MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 30, NO. 4, 2023
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Keyphrase generation for the Russian-language scientific texts using mT5

A. V. Glazkova[1,3], D. A. Morozov[2,3], M. S. Vorobeva[1], A. A. Stupnikov[1]

DOI: 10.18255/1818-1015-2023-4-418-428

[1]University of Tyumen, 6 Volodarskogo str., Tyumen, 625003, Russia.
[2]Novosibirsk National Research State University, 1 Pirogova str., Novosibirsk, 630090, Russia.
[3]Institute for Information Transmission Problems (Kharkevich Institute), 19/1 Bol'shoj Karetnyj pereulok str., Moscow, 127051, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received November 13, 2023
After revision November 22, 2023
Accepted November 29, 2023

In this work, we applied the multilingual text-to-text transformer (mT5) to the task of keyphrase generation for Russian scientific texts using the Keyphrases CS&Math Russian corpus. The automatic selection of keyphrases is a relevant task of natural language processing since keyphrases help readers find the article easily and facilitate the systematization of scientific texts. In this paper, the task of keyphrase selection is considered as a text summarization task. The mT5 model was fine-tuned on the texts of abstracts of Russian research papers. We used abstracts as an input of the model and lists of keyphrases separated with commas as an output. The results of mT5 were compared with several baselines, including TopicRank, YAKE!, RuTermExtract, and KeyBERT. The results are reported in terms of the full-match F1-score, ROUGE-1, and BERTScore. The best results on the test set were obtained by mT5 and RuTermExtract. The highest F1-score is demonstrated by mT5 (11,24 %), exceeding RuTermExtract by 0,22 %. RuTermextract shows the highest score for ROUGE-1 (15,12 %). According to BERTScore, the best results were also obtained using these methods: mT5 — 76,89 % (BERTScore using mBERT), RuTermExtract — 75,8 % (BERTScore using ruSciBERT). Moreover, we evaluated the capability of mT5 for predicting the keyphrases that are absent in the source text. The important limitations of the proposed approach are the necessity of having a training sample for fine-tuning and probably limited suitability of the fine-tuned model in cross-domain settings. The advantages of keyphrase generation using pre-trained mT5 are the absence of the need for defining the number and length of keyphrases and normalizing produced keyphrases, which is important for flective languages, and the ability to generate keyphrases that are not presented in the text explicitly.

Keywords: automatic text summarization; selecting keyphrases; mT5

INFORMATION ABOUT THE AUTHORS

Anna V. Glazkova corresponding author	orcid.org/0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru Candidate of Technical Sciences, Associate Professor, University of Tyumen; Senior Researcher, Institute for Information Transmission Problems (Kharkevich Institute).
Dmitry A. Morozov	orcid.org/0000-0003-4464-1355. E-mail: morozowdm@gmail.com Junior Researcher, Novosibirsk State University and Institute for Information Transmission Problems (Kharkevich Institute).
Marina S. Vorobeva	orcid.org/0000-0002-1508-4089. E-mail: m.s.vorobeva@utmn.ru Candidate of Technical Sciences, Head of the Department of Software, University of Tyumen, Docent.
Andrey A. Stupnikov	orcid.org/0000-0001-5201-1260. E-mail: a.a.stupnikov@utmn.ru Candidate of Physical and Mathematical Sciences, Associate Professor, University of Tyumen, Docent.

Funding: The work was carried out within the framework of project No. MK-3118.2022.4, supported by a grant from the President of the Russian Federation for young scientists — candidates of science.

For citation: A. V. Glazkova, D. A. Morozov, M. S. Vorobeva, and A. A. Stupnikov, "Keyphrase generation for the Russian-language scientific texts using mT5", *Modeling and analysis of information systems*, vol. 30, no. 4, pp. 418-428, 2023.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 30, № 4, 2023
сайт журнала: www.mais-journal.ru

Генерация ключевых слов для русскоязычных научных текстов с помощью модели mT5

А. В. Глазкова[1,3], Д. А. Морозов[2,3], М. С. Воробьева[1], А. А. Ступников[1]

DOI: 10.18255/1818-1015-2023-4-418-428

[1]Тюменский государственный университет, ул. Володарского, д. 6, г. Тюмень, 625003, Россия.
[2]Новосибирский национальный исследовательский государственный университет, ул. Пирогова, д. 1, г. Новосибирск, 630090, Россия.
[3]Институт проблем передачи информации РАН им. А. А. Харкевича, Большой Каретный переулок, д. 19, стр. 1, г. Москва, 127051, Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 13 ноября 2023 г.
После доработки 22 ноября 2023 г.
Принята к публикации 29 ноября 2023 г.

Авторами предлагается подход к генерации ключевых слов для русскоязычных научных текстов с помощью модели mT5 (multilingual text-to-text transformer), дообученнной на материале текстового корпуса Keyphrases CS&Math Russian. Автоматический подбор ключевых слов является актуальной задачей обработки естественного языка, поскольку ключевые слова помогают читателям осуществлять поиск статей и облегчают систематизацию научных текстов. В данной работе задача подбора ключевых слов рассматривается как задача автоматического реферирования текстов. Дообучение mT5 осуществлялась на текстах аннотаций русскоязычных научных статей. В качестве входных и выходных данных выступали тексты аннотаций и списки ключевых слов, разделенных запятыми, соответственно. Результаты, полученные с помощью mT5, были сравнены с результатами нескольких базовых методов: TopicRank, YAKE!, RuTermExtract, и KeyBERT. Для представления результатов использовались следующие метрики: F-мера, ROUGE-1, BERTScore. Лучшие результаты на тестовой выборке были получены с помощью mT5 и RuTermExtract. Наиболее высокое значение F-меры продемонстрировала модель mT5 (11.24 %), превзойдя RuTermExtract на 0.22 %. RuTermExtract показал лучший результат по метрике ROUGE-1 (15.12 %). Лучшие результаты по BERTScore также были достигнуты этими двумя методами: mT5 — 76.89 % (BERTScore, использующая модель mBERT), RuTermExtract — 75.8 % (BERTScore на основе ruSciBERT). Также авторами была оценена возможность mT5 генерировать ключевые слова, отсутствующие в исходном тексте. К ограничениям предложенного подхода относятся необходимость формирования обучающей выборки для дообучения модели и, вероятно, ограниченная применимость дообученной модели для текстов других предметных областей. Преимущества генерации ключевых слов с помощью mT5 — отсутствие необходимости задавать фиксированные значения длины и количества ключевых слов, необходимости проводить нормализацию, что особенно важно для флективных языков, и возможность генерировать ключевые слова, в явном виде отсутствующие в тексте.

Ключевые слова: автоматическое реферирование; подбор ключевых слов; mT5

ИНФОРМАЦИЯ ОБ АВТОРАХ

Анна Валерьевна Глазкова автор для корреспонденции	orcid.org/0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru доцент каф. программного обеспечения ТюмГУ; с. н. с. ИППИ РАН, к. т. н.
Дмитрий Алексеевич Морозов	orcid.org/0000-0003-4464-1355. E-mail: morozowdm@gmail.com м. н. с. Лаборатории ПЩТ НГУ и Лаборатории КЛ ИППИ РАН.
Марина Сергеевна Воробьева	orcid.org/0000-0002-1508-4089. E-mail: m.s.vorobeva@utmn.ru доцент, заведующий каф. программного обеспечения ТюмГУ, к. т. н.
Андрей Анатольевич Ступников	orcid.org/0000-0001-5201-1260. E-mail: a.a.stupnikov@utmn.ru доцент каф. программного обеспечения ТюмГУ, к. ф.-м. н.

Финансирование: Работа выполнена в рамках проекта № МК-3118.2022.4, поддержанного грантом Президента Российской Федерации для молодых ученых — кандидатов наук.

Для цитирования: A. V. Glazkova, D. A. Morozov, M. S. Vorobeva, and A. A. Stupnikov, “Keyphrase generation for the Russian-language scientific texts using mT5”, *Modeling and analysis of information systems*, vol. 30, no. 4, pp. 418-428, 2023.

Введение

Ключевые слова являются важным элементом научного текста. Использование ключевых слов позволяет облегчить поиск статей, улучшить систематизацию научных текстов и резюмировать содержание статей для читателя. Автоматизация подбора ключевых слов представляет собой актуальную задачу в условиях большого количества информационных ресурсов. В настоящее время большинство методов подбора ключевых слов протестировано на англоязычных текстовых корпусах, в то время как для анализа русскоязычных текстов используется достаточно узкий набор методов выделения ключевых слов [1]. Существует несколько подходов к подбору ключевых слов: извлечение ключевых слов непосредственно из текста, подбор ключевых слов из заранее определенного перечня тематик или рубрик и генерация ключевых слов на основе семантики текста путем его обобщения и перефразирования [2, 3]. В последнем случае задача подбора ключевых слов схожа с задачей автоматического абстрактного реферирования текстов.

Большая часть широко используемых подходов к извлечению ключевых слов основана на выделении из текста наиболее значимых слов и словосочетаний по принципу обучения без учителя (unsupervised learning). К таким подходам относятся, в частности, статистические алгоритмы, такие как YAKE! [4] и KP-Miner [5], графовые (TopicRank [6], TextRank [7]) и ряд алгоритмов, основанных на применении методов машинного обучения и современных лингвистических моделей (KEA [8], KeyBERT [9]). Несмотря на впечатляющие результаты для ряда текстовых корпусов, алгоритмы, основанные на извлечении ключевых слов, обладают некоторыми ограничениями. В частности, они не способны определять количество ключевых слов автоматически и генерировать ключевые слова, отсутствующие в тексте в явном виде. На практике же списки ключевых слов обычно включают в себя как слова и словосочетания, встречающиеся в тексте непосредственно, так и ключевые слова, семантически связанные с содержанием текста, но не упомянутые в нем явно [10]. Данные ограничения могут быть преодолены при помощи нейросетевых моделей, в том числе современных лингвистических моделей для генерации текстов. В работах [11—13] описаны модели для генерации ключевых слов с рекуррентной архитектурой, позволяющие генерировать как присутствующие, так и отсутствующие в тексте ключевые слова. В работах [14, 15] предлагаются модели для генерации ключевых слов, основанные на архитектуре Transformer [16]. В статьях [17—20] представлены эксперименты по дообучению лингвистических моделей для генерации ключевых слов. Все перечисленные подходы протестированы на англоязычных корпусах.

К настоящему моменту опубликован ряд исследований, применяющих алгоритмы обучения без учителя к русскоязычным текстам [21—25]. В статье [26] предложен подход к извлечению ключевых слов с помощью комбинации данных алгоритмов и нейросетевых моделей. Подход протестирован на мультиязычном корпусе, включающем в себя русскоязычные тексты. Более подробное исследование нейросетевых моделей затруднено ввиду недостатка русскоязычных текстовых корпусов для подбора ключевых слов. Насколько нам известно, единственными корпусами текстов для данной задачи, находящимся в открытом доступе, являются корпус аннотаций научных статей [27], описанный в работе [25], и мультиязычный корпус новостных текстов, представленный в работе [26].

Данная работа направлена на преодоление пробела в использовании современных лингвистических моделей для генерации ключевых слов для русскоязычных научных текстов. В статье представлены результаты экспериментов по генерации списка ключевых слов как последовательности токенов на примере модели mT5 [28]. Выбор модели обусловлен ее широким использованием для автоматического реферирования и, в частности, для реферирования русскоязычных текстов (например, в работах [29, 30]). Дообучение mT5 выполнялось на русскоязычном корпусе для подбора ключевых слов аналогично тому, как выполняется дообучение для задачи автоматического реферирования. Проанализированы преимущества и недостатки данного подхода. Проведено сравнение

Table 1. The characteristics of the dataset **Таблица 1.** Характеристики набора данных

Характеристика	Обучающая выборка	Тестовая выборка
Количество текстов	5 844	2 504
Средняя длина текста (токенов)	102.97±86.05	62.41±39.14
Среднее количество ключевых слов	4.39	4.2
% ключевых слов, отсутствующих в исходном тексте	53.17	54.8

полученных результатов с результатами нескольких широко используемых подходов к извлечению ключевых слов.

Статья организована следующим образом. В разделе 1 представлены основные характеристики корпуса текстов, используемого в экспериментах. Раздел 2 содержит описание параметров используемой модели для генерации ключевых слов и перечень базовых методов ключевых слов, которые были использованы для сравнения. В разделе 3 приводятся и обсуждаются полученные результаты. Выводы к работе представлены в заключении.

1. Корпус текстов

В работе использован корпус аннотаций и соответствующих им списков ключевых слов, собранный из интернет-ресурсов «Киберленинка» и MathNet и описанный в работе [25]. «Киберленинка» комплектуется научными статьями различной тематики, в то время как публикации, размещенные на MathNet, включают в себя исследования по математике, автоматике и вычислительной технике, информатике и кибернетике [31]. Для исследования были отобраны русскоязычные аннотации, которым соответствуют не менее трех ключевых слов. Дубликаты текстов были удалены. Полученный датасет был разделен на обучающую и тестовую выборки в соотношении 70:30. Характеристики полученного набора данных представлены в таблице 1. Средняя длина текстов представляет собой среднее количество токенов, полученное с помощью токенизатора модели RuBERT-base-cased[1] [32]. Доля ключевых слов, отсутствующих в тексте, характеризует долю ключевых слов, которые не встречаются в исходном тексте (тексте аннотации) в явном виде. При подсчете данной характеристики текст и ключевые слова подвергались предварительной нормализации. В обучающей и тестовой выборках содержится значительная доля ключевых слов, отсутствующих в исходном тексте (53.17 % и 54.8 % соответственно). Таким образом, списки ключевых слов часто содержат слова и словосочетания, семантически связанные с содержанием аннотации, но непосредственно не присутствующие в тексте (например, обобщающие ключевые слова, описывающие предметную область, синонимы и гиперонимы слов, встречающихся в тексте и так далее).

2. Методы подбора ключевых слов

Для генерации ключевых слов была использована модель mT5 (multilingual text-to-text transformer) [28], предварительно обученная на текстах на 101 языке. Архитектура mT5 в целом повторяет архитектуру модели T5 [33], разработанную ранее для английского языка, и использует подход Text-to-Text. Суть данного подхода заключается в том, что данные для каждой задачи, на которой проходило предварительное обучение, были представлены в текстовом виде. В данной работе была использована модель mT5-base[2], имеющая 580 млн параметров.

Дообучение mT5 для задачи генерации ключевых слов выполнялось на обучающей выборке в течение 10 эпох с использованием следующих параметров: максимальная длина входной последовательности — 256 токенов, скорость обучения — $4 \cdot 10^{-5}$, размер батча — 8. Параметр repetition penalty, влияющий на размер ошибки модели в случае генерации повторяющихся токенов, варьировался в диапазоне [1;2] с шагом 0.1. В результате экспериментов было выбрано значение repetition

[1] https://huggingface.co/DeepPavlov/rubert-base-cased
[2] https://huggingface.co/google/mt5-base

penalty, равное 1.4. В модель подавались тексты аннотаций в качестве исходных текстов и списки ключевых слов, разделенных запятыми, в качестве выходных. Таким образом, дообучение выполнялось аналогично дообучению для задачи автоматического реферирования, однако в роли рефератов для исходных текстов выступали списки соответствующих им ключевых слов. Тестирование модели выполнялось на тестовой выборке.

Результаты модели mT5 были сравнены с результатами нескольких базовых методов извлечения ключевых слов:

1. TopicRank [6], графовый алгоритм, в ходе работы которого все последовательности идущих подряд прилагательных и существительных в тексте обозначаются как потенциальные ключевые фразы, а затем объединяются в семантические группы с использованием алгомеративной иерархической кластеризации. Среди получившихся кластеров выбираются наиболее значимые при помощи алгоритма PageRank [34], лежащего в основе поисковой системы Google.
2. YAKE! (Yet Another Keyword Extractor) [4], статистический алгоритм, основанный на применении ряда вычислимых эвристик: вероятности написания слова с заглавной буквы, наиболее вероятной позиции слова в тексте, доли предложений, содержащих слов и т. д.
3. RuTermExtract[3], алгоритм, основанный на анализе морфологических характеристик слов и словосочетаний и набора правил для извлечения ключевых слов. Алгоритм представляет собой адаптированную для русского языка версию алгоритма TermExtract[4]. Для морфологического анализа в русскоязычной версии используется библиотека PyMorphy2 [35].
4. KeyBERT [9], нейросетевой алгоритм на основе Bidirectional Encoder Representations from Transformers (BERT) [36], выбирающий из текста слова и словосочетания, семантические вектора которых наиболее схожи с векторным представлением всего текста.

В данной работе была использована реализация алгоритмов TopicRank и YAKE! из библиотеки PKE [37] с подключением spacy[5]-модели ru_core_news_sm. В качестве основы для алгоритма KeyBERT использовалась модель ruSciBERT[6] [38]. С помощью каждого из базовых алгоритмов были извлечены по 5, 10 и 15 ключевых слов. Для TopicRank и YAKE! также варьировалась максимальная длина n-грамм ($N = 1$ и $N = 3$, N — максимальная длина ключевого слова). Тестирование базовых алгоритмов также выполнялось на тестовой выборке.

3. Результаты

Для представления результатов были использованы четыре метрики: F-мера, ROUGE-1 [39] и два варианта BERTScore [40]. F-мера рассчитывается на основании показателей точности (precision) и полноты (recall) двух списков ключевых слов: представленного в корпусе текстов и полученного машинным способом. Ключевые слова при этом предварительно нормализуются с помощью библиотеки PyMorphy2 [35]. ROUGE-1 отражает количество совпадающих униграмм в исходном и сгенерированном списках ключевых слов. Метрика BERTScore показывает семантическое сходство списков ключевых слов, оценивая степень близости векторных представлений входящих в них токенов. Векторные представления токенов могут быть получены из предварительно обученных лингвистических моделей. Метрика BERTScore рассчитана в данной работе двумя способами. В первом случае используется многоязычная модель BERT (mBERT) [36], так как данная модель применялась для оценки близости неанглоязычных текстов в оригинальной статье [40]. Во втором варианте близость списков ключевых слов оценивается с помощью ruSciBERT [38], обученной на русскоязычных научных текстах.

[3] https://github.com/igor-shevchenko/rutermextract
[4] https://pypi.org/project/topia.termextract
[5] https://spacy.io/usage/models
[6] https://huggingface.co/ai-forever/ruSciBERT

Table 2. Results, %

Таблица 2. Результаты, %

Метод	Все ключевые слова				Ключевые слова, отсутствующие в тексте	
	F	R	BS	$BS(S)$	BS	$BS(S)$
TopicRank ($N = 1$, $k = 5$)	3.86	7.62	71.65	71.09	68.81	66.26
TopicRank ($N = 1$, $k = 10$)	3.84	8.08	70.89	72.03	66.75	66.29
TopicRank ($N = 1$, $k = 15$)	3.72	7.99	70.49	71.92	65.82	65.94
TopicRank ($N = 3$, $k = 5$)	4.79	6.30	73.48	70.6	69.47	65.88
TopicRank ($N = 3$, $k = 10$)	4.96	6.64	73.52	71.24	68.81	65.85
TopicRank ($N = 3$, $k = 15$)	4.92	6.66	73.48	71.27	68.74	65.77
YAKE! ($N = 1$, $k = 5$)	3.75	7.25	71.47	71.04	68.73	66.29
YAKE! ($N = 1$, $k = 10$)	3.92	8.38	70.87	72.30	66.72	66.54
YAKE! ($N = 1$, $k = 15$)	3.79	8.09	70.45	72.19	65.75	66.13
YAKE! ($N = 3$, $k = 5$)	2.82	5.17	69.30	67.10	66.15	63.09
YAKE! ($N = 3$, $k = 10$)	5.37	6.41	68.27	66.64	64.46	61.50
YAKE! ($N = 3$, $k = 15$)	6.39	6.47	67.01	65.15	62.81	59.64
RuTermExtract ($k = 5$)	9.75	14.42	75.85	74.98	70.73	68.31
RuTermExtract ($k = 10$)	11.02	**15.12**	75.95	**75.80**	70.25	68.16
RuTermExtract ($k = 15$)	10.86	14.87	75.86	75.71	70.04	67.84
KeyBERT ($k = 5$)	5.27	4.24	70.63	68.01	67.08	63.70
KeyBERT ($k = 10$)	6.43	5.31	70.00	67.85	65.70	62.74
KeyBERT ($k = 15$)	6.53	5.65	69.36	66.61	64.85	61.16
mT5	**11.24**	13.10	**76.89**	75.06	**72.85**	**68.95**

Результаты сравнения методов подбора ключевых слов представлены в таблице 2. Для каждого метода были рассчитаны метрики для всех ключевых слов из списка ключевых слов, представленного в корпусе («Все ключевые слова»), а также только для тех ключевых слов из списка, представленного в корпусе, которые не встречаются в исходном тексте в явном виде («Ключевые слова, отсутствующие в тексте»). Во втором случае сравнивался список ключевых слов, представленный в корпусе, за исключением ключевых слов, встречающихся в тексте, и список ключевых слов, полученный моделью. Поскольку методы, выполняющие извлечение ключевых слов из исходного текста, не способны генерировать ключевые слова, отсутствующие в тексте, во втором случае оценка с помощью F-меры и ROUGE не проводилась. Результаты были оценены с позиции семантического сходства (BERTScore). В таблице 2 используются следующие сокращения: F — F-мера, R — ROUGE-1, BS — BERTScore, основанная на mBERT, $BS(S)$ — BERTScore, основанная на ruSciBERT, N — максимальная длина ключевого слова (длина n-граммы), k — количество ключевых слов. Лучший результат в каждом столбце подчеркнут и выделен полужирным шрифтом.

На рисунке 1 показана зависимость результатов модели mT5 от величины параметра repetition penalty, влияющего на размер ошибки модели в случае генерации повторяющихся токенов. График построен на основе результатов моделей на полной тестовой выборке. Лучший результат для каждой метрики выделен пятиугольным маркером. Поскольку для трех метрик из четырех лучшие результаты были получены при значении repetition penalty, равном 1.4, в таблице 2 указаны результаты этой модели.

Результаты, полученные с помощью RuTermExtract и mT5, в целом выше, чем результаты TopicRank, YAKE! и KeyBERT. При этом наиболее высокий показатель по F-мере демонстрирует mT5 (11.24 %), превосходя RuTermExtract на 0.22 %. RuTermExtract показывает наиболее высокие по-

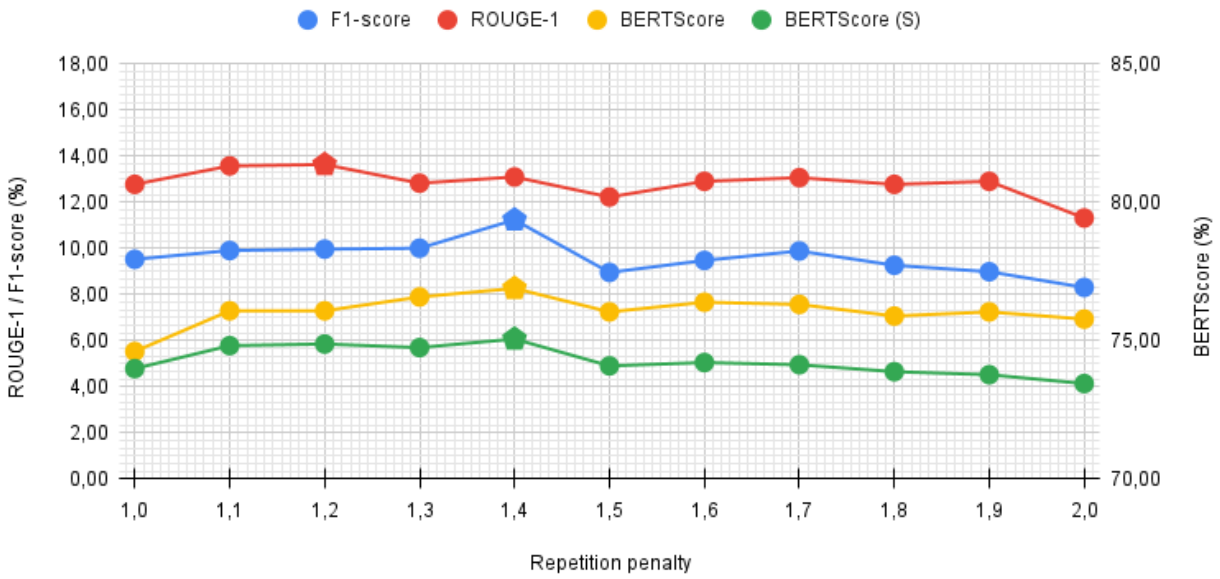

Fig. 1. The influence of the repetition penalty parameter on keyphrase generation performance

Рис. 1. Влияние параметра repetition penalty на качество генерации ключевых слов

казатели по метрике ROUGE-1 (15.12 %). По BERTScore лучшие результаты также получены с помощью данных методов: mT5 — 76.89 % (BS), RuTermExtract — 75.8 % ($BS(S)$). mT5 демонстрирует лучшие результаты при генерации ключевых слов, отсутствующих в исходном тексте. Генерация таких ключевых слов является технически более сложной задачей, чем извлечение ключевых слов, поскольку требует от модели «понимания» семантики текста и способности к реферированию и обобщению. Мы дополнительно рассчитали показатели F-меры и полноты (recall) для генерации ключевых слов, отсутствующих в тексте, и получили показатели 1.33 % и 1.7 % соответственно.

В таблице 3 представлены примеры ключевых слов, полученных с помощью разных алгоритмов. Для удобства сравнения приведены только результаты при $k = 5$. Приведенные примеры показывают, что ключевые слова, сгенерированные с помощью mT5, носят ожидаемо более общий характер, чем ключевые слова, выделенные другими методами. Количество ключевых слов, сгенерированных mT5, невелико и в целом близко среднему количеству ключевых слов для текста в используемом корпусе. В приведенных примерах mT5 генерирует грамматически корректные словосочетания и не требует проведения дополнительной нормализации, что является преимуществом данного подхода.

Заключение

В данной работе задача подбора ключевых слов рассматривается как задача автоматического реферирования текста на естественном языке. Мы описываем результаты экспериментов по генерации списка ключевых слов в виде одной строки для аннотаций научных текстов на русском языке с помощью предобученной лингвистической модели mT5. Результаты сравниваются с результатами ряда широко используемых методов извлечения ключевых слов.

Среди преимуществ генерации ключевых слов с помощью предобученной лингвистической модели можно назвать отсутствие необходимости проводить нормализацию и задавать ограничения на количество и длину ключевых слов, возможность генерировать ключевые слова, которые не упомянуты в исходном тексте в явном виде. С другой стороны, указанные свойства могут быть также ограничениями указанного подхода. Дообучение рассмотренной модели требует наличия обуча-

Table 3. The examples of keyphrases **Таблица 3.** Примеры ключевых слов

Метод	Ключевые слова
Пример 1. Рассмотрено современное состояние разработок газовых сенсоров на основе оксидов металлов, как наиболее перспективных. Для исследования сенсоров разработана информационная система, обладающая широкими возможностями по обеспечению их эффективного функционирования, а также передачи информации с использованием сетевых технологий.	
Список ключевых слов, представленный в корпусе	газовый анализ, информационная система, электронный нос, нейронные сети
TopicRank ($N = 1$, $k = 5$)	сенсор, возможность, система, обеспечение, исследование
TopicRank ($N = 3$, $k = 5$)	обеспечению, широкими возможностями, информационная система, исследования сенсоров, эффективного функционирования
YAKE! ($N = 1$, $k = 5$)	состояние, технология, сенсор, разработка, основа
YAKE! ($N = 3$, $k = 5$)	рассмотрено современное состояние, современное состояние разработок, состояние разработок газовых, основе оксидов металлов, разработок газовых сенсоров
RuTermExtract ($k = 5$)	современное состояние разработок, основа оксидов металлов, обладающая широкие возможности, эффективное функционирование, сетевые технологии
KeyBERT ($k = 5$)	рассмотрено современное, разработана информационная, система обладающая, эффективного функционирования, современное состояние
mT5	оксиды металлов, информационная система, обработка информации
Пример 2. Впервые выявляются общие и особенные черты методики аннотирования машиночитаемых документов. Проанализирована практика аннотирования электронных документов локального доступа и ресурсов удаленного доступа. Выделены источники сведений об электронных документах, их идентификационные признаки, значимые для методики аннотирования. Названы причины, которые осложняют процесс аннотирования электронных информационных ресурсов удаленного доступа. Показаны пути создания работоспособных методик аннотирования машиночитаемых документов.	
Список ключевых слов, представленный в корпусе	аннотирование, электронные документы, аспектная схема
TopicRank ($N = 1$, $k = 5$)	аннотирование, документ, доступ, методика, ресурс
TopicRank ($N = 3$, $k = 5$)	доступа, значимые, идентификационные признаки, методики аннотирования, электронных документах
YAKE! ($N = 1$, $k = 5$)	аннотирование, документ, методика, доступ, ресурс
YAKE! ($N = 3$, $k = 5$)	впервые выявляются общие, особенные черты методики, впервые выявляются, выявляются общие, особенные черты
RuTermExtract ($k = 5$)	электронные документы, удалённый доступ, машиночитаемые документы, особенные черты методики аннотирования, электронные информационные ресурсы
KeyBERT ($k = 5$)	аннотирования машиночитаемых, машиночитаемых документов, методик аннотирования, методики аннотирования, машиночитаемых
mT5	машиночитаемый документ, удаленный доступ, аннотирование, методы аннотирования

ющей выборки и, вероятно, дообученная модель ограниченно пригодна для генерации ключевых слов для текстов других предметных областей. Кроме того, эффективность предложенного подхода и значения метрик зависят от специфики корпуса текстов, используемого для экспериментов. В рассмотренном корпусе доля ключевых слов, не встречающих в тексте в явном виде, составляет 53.17 % и 54.8 % для обучающей и тестовой выборок соответственно. Поскольку подходы, осуществляющие извлечение, а не генерацию ключевых слов, не способны генерировать ключевые слова данного типа, модели генерации текста, подобные mT5, имеют преимущество на таких корпусах.

В дальнейшей работе будут исследованы другие лингвистические модели и тексты, относящиеся к различным предметным областям (например, новостные). Также будет изучена возможность генерации заданного пользователем количества ключевых слов и введения других ограничений на параметры списка ключевых слов, генерируемого моделью. Отдельным направлением для дальнейших исследований является генерация ключевых слов, в явном виде отсутствующих в исходном тексте. Вероятно, полученные в работе показатели F-меры и полноты при генерации ключевых слов, отсутствующих в тексте, были бы выше, если бы дообучение модели проводилось не на списках всех ключевых слов, а только на ключевых словах, отсутствующих в тексте. Также может быть проведена типизация таких ключевых слов (синонимы, гиперонимы и так далее), и для генерации ключевых слов каждого типа может быть проведена отдельная оценка эффективности подходов.

References

[1] N. S. Lagutina, K. V. Lagutina, A. S. Adrianov, and I. V. Paramonov, “Russian language thesauri: Automated construction and application for natural language processing tasks”, *Modeling and Analysis of Information Systems*, vol. 25, no. 4, pp. 435–458, 2018, in Russian.

[2] S. Beliga, *Keyword extraction: A review of methods and approaches*, https://api.semanticscholar.org/CorpusID:6834431, 2014.

[3] E. Çano and O. Bojar, “Keyphrase generation: A multi-aspect survey”, in *25th Conference of Open Innovations Association (FRUCT)*, IEEE, 2019, pp. 85–94.

[4] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, “YAKE! keyword extraction from single documents using multiple local features”, *Information Sciences*, vol. 509, pp. 257–289, 2020.

[5] S. R. El-Beltagy and A. Rafea, “KP-Miner: A keyphrase extraction system for English and Arabic documents”, *Information systems*, vol. 34, no. 1, pp. 132–144, 2009.

[6] A. Bougouin, F. Boudin, and B. Daille, “TopicRank: Graph-based topic ranking for keyphrase extraction”, in *International joint conference on natural language processing (IJCNLP)*, 2013, pp. 543–551.

[7] R. Mihalcea and P. Tarau, “TextRank: Bringing order into text”, in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[8] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, “KEA: Practical automatic keyphrase extraction”, in *Proceedings of the fourth ACM conference on Digital libraries*, 1999, pp. 254–255.

[9] M. Grootendorst, *KeyBERT: Minimal keyword extraction with BERT*, version v0.3.0, 2020. DOI: 10.5281/zenodo.4461265. [Online]. Available: https://github.com/MaartenGr/KeyBERT.

[10] F. Boudin and Y. Gallina, “Redefining absent keyphrases and their effect on retrieval effectiveness”, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 4185–4193.

[11] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, "Deep keyphrase generation", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 582–592.

[12] E. Cano and O. Bojar, "Keyphrase generation: A text summarization struggle", in *Proceedings of NAACL-HLT*, 2019, pp. 666–672.

[13] J. Zhao and Y. Zhang, "Incorporating linguistic constraints into keyphrase generation", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5224–5233.

[14] R. Liu, Z. Lin, and W. Wang, *Keyphrase prediction with pre-trained language model*, 2020. arXiv: 2004.10462 [cs.CL].

[15] M. Kulkarni, D. Mahata, R. Arora, and R. Bhowmik, "Learning rich representation of keyphrases from text", in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 891–906.

[16] A. Vaswani *et al.*, "Attention is all you need", in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[17] M. F. M. Chowdhury, G. Rossiello, M. Glass, N. Mihindukulasooriya, and A. Gliozzo, *Applying a generic sequence-to-sequence model for simple and effective keyphrase generation*, 2022. arXiv: 2201.05302 [cs.CL].

[18] A. Glazkova and D. Morozov, "Applying transformer-based text summarization for keyphrase generation", *Lobachevskii Journal of Mathematics*, vol. 44, no. 1, pp. 123–136, 2023.

[19] A. Glazkova and D. Morozov, "Multi-task fine-tuning for generating keyphrases in a scientific domain", in *IX International Conference on Information Technology and Nanotechnology (ITNT)*, IEEE, 2023, pp. 1–5.

[20] D. Wu, W. U. Ahmad, and K.-W. Chang, *Pre-trained language models for keyphrase generation: A thorough empirical study*, 2022. arXiv: 2212.10233 [cs.CL].

[21] E. G. Sokolova and O. Mitrofanova, "Automatic keyphrase extraction by applying KEA to Russian texts", in *Computational linguistics and computing ontologies*, in Russian, 2017, pp. 157–165.

[22] M. V. Sandul and E. G. Mikhailova, "Keyword extraction from single Russian document", in *Proceedings of the Third Conference on Software Engineering and Information Management*, 2018, pp. 30–36.

[23] E. Sokolova, A. Moskvina, and O. Mitrofanova, "Keyphrase extraction from the Russian corpus on linguistics by means of KEA and RAKE algorithms", in *Data analytics and management in data-intensive domains*, 2018, pp. 369–372.

[24] O. A. Mitrofanova and D. A. Gavrilic, "Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts", *Terra Linguistica*, vol. 50, no. 4, pp. 22–40, 2022, in Russian.

[25] D. Morozov, A. Glazkova, M. Tyutyulnikov, and B. Iomdin, "Keyphrase generation for abstracts of the Russian-language scientific articles", *NSU Vestnik. Series: Linguistics and Intercultural Communication*, vol. 21, no. 1, pp. 54–66, 2023, in Russian.

[26] B. Koloski, S. Pollak, B. Škrlj, and M. Martinc, "Extending neural keyword extraction with TF-IDF tagset matching", in *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 2021, pp. 22–29.

[27] D. Morozov and A. Glazkova, *Keyphrases CS&Math Russian*, version v1, 2022. DOI: 10.17632/dv3j9wc59v.1. [Online]. Available: https://data.mendeley.com/datasets/dv3j9wc59v/1.

[28] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 483–498.

[29] K. Grashchenkov, A. Grabovoy, and I. Khabutdinov, "A method of multilingual summarization for scientific documents", in *Ivannikov Ispras Open Conference (ISPRAS)*, IEEE, 2022, pp. 24–30.

[30] A. Gryaznov, R. Rybka, I. Moloshnikov, A. Selivanov, and A. Sboev, "Influence of the duration of training a deep neural network model on the quality of text summarization task", *AIP Conference Proceedings*, vol. 2849, no. 1, p. 400 006, 2023.

[31] A. A. Pechnikov, "Comparative analysis of scientometrics indicators of journals Math-Net.ru and Elibrary.ru", *Vestnik Tomskogo gosudarstvennogo universiteta*, no. 56, pp. 112–121, 2021, in Russian.

[32] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language", in *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, in Russian, 2019, pp. 333–339.

[33] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer", *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[34] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web: Stanford InfoLab", 1999, p. 1 508 503.

[35] M. Korobov, "Morphological analyzer and generator for Russian and Ukrainian languages", in *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4*, Springer, 2015, pp. 320–332.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[37] F. Boudin, "PKE: An open source python-based keyphrase extraction toolkit", in *Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations*, 2016, pp. 69–73.

[38] N. A. Gerasimenko, A. S. Chernyavsky, and M. Nikiforova, "ruSciBERT: A transformer language model for obtaining semantic embeddings of scientific texts in Russian", in *Doklady Mathematics*, Springer, vol. 106, 2022, S95–S96.

[39] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries", in *Text summarization branches out*, 2004, pp. 74–81.

[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *BERTScore: Evaluating text generation with BERT*, 2020. arXiv: 1904.09675 [cs.CL].