DOI: https://doi.org/10.15688/jvolsu2.2024.5.4

UDC 81’322.2
LBC 81.112

Submitted: 28.03.2024
Accepted: 20.08.2024

EXPLORING AUTOMATED SUMMARIZATION: FROM EXTRACTION TO ABSTRACTION

Svetlana G. Sorokina

I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Abstract. This paper provides a review of AI-powered automated summarization models, with a focus on two principal approaches: extractive and abstractive. The study aims to evaluate the capabilities of these models in generating concise yet meaningful summaries and analyze their lexical proficiency and linguistic fluidity. The compression rates are assessed using quantitative metrics such as page, word, and character counts, while language fluency is described in terms of ability to manipulate grammar and lexical patterns without compromising meaning and content. The study draws on a selection of scientific publications across various disciplines, testing the functionality and output quality of automated summarization tools such as Summate.it, WordTune, SciSummary, Scholarcy, and OpenAI ChatGPT-4. The findings reveal that the selected models employ a hybrid strategy, integrating both extractive and abstractive techniques. Summaries produced by these tools exhibited varying degrees of completeness and accuracy, with page compression rates ranging from 50 to 95%, and character count reductions reaching up to 98%. Qualitative evaluation indicated that while the models generally captured the main ideas of the source texts, some summaries suffered from oversimplification or misplaced emphasis. Despite these limitations, automated summarization models exhibit significant potential as effective tools for both text compression and content generation, highlighting the need for continued research, particularly from the perspective of linguistic analysis. Summaries generated by AI models offer new opportunities for analyzing machine-generated language and provide valuable data for studying how algorithms process, condense, and restructure human language.

Key words: automated summarization, extractive summarization, abstractive summarization, artificial intelligence, neural networks, interdisciplinary research.

Citation. Sorokina S.G. Exploring Automated Summarization: From Extraction to Abstraction. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2024, vol. 23, no. 5, pp. 47-59. DOI: https://doi.org/10.15688/jvolsu2.2024.5.4

УДК 81’322.2
ББК 81.112

Дата поступления статьи: 28.03.2024
Дата принятия статьи: 20.08.2024

АВТОМАТИЗИРОВАННОЕ РЕЗЮМИРОВАНИЕ: ОТ МЕТОДОВ ИЗВЛЕЧЕНИЯ К АБСТРАКТНОМУ ОБОБЩЕНИЮ

Светлана Геннадьевна Сорокина

Первый Московский государственный медицинский университет им. И.М. Сеченова
(Сеченовский университет), г. Москва, Россия

Аннотация. В статье представлен обзор моделей автоматизированного резюмирования текста, основанных на технологиях искусственного интеллекта и использующих два основных подхода: экстрактивный (извлекающий) и абстрактивный (обобщающий). Цель исследования заключается в оценке компрессионных возможностей этих моделей и их языковой компетентности. Степень сжатия оценивается при помощи количественных показателей: количество страниц, слов и символов. Для оценки языковой компетентности принимается во внимание способность моделей применять разнообразные грамматические и лексические конструкции без искажения смысла и содержания. Для оценки потенциала автоматизированного резюмирования были выбраны модели OpenAI Summate.it, WordTune, SciSummary, Scholarcy и OpenAI ChatGPT-4, материалом для анализа послужили тексты публикаций по разным научным дисциплинам. Результаты позволили установить, что выбранные модели с опорой на гибридную стратегию интегрируют как экстрактивные, так и абстрактивные технологии. Тексты, созданные этими инструментами, варьировались по степени полноты и точности, при этом степень сжатия страниц составила от 50 до 95 %, а сокращение количества символов достигло 98 %. Качественная оценка показала, что, хотя модели в целом обладают способностью точно передавать основные идеи исходных текстов, некоторые резюме отличаются излишним упрощением или неверными смысловыми акцентами. Несмотря на эти ограничения, модели автоматического резюмирования обладают значительным потенциалом не только как инструменты для сжатия текста, но и как генераторы нового контента, который может стать ценным объектом для лингвистического анализа, способствуя изучению процессов машинного порождения языка и смысловой переработки текстов.

Ключевые слова: автоматизированное резюмирование, экстрактивное резюмирование, абстрактивное резюмирование, искусственный интеллект, нейронные сети, междисциплинарные исследования.

Цитирование. Сорокина С. Г. Автоматизированное резюмирование: от методов извлечения к абстрактному обобщению // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2024. – Т. 23, № 5. – С. 47–59. – (На англ. яз.). – DOI: https://doi.org/10.15688/jvolsu2.2024.5.4

Introduction

The rise of digital technologies and expansion of internet resources has led to information overload [Bawden, Robinson, 2020], posing certain challenges for researchers in managing vast data, identifying reliable information sources and making decisions [Vertinova et al., 2022]. Automated text summarization, a key technique in Natural Language Processing (NLP), addresses this issue by condensing large texts into concise, essential content helping researchers quickly assess the relevance of publications [Sorokina, 2024].

Furthermore, by focusing on key ideas and eliminating redundancies, summarization simplifies information management and improves the absorption of scientific content, ensuring critical details are retained [Sorokina, 2023].

This paper aims to review AI-driven summarization models and technologies and evaluate their effectiveness in generating concise texts without losing essential content. Summaries and source texts are compared using quantitative and qualitative measures, focusing on scientific publications due to their complexity and technical nature [Sorokina, 2016]. Summarizers must be sensitive to details and capable of discerning core information from peripheral content. Furthermore, the unique structure of scientific texts further challenges the extraction process [Sorokina, 2016]. The texts produced through the summarization process also offer valuable insights for linguistic analysis, emphasizing the need for consistency and relevance in summarization.

Overview of the main paradigms of automated summarization

A review of the current literature shows that automated text summarization can be categorized on the basis of various approaches. Depending on the number of source texts involved there is single-document summarization, which focuses on condensing a single document [Lamsiyah et al., 2020], and multi-text summarization, which aims to distill key information from a collection of documents related by common themes [Thaiprayoon, Unger, Kubek, 2021; Khan, Salim, Jaya Kumar, 2015]. Besides, summarization processes yield two primary types of summaries: extractive and abstractive.

Extractive summarization

Extractive summarization condenses a text by selecting key sentences or phrases directly from the original content without altering them significantly [Bhargava, Sharma, 2020; Collins, Augenstein, Riedel, 2017]. This method preserves the text's original style and meaning using algorithms to evaluate sentence importance based on factors like sentence length, word frequency, and keywords related to the title or main topic.

Other important markers include proper names, named entities like organizational or geographical names, toponyms reflecting basic concepts of reality, and unique terms that are essential for understanding the text content. Additionally, importance markers, linking words, i.e. phrases like *consequently*, *in conclusion*, *as a result*, *furthermore* and *thus* indicate the sentences containing key ideas or conclusions. Text formatting features such as italics, bolding, or underlining are also considered, as they can denote emphasis or key ideas. The length of sentences is another factor in assessing importance; longer sentences are often seen as more informative.

Abstractive summarization

Abstractive summarization represents a more complex aspect of natural language processing. Unlike extractive summarization, where phrases are pulled directly from the text, abstractive summarization generates entirely new content that maintains semantic consistency and syntactic coherence with the original text [Arana-Catania et al., 2021; Gupta S., Gupt S.K., 2019]. This process involves advanced neural networks that can retain key terms, the grammatical case, and the emotional neutrality typical of scientific texts while also creating novel sentence constructions not found in the original material.

Technological framework

The technological framework for automated summarization includes a variety of algorithms that are described here in simplified, non-mathematical terms to provide a general understanding of how elements from the source texts are chosen for the summarized output: centroid-based methods [Thaiprayoon, Unger, Kubek, 2021], graph-based methods [Polyakova, Zaitsev, 2022; Yadav et al., 2023], term-based methods [Orasan, Pekar, Hasler, 2004], bottom-up attention [Gehrmann, Deng, Rush, 2018], ontology-based methods [Mohan et al., 2016].

Centroid-based methods determine the importance of text elements by comparing them to a ‘centroid’, a central point that epitomizes key elements in the data set [Puduppully et al., 2023; Thaiprayoon, Unger, Kubek, 2021]. Graph-based methods visualize and analyze relationships between text elements [Belwal, Rai, Gupta, 2021; Polyakova, Zaitsev, 2022; Yadav et al., 2023]. Term-based methods focus on the term frequency (TF) and inverse document frequency (IDF), or term weights [Mishra, Naruka, Tiwari, 2023; Orasan, Pekar, Hasler, 2004]. Bottom-up attention methods prioritize keywords and proper names to grasp the document’s theme.

These methods of primarily extracting sentences characterize extractive summarization. Abstractive summarization uses ontology-based methods, employing structured knowledge representations to understand deeper semantic aspects like synonymy and polysemy.

Materials and methods

To illustrate the process of automated summarization, four articles covering a spectrum of topics were analyzed. The article *The Emergence of Convergence* (hereinafter – T1; Convergence) discusses the transdisciplinary nature of contemporary science, blending elements of social philosophy and scientific methodology. The article *The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture* (hereinafter – T2; Agriculture) delves into the advancements and challenges of precision agriculture, highlighting the innovations in sustainable farming practices. The article *Solar Energy Technology and Its Roles in Sustainable Development* (hereinafter – T3; Solar Energy) explores the applications of solar technology in promoting economic and environmental sustainability. The linguistic publication *Corpus-Based Studies of Metaphor: An Overview* (hereinafter – T4; Metaphor) focuses on the corpus-based analysis of metaphors.

To reduce volume and minimize informational noise, these texts were pre-processed. This included the removal of bibliographic lists, in-text references, footnotes, figures, and graphs. Furthermore, to avoid the influence of centroids, titles, abstracts, headings, and subheadings were also excluded.

Automated summarization tools

For the purposes of the research, four summarization models were selected, namely *WordTune*, *SciSummary*, *Scholarcy* and *OpenAI Summate.it* based on their user-friendly interfaces and proven effectiveness in handling unstructured language data. In recent years, generative pre-trained transformer models, particularly the GPT series, have gained significant attention due to their ability to process and accurately summarize large volumes of text. This paper addresses the performance and summarization potential of *OpenAI ChatGPT-4*.

Results and discussion

This section presents the results of testing the above listed summarization tools, focusing on their efficiency in text compression, accuracy and lexical proficiency.

OpenAI Summate.it

As already noted, a key expected benefit of automated summarization is the significant reduction in the source text size, measured by metrics like page, word, and character count. Among the models tested, *OpenAI Summate.it* is particularly effective requiring only a hyperlink to access to the source text. Table 1 illustrates the substantial text volume reduction achieved by this tool.

Table 1 shows that the resulting summaries typically fit on one page, though the actual compressed text often occupies less than one full page of the summary document. In practice, summaries consist of about five paragraphs, including the title, with each paragraph having one or two sentences. As a result of the intense compression, the number of words, characters, and lines reduced to 1–2% of the original volume. However, while key ideas are reflected, the summaries can be schematic and less informative.

As a matter of fact, the model combines extractive copy-and-paste methods with synonym substitutions typical of abstractive techniques. To illustrate these points, examine some specific examples of the sentences generated by *Summate.it* compared to those in the source texts.

The first example shows only minor syntactic alterations, namely the relocation of the adverbial modifier of time ‘in 2016’. (T1) ***In 2016**, the U.S. National Science Foundation launched an initiative prioritizing support for convergence research* (source text). – *The U.S. National Science Foundation (NSF) launched an initiative* ***in 2016*** *prioritizing support for convergence research* (summary). It is important to acknowledge that placing the adverbial modifier at the start of the sentence appears more appropriate than its position in the revised sentence. The present participle ‘*prioritizing*’ functions adjectivally and should immediately follow the noun ‘initiative’ it modifies, to avoid sentence semantic shifts.

Another example presents the combining of two simple sentences into a compound one without a loss of meaning: (T3) *It plays a significant role in achieving sustainable development energy solutions* and <...> *...The massive amount of solar energy attainable daily makes it a very attractive resource for generating electricity* (source text). – *Solar energy plays a substantial role in achieving sustainable development energy solutions and is a very attractive resource for generating electricity* (summary).

Furthermore, *Summate.it* proves to be aware of various grammatical constructions and the ways to handle them. Thus, in the following case a gerundial phrase of the original sentence is transformed into an attributive subordinate clause in the summary. Source text (T2) reads: *Precision agriculture is a management strategy* ***for addressing geographical and temporal***

Table 1. **Quantitative changes in source text volume using the *OpenAI Summate.it* model**

Text		Pages	Words	Characters (no spaces)	Characters (with spaces)	Paragraphs	Lines
T1	Source text	10	7,487	43,512	50,944	43	529
	Summary	1	82	573	654	5	9
Summary / Sourcetext, %		10	1.1	1.3	1.3	12	1,7
T2	Source text	14	9,366	59,068	68,382	57	738
	Summary	1	168	1,084	1,251	5	15
Summary / Sourcetext, %		7	1.8	1.8	1.8	9	2
T3	Source text	8	3,958	22,617	26,523	50	321
	Summary	1	78	482	559	5	8
Summary / Source text, %		12,5	2	2	2	10	2,5
T4	Source text	6	3,845	21,683	25,485	44	287
	Summary	1	94	548	641	5	9
Summary / Source text, %		16	2.4	2.5	2.5	11	3

variabilities in agricultural fields (source text). – *Precision agriculture is a management strategy* ***that addresses variabilities in agricultural fields*** (summary).

Following a similar pattern but in a reverse direction, another example shows the alteration of an attributive subordinate clause – *that involves data and contemporary technologies* (source text) into a gerundial phrase *by utilizing data and contemporary technologies* (summary).

Particularly noteworthy are sentences that integrate the information from several sentences. Thus, the original sentence (T2) *The integration of digital technologies into agriculture* ***has opened up*** *new opportunities and possibilities,* ***revolutionizing*** *the way farmers manage their crops, resources, and operations* (source text) gets extended and combines a few sentences into the following one: *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* (summary). Here we can note the alteration in the predicate tense form, in fact, the transformation of the participle construction into the predicate, though such changes are self-evident during the transformation process.

Similarly, sentences compile data from multiple sections of the source text into parallel syntactic structures, as in the following example: (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*** (summary).

As seen in the following examples, the obtained summaries demonstrate the model's ability to handle synonymic substitutions: (T4) *The* ***analysis*** *of metaphor-related research studies published between 2015 and 2020* ***revealed****...* (source text). – *A systematic* ***review*** *of metaphor-related research studies published between 2015 and 2020* ***found****...* (summary); (T4) <...> *...The thematic analysis* ***unearthed*** *potential gaps and under-researched areas* (source text). – *The thematic analysis* ***identified*** *gaps and under-researched areas* (summary).

WordTune

Another AI-powered writing assistant tested in the current research is *WordTune* by Israeli AI company *AI21 Labs.* The assistant helps users refine their writing, particularly benefiting non-native English speakers and professionals. *WordTune* integrates with web browsers and word processors, offering real-time suggestions to rewrite sentences, improve word choice, and adjust tone. *WordTune* also includes a text summarization feature, using natural language processing to understand context and meaning.

The *WordTune* model employs both extractive and abstractive summarization algorithms. In the resulting summaries, some sentences are directly extracted with little alteration. As shown in Table 2, there are significant reductions in various metrics. For instance, document (T1), saw a page reduction from 10 to 3, while the other three texts decreased by half.

Table 2. **Quantitative changes in source text volume using the *WordTune* model**

Text		Pages	Words	Characters (no spaces)	Characters (with spaces)	Paragraphs	Lines
T1	Source text	10	7,487	43,512	50,944	43	529
	Summary	3	1,601	9,971	11,535	37	136
Summary / Source text, %		30	21	23	23	86	25
T2	Source text	14	9,366	59,068	68,382	57	738
	Summary	7	2,650	16,897	19,460	87	279
Summary / Source text, %		50	28	29	28	**153**	38
T3	Source text	8	3,958	22,617	26,523	50	321
	Summary	4	1,209	7,017	8,184	42	121
Summary / Source text, %		50	30	31	31	84	38
T4	Source text	6	3,845	21,683	25,485	44	287
	Summary	3	1,031	5,769	6,768	32	103
Summary / Source text, %		50	27	27	27	73	37

The number of printed characters with spaces has also significantly decreased: by a factor of 4 in (T1), 3.5 in (T2), 3 in (T3), and 3.8 in (T4).

Similar reductions are observed in the metrics for "number of printed characters without spaces" and "number of lines." The relatively smaller decrease in paragraphs is due to their role in structuring content and signaling completeness. Consequently, almost every paragraph is viewed by the neural network as a source of meaningful information. Notably, in document (T2), the number of paragraphs even increases (see Table 2 **in bold**), indicating the model's ability to highlight semantic nuances.

Though some sentences, or rather phrases, of the source text go to the summary without alterations, a closer analysis of the summary qualitative indicators shows the presence of changes in syntax and semantics, as well assynonymous replacement. Let's consider a few examples. In (T1), *convergence research* ***may provide opportunities*** *to confront and navigate Arctic change* (source text) becomes *convergence research* ***can help*** *confront and navigate Arctic change* (summary). Similarly, in (T3) the phrase *the global community* ***is starting to shift towards*** *utilizing* ***sustainable energy sources*** *and reducing dependence on traditional fossil fuels as a source of energy* (source text) is summarized as *decision-makers* ***are switching to renewable energy sources*** *and reducing dependence on traditional fossil fuels* (abstract).

Regarding the way the sentence volume is reduced, the summaries produced by *WordTune* show sentence shortening with no transformation, with minor transformation and significant content simplification. Thus, an example from the original text reads *Alternative metrics, such as a potential indicator of creativity, are needed, but these may be more difficult to assess because they will be less tangible* (source text) is succinctly summarized to *Alternative metrics are needed, such as a potential measure of creativity* (summary). It is clear that the sentence is compressed by excluding some words from it, the essence of the content is efficiently maintained.

Minor syntactic transformations can be illustrated by the following example. In text (T1), we read *Problems requiring a convergent approach are problems with nonlinearity* that is transformed to *Convergent approaches are used to solve problems with nonlinearity.* The summary shows grammar changes: the active voice in the source sentence shifts to passive, and the singular noun *approach* becomes plural. There are also pragmatic meaning shifts in the summary. The original sentence focuses on *problems* with the rhema defining them, while the summary shifts to *convergent approaches* as the subject, with the rhema explaining their function. In the next example, a compound sentence is simplified into a shorter simple sentence, reducing word count while retaining the content and pragmatic meaning: (T3) *Solar cells are devices that convert sunlight directly into electricity; typical semiconductor materials are utilized to form a PV solar cell device* (source text). – *Solar cells use semiconductor materials to convert sunlight directly into electricity* (summary).

A significant content and sintactic simplification is evident in cases where complex descriptions are streamlined, as in the reduction of a lengthy discussion in text (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 proces****s* (source text). – *The problem identification process in a convergence process should be flexible and allow for moving targets* (summary). Only highlighted words of the source sentence are present in the summary, yet in a new combination.

An important issue is the summary's pragmatic inaccuracy, as it becomes too generalized and omits a key point, namely the fact that *the application of technologies* in precision agriculture can *improve performance and environmental quality*, which is the focus of article (T2).

Finally, *WordTune* effectively condenses multiple complex issues into concise summaries without apparent modifications. For instance, a lengthy exposition on the challenges facing precision agriculture is compressed into (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*** (source text). – *In the current status of precision agriculture, several issues prevent achieving efficiency, productivity, and sustainability, and escalate unintended impacts on ecosystems* (summary). The summary sentence is the result of the extactive algorythm performance.

Scholarcy

Developed by Phil Gooch, *Scholarcy* simplifies information processing in academic settings [Gooch,Warren-Jones, 2020]. Analysis shows similar quantitative text reductions as seen with the *WordTune* tool (Table 3).

Scholarcy tool belongs to the extractive summarisation model, based on the PageRank algorithm which identifies key sentences and generates structured summaries resembling sections of a scientific paper.

Note that for the purpose of the current research the article titles, subheadings, and structural indicators were removed from the source texts. Nevertheless, in each received text, the tool assigned the first sentence as the title, but this often did not align with the original title or the text's main idea.

As a matter of fact, the selected sentences neither structurally nor semantically can function as scientific article titles, whose role is to concisely reflect the content and engage readers [Sorokina, Ulanova, 2020]. These pseudo-titles shown in Table 4 alongside with the original titles, highlight and the limitations of the Scholarcy model.

Following the title, every summary features a colour-coded and italicized sentence. It resembles a news lead aiming to attract attention, that is commonplace in journalistic practices, though rare in scientific publications.

Analysis reveals that these introductory sentences are often extracted from the article's final part and fail to serve their intended purpose, appearing either meaningless or misleading. This likely occurs because the final part of a scientific text often contains the core idea, synthesised and explained [Sorokina, 2016], which the summarizer

Table 3. **Quantitative changes in source text volume using the *Scholarcy* model**

Text		Pages	Words	Characters (no spaces)	Characters (with spaces)	Paragraphs	Lines
T1	Source text	10	7,487	43,512	50,944	43	529
	Summary	3	1,779	10,437	12,166	56	150
Summary / Source text, %		30	24	24	24	130	28
T2	Source text	14	9,366	59,068	68,382	57	738
	Summary	3	1,124	7,281	8,373	38	106
Summary / Source text, %		21	12	12	12	67	14
T3	Source text	8	3,958	22,617	26,523	50	321
	Summary	2	1,002	5,652	6,620	39	92
Summary / Source text, %		25	25	25	25	78	29
T4	Source text	6	3,845	21,683	25,485	44	287
	Summary	2	1,058	5,820	6,845	39	88
Summary / Source text, %		33	28	29	27	87	31

Table 4. **Comparative analysis of the source text titles and the summary titles defined by *Scolarcy***

Texts	The title of the source text	"The title sentence" of the summary defined by *Scolarcy*
T1	The Emergence of Convergence	Science is increasingly a collaborative pursuit
T2	The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture	Precision agriculture is a management strategy for addressing geographical and temporal variabilities in agricultural fields
T3	Solar Energy Technology and Its Roles in Sustainable Development	With reference to the recommendations of the UN, the Climate Change Conference, COP26, was held in Glasgow, UK, in 2021
T4	Corpus-Based Studies of Metaphor: An Overview	The classical theorists of metaphors believed that metaphor functions as a literary device to create an artistic effect

attempts to extract, though unsuccessfully, as shown in the following examples.

(T1) Lead: *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* (summary). The phrase aligns with the first sentence of the article's conclusion following academic writing conventions, where paragraphs present a clear and logical progression [Sorokina, 2016]. Typically, each paragraph starts with a Topic Sentence that introduces the central idea or theme, sets the tone, and links to the main thesis.

The algorithms of neural models are calibrated to recognize such structural elements in scientific texts. However, since the text under analysis is written by a biological author, not bound by algorythms, the semantic weight is carried by the Concluding Sentence, which is also common in academic writing.

(T2) Lead: *This study examined the rheological properties and printing performances of edible inks made from soy protein isolate, wheat gluten, and rice protein* (summary) is absolutely misleading to the potential reader. In fact, the resource text analyses precision agriculture proposals and describes their applications: *Throughout this review, successful precision agriculture proposals and real-world implementations are analysed* (source text). The article explains how this field can continually evolve to support sustainable farming practices: *we aim to provide a comprehensive understanding of how this field can continually evolve to support sustainable farming practices* (source text).

(T3) Lead: *The Paris Climate Accords is a worldwide agreement on climate change signed in 2015, which addressed the mitigation of climate change, adaptation and finance* (summary). This lead is extracted from the introduction section of the source text and fails to convey the factual aim or significance of the entire article. The actual research purpose is stated later in the introduction: *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* (source text).

Building on the postulate that the semantic node of a paper is often encapsulated towards the end [Sorokina, 2016], a more suitable leading idea can be recognised in the final section better reflecting the publication's central ideas: *This paper highlights the significance of sustainable energy development. Solar energy would help steady energy prices and give numerous social, environmental and economic benefits* (source text). This observation again leads us to a conclusion about the limitations in the summarizing capabilities of *Scholarcy*. In fact, Scholarcy failed to identify the most important issue that the source article focuses on and instead selected an insignificant statement for the lead.

(T4) Lead: *The findings revealed that the overall mean of 3.83 research studies related to metaphor using the corpus approach per year seems low for six years* (summary) can be found in the Findings section of the original article. To a certain extent, this statement addresses one of the research questions: *What is the trend of metaphor study that uses the corpus approach in the last six years (2015–2020)?* (source text). However, the article also discusses more interesting questions such as *What are the potential gaps and under-researched areas in the analyzed literature?* (source text).

The obtained summaries featured highlighted sections like the Abstract, Scholarcy Highlights, and Scholarcy Summary, including marked parts such as the Introduction, Objectives, Results, Conclusion, and Future Work. The Scholarcy model selected full sentences, multiple sentences, or entire paragraphs without modification. Despite this, the model's structured approach ensures the summaries are both readable and effectively convey the key elements of the source text.

SciSummary

The next automated abstracting tool, *SciSummary*, employs algorithms specifically designed for generating summative abstracts. The generated abstracts not only exhibit significant compression of length (Table 5) but also demonstrate characteristics of a successful academic abstract that provides an overview of the main points, findings, and conclusions of the text.

Table 5. **Quantitative changes in source text volume using the *SciSummary* model**

Text		Pages	Words	Characters (no spaces)	Characters (with spaces)	Paragraphs	Lines
T1	Source text	10	7,487	43,512	50,944	43	529
	Summary	0.5	255	1,591	1,838	9	24
Summary / Source text, %		5	3	4	4	21	5
T2	Source text	14	9,366	59,068	68,382	57	738
	Summary	0.5	406	2,730	3,134	3	34
Summary / Source text, %		3	4	5	5	5	5
T3	Source text	8	3,958	22,617	26,523	50	321
	Summary	0.5	239	1,505	1,739	6	22
Summary / Source text, %		6	6	7	7	12	7
T4	Source text	6	3,845	21,683	25,485	44	287
	Summary	0.5	167	989	1 150	4	16
Summary / Source text, %		8	4	5	5	9	6

Note. T1 – Convergence; T2 – Agriculture; T3 – Solar Energy; T4 – Linguistics.

The examples are excerpts from each summary received:

(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...*

Even a cursory glance at these summaries detects that they follow the structure of a scientific article abstract.

OpenAI ChatGPT-4 *(Generative Pretrained Transformer)*

Today, both users and developers unanimously acknowledge that generative pre-trained transformer models, like *OpenAI ChatGPT-4,* excel at processing unstructured linguistic data and generating concise summaries. However, due to their brevity and lack of paragraph division, details on page, paragraph, and line counts are omitted (see Table 6). In terms of word and character count, *OpenAI ChatGPT-4* demonstrated the most substantial reduction among all tested AI summarization models.

These summaries successfully encapsulated the main ideas of the source texts, namely (T1) – *convergent research, social-ecological challenges, transdisciplinary approach, the Arctic, novel scientific approaches and solutions*; (T2) – *analysis of the evolution and impact of precision agriculture, its current state, challenges, and future directions, the incorporation of technologies such as IoT, AI, ML, robotics, and blockchain in precision agriculture*; (T3) – *the potential and challenges of solar energy technology, historical development of solar energy, the advancement of photovoltaic and concentrated solar power technologies, the economic, social, and environmental impacts of solar energy deployment*; (T4) – *the use of metaphors in*

Table 6. **Quantitative changes in source text volume using *OpenAI ChatGPT-4***

Text		Words	Characters (no spaces)	Characters (with spaces)
T1	Source text	7,487	43,512	50,944
	Summary	109	743	850
Summary / Source text, %		1.5	1.7	1.7
T2	Source text	9,366	59,068	68,382
	Summary	160	1 013	1172
Summary / Source text, %		1.7	1.7	1.7
T3	Source text	3,958	22,617	26,523
	Summary	127	788	914
Summary / Source text, %		3	3.4	3.4
T4	Source text	3,845	21,683	25,485
	Summary	93	559	652
Summary / Source text, %		2.4	2.5	2.5

various discourses, the importance of metaphors in language and communication, a corpus-based approach, areas that require further exploration. Furthermore, the model skillfully utilizes clichés and indirect speech verbs such as *provides an in-depth analysis*, *focuses on*, and *examines* to mirror academic rhetorical structures. Besides, when analyzing the linguistic components of these summaries, it becomes apparent that ChatGPT often generates novel word combinations that were not present in the original texts. For example, while the source text (T1) mentions *the concept of ecological resilience*, *the concept of a solution*, the summary refers to *the concept of convergence research*, introducing a new collocation; (T2) describes different types of research, such as *statistical analysis*, *image analysis*, *soil analysis*, *on-site analysis*, *real-time analysis*, the summary gives a qualitative characteristic of an in-depth analysis. Similarly, in (T2) there are the following word combinations *a comprehensive understanding*, *comprehensive frameworks*, with the descriptor *comprhensive*. In the summary text we find *a comprehensive examination.*

Additionally, the model's ability to infer central ideas and rephrase them with different words is woth commenting on. This capability highlights its proficiency in generating paraphrased content that preserves the original meaning while presenting it in a new form. Consider the following examples.

(T1) analyses the possibilities of *a convergent approach* to solving complex problems, while the summarized text indicates that the source article describes the need for *broad-based research support mechanisms*, but the adjective *broad-based* is not used in the source text. Both the Cambrige and Oxford Dictionaries agee that *convergent* means *coming closer, meeting, becoming similar*; *something that will affect many different places, activities, etc*. However, there is a certain inconsistency in the definition of the compound adjective *broad-based*. The Cambridge Dictionary states that *broad-based* is *used to describe something that will affect many different places, activities, etc.*, while the Oxford dictionary defines the adjective as *based on a wide variety of people, things or ideas; not limited.*

As for OpenAI ChatGPT-4 itself, it interpretes these adjectives in a similar way and describes convergent research as *a multidisciplinary approach that integrates knowledge, methods, and expertise from various scientific fields to address complex problems. This type of research aims to* ***bring together different disciplines*** *to converge on a* ***shared goal or solution*** (*OpenAI ChatGPT-4*). According to OpenAI ChatGPT-4, broad-based research ***encompasses a wide range of subjects, disciplines, or methods*** *to provide* ***comprehensive insights or solutions*** (*OpenAI ChatGPT-4*). Thus, both convergent research and broad-based research are considered interdisciplinary or multidisciplinary approaches. They aim to leverage insights from multiple fields to enhance the depth and breadth of understanding complex issues. These approaches recognize the value of integrating diverse perspectives to tackle broad, multifaceted problems.

The summary (T3) identifies *intermittency* as a challenge of solar energy though the specific term *intermittenc*y is not used in the source text. The neural network likely derived this conclusion from the following sentence: *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* (source text). Looking up the lexeme *intermittency* in dictionaries, we find out that both the Cambrige and Oxford Dictionaries define it as *the fact of stopping and starting repeatedly or with periods of time in between. OpenAI ChatGPT-4* shares the point and gives the following definition: *Intermittency is a noun that describes the quality or state of being intermittent – occurring at irregular intervals, not continuous or steady. It is commonly used to refer to phenomena that start and stop at intervals rather than proceeding continuously* (OpenAI ChatGPT-4). The model elaborates on the context associated with this noun, particularly in electricity supply, noting the intermittent nature of power sources like wind and solar, which depend on conditions. Given the article's key idea (T3), *intermittency* accurately reflects how solar energy is supplied. This highlights ChatGPT's skill in simplifying complex information and occasionally introducing new vocabulary, demonstrating both its strengths and challenges in AI-driven summarization.

Conclusion

In reviewing automated summarization methods that integrate advancements in computer science, artificial intelligence, and linguistics, it becomes evident that these techniques hold significant potential for processing large volumes of textual data. Algorithms that analyze and generate text, while preserving its core meaning and structure, offer a robust tool for scientific text processing. Extractive methods excel in maintaining accuracy by capturing key phrases and sentences, thereby preserving the original language and style of the text, which makes them particularly effective for factual and informational content. However, due to their reliance on formal determinants, extractive summarization algorithms might overlook subtle nuances, potentially omitting critical aspects of the text. Additionally, such automatic summarizations do not always yield coherent and logical outcomes.

Contrastingly, abstractive techniques can produce more coherent and succinct summaries that capture the essence of the source material more effectively. Yet, while these methods can enhance the clarity and brevity of summaries, they may introduce alterations that hypothetically distort the original information. Therefore, it is crucial to acknowledge and understand the limitations inherent to each model, including those related to linguistic and contextual factors. As we continue to explore these technologies, refining their capabilities will be essential for enhancing their accuracy and reliability in various applications, ensuring that they serve as invaluable assets in the realm of informational synthesis.

REFERENCES

Arana-Catania M., Procter R., He Y., Liakata M., 2021. Evaluation of Abstractive Summarisation Models with Machine Translation in Deliberative Processes. *ArXiv* (*Cornell University*). DOI: https://doi.org/10.48550/arxiv.2110.05847

Bawden D., Robinson L. 2020. Information Overload: An Overview. *Oxford Encyclopedia of Political Decision Making*. Oxford, Oxford University Press. DOI: 10.1093/acrefore/9780190228637.013.1360

Belwal R.C., Rai S., Gupta A., 2021. A New Graph-Based Extractive Text Summarization Using Keywords or Topic Modeling. *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 8975-8990. DOI: https://doi.org/10.1007/s12652-020-02591-x

Bhargava R., Sharma Y., 2020. Deep Extractive Text Summarization. *Procedia Computer Science*, no. 167, pp. 138-146. DOI: https://doi.org/10.1016/j.procs.2020.03.191

Collins E., Augenstein I., Riedel S., 2017. A Supervised Approach to Extractive Summarisation of Scientific Papers. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017).* Association for Computational Linguistics, pp. 195-205. DOI: https://doi.org/10.18653/v1/K17-1021

Gehrmann S., Deng Y., Rush A., 2018. Bottom-Up Abstractive Summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. DOI: https://doi.org/10.18653/v1/d18-1443

Gooch P., Warren-Jones E., 2020. *A Study's Got to Know Its Limitations.* DOI: 10.1101/2020.04.29.067843.

Gupta S., Gupta S.K., 2019. Abstractive Summarization: An Overview of the State of the Art. *Expert Systems with Applications*, no. 121, pp. 49-65. DOI: https://doi.org/10.1016/j.eswa.2018.12.011

Khan A., Salim N., Jaya Kumar Y., 2015. A Framework for Multi-Document Abstractive Summarization Based on Semantic Role Labelling. *Applied Soft Computing*, no. 30, pp. 737-747. DOI: https://doi.org/10.1016/j.asoc.2015.01.070

Lamsiyah S., El Mahdaouy A., El Alaoui S.O., Espinasse B., 2020. A Supervised Method for Extractive Single Document Summarization Based on Sentence Embeddings and Neural Networks. *Advances in Intelligent Systems and Computing*, vol. 1105, pp. 75-88. DOI: https://doi.org/10.1007/978-3-030-36674-2_8

Mishra A.R., Naruka M.S., Tiwari S., 2023. Extraction Techniques and Evaluation Measures for Extractive Text Summarisation. *Sustainable Computing: Transforming Industry 4.0 to Society*. Springer EBooks, pp. 279-290. DOI: https://doi.org/10.1007/978-3-031-13577-4_17

Mohan M.J., Sunitha C., Ganesh A., Jaya A., 2016. A Study on Ontology Based Abstractive Summarization. *Procedia Computer Science*, no. 87, pp. 32-37. DOI: https://doi.org/10.1016/j.procs.2016.05.122

Orasan C., Pekar V., Hasler, L., 2004. A Comparison of Summarisation Methods Based on Term Specificity Estimation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, European Language Resources Association (ELRA), pp. 1037-1040.

Polyakova I.N., Zaitsev I.O., 2022. Modification of the Graph Method for Automatic Summarization Tasks Taking into Account Synonymy. *International Journal of Open Information Technologies*, vol. 10, no. 4, pp. 45-54.

Puduppully R.S., Jain P., Chen N., Steedman M., 2023. *Multi-Document Summarization with Centroid-Based Pretraining. Edinburgh Research Explorer (University of Edinburgh)*. DOI: https://doi.org/10.18653/v1/2023.acl-short.13

Sorokina S.G., 2016. *Ispolzovaniye rekurentnosti kak sredstva argumentatsii pri postroyenii tekstov nauchnogo soderzhaniya: dis. ... kand. filol. nauk* [Use of Recurrence as a Means of Argumentation in the Construction of Texts of Scientific Content. Cand. philol. sci. diss.]. Moscow. 196 p.

Sorokina S.G., 2023. Iskusstvennyy intellekt v kontekste mezhdistsiplinarnykh issledovaniy yazyka [Artificial Intelligence in Interdisciplinary Linguistics]. *Vestnik Kemerovskogo gosudarstvennogo universiteta. Seriya: Gumanitarnye i obshchestvennye nauki* [Bulletin of Kemerovo State University. Series: Humanities and Social Science], vol. 7, no. 3, pp. 267-280. DOI: https://doi. org/10.21603/2542-1840-2023-7-3-267-280

Sorokina S.G., 2024. Osobennosti primeneniya tekhnologii avtomaticheskoy summarizatsii k nauchnym publikatsiyam [Applying Automatic Summarization Technology to Academic Publications]. *Tri «l» v paradigme sovremennogo gumanitarnogo znaniya: lingvistika, literaturovedenie, lingvodidaktika: sb. nauch. st.* [Three L's in the Paradigm of Modern Humanitarian Knowledge: Linguistics, Literary Criticism, Linguodidactics. Collection of Scientific Articles]. Moscow, Yaz. narodov mira Publ., pp. 132-138.

Sorokina S.G., Ulanova K.L., 2020. Implementatsiya kategorii tozhdestva v nazvaniyakh publitsisticheskikh i nauchnykh tekstov [Role of Article Title in Implementing the Category of Identity]. *Sovremennoe pedagogicheskoe obrazovanie* [Modern Pedagogical Education], no. 2, pp. 202-207.

Thaiprayoon S., Unger H., Kubek M., 2021. Graph and Centroid-Based Word Clustering. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pp. 163-168. DOI: https://doi.org/10.1145/3443279.3443290

Vertinova A.A., Pashuk N.R., Makogonova P.V., Kosheleva A.I., 2022. Otsenka vliyaniya informatsionnogo shuma na prinyatiye resheniy [Assessing the Infoglut Impact on Decision-Making]. *Liderstvo i menedzhment* [Leadership and Management], vol. 9, no. 3, pp. 877-890. DOI: https://doi.org/10.18334/lim.9.3.116218

Yadav A.K., Ranvijay N., Yadav R.S., Maurya A.K., 2023. Graph-Based Extractive Text Summarization Based on Single Document. *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 18987-19013. DOI: https://doi.org/10.1007/s11042-023-16199-8

SOURCES

T1 – Sundstrom S.M., Angeler D.G., Ernakovich J.G., García J., Hamm J.A., Huntington O., Allen C.R. The Emergence of Convergence. *Elementa*, 2023, vol. 11, no. 1. DOI: https://doi.org/10.1525/elementa.2022.00128

T2 – Karunathilake E.M.B.M., Le A.T., Heo S., Chung Y.S., 2023. The Path to Smart Farming: Innovations and Opportunities in Precision

Agriculture. *Agriculture*, vol. 13, no. 8, p. 1593. DOI: 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, vol. 6, no. 3, pp. 476-483. DOI: https://doi.org/10.1093/ce/zkac023

T4 – Abdul Malik N., Syafiq Ya Shak M., Mohamad F., Joharry S.A. Corpus-Based Studies of Metaphor: An Overview. *Arab World English Journal*, 2022, vol. 13, no. 2, pp. 512-528. DOI: https://doi.org/10.24093/awej/vol13no2.36

DICTIONARIES

Cambridge Dictionary. URL: https://dictionary.cambridge.org/ru/

Oxford English Dictionary. URL: https://www.oed.com/?tl=true

Information About the Author

Svetlana G. Sorokina, Candidate of Sciences (Philology), Associate Professor, Institute of Linguistics and Intercultural Communication, I.M. Sechenov First Moscow State Medical University, Trubetskaya St, 8, Bld. 2, 119048 Moscow, Russia, lana40ina@mail.ru, https://orcid.org/0000-0002-8667-6743

Информация об авторе

Светлана Геннадьевна Сорокина, кандидат филологических наук, доцент Института лингвистики и межкультурной коммуникации, Первый Московский государственный медицинский университет им. И.М. Сеченова (Сеченовский университет), ул. Трубецкая, 8, стр. 2, 119048 г. Москва, Россия, lana40ina@mail.ru, https://orcid.org/0000-0002-8667-6743