УДК 004.89:02+025.5:004.8
https://doi.org/10.33186/1027-3689-2023-10-99-120

Анализ возможностей автоматического реферирования статей на примере источников базы данных «Экология: наука и технологии» ГПНТБ России

Е. Ф. Бычкова[1], К. А. Колосов[2]

[1, 2]ГПНТБ России, Москва, Российская Федерация

[1]bef@gpntb.ru
[2]kolosov@gpntb.ru

Аннотация. Рассмотрена возможность автоматического реферирования публикаций с использованием моделей генерирующего реферирования. Приводится обзор подходов к автоматическому регулированию, в том числе с использованием нейронных сетей. Дан обзор распространённых программных сред, а также проведён анализ, в результате которого определены их преимущества и недостатки при автореферировании. Проблема создания рефератов статей с использованием технологий автореферирования актуальна и позволяет увеличить доступность публикаций, прежде всего, не представленных в открытом доступе, при снижении затрат на их библиографическую обработку. По мнению авторов, создание расширенной библиографической записи (БЗ), снабжённой аннотацией или рефератом, очень важно при предоставлении информации о новых экологичных технологиях. Тогда как создание грамотного реферата требует не только интеллектуальных усилий и времени сотрудников, но и специальных знаний. В качестве объекта исследования была взята база данных «Экология: наука и технологии» в целом и содержащиеся в ней публикации, посвящённые внедрению новых природо- и ресурсосберегающих технологий. Сделан вывод о том, что автореферирование, в отличие от ручного реферирования, не требует наличия специалистов высокой квалификации по тематике обрабатываемых документов. При этом качество формируемых рефератов получается достаточно высоким даже при использовании типовых наборов данных (датасетов).

Статья подготовлена в рамках Государственного задания ГПНТБ России от 29.12.2022 № 075-01235-23-00 по выполнению работы «Развитие электронного библиотековедения как научной и учебной дисциплины в условиях

трансформации библиотечных фондов, справочно-библиографического и документного обслуживания в цифровой среде (FNEG-2022-0004)», реестровый № 720000Ф.99.1БН60АВ03000 на 2022–2024 гг.

Ключевые слова: экология, реферат, база данных, нормирование труда, автоматическое реферирование, нейросетевой машинный перевод, генерирующая модель mBARTru, датасет

Для цитирования: Бычкова Е. Ф., Колосов К. А. Анализ возможностей автоматического реферирования статей на примере источников базы данных «Экология: наука и технологии» ГПНТБ России // Научные и технические библиотеки. 2023. № 10. С. 99–120. https://doi.org/10.33186/1027-3689-2023-10-99-120

UDC 004.89:02+025.5:004.8
https://doi.org/10.33186/1027-3689-2023-10-99-120

Analyzing the prospects for computerized article abstracting as a case study of RNPLS&T's database "Ecology: Science and technology"

Elena F. Bychkova[1] and Kirill A. Kolosov[2]

[1, 2]Russian National Public Library for Science and Technology, Moscow, Russian Federation

[1]bef@gpntb.ru
[2]kolosov@gpntb.ru

Abstract. The authors examine the possibility of computerized abstracting of publications based on generative abstracting models. They review the approaches toward automatic control including that of neural networks, characterize popular software environments, and discuss their advantages and disadvantages for computerized abstracting. The problems under discussion are relevant as the technology of computerized abstracting increases the accessibility of publications, in particular, those out of the open access while decreases bibliographic processing costs. The authors insist that augmented bibliographic record supplied with anno-

tation or abstract is very important for providing information on new environmental technologies. At the same time, the appropriate abstract consumes intellectual efforts and time of competent professionals. The database "Ecology: Science and Technologies" was chosen as the study object; the database comprises publications on implementation of new environmental-friendly and resource-saving technologies. The authors conclude that computerized abstracting as opposed to similar manual process requires no specialists highly qualified in the discipline of the document being processed while the quality of the abstracts is rather high even when the standard datasets are used.

The article is prepared within the framework of the Government Order to RNPL&T No. 075-01235-23-00 of December 29, 2022 "Developing e-librarianship as a scientific and academic discipline the circumstances of transforming library collections, reference bibliographic and document services in the digital environment" (FNEG-2022-0004), Register No. 720000F.99.1BN60AV03000 for the years 2022–2024.

Keywords: ecology, abstract, database, work standardization, computerized abstracting, neural machine translation, mBARTru generative model, dataset

Cite: Bychkova E. F., Kolosov K. A. Analyzing the prospects for computerized article abstracting as a case study of RNPLS&T's database "Ecology: Science and technology" // Scientific and technical libraries. 2023. No. 10. P. 99–120. https://doi.org/10.33186/1027-3689-2023-10-99-120

Введение

Библиографическая деятельность библиотеки направлена на раскрытие информационных ресурсов для полноценного удовлетворения информационных потребностей пользователей [**1**]. Особое значение в системе сведений о текстовом документе имеют аннотация и реферат. Различие между этими жанрами состоит в том, что «реферат включает краткое, максимально свёрнутое изложение содержания публикации, а аннотация – краткую её характеристику» [**2**]. В требованиях к публикациям в современных научных журналах, сборникам научных статей и материалам конференций оговаривается, что авторы должны сами составлять аннотации к публикуемым материалам. Реферат призван в лаконичной форме ответить на вопрос «Что именно сообщается

в первичном документе?», в отличие от аннотации, отвечающей, как правило, на вопрос «О чём сообщается в первичном документе?». Соответственно, ведущим свойством реферата является информативность – способность кратко передать смысл первичного документа в отличие от свойства индикативности (указательности), которым в большей мере обладают аннотации [**3**]. Написание реферата – сложный диалектико-логический процесс, составляющими которого являются отбор и описание, оценка, систематизация и обобщение фактографической информации. По некоторым оценкам, если просмотр ограничен только библиографическими описаниями, то он в 30–50% случаев приводит к ошибочному решению: читатель либо не находит нужный ему документ, либо тратит время на просмотр ненужной информации. Наличие реферата снижает долю ошибочных решений до 8–12% [Там же].

В настоящее время в библиотеках намечаются перспективные тренды, влияние которых будет только нарастать. В пленарном докладе на конференции «LIBCOM–2022» научный руководитель ГПНТБ России, профессор Я. Л. Шрайберг отметил, что главным проявлением цифровой трансформации, помимо массовой генерации и использования оцифрованных продуктов, является четвёртая промышленная революция и активно развивающаяся система искусственного интеллекта (ИИ), включая нейронные сети [**4**]. Одно из важных направлений использования технологий ИИ в библиотеках – автоматическое реферирование, прежде всего текстов научных публикаций.

База данных «Экология: наука и технологии» в ГПНТБ России

База данных (БД) ведётся в ГПНТБ России с 1998 г., с мая 2003 г. БД носит название: «Экология: наука и технологии» и зарегистрирована в ФГБУ НТЦ «Информрегистр», в 2020 г. получено свидетельство о государственной регистрации БД в федеральной службе интеллектуальной собственности.

БД обеспечивает доступность информации по экологической тематике для читателей научно-технической библиотеки. Профиль библиотеки определяет содержание БД. Порядок формирования и особенности БД неоднократно анализировались и рассматривались

в публикациях сотрудников ГПНТБ России, были представлены на конференциях и в публикациях в профессиональной прессе [**5** – **8**].

Источниками формирования БД «Экология: наука и технологии» являются книги и статьи из всех периодических и продолжающихся изданий, которыми комплектуется фонд ГПНТБ России. В соответствии с отчётом ГПНТБ России за 2022 г. [**9**] всего в библиотеку поступило (включая дары) и, соответственно, было просмотрено для БД 2 118 названий отечественных журналов (1 924 в 2021 г.). Всего же с 2002 г. по март 2023 г. включительно для БД были отобраны статьи из 1 279 журналов.

В табл. 1 приведён список журналов, публикации из которых наиболее значимы для наполнения БД.

Таблица 1

Источники формирования БД

Периодическое издание	Количество статей, отражённых в БД	Процент от общего числа библиографических записей в БД
«Экология и промышленность России. ЭКиП»	2 024	5.0
«Экология производства»	1 471	4.0
«Экологические системы и приборы»	791	2.0
«Твёрдые бытовые отходы»	789	2.0
«Защита окружающей среды в нефтегазовом комплексе»	765	2.0
«Безопасность жизнедеятельности»	752	2.0
«Горный информационно-аналитический бюллетень»	646	1.0
«Проблемы региональной экологии»	518	1.0
«Успехи современного естествознания»	499	1.0
«Экология промышленного производства»	460	1.0
«Вода: химия и экология»	458	1.0
«Водные ресурсы»	457	1.0

Окончание таблицы 1

Периодическое издание	Количество статей, отражённых в БД	Процент от общего числа библиографических записей в БД
«Водоснабжение и санитарная техника: ВСТ»	435	1.0
«Экология урбанизированных территорий»	420	1.0
«Теоретическая и прикладная экология»	420	1.0
«География и природные ресурсы»	378	1.0
«Проблемы региональной экологии»	373	1.0
«Использование и охрана природных ресурсов в России»	366	1.0

БД привлекает внимание читателей библиотеки к новым инновационным технологиям, позволяющим оптимизировать процесс производства и снизить ущерб для окружающей среды.

Задачи и цели исследования

Задачи данного исследования авторы видят в следующем:

оценить наполнение БД за условный период с точки зрения соответствия публикаций её тематике;

определить примерное количество статей, в которых рассматриваются новые технологии и возможности их внедрения в производственных условиях;

оценить возможность доступа к полным текстам отобранных публикаций из сети интернет и, как следствие, обосновать целесообразность их реферирования в случае ограниченного доступа;

рассмотреть возможности библиотеки по реферированию статей с использованием технологии автоматического реферирования.

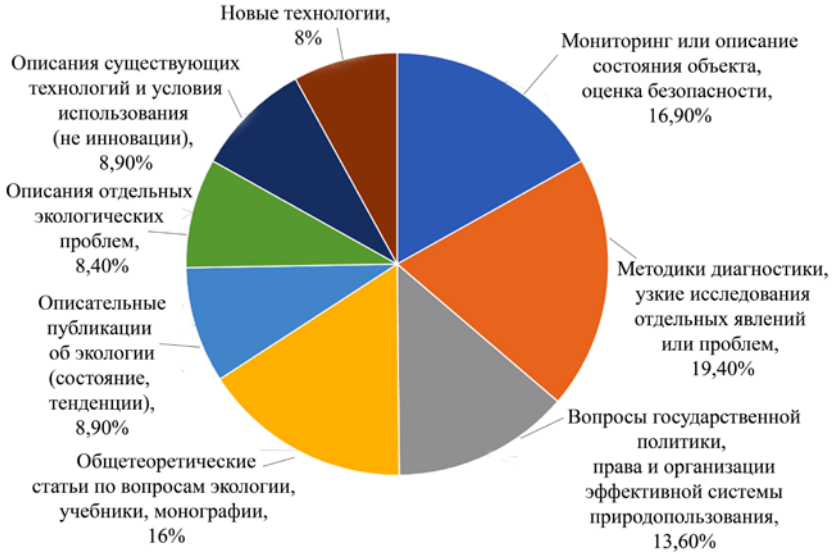
Цель данного исследования – обоснование применения технологии автореферирования актуальных тематических статей в работе библиотеки на примере публикаций в БД «Экология: наука и технологии».

В качестве объекта исследования рассматривается часть БД: 1 000 библиографических записей, последовательно введённых за период с октября 2022 г. по январь 2023 г.

Результаты анализа БД

Результат анализа показал, что публикации, отражающие общие вопросы экологии, как-то: теоретические вопросы, описания экологических проблем или явлений и т. п., составляют 33,3% от общего числа библиографических записей. К публикациям, отражающим фактические данные: методики диагностики, узкие исследования отдельных явлений или проблем, мониторинг или описание состояния объекта, а также оценку состояния объекта, можно отнести 36,3% библиографических записей. Вопросы государственной политики, права и организации эффективной системы природопользования освещаются примерно в 13,6% изданий, включённых в БД. И, наконец, описания новых природосберегающих технологий, уже внедрённых в производство или только разработанных и находящихся на стадии внедрения (в том числе включающих оценку их эффективности), составляют соответственно 8,9% и 8%.

Графически эти данные представлены на рисунке.



Тематика публикаций в БД за анализируемый период

Если теоретические вопросы и описания экологических проблем важны при изучении экологии как науки, для экологического просвещения и образования, а проблемы мониторинга и диагностики состояния объектов и систем – для безопасности жизни, то внедрение и использование природо-, энерго-, ресурсосберегающих и др. инновационных технологий необходимы для развития так называемой «зелёной экономики». Именно эти технологии нуждаются в активной популяризации, и описания их должны быть доступными широкому кругу читателей.

В фокусе внимания данного исследования находятся прежде всего статьи, в которых рассмотрены **инновационные технологические процессы**. Из 1 000 рассмотренных записей была выделена 81 такая публикация. Выборка делалась на основе анализа заглавий, ключевых слов и аннотаций.

Дальнейший анализ полных текстов отобранных публикаций показал очевидное: выборка, сделанная из достаточно подробных библиографических записей, включающих аннотации, не всегда корректна. Часть статей не содержала описания инновационных технологий и способов их внедрения, в иных случаях описание новых технологий не было отражено в библиографической записи.

На следующем этапе исследования было выявлено, что значительная часть отобранных статей (53 из 81, то есть около 65%) представлены в открытом доступе. Для достижения поставленной цели – продвижения тематических публикаций – важно дополнить библиографические записи в БД «Экология: наука и технологии» ссылками на полные тексты в интернете.

Относительно публикаций, полные тексты которых недоступны, сделан вывод о необходимости создания реферата статьи, подробно раскрывающего содержание и значение внедрения описываемой технологии.

Написание реферата – трудоёмкая задача, требующая понимания тематики статьи. По сути это научное изучение содержания первоисточника информации, представляющее собой довольно сложный творческий процесс. Задача получения реферата, семантически адекватного первоисточнику, реализуется при условии научно-информационного анализа первоисточника, то есть путём научного отбора и характеристики только новой (ценной и полезной) информации [**3**].

Можно ли оптимизировать этот труд, сократить количество возможных ошибок, связанных с субъективным или непрофессиональным подходом к содержанию статьи? Перспективным вариантом для решения этой проблемы может стать обращение к программам автореферирования статей. Эта категория программ на основе технологий ИИ быстро развивается.

Автоматическое реферирование информации

Системы автоматического реферирования информации позволяют уменьшить время на изучение первоисточника при составлении аннотаций и рефератов. В настоящее время востребованность таких систем возрастает, особенно в связи с широким распространением новых решений, использующих нейронные сети.

Пока не существует общепринятого эффективного способа автоматической оценки систем автореферирования, поэтому результаты ручного и автоматического составления рефератов сопоставляются на основании экспертных оценок [**10**].

По способу построения текста методы автоматического реферирования делятся на две группы: извлекающие (квазиреферирование, Sentence extraction) и генерирующие (генерация реферата с порождением нового текста, Abstraction). Извлекающие методы выделяют фрагменты из текста в том порядке и виде, в которых они приведены в исходном документе. Генерирующие методы предполагают наличие лингвистической БД, с использованием которой генерируется новый текст, не представленный явно в исходном документе. Извлекающие методы в литературе также называют поверхностными, а генерирующие – глубинными. Как отмечается в [**11**], некоторые авторы выделяют пять различных подходов к автореферированию: статистический, когерентный, алгебраический, графовый, а также подход, основанный на машинном обучении.

Генерирующие модели реферирования с использованием нейронных сетей

Генерирующие модели реферирования, в отличие от извлекающих, позволяют создавать новые тексты, используя редактирование предложений, синонимы, обобщения, что делает результаты их работы

более интересными с практической точки зрения. Эти методы широко используют технологию нейросетевого машинного перевода [**12**].

Наиболее широко упоминаемыми моделями такого типа является семейство моделей *GPT (Generative pre-trained transformer) (GPT, GPT-2, GPT-3)* [**13**]. У широкой публики аббревиатура *GPT* ассоциируется с чатботом *ChatGPT* [**14**], который позиционируется в качестве одной из первых моделей ИИ. В основе работы моделей *GPT* лежит предварительное обучение трансформера-декодировщика языковому моделированию. Задача заключается в предсказании текста слева направо, используется предварительное обучение на огромном обработанном текстовом массиве информации.

Для обучения модели *GPT-3* использовался набор данных из более 570 Гб текстов, включающий данные проекта Common Crawl, английскую Википедию, два датасета (набора данных) с книгами и датасет WebText2 с текстами веб-страниц. Лишь 0,11% документов, входящих в набор данных для обучения, были на русском языке. Обучение модели происходило на суперкомпьютере Microsoft Azure AI, специально построенном для компании OpenAI. На обучение, по некоторым оценкам, могло уйти 4,6 млн долларов США [**15**].

Для русского языка существует адаптация *GPT-3* от «Сбера». За последние несколько лет был обучен и выложен в открытый доступ ряд русскоязычных и мультиязычных генеративных моделей. Это *ruGPT-3, ruT5, mGPT, FRED-T5* и др. Как отмечено в [**16**], общий размер датасета около 300 Гб – это Википедия, книги, новости на русском и английском языках, разговорная речь, научные статьи и т. д. Обучение модели заняло около полутора месяцев.

Ещё одной моделью, используемой для генерирующего реферирования, является *BART* [**17**]. Данная модель предобучается реконструкции испорченного зашумлённого текста, причём сразу на генерации текста, и поэтому лучше подходит для автоматического реферирования. Кроме английской версии *BART*, была обучена ещё и многоязычная версия этой модели, *mBART*. Она обучалась на подмножестве Common Crawl из 25 языков, в котором русский язык является вторым по степени представленности после английского [**12**].

Выбор технологии и программной среды для проведения исследования

Google Colab (https://colab.research.google.com) – это инструмент, позволяющий писать, запускать и публиковать код Python просто в браузере. *Google Colab* также имеет множество функций, которые делают его популярным инструментом для анализа данных, машинного обучения и ИИ. Многие учебники по машинному обучению, которые можно найти в интернете, написаны в *Google Colab*. Программным кодом, созданным в *Google Colab*, можно поделиться, используя гиперссылки. Любой пользователь может запустить этот код в своём собственном браузере без какой-либо настройки.

В *Google Colab* есть возможность использования трёх типов сред выполнения, к которым можно подключиться [**18**]. Первая, используемая по умолчанию, – это среда выполнения «Центральный процессор» (ЦП). Это лучший вариант для работы с Python или с небольшой моделью машинного обучения.

Второй тип среды выполнения использует графические процессоры. Хотя они изначально были разработаны для повышения производительности видеоигр, впоследствии стали стандартным способом запуска кода машинного обучения благодаря своей эффективности и результативности.

Третий тип – это Tensor Processing Unit (TPU), использующий микросхемы обработки данных разработки Google с целью значительного ускорения кода машинного обучения.

Анализ отдельных моделей генерирующего реферирования

Выбор модели генерирующего реферирования осуществлялся из моделей, представленных в свободном доступе, описание которых приведено в статье [**12**]:

mBARTru,
T5-baseru,
GPT3-medium,
mT5-base.

В качестве тестового текста научной публикации использовался следующий фрагмент:

«На основании результатов исследований выявлено увеличение производительности разделения модельной эмульсии АЦ мембранами, обработанными при напряжении коронного разряда = 5 и 25 кВ, по сравнению с исходными фильтр-элементами. Пиковая производительность разделения эмульсии у модифицированных мембран увеличивается до 5 раз, а эффективность процесса разделения составила 99,22% с содержанием УВ в фильтрате 627,3 мг/дм при напряжении 25 кВ и продолжительности т = 5 мин времени обработки коронным разрядом. Обработка коронным разрядом способствует образованию потенциала поверхности и напряжённости электрического поля мембран, снижению шероховатости и увеличению их гидрофильности. Таким образом, показана возможность уменьшения загрязнения окружающей среды при разделении ВНЭ пластовой воды в результате обработки АЦ мембран коронным разрядом с интенсификацией производительности на 51% и эффективностью более 99%».

Полученные результаты автоматического реферирования приведены в табл. 2.

Таблица 2

Результаты автоматического реферирования, полученные для некоторых моделей генерирующего реферирования

Тип модели	Результат автоматического реферирования
mBARTru	Снижение производительности разделения модельной эмульсии АЦ мембранами, обработанными при напряжении коронного разряда = 5 и 25 кВ, по сравнению с исходными фильтр-элементами. Таким образом, показана возможность уменьшения загрязнения окружающей среды при разделении ВНЭ пластовой воды в результате обработки мембран коронным разрядом с интенсификацией производительности на 51% и эффективностью более 99%.
T5-baseru	У модифицированных мембран увеличивается производительность разделения эмульсии АЦ.
GPT3-medium	Фильтр-эмульсия АЦ, обработанная при напряжении 5 кВ и длительности т = 25 мин, показала увеличение производительности разделения на 51% по сравнению с эмульсией, обработанной при напряжении 10 кВ.
mT5-base	Обработка коронным разрядом пластовой воды может увеличить эффективность процесса разделения эмульсии. Об этом говорится в новом исследовании, опубликованном в журнале Nature.

Смысловой анализ сгенерированных аннотаций показывает, что наиболее интересные результаты формируются на модели *mBARTru*, представляющей собой генерирующую модель реферирования многоязычной версии *BART*, в которой русский язык является вторым по степени представленности. Эта модель была выбрана в качестве базовой для проведения исследования. Приведённая оценка качества результатов работы моделей является субъективной, поскольку модели постоянно совершенствуются, а процесс их обучения продолжается. Кроме того, следует учитывать, что модели можно обучать на различных наборах данных (датасетах), а используемые общедоступные версии базируются не на массиве научных текстов, а на текстах общего характера.

Технологические этапы и оценка времени обработки статей с использованием технологии автоматического реферирования

Как было отмечено в целях исследования, для автоматического реферирования отбирались статьи, полные тексты которых недоступны в сети интернет, или доступ к которым затруднён для индексирования поисковыми машинами. В этом случае технологический процесс включал следующие этапы:

сканирование и распознавание текста;

формирование реферата статьи с использованием модели генерирующего реферирования;

редактирование полученного текста с добавлением, в случае необходимости, авторских таблиц и рисунков.

Для обоснования целесообразности использования технологии автореферирования в библиотечной практике имеет смысл оценить трудозатраты каждого технологического этапа, чтобы сопоставить их с трудозатратами сотрудников при ручной обработке. Для этой цели в качестве объекта исследования была выбрана статья [**19**] объёмом три страницы.

На первом этапе было произведено сканирование на простом офисном сканере, с сохранением документа в формате PDF. При этом важно было обеспечить сохранность журнала при сканировании, с одной стороны, и получить чёткое изображение полного разворота страницы, необходимое для корректного распознавания текста, – с другой.

Поэтому подготовка каждой страницы (разворот журнала, обрезка и т. п.) потребовала дополнительного времени. По нашим оценкам, сканирование одной страницы заняло от 1,5 до 2,5 минут. Всего на сканирование статьи из трёх страниц на простом офисном сканере потребовалось от 6 до 7,5 минут. Дальнейшее распознавание отсканированного документа с целью получения текстового файла осуществлялось с помощью программного пакета Content Reader PDF 15 Standard. Обработка трёхстраничной статьи заняла около 5 минут, включая временные затраты на обрезку рисунков, таблиц и списка литературы.

Формирование реферата статьи с использованием модели генерирующего реферирования проводилось путём подстановки последовательных фрагментов распознанного файла документа в строку входного текста модели *mBARTru* на платформе Google Colab. Первоначально мы пробовали загружать документ целиком, без разбивки на части, но результирующий реферат получался хотя и неплохим по качеству, но слишком коротким, что связано с ограничениями используемой модели. Наилучшие результаты достигались при разбивке исходного документа на фрагменты в соответствии со структурой документа, например: «введение», «актуальность», «описание технологии», «выводы». Обработка одного фрагмента после загрузки среды обработки занимала не более 30 секунд, включая время на копирование исходного текста и результатов. Среда обработки загружается один раз в начале работы модели на платформе Google Colab и занимает от 1 до 2 минут.

Получаемый в результате обработки реферат требует прочтения библиотекарем для сопоставления с исходным документом, так как из-за стилевых особенностей текста документа возможно появление грубых смысловых ошибок.

Например, в качестве исходного фрагмента был использован следующий текст:

«Проблема очистки сточных вод и уменьшения количества отходов по-прежнему актуальна для многих отраслей промышленности. В связи с этим особое значение приобретают вопросы как очистки сточных вод, так и уменьшения количества отходов, в том числе за счёт использования их в качестве различных сорбентов. В данной статье приведена сравнительная оценка эффективности очистки сточ-

ных вод от ионов железа, меди и никеля сорбентом, предложенным нами ранее. Сорбент получен из шлама с участка водоподготовки теплоэлектростанций и после соответствующей термической обработки позволяет удалять ионы тяжёлых металлов из промышленных сточных вод».

Автоматическое реферирование сформировало следующий результат:

«Российский сорбент ~~от ионов железа, меди и никеля~~ позволяет удалять ионы тяжёлых металлов из промышленных сточных вод. В связи с этим особое значение приобретают вопросы как очистки сточных вод, так и уменьшения количества отходов, в том числе за счёт использования их в качестве различных сорбентов».

После сравнения с исходным текстом библиограф удалил в полученном реферате слова, противоречащие смысловому содержанию документа (в примере выше – зачёркнуты).

Время обработки печатных версий научных статей с использованием технологии автоматического реферирования складывается из времени на сканирование документа, распознавание текста, формирование отдельных частей реферата генерирующей моделью, смысловой анализ фрагментов и формирование сводного реферата. В зависимости от размера публикации и сложности компоновки печатной версии (рисунки, графики, таблицы и т. п.) суммарное время формирования реферата с использованием модели генерирующего реферирования составляло от 20 до 45 минут.

Выводы по качеству
полученных результатов реферирования

В процессе исследования было обработано 10 научных статей из БД «Экология: наука и технологии» по тематике инновационных технологий и способов их внедрения. Пример реферата, полученного с использованием технологии автоматического реферирования и реферата, составленного библиографом, указан в расширенных данных к статье (https://disk.yandex.ru/i/29oT3-v1WMbn7Q). Приведём общие выводы, сделанные нами по результатам сравнительного анализа полученных рефератов:

Автореферирование сформировало связный текст объёмом 2 420 знаков без пробелов, в целом отражающий содержание статьи.

Реферат, составленный библиографом, содержит текст объёмом 2 548 знаков без пробелов.

Преимущества текста, полученного с помощью автореферирования:

требуемый для реферата объём;

описывает содержание технологического процесса;

отсутствуют неправильные согласования или бессмысленные предложения.

К недостаткам текста, полученного автореферированием, можно отнести:

не всегда расшифрованы аббревиатуры, использованные в статье;

необходимо минимальное редактирование;

целесообразно добавление в реферат иллюстраций или таблиц, хотя это вопрос спорный (в аннотации, сделанной традиционным способом, таблицы и рисунки также не приведены).

Нормы трудозатрат при ручном реферировании научных статей

Для обоснования целесообразности внедрения и использования технологии автореферирования имеет смысл рассмотреть нормы труда на подготовку аннотаций и рефератов статей в библиотеках и сравнить их с реальными трудозатратами сотрудников.

Нормы труда для библиотекарей устанавливаются в соответствии с Приложением к приказу Министерства культуры РФ от 30 декабря 2014 г. № 2477 [**20**].

На основании этого документа библиотеками были подготовлены соответствующие методические рекомендации, например издания [**21**] и [**22**].

В соответствии с представленными в вышеперечисленных документах нормами составление аннотации, изучение документа, написание текста составляют 20 минут для книги и 20 минут для статьи [**21**]. В соответствии с Приложением в материалах РГБ на аннотирование книги отводится от 270 минут (раздел 4.8.3. Информационная работа) до 120 минут (раздел 3.15. Справочная и информационная работа.

Организация справочно-библиографического аппарата (СБА)) и, соответственно, от 160 до 60 минут (те же разделы) для статей.

Подготовка реферата, включающего изучение и анализ документа, составление текста, компьютерный набор и внесение исправлений, составляет 6 600 мин (110 часов) (раздел 3.17. Справочная и информационная работа. Библиографическое информирование) или 5 920 минут (около 98,7 часа) (раздел 4.8.3. Информационная работа) на один авторский лист (40 тыс. печатных знаков с пробелами).

ГОСТом не оговаривается размер реферата, но практический опыт его написания для научной статьи по тематике комплектования БД «Экология: наука и технологии» составляет от 1 000 до 4 500 знаков в зависимости от её объёма и сложности.

Таким образом, в соответствии с нормативами, составление реферата к статье составляет более 148 минут (2 часа 28 минут для реферата объёмом 1 тыс. знаков).

Заключение

Развитие моделей генерирующего реферирования, использующих технологии нейронных сетей, открывает для библиотек новые возможности по подготовке рефератов научных статей сложной научно-технической тематики. Использование технологий автореферирования не требует наличия специалистов высокой квалификации по тематике обрабатываемых документов. При этом качество формируемых рефератов получается достаточно высоким даже при использовании типовых наборов данных (датасетов). Поскольку технологии нейронных сетей включают обучающие модули и возможности пополнения датасетов, это открывает широкие возможности для дальнейшего повышения качества результатов автореферирования.

В проведённом нами исследовании применялись общедоступные технические и программные средства. Для использования программы распознавания текстов Content Reader требуется оплата лицензии, а программная среда Google Colab снизила производительность после нескольких десятков обращений, предложив перейти в платный режим. Для эффективного использования технологии автореферирования целесообразно выделить отдельный физический сервер высокой производительности с большим объёмом памяти, так как это позволит избе-

жать затрат на сторонние сетевые ресурсы. Программное обеспечение моделей и наборы данных доступны через интернет.

Время на составление реферата научной статьи библиографом может различаться в зависимости от сложности и объёма обрабатываемого документа, но в среднем его значение превышает время автоматической обработки.

В контексте всех вышеперечисленных расчётов создание реферата с использованием специальных программ автоматического реферирования представляется целесообразным, а затраты времени на сканирование, распознавание и обработку текста не превышают существующих норм ручного реферирования статей.

Расширенные данные к статье: https://disk.yandex.ru/i/29oT3-v1WMbn7Q

Список источников

1. **Саломатова О. И.**, **Зеленина Г. Н.** Возможности проекта МАРС (Межрегиональная аналитическая роспись статей) в информационно-библиографической работе библиотеки // Библиотеки вузов Урала: проблемы и опыт работы. 2003. Вып. 4. 2003.

2. **Липницкий С. Ф.**, **Мамчич А. А.**, **Сорудейкина С. А.** Веб-поиск и аннотирование научно-технической информации на основе тематических корпусов текстов // Информатика. 2018. Т. 1. № 2 (22). С. 114–125.

3. **Мукамбетова Г. И.** Аннотирование и реферирование документов: проблемы изучения при подготовке специалистов для библиотек // Вестник Бишкекского гуманитарного университета. 2009. № 2. С. 242–244.

4. **Шрайберг Я. Л.** Библиотечно-информационная сфера в современных условиях нарастающей цифровизации, постпандемийной обстановки и новых социально-политических реалий: главные результаты : пленарный доклад Председателя Оргкомитета Двадцать шестой Международной конференции и выставки «LIBCOM–2022». Москва : ГПНТБ России, 2022. 27 с. : ил. Библиогр.: с. 26–27 (17 назв.). 350 экз. ISBN 978-5-85638-253-1. doi: 10.33186/978-5-85638-253-1-2022

5. **Соловьёва Л. С.** База данных «ЭКО»: особенности формирования и обслуживания пользователей // Научные и технические библиотеки. 2003. № 4. С. 51–53. URL: http://ellib.gpntb.ru/subscribe/index.php?journal=ntb&year=2003&num=4&art=8 (дата обращения: 11.11.2019).

6. **Бычкова Е. Ф.** Реферативная БД «Экология: наука и технологии» – важная часть Электронной библиотеки ГПНТБ России по экологии // Научные и технические библиотеки. 2008. № 2. С. 77–84.
URL: http://ellib.gpntb.ru/subscribe/index.php?journal=ntb&year=2008& num=2&art=13 (дата обращения: 27.01.2020).

7. **Боргоякова К. С.** Библиометрический анализ научных публикаций по экологии на основе реферативной базы данных «Экология: наука и технологии» ГПНТБ России // Научные и технические библиотеки. 2017. № 10. С. 54–68.
URL: https://www.gpntb.ru/ntb/ntb/2017/10/NTB10_2017_A5_6.pdf
(дата обращения: 20.12.2019).

8. **Бычкова Е. Ф.** Отражение публикаций по теме аварии на Чернобыльской АЭС и смежным с ней вопросам в БД ГПНТБ России «Экология: наука и технологии» // Чернобыль 35 лет спустя : материалы Межгосударственной научно-практической конференции (22 апреля 2021 г.). Брянск, 2021. С. 30–37.

9. **Краткий** отчёт о деятельности ГПНТБ России за 2021 год. URL: https://www.gpntb.ru/ofitsialnye-dokumenty/84--12/ofitsialnye-dokumenty/9672-kratkij-otchet-o-deyatelnosti-gpntb-rossii-za-2022-god.html (дата обращения: 07.09.2023). URL: свободный.

10. **Das D., Martins A. A.** Survey on Automatic Text Summarization : Technical report // Literature Survey for the Language and Statistics II course at Carnegie Mellon University. Pittsburgh, US, 2007. P. 192–195.

11. **Батура Т. В., Бакиева А. М.** Методы и системы автоматического реферирования текстов. Новосибирск : ИПЦ НГУ, 2019.

12. **Секреты** генерирующего реферирования текстов.
URL: https://habr.com/ru/articles/596481 (дата обращения: 12.05.2023).

13. **ChatGPT.** URL: https://ru.wikipedia.org/wiki/ChatGPT (дата обращения: 12.05.2023).

14. **Liu Y., Lapata M.** Text summarization with pretrained encoders // arXiv preprint arXiv:1908.08345. 2019. URL: https://arxiv.org/pdf/1908.08345v2.pdf
(дата обращения: 12.05.2023).

15. **GPT-3.** URL: https://ru.wikipedia.org/wiki/GPT-3 (дата обращения: 12.05.2023).

16. **RED-T5.** Новая SOTA модель для русского языка от SberDevices.
URL: https://habr.com/ru/companies/sberdevices/articles/730088
(дата обращения: 12.05.2023).

17. **Lewis M. et al.** Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // arXiv preprint arXiv:1910.13461. 2019.

18. **Лучшее** место для начала работы с искусственным интеллектом: руководство по Google Colab для начинающих. URL: https://digitrain.ru/articles/156113
(дата обращения: 12.05.2023).

19. **Кирпичникова И. М.** Утилизация выбросов CO_2 на электростанциях с использованием биореакторов // Энергосбережение и водоподготовка. 2022. № 5. С. 15–18.

20. **Приказ** Министерства культуры РФ от 30 декабря 2014 г. № 2477 «Об утверждении типовых отраслевых норм труда на работы, выполняемые в библиотеках». URL: https://www.garant.ru/products/ipo/prime/doc/70921222 (дата обращения: 12.05.2023).

21. **Тикунова И. П.** Организация нормирования труда в библиотеке : сборник нормативных, методических и информационных материалов; Российская государственная библиотека [и др.]. Москва : Пашков дом, 2017. 454 с.

22. **Нормы** труда на работы, выполняемые в библиотеках Корпоративной сети общедоступных библиотек Санкт-Петербурга (КСОБ СПб) / Центральная городская публичная библиотека им. В. В. Маяковского; составители: Марина Николаевна Сухарева [и др.]. 2-е изд., испр. и доп. Санкт-Петербург : ЦГПБ, 2021. 80 с. : ил. (Общедоступные библиотеки Санкт-Петербурга). Библиогр.: 77–80 (57 назв.).

References

1. **Salomatova O. I., Zelenina G. N.** Vozmozhnosti proekta MARS (Mezhregional`naia analiticheskaia rospis` statei`) v informatcionno-bibliograficheskoi` rabote biblioteki // Biblioteki vuzov Urala: problemy` i opy`t raboty`. 2003. Vy`p. 4. 2003.

2. **Leepnitckii` S. F., Mamchich A. A., Sorudei`kina S. A.** Veb-poisk i annotirovanie nauchno-tekhnicheskoi` informatcii na osnove tematicheskikh korpusov tekstov // Informatika. 2018. T. 1. № 2 (22). S. 114–125.

3. **Mukambetova G. I.** Annotirovanie i referirovanie dokumentov: problemy` izucheniia pri podgotovke spetcialistov dlia bibliotek // Vestneyk Bishkekskogo gumanitarnogo universiteta. 2009. № 2. S. 242–244.

4. **Shrai`berg Ia. L.** Bibliotechno-informatcionnaia sfera v sovremenny`kh usloviiakh narastaiushchei` tcifrovizatcii, postpandemii`noi` obstanovki i novy`kh sotcial`no-politicheskikh realii`: glavny`e rezul`taty` : plenarny`i` doclad Predsedatelia Orgkomiteta Dvadtcat` shestoi` Mezhdunarodnoi` konferentcii i vy`stavki «LIBCOM–2022». Moskva : GPNTB Rossii, 2022. 27 s. : il. Bibliogr.: s. 26–27 (17 nazv.). 350 e`kz. ISBN 978-5-85638-253-1. doi: 10.33186/978-5-85638-253-1-2022

5. **Solov`yova L. S.** Baza danny`kh «E`KO»: osobennosti formirovaniia i obsluzhivaniia pol`zovatelei` // Nauchny`e i tekhnicheskie biblioteki. 2003. № 4. S. 51–53. URL: http://ellib.gpntb.ru/subscribe/index.php?journal=ntb&year=2003&num=4&art=8 (data obrashcheniia: 11.11.2019).

6. **By`chkova E. F.** Referativnaia BD «E`kologiia: nauka i tekhnologii» – vazhnaia chast` E`lektronnoi` biblioteki GPNTB Rossii po e`kologii // Nauchny`e i tekhnicheskie biblioteki. 2008. № 2. S. 77–84. URL: http://ellib.gpntb.ru/subscribe/index.php?journal=ntb&year=2008& num=2&art=13 (data obrashcheniia: 27.01.2020).

7. **Borgoiakova K. S.** Bibliometricheskii` analiz nauchny`kh publikatcii` po e`kologii na osnove referativnoi` bazy` danny`kh «E`kologiia: nauka i tekhnologii» GPNTB Rossii // Nauchny`e i tekhnicheskie biblioteki. 2017. № 10. S. 54–68. URL: https://www.gpntb.ru/ntb/ntb/2017/10/NTB10_2017_A5_6.pdf (data obrashcheniia: 20.12.2019).

8. **By`chkova E. F.** Otrazhenie publikatcii` po teme avarii na Chernoby`l`skoi` AE`S i smezhny`m s nei` voprosam v BD GPNTB Rossii «E`kologiia: nauka i tekhnologii» // Chernoby`l` 35 let spustia : materialy` Mezhgosudarstvennoi` nauchno-prakticheskoi` konferentcii (22 aprelia 2021 g.). Briansk, 2021. S. 30–37.

9. **Kratkii`** otchyot o deiatel`nosti GPNTB Rossii za 2021 god. URL: https://www.gpntb.ru/ofitsialnye-dokumenty/84--12/ofitsialnye-dokumenty/9672-kratkij-otchet-o-deyatelnosti-gpntb-rossii-za-2022-god.html (data obrashcheniia: 07.09.2023). URL: свободный.

10. **Das D., Martins A. A.** Survey on Automatic Text Summarization : Technical report // Literature Survey for the Language and Statistics II course at Carnegie Mellon University. Pittsburgh, US, 2007. P. 192–195.

11. **Batura T. V., Bakieva A. M.** Metody` i sistemy` avtomaticheskogo referirovaniia tekstov. Novosibirsk : IPTC NGU, 2019.

12. **Sekrety`** generiruiushchego referirovaniia tekstov. URL: https://habr.com/ru/articles/596481 (data obrashcheniia: 12.05.2023).

13. **ChatGPT.** URL: https://ru.wikipedia.org/wiki/ChatGPT (data obrashcheniia: 12.05.2023).

14. **Liu Y., Lapata M.** Text summarization with pretrained encoders // arXiv preprint arXiv:1908.08345. 2019. URL: https://arxiv.org/pdf/1908.08345v2.pdf (data obrashcheniia: 12.05.2023).

15. **GPT-3.** URL: https://ru.wikipedia.org/wiki/GPT-3 (data obrashcheniia: 12.05.2023).

16. **RED-T5.** Новая SOTA модель для русского языка от SberDevices. URL: https://habr.com/ru/companies/sberdevices/articles/730088 (data obrashcheniia: 12.05.2023).

17. **Lewis M. et al.** Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // arXiv preprint arXiv:1910.13461. 2019.

18. **Luchshee** mesto dlia nachala raboty` s iskusstvenny`m intellektom: rukovodstvo po Google Colab dlia nachinaiushchikh. URL: https://digitrain.ru/articles/156113 (data obrashcheniia: 12.05.2023).

19. **Kirpichnikova I. M.** Utilizatciia vy`brosov $CO_2$ na e`lektrostantciiakh s ispol`zovaniem bioreaktorov // E`nergosberezhenie i vodopodgotovka. 2022. № 5. S. 15–18.

20. **Prikaz** Ministerstva kul`tury` RF ot 30 dekabria 2014 g. № 2477 «Ob utverzhdenii tipovy`kh otraslevy`kh norm truda na raboty`, vy`polniaemy`e v bibliotekakh». URL: https://www.garant.ru/products/ipo/prime/doc/70921222 (data obrashcheniia: 12.05.2023).

21. **Tikunova I. P.** Organizatciia normirovaniia truda v biblioteke : sbornik normativny`kh, metodicheskikh i informatcionny`kh materialov; Rossii`skaia gosudarstvennaia biblioteka [i dr.]. Moskva : Pashkov dom, 2017. 454 s.

22. **Normy`** truda na raboty`, vy`polniaemy`e v bibliotekakh Korporativnoi` seti obshchedostupny`kh bibliotek Sankt-Peterburga (KSOB SPb) / Central`naia gorodskaia publichnaia biblioteka im. V. V. Maiakovskogo; sostaviteli: Marina Nicolaevna Suhareva [i dr.]. 2-e izd., ispr. i dop. Sankt-Peterburg : TCGPB, 2021. 80 s. : il. (Obshchedostupny`e biblioteki Sankt-Peterburga). Bibliogr.: 77–80 (57 nazv.).

Информация об авторах / Information about the authors

Бычкова Елена Феликсовна – ведущий научный сотрудник, руководитель группы развития проектов в области экологии и устойчивого развития ГПНТБ России, Москва, Российская Федерация

bef@gpntb.ru

Колосов Кирилл Анатольевич – канд. техн. наук, ведущий научный сотрудник ГПНТБ России, доцент Московского государственного лингвистического университета, Москва, Российская Федерация

kolosov@gpntb.ru

Elena F. Bychkova – Cand. Sc. (Pedagogy), Leading Researcher, Head, Ecology and Sustainable Development Group of Academic Secretary Department, Russian National Public Library for Science and Technology, Moscow, Russian Federation

bef@gpntb.ru

Kirill A. Kolosov – Cand. Sc. (Engineering), Leading Researcher, Russian National Public Library for Science and Technology; Associate Professor, Moscow State Linguistic University, Moscow, Russian Federation

kolosov@gpntb.ru