ML) and deep learning (DL) in modern agriculture* <…>.
- (T3) ***The research paper discusses*** *the importance of sustainable energy development,* ***particularly focusing on*** *solar energy applications in* <…>. ***It highlights*** *key international agreements such as* <…>. ***The paper emphasizes*** *the increasing demand for* <…>.
- (T4) ***This review paper discusses*** *the use of metaphors,* ***particularly*** *in everyday language, and* <…>. ***The paper presents*** *a systematic review of* <…>. ***The review emphasizes*** *the trends, gaps, and under-researched areas in the analyzed literature.* ***It showcases*** *the distribution of published research* <…>.

Даже беглый взгляд на эти суммарные тексты позволяет сделать вывод, что они созданы по принципу аннотации научной статьи и используют глаголы косвенной речи, или глаголы отчетности, а также в той или иной степени повторяют структуру друг друга.

Ограничения в работе автоматизированных суммаризаторов

Будучи обученными действовать в рамках определенных алгоритмов, автоматизированные суммаризаторы создают резюме определенного объема и структуры. Так, ИИ-инструментом *Scholarcy* в каждом формируемом резюме были выделены и маркированы такие структурные компоненты, как: аннотация (*Abstract*); ключевые аспекты (*Scholarcy Highlights*); резюме (*Scholarcy Summary*), внутри которого были отмечены введение (*Introduction*), цели (*Objectives*), результаты (*Results*), выводы (*Conclusion*), направления дальнейших исследований (*Future Work*).

Кроме того, для каждого резюме было определено название[5] по формальному признаку, т. е. за него суммаризатор принял первое предложение в исходном тексте, что не просто не имеет ничего общего с оригинальным названием статьи, но и абсолютно не отражает смысловой идеи текста. Выбранные предложения ни структурно, ни семантически не могут рассматриваться в качестве названия научной статьи, роль которого – сжато представить содержание текста и привлечь внимание потенциального читателя [Сорокина, Уланова 2020]. Эти псевдоназвания приведены в таблице в сопоставлении с оригинальными названиями исходных текстов, что наглядно демонстрирует абсурдность первых и определенные ограничения в работе данной суммаративной модели.

[3] Cambridge Dictionary. URL: https://dictionary.cambridge.org/ (accessed 3 May 2024).

[4] Здесь и далее по тексту перевод выполнен автором статьи.

[5] Хотя в ходе подготовки исходных текстов для тестовой суммаризации подзаголовки и другие указатели на структурные части были удалены из резюме.

Табл. Сопоставление названий в исходных и суммарных текстах
Tab. Comparative analysis of titles: reference texts vs. summaries

Статьи	Исходный текст	Суммарный текст
T1	*The emergence of convergence* (зарождение конвергенции)	*Science is increasingly a collaborative pursuit* (Наука все чаще становится совместным занятием)
T2	*The path to smart farming: innovations and opportunities in precision agriculture* (Путь к умному земледелию: инновации и возможности точного земледелия)	*Precision agriculture is a management strategy for addressing geographical and temporal variabilities in agricultural fields* (Точное земледелие – это стратегия управления, направленная на устранение географических и временных различий в сельскохозяйственных полях)
T3	*Solar energy technology and its roles in sustainable development* (Технология солнечной энергетики и ее роль в устойчивом развитии)	*With reference to the recommendations of the UN, the Climate Change Conference, COP26, was held in Glasgow, UK, in 2021* (в соответствии с рекомендациями ООН в 2021 г. в Глазго (Великобритания) прошла конференция по изменению климата COP26)
T4	*Corpus-based studies of metaphor: an overview* (Обзор корпусных исследований метафоры)	*The classical theorists of metaphors believed that metaphor functions as a literary device to create an artistic effect* (Классические теоретики метафоры считали, что метафора функционирует как литературный прием для создания художественного эффекта)

Сразу за фразой, выполняющей роль названия, следует предложение, которое, вероятно, должно считаться вводной фразой-лидом (*Lead*). Оно, подобно оформлению новостной статьи, маркировано цветом и курсивом. В журналистике задачей такого одного предложения или мини-абзаца является привлечение внимания и анонсирование основных идей самого текста [Ленкова 2023].

В научных публикациях лид используется крайне редко. Анализ этих вводных предложений в полученных резюме показывает, что они без изменений экстрагированы из заключительной части статьи или раздела близкого к концу. К сожалению, данные предложения не отвечают используемой цели, более того, они или выглядят бессмысленными или вовсе ввводят в заблуждение. Вероятно, обращение к заключительным частям исходного текста обусловлено тем, что именно конец любого текста, в том числе и научной публикации, формирует смысловой узел всего замысла [Сорокина, Уланова 2020], который суммаризатор и попытался вычленить в ходе анализа. Но, как можно видеть в приведенных ниже примерах, результат не был успешен:

- (T1) Лид *The three transcendent-style workshops undertaken in the New Arctic convergence workshops each represented a broadening of the problem definition and the voices and disciplines represented in the room* (В рамках семинаров по конвергенции в Новом Арктическом регионе были проведены три трансцендентные мастерские, каждая из которых способствовала расширению понимания проблематики и вовлечению большего числа разнообразных мнений и научных дисциплин) соответствует первому предложению заключительной части статьи.

Такой выбор можно объяснить конвенцией академического письма, структура абзаца которого следует четко определенному порядку, помогающему читателю понять аргументацию автора и следовать ей [Сорокина 2016]. Согласно этому порядку, каждый абзац должен начинаться с тематического предложения *(Topic Sentence*), ясно указывающего на основную идею или тему абзаца, задающего тон всему абзацу и часто устанавливающего связь с тезисом или главной идеей текста.

С большой долей вероятности алгоритмы данной нейронной модели были настроены на распознавание именно таких взаимосвязей научного текста. Но анализируемый текст написан биологическим автором, который не всегда действует в рамках жестких правил, заданных алгоритмами, и в данном абзаце смысловая нагрузка содержится в заключительном предложении (*Concluding Sentence*), что характерно и для академического письма: *We argue that convergence research will benefit from the purposeful movement between focused and transcendent science*,

 https://doi.org/10.21603/2782-4799-2024-3-3-203-222

and that these processes will take time with success measured by an expanded set of non-traditional and traditional metrics (Мы утверждаем, что исследования конвергенции выиграют от целенаправленного взаимодействия между узконаправленной и трансцендентной наукой, и что эти процессы потребуют времени, а успех их реализации будет измеряться с использованием расширенного набора как традиционных, так и иновационных метрик).

- (T2) Лид *This study examined the rheological properties and printing performances of edible inks made from soy protein isolate, wheat gluten, and rice protein* (В этом исследовании были изучены реологические свойства и характеристики печати съедобных чернил, изготовленных из соевого изолята, пшеничного глютена и рисового белка) вводит в заблуждение потенциального читателя, т.к. в представленном обзоре анализируются предложения по точному земледелию, описываются области их применения и объясняется, какими способами каждая из этих областей может постоянно развиваться, чтобы поддерживать методы устойчивого ведения сельского хозяйства: *Throughout this review, successful precision agriculture proposals and real-world implementations are analyzed* (В этом обзоре проанализированы успешные предложения и реальные примеры внедрения точного земледелия); *we aim to provide a comprehensive understanding of how this field can continually evolve to support sustainable farming practices* (мы стремимся обеспечить всестороннее понимание того, как эта область может непрерывно развиваться для поддержки устойчивых методов ведения сельского хозяйства).
- (T3) Лид *The Paris Climate Accords is a worldwide agreement on climate change signed in 2015, which addressed the mitigation of climate change, adaptation and finance* (Парижское соглашение по климату – это всемирное соглашение по изменению климата, подписанное в 2015 г., которое охватывает вопросы смягчения последствий изменения климата, адаптации и финансирования) экстрагирован из введения, не несет никакой значимой смысловой нагрузки для основных идей публикации и, безусловно, несколько девиантен по отношению к истинным целям и задачам публикации, которые легко обнаруживаются далее во вводной части исходной статьи: *The significance of this paper is to highlight solar energy applications to ensure sustainable development; thus, it is vital to researchers, engineers and customers alike. The article's primary aim is to raise public awareness and disseminate the culture of solar energy usage in daily life, since moving forward, it is the best.*

Следуя постулату о смысловом узле произведения, в самом начале заключительной части исходной статьи легко находим возможную опцию для лида, которая лучше всего отражает идеи данной публикации: *This paper highlights the significance of sustainable energy development. Solar energy would help steady energy prices and give numerous social, environmental and economic benefits.*

Заключение

Обзор автоматизированных методов суммаризации, основанных на синтезе достижений в области лингвистики, компьютерных технологий и ИИ, показывает, что современные алгоритмы суммаризации опираются на знания о структуре языка и активно используют лингвистические приемы сжатия текста, позволяющие моделям создавать резюме без потери смысла и когерентности исходного текста.

Экстрактивные технологии обеспечивают точность в извлечении ключевых фраз и предложений, сохраняют оригинальный язык и стиль исходного текста, что делает их особенно эффективными для работы с фактическими и информационными материалами. Однако, следуя жестко заданным алгоритмам, экстрактивные модели часто упускают важные нюансы и создают резюме, которые могут быть несвязными и нелогичными. В свою очередь, абстрактивные технологии используют более сложные лингвистические приемы, такие как перефразирование, синонимическая замена и создание новых синтаксических конструкций, помогающие в генерации более связных и осмысленных резюме. Однако способность модели к созданию нового текста может приводить к искажению исходного содержания, что требует внимательного подхода к выбору модели суммаризации в зависимости от специфики задачи и сложности текста.

Интересным аспектом для дальнейших исследований является анализ созданных резюме, которые представляют собой уникальный лингвистический продукт. Во-первых, такие резюме демонстрируют, как алгоритмы обрабатывают и трансформируют исходный текст, что позволяет изучать закономерности машинного понимания языка и его

ограничения. Во-вторых, созданные резюме дают возможность исследовать новые формы выражения ключевых идей текста, выявляя неочевидные для человека оригинальные синтаксические и лексические решения. Эти резюме могут служить материалом для анализа специфических лингвистических особенностей (использование синонимии, изменение структуры предложений и креативные приемы генерации).

Итак, автоматизированные методы суммаризации не только являются эффективным инструментом для обработки текстов, но и предоставляют уникальный материал для лингвистического анализа, способствующего углубленному пониманию процессов генерации и трансформации текстовой информации.

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

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

Материал для тестирования ИИ-суммаризаторов / Material for testing AI-powered summarizers

(T1) Sundstrom Sh. M., Angeler D. G., Ernakovich J. G., García J. H., Hamm J. A., Huntington O., Allen C. R. The emergence of convergence. *Elementa: Science of the Anthropocene*, 2023, 11(1). https://doi.org/10.1525/elementa.2022.00128

(T2) Karunathilake E. M. B. M., Le A. T., Heo S., Chung Y. S., Mansoor Sh. The path to smart farming: Innovations and opportunities in precision agriculture. *Agriculture*, 2023, 13(8). https://doi.org/10.3390/agriculture13081593

(T3) Maka A. O. M., Alabid J. M. Solar energy technology and its roles in sustainable development. *Clean Energy*, 2022, 6(3): 476–483. https://doi.org/10.1093/ce/zkac023

(T4) Abdul Malik N., Ya Shak M. S., Mohamad F., Joharry S. A. Corpus-based studies of Metaphor: An overview. *Arab World English Journal*, 2022, 13(2): 512–528. https://dx.doi.org/10.24093/awej/vol13no2.36

Литература / References

Арефьева Е. С. Статья как основной жанр современного научного стиля. *Современные лингвокоммуникативные практики*, отв. ред. Д. А. Розеватов. Саратов, 2018. Вып. 1. С. 14–19. [Arefeva E. S. Article as the main genre of modern scientific style. *Modern linguistic and communicative practices*, ed. Rozavatov D. A. Saratov, 2018, iss. 1, 14–19. (In Russ.)] https://elibrary.ru/xvzowd

Безлепкин Е. А., Зайкова А. С. Нейрофилософия, философия нейронаук и философия искусственного интеллекта: проблема различения. *Философские науки*. 2021. Т. 64. № 1. С. 71–87. [Bezlepkin E. A., Zaykova A. S. Neurophilosophy, philosophy of neuroscience, and philosophy of artificial intelligence: The problem of distinguishing. *Russian Journal of Philosophical Sciences*, 2021, 64(1): 71–87. (In Russ.)] https://doi.org/10.30727/0235-1188-2021-64-1-71-87

Белов С. Д., Зрелова Д. П., Зрелов П. В., Кореньков В. В. Обзор методов автоматической обработки текстов на естественном языке. *Системный анализ в науке и образовании*. 2020. № 3. С. 8–22. [Belov S. D., Zrelova D. P., Zrelov P. V., Korenkov V. V. Overview of methods for automatic natural language text processing. *System Analysis in Science and Education*, 2020, (3): 8–22. (In Russ.)] https://doi.org/10.37005/2071-9612-2020-3-8-22

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

Вертинова А. А., Пашук Н. Р., Макогонова П. В., Кошелева А. И. Оценка влияния информационного шума на принятие решений. *Лидерство и менеджмент*. 2022. Т. 9. № 3. С. 877–890. [Vertinova A. A., Pashuk N. R., Makogonova P. V., Kosheleva A. I. Assessing the infoglut impact on decision-making. *Liderstvo i menedzhment*, 2022, 9(3): 877–890 (In Russ.)] https://doi.org/10.18334/lim.9.3.116218

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. Т. 4: Искусственный интеллект. [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, vol. 4: Artificial intelligence. (In Russ.)] URL: http://repo.ssau.ru/handle/Informacionnye-tehnologii-i-nanotehnologii/Avtomaticheskoe-referirovanie-tekstov-100191 (дата обращения: 03.05.2024). 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. 11 Intern. Sci.-Prac. Conf., Moscow, 12 May 2023. St. Petersburg: Pechatnyy tsekh, 2023, 43–52. (In Russ.)] https://elibrary.ru/nonvjs

Гринев-Гриневич С. В., Сорокина Э. А., Молчанова М. А. Еще раз к вопросу об определении термина. *Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика*. 2022. Т. 13. № 3. С. 710–729. [Grinev-Griniewicz S. V., Sorokina E. A., Molchanova M. M. Reconsidering the definition of the term. *RUDN Journal of Language Studies, Semiotics and Semantics*, 2022, 13(3): 710–729. (In Russ.)] https://doi.org/10.22363/2313-2299-2022-13-3-710-729

Гурьева Н. Н. Этапы и аспекты изучения парцеллированных конструкций в отечественном языкознании. *Вестник Тверского государственного университета. Серия: Филология*. 2020. № 1. С. 109–114. [Guryeva N. N. Stages and aspects of the study of parceled constructions in Russian linguistics. *Vestnik Tverskogo gosudarstvennogo universiteta. Seriia: Filologiia*, 2020, (1): 109–114. (In Russ.)] https://elibrary.ru/xjljuw

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

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

Ивановская О. И., Криводерева Л. В., Харченко В. А. Об одном из приемов сжатия текста. *Вестник научных конференций*. 2021. № 7-2. С. 57–58. [Ivanovskaia O. I., Krivodereva L. V., Kharchenko V. A. A text compression method. *Vestnik nauchnykh konferentsii*, 2021, (7-2): 57–58. (In Russ.)] https://elibrary.ru/hptaxm

Иванюкович В. А., Борковский Н. Б., Лефанова И. В. Применение нейросетевых технологий при обработке неструктурированной информации. *Управление информационными ресурсами*: XIX Междунар. науч.-практ. конф. (Минск, 22 марта 2023 г.) Мн.: АУ РБ, 2023. С. 277–279. [Ivaniukovich V. A., Borkovskii N. B., Lefanova I. V. Application of neural network technologies in processing unstructured information. *Information resource management*: Proc. XIX Intern. Sci.-Prac. Conf., Minsk, 22 Mar 2023. Minsk: AHA RB, 2023, 277–279. (In Russ.)] https://elibrary.ru/funmfv

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. *Sovremennye problemy nauki i obrazovaniia*, 2021, (2). (In Russ.)] https://doi.org/10.17513/spno.30665

Ленкова Т. А. Лид – структурный элемент статьи и самодостаточный текст. *Филология и человек*. 2023. № 1. С. 179–191. [Lenkova T. A. The lead paragraph is a structural element of the article and a self-contained text. *Filologiya i chelovek*, 2023, (1): 179–191. (In Russ.)] https://elibrary.ru/zxfpzg

Малышева Е. Ю., Лычагина В. А. Математические методы исследования лингвистики. *Язык и культура в эпоху интеграции научного знания и профессионализации образования*. 2022. № 3-1. С. 170–177. [Malisheva E. Yu., Lichagina V. A. Mathematical methods in linguistic research. *Iazyk i kultura v epokhu integratsii nauchnogo znaniia i professionalizatsii obrazovaniia*, 2022, (3-1): 170–177. (In Russ.)] https://elibrary.ru/pxlqjx

Моисеенко И. М., Мальцева-Замковая Н. В., Чуйкина Н. В. Смысловое сжатие текста как компонент коммуникативной компетенции. *Коммуникативные исследования*. 2020. Т. 7. № 2. С. 439–458.

[Moiseenko I. M. Maltseva-Zamkovaja N. V., Tšuikina N. V. Conceptual compression of a text as a component of communicative competence. *Communication Studies*, 2020, 7(2): 439–458. (In Russ.)] https://doi.org/10.24147/2413-6182.2020.7(2).439-458

Мусаев А. А., Григорьев Д. А. Обзор современных технологий извлечения знаний из текстовых сообщений. *Компьютерные исследования и моделирование*. 2021. Т. 13. № 6. С. 1291–1315. [Musaev A. A., Grigoriev D. A. Extracting knowledge from text messages: Overview and state-of-the-art. *Computer Research and Modeling*, 2021, 13(6): 1291–1315. (In Russ.)] https://doi.org/10.20537/2076-7633-2021-13-6-1291-1315

Пенцова М. М. Лингвосемиотика скандинавской топонимии Шотландии. *Язык. Культура. Перевод. Коммуникация*, науч. ред. В. З. Демьянков. М.: Тезаурус, 2015. С. 533–537. [Pentsova M. M. Linguistic semiotics of Scandinavian place-names in Scotland. *Language. Culture. Translation. Communication,* ed. Demyankov V. Z. Moscow: Tezaurus, 2015, 533–537. (In Russ.)] https://elibrary.ru/ynpdqd

Перелетов К. С. Обзор методов суммаризации текстов и области их применения. *Высшая школа: научные исследования*: Межвуз. Междунар. конгресс (Москва, 10 июня 2021 г.) М.: Инфинити, 2021. С. 147–156. [Pereletov K. S. Review of methods for summarizing texts and their areas of application. *Higher school: Scientific research*: Proc. Interuniv. Intern. Congress, Moscow, 10 Jun 2021. Moscow: Infiniti, 2021, 147–156. (In Russ.)] https://elibrary.ru/xipzom

Полонский Д. А., Федосова А. О. Предобработка текста для решения NLP (Natural Language Processing). *Мавлютовские чтения*: XV Всерос. науч. конф. (Уфа, 26–28 октября 2021 г.) Уфа: УГАТУ, 2021. Т. 4. С. 798–802. [Polonsky D. A., Fedosova A. O. Text preprocessing for solving NLP (Natural Language Processing). *Mavlyutov Readings*: Proc. XV All-Russian Sci. Conf., Ufa, 26–28 Oct 2021. Ufa: USATU, 2021, vol. 4, 798–802. (In Russ.)] https://elibrary.ru/autkfl

Полякова И. Н., Зайцев И. О. Модификация графового метода для задач автоматического реферирования с учетом синонимии. *International Journal of Open Information Technologies*. 2022. Т. 10. № 4. С. 45–54. [Polyakova I. N., Zaitsev I. O. Modification of the graph method for automatic abstraction tasks taking into account synonymy. *International Journal of Open Information Technologies*, 2022, 10(4): 45–54. (In Russ.)] https://elibrary.ru/chvbat

Соколова Ю. В., Чалова О. А. Особенности формирования и развития навыков самостоятельной работы на начальных этапах высшего профессионального образования. *Мир науки. Педагогика и психология*. 2020. Т. 8. № 2. [Sokolova Yu. V., Chalova O. A. Formation and development Features of independent work skills at the initial stages of higher professional education. *World of Science. Pedagogy and psychology*, 2020, 8(2). (In Russ.)] https://doi.org/10.15862/81PDMN220

Сорокина С. Г. Искусственный интеллект в контексте междисциплинарных исследований языка. *Вестник Кемеровского государственного университета. Серия: Гуманитарные и общественные науки*. 2023. Т. 7. № 3. С. 267–280. [Sorokina S. G. Artificial intelligence in interdisciplinary linguistics. *Vestnik Kemerovskogo gosudarstvennogo universiteta. Seriia: Gumanitarnye i obshchestvennye nauki*, 2023. 7(3): 267–280. (In Russ.)] https://doi.org/10.21603/2542-1840-2023-7-3-267-280

Сорокина С. Г. Использование рекуррентности как средства аргументации при построении текстов научного содержания: дис. … канд. филол. наук. М., 2016. 196 с. [Sorokina S. G. *Recurrence as a means of argumentation in the construction of texts of scientific content*. Cand. Philol. Sci. Diss. Moscow, 2016, 196. (In Russ.)] https://elibrary.ru/zejqeb

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, 2024. С. 132–138. [Sorokina S. G. Applying automatic summarization technology to academic publications. *The three "L's" in the paradigm of modern humanities: Linguistics, literary studies, linguodidactics*: Proc. All-Russian Sci.-Prac. Conf., Moscow, 23 Nov 2023. Moscow: Yazyki Narodov Mira, 2024, 132–138. (In Russ.)] https://elibrary.ru/duydpi

Сорокина С. Г., Уланова К. Л. Имплементация категории тождества в названиях публицистических и научных текстов. *Современное педагогическое образование*. 2020. № 2. С. 202–207. [Sorokina S. G., Ulanova K. L. The role of article title in implementing the category of identity. *Sovremennoe pedagogicheskoe obrazovanie*, 2020, (2): 202–207. (In Russ.)] https://elibrary.ru/aqclzy

Степанюк Ю. В. К проблеме классификации способов лингводидактической адаптации иноязычных текстов. *Язык и действительность. Научные чтения на кафедре романских языков им. В. Г. Гака*: VI Междунар. конф. (Москва, 22–26 марта 2021 г.) М.: Спутник+, 2021. С. 411–417. [Stepanyuk Yu. V. Classifying methods of linguadidactic adaptation of foreign language texts. *Language and reality. Scientific readings at the V. G. Gak Department of Romance Language*s: Proc. VI Intern. Conf., Moscow, 22–26 Mar 2021. Moscow: Sputnik+, 2021, vol. 6, 411–417. (In Russ.)] https://elibrary.ru/hkllet

Толстых О. М. Использование образовательной электронной среды Moodle для оптимизации образовательного процесса по иностранному языку студентов неязыковых специальностей. *Омские научные чтения*: Всерос. науч.-практ. конф. (Омск, 11–16 декабря 2017 г.) Омск: ОмГУ, 2017. С. 442–443. [Tolstykh O. M. The usage of the educational electronic environment Moodle for optimisation of the educational process in teaching a foreign language of non-linguistic students. *Omsk Scientific Readings*: Proc. All-Russian Sci.-Prac. Conf., Omsk, 11–16 Dec 2017. Omsk: OmSU, 2017, 442–443. (In Russ.)] https://elibrary.ru/otgrhl

Чернышкова Е. В., Родионова Т. В., Веретельникова Ю. Я. Особенности обучения студентов медицинского профиля реферированию и аннотированию иноязычных текстов по специальности. *Педагогическое взаимодействие: возможности и перспективы*: V Междунар. науч.-практ. конф. (Саратов, 28–29 апреля 2023 г.) Саратов: СГМУ, 2023. С. 231–241. [Chernyshkova E. V. Rodionova T. V., Veretelnikova Yu. Ya. Teaching medical students to summarize and annotate foreign language texts. *Pedagogical interaction: Opportunities and prospects*: Proc. V Intern. Sci.-Prac. Conf., Saratov, 28–29 Apr 2023. Saratov: SSMU, 2023, 231–241. (In Russ.)] https://elibrary.ru/tqwzlg

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

Alam H., Kumar A., Nakamura M., Rahman F., Tarnikova Y., Wilcox Che. Structured and unstructured document summarization: Design of a commercial summarizer using Lexical chains. *ICDAR'03*: Proc. 7 Intern. Conf., Edinburgh, 6 Aug 2003. IEEE, 2003, 1147–1152. https://doi.org/10.1109/ICDAR.2003.1227836

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

Al-Thanyyan S. S., Azmi A. M. Automated text simplification: A survey. *ACM Computing Surveys*, 2021, 54(2): 1–36. https://doi.org/10.1145/3442695

Arana-Catania M., Procter R., He Yu., Liakata M. Evaluation of abstractive summarisation models with machine translation in deliberative processes. *Proceedings of the Third Workshop on New Frontiers in Summarization, online*, 2021. Stroudsburg: ACL, 2021, 57–64. https://doi.org/10.18653/v1/2021.newsum-1.7

Aydın Ö., Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, 2023, 11(3): 118–134. https://doi.org/10.2139/ssrn.4341500

Azaria A. ChatGPT: Usage and limitations. 2022. https://doi.org/10.31219/osf.io/5ue7n

Belwal R. C., Rai S., Gupta A. A new graph-based extractive text summarization using keywords or topic modeling. *Journal of Ambient Intelligence and Humanized Computing*, 2022, 12: 8975–8990. https://doi.org/10.1007/s12652-020-02591-x

Bhargava R., Sharma Ya. Deep extractive text summarization. *Procedia Computer Science*, 2020, 167: 138–146. https://doi.org/10.1016/j.procs.2020.03.191

Bhat I. K., Mohd M., Hashmy R. SumItUp: A hybrid single-document text summarizer. *Soft computing: Theories and applications. Advances in intelligent systems and computing*, eds. Pant M., Ray K., Sharma T., Rawat S., Bandyopadhyay A. Singapore: Springer, 2018, 619–634. https://doi.org/10.1007/978-981-10-5687-1_56

Cao M., Zhuge H. Automatic evaluation of text summarization based on semantic link network. *SKG 2019*: Proc. 15 Intern. Conf., Guangzhou, 17–18 Sep 2019. IEEE, 2020, 107–114. https://doi.org/10.1109/SKG49510.2019.00026

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*, Hong Kong, 2019. ACL, 2019, 41–50. https://doi.org/10.18653/v1/D19-5106

Dehru V., Tiwari P. K., Aggarwal G., Joshi B., Kartik P. Text summarization techniques and applications. *ASCI 2020*: Proc. Intern. Conf., Jaipur, 22–23 Dec 2020. IOP, 2021, vol. 1099. https://doi.org/10.1088/1757-899X/1099/1/012042

Dönicke T., Gödeke L., Varachkina H. Annotating quantified phenomena in complex sentence structures using the example of generalising statements in literary texts. *Proceedings of the 17th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, online,* 2021. ACL, 2021, 20–32.

Fabbri A. R., Kryściński W., McCann B., Xiong C., Socher R., Radev D. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 2021, 9: 391–409. https://doi.org/10.1162/tacl_a_00373

Ganesh A., Jaya A., Sunitha C. An overview of semantic based document summarization in different languages. *ECS Transactions*, 2022, 107(1): 6007–6017. https://doi.org/10.1149/10701.6007ecst

Gao Y., Xu Y., Huang H., Liu Q., Wei L., Liu L. Jointly learning topics in sentence embedding for document summarization. *IEEE Transactions on Knowledge and Data Engineering*, ed. Chen L. Piscataway: IEEE, 2020, 32(4): 688–699. https://doi.org/10.1109/TKDE.2019.2892430

Gehrmann S., Deng Y., Rush A. M. Bottom-up abstractive summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 Oct – 4 Nov 2018. ACL, 2018, 4098–4109. https://doi.org/10.18653/v1/D18-1443

Ghadimi A., Beigy H. Hybrid multi-document summarization using pre-trained language models. *Expert Systems with Applications*, 2022, 192. https://doi.org/10.1016/j.eswa.2021.116292

Ghodratnama S., Zakershahrak M., Sobhanmanesh F. *Adaptive summaries: A personalized concept-based summarization approach by learning from users' feedback*, 2021. https://doi.org/10.48550/arXiv.2012.13387

Goldstein J., Mittal V., Carbonell J., Kantrowitz M. Multi-document summarization by sentence extraction. *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarizatio*n, Seattle, 30 Apr 2000. Stroudsburg: ACL, 2000, 4: 40–48. https://doi.org/10.3115/1117575.1117580

Guadalupe Ramos J., Navarro-Alatorre I., Flores Becerra G., Flores-Sánchez O. A formal technique for text summarization from web pages by using latent semantic analysis. *Research in Computing Science*, 2019, 148(3): 11–22. https://doi.org/10.13053/rcs-148-3-1

Gupta S., Gupta S. K. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 2019, 121: 49–65. https://doi.org/10.1016/j.eswa.2018.12.011

Gupta H., Kottwani A., Gogia S., Chaudhari Sh. Text analysis and information retrieval of text data. *WiSPNET 2016*: Proc. Intern. Conf., Chennai, 23–25 Mar 2016. IEEE, 2016, 788–792. https://doi.org/10.1109/WiSPNET.2016.7566241

Gupta H., Patel M. Study of extractive text summarizer using the elmo embedding. *I-SMAC 2020*: Fourth Intern. Conf., Palladam, 7–9 Oct 2020. IEEE, 2020, 829–834. https://doi.org/10.1109/I-SMAC49090.2020.9243610

Gupta S., Sharaff A., Nagwani N. K. Frequent item-set mining and clustering based ranked biomedical text summarization. *The Journal of Supercomputing*, 2023, 79: 139–159. https://doi.org/10.1007/s11227-022-04578-1

Hovy E., Lin Ch.-Y. Automated Text Summarization and the summarist system. *Proceedings of a Workshop held at Baltimore*, Baltimore, 13–15 Oct 1998. ACL, 1998, 197–214. https://doi.org/10.3115/1119089.1119121

Huang D., Cui L., Yang S., Bao G., Wang K., Xie J., Zhang Y. What have we achieved on text summarization? *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, online, 16–20 Nov 2020. ACL, 2020, 446–469. https://doi.org/10.18653/v1/2020.emnlp-main.33

Hupkes D., Dankers V., Mul M., Bruni E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 2020, 67: 757–795. https://doi.org/10.1613/jair.1.11674

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

Jalilifard A., Caridá V. F., Mansano A. F., Cristo R. S., da Fonseca F. P. C. Semantic sensitive TF-IDF to determine word relevance in documents. *Advances in Computing and Network Communications*, eds. Thampi S. M., Gelenbe E., Atiquzzaman M., Chaudhary V., Li K. C. Singapore: Springer, 2021. https://doi.org/10.1007/978-981-33-6987-0_27

Ježek K., Steinberger J. Automatic summarizing (The state-of-the-art 2007 and new challenges). *Znalosti*, 2008, 1–12.

Khan A., Salim N., Kumar Y. J. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 2015, 30: 737–747. https://doi.org/10.1016/j.asoc.2015.01.070

Khurana D., Koli A., Khatter K., Singh S. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 2023, 82: 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kutlu M., Ciğir C., Cicekli I. Generic text summarization for Turkish. *The Computer Journal*, 2010, 53(8): 1315–1323. https://doi.org/10.1093/comjnl/bxp124

Lamsiyah S., El Mahdaouy A., El Alaoui S. O., Espinasse B. A supervised method for extractive single document summarization based on sentence embeddings and neural networks. *AI2SD'2019*: Proc. Conf., Marrakech, 8–11 Jul 2019. Cham: Springer, 2020, 1105: 75–88. https://doi.org/10.1007/978-3-030-36674-2_8

Linhares Pontes E., Moreno J. G., Doucet A. Linking named entities across languages using multilingual word embeddings. *JCDL'20*: Proc. Conf., Wuhan, 1–5 Aug 2020. NY: ACL, 2020, 329–332. https://doi.org/10.1145/3383583.3398597

Lubis A. R., Nasution M. K., Sitompul O. S., Zamzami E. M. The effect of the TF-IDF algorithm in times series in forecasting word on social media. *Indonesian Journal of Electrical Engineering and Computer Science*, 2021, 22(2): 976–984. https://doi.org/10.11591/ijeecs.v22.i2.pp976-984

Maddela M., Alva-Manchego F., Xu W. Controllable text simplification with explicit paraphrasing. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, online, 6–11 Jun 2021. ACL, 2021, 3536–3553. https://doi.org/10.18653/v1/2021.naacl-main.277

Mihalcea R. Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Barcelona, 21–26 Jul 2004. Stroudsburg: ACL, 2004. https://doi.org/10.3115/1219044.1219064

Mishra A. R., Naruka M. S., Tiwari S. Extraction techniques and evaluation measures for extractive text summarisation. In: *Sustainable Computing. Transforming Industry 4.0 to Society 5.0*, eds. Awasthi S., Sanyal G., Travieso-Gonzalez C. M., Srivastava P. K., Singh D. K., Kant R. Cham: Springer, 2023, 279–290. https://doi.org/10.1007/978-3-031-13577-4_17

Mohammed Badry R., Sharaf Eldin A., Saad Elzanfally D. Text summarization within the latent semantic analysis framework: Comparative study. *International Journal of Computer Applications*, 2013, 81(11): 40–45. https://doi.org/10.5120/14060-2366

Mohan M. J., Sunitha C., Ganesh A., Jaya A. A study on ontology based abstractive summarization. *Procedia Computer Science*, 2016, 87: 32–37. https://doi.org/10.1016/j.procs.2016.05.122

Mutlu B., Sezer E. A., Ali Akcayol M. Multi-document extractive text summarization: A comparative assessment on features. *Knowledge-Based Systems*, 2019, 183. https://doi.org/10.1016/j.knosys.2019.07.019

Orăsan C., Pekar V., Hasler L. a comparison of summarisation methods based on term specificity estimation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC'04). Lisbon: ELRA, 2004, 1037–1040. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/362.pdf (3 May 2024).

Pramita Widyassari A., Rustad S., Fajar Shidik G., Noersasongko E., Syukur A., Affandy A., Rosal Ignatius Moses Setiadi D. Review of automatic text summarization techniques & methods. *Journal of King Saud University – Computer and Information Sciences*, 2022, 34(4): 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

Puduppully R. S., Jain P., Chen N., Steedman M. Multi-document summarization with centroid-based pretraining. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, 9–14 Jul 2023. ACL, 2023, 128–138. https://doi.org/10.18653/v1/2023.acl-short.13

Saadany H., Orasan C. BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. *TRITON 2021*: Proc. Conf., online, 5–7 Jul 2021. 2021, 48–56. https://doi.org/10.26615/978-954-452-071-7_006

Saggion H., Lapalme G. Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 2002, 28(4): 497–526. https://doi.org/10.1162/089120102762671963

Sharma G., Sharma D. Automatic text summarization methods: A comprehensive review. *SN Computer Science*, 2022, 4(1). https://doi.org/10.1007/s42979-022-01446-w

Shinde M., Mhatre D., Marwal G. Techniques and research in text summarization – a survey. *2021 ICACITE*: Proc. Intern. Conf., Greater Noida, 4–5 Mar 2021. IEEE, 2021, 260–263. https://doi.org/10.1109/ICACITE51222.2021.9404670

Sri S. H. B., Dutta S. R. A survey on automatic text summarization techniques. *Journal of Physics: Conference Series*, 2021, 2040(1). https://doi.org/10.1088/1742-6596/2040/1/012044

Supriyono, Wibawa A. P., Suyono, Kurniawan F. A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*, 2024, 7. https://doi.org/10.1016/j.nlp.2024.100070

Thaiprayoon S., Unger H., Kubek M. Graph and centroid-based word clustering. *NLPIR'20*: Proc. 4 Intern. Conf., Seoul, 18–20 Dec 2020. NY: ACL, 2021, 163–168. https://doi.org/10.1145/3443279.3443290

Uçkan T., Karci A. Extractive multi-document text summarization based on graph independent sets. *Egyptian Informatics Journal*, 2020, 21(3): 145–157. https://doi.org/10.1016/j.eij.2019.12.002

Wilber M., Timkey W., Van Schijndel M. To point or not to point: Understanding how abstractive summarizers paraphrase text. *Findings of ACL: ACL-IJCNLP 2021*, eds. Zong Ch., Xia F., Li W., Navigli R. Stroudsburg: ACL, 2021, 3362–3376. https://doi.org/10.18653/v1/2021.findings-acl.298

Wolhandler R., Cattan A., Ernst O., Dagan I. How "multi" is multi-document summarization? *EMNLP 2022*: Proc. Conf., Abu Dhabi, 7–11 Dec 2022. Stroudsburg: ACL, 2022, 5761–5769. https://doi.org/10.18653/v1/2022.emnlp-main.389

Xiao L., Wang L., He H., Jin Y. Copy or rewrite: Hybrid summarization with hierarchical reinforcement learning. *AAAI-20*: Proc. 34 Conf., New York, 7–12 Feb 2020. Palo Alto: AAAI Press, 2020, 34(5): 9306–9313. https://doi.org/10.1609/aaai.v34i05.6470

Yadav D., Desai J., Yadav A. K. *Automatic text summarization methods: A comprehensive review*, 2022. https://doi.org/10.48550/arXiv.2204.01849

Yadav A. K., Maurya A. K., Ranvijay R. S., Yadav R. Sh. Extractive text summarization using recent approaches: A survey. *International Information and Engineering Technology Association*, 2021, 26(1): 109–121. https://doi.org/10.18280/isi.260112

Yadav A. K., Ranvijay R. S., 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

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