Научная статья

УДК 004.89
DOI 10.25205/1818-7935-2023-21-1-54-66

Генерация ключевых слов для аннотаций русскоязычных научных статей

Дмитрий Алексеевич Морозов[1], Анна Валерьевна Глазкова[2]
Михаил Андреевич Тютюльников[3], Борис Леонидович Иомдин[4]

[1,3]Новосибирский государственный университет
Новосибирск, Россия

[2]Тюменский государственный университет
Тюмень, Россия

[4]Институт русского языка им. В. В. Виноградова РАН
Москва, Россия

[1]morozowdm@gmail.com, https://orcid.org/0000-0003-4464-1355
[2]a.v.glazkova@utmn.ru, https://orcid.org/0000-0001-8409-6457
[3]mishatyty@gmail.com, https://orcid.org/0000-0001-5367-3944
[4]iomdin@ruslang.ru, https://orcid.org/0000-0002-1767-5480

Аннотация

В этой работе мы попробовали адаптировать различные известные механизмы генерации ключевых слов к весьма специфичному корпусу: аннотациям русскоязычных научных статей из области математики и компьютерных наук. В такой постановке сразу несколько сложностей: отсутствие масштабных исследований механизмов генерации для русского языка, отсутствие крупных корпусов аннотаций и в целом длина аннотаций: если для полного текста ключевые слова обычно встречаются в статье и достаточно лишь выделить их, для аннотаций характерно отсутствие ключевых слов в тексте в явном виде. При этом в открытый доступ попадают обычно именно аннотации, и автоматическая генерация ключевых слов для них позволила бы существенно улучшить возможности поиска по статьям. Причем генерировать слова стоит и для тех статей, в которых авторы сами их указали, так как в ходе исследования мы обнаружили, что используемые ключевые слова нередко уникальны для конкретной статьи, а значит, по таким словам невозможно сформировать подкорпус статей по заданной тематике. Для визуализации результатов работы мы создали ресурс keyphrases.mca.nsu.ru, на котором начинающие исследователи могут сформировать приблизительный список слов для своей первой публикации.

Ключевые слова

статье, отсутствие крупных корпусов аннотаций, слова, целом длина аннотаций, ключевые

Благодарности

Работа выполнена в рамках проекта № МК-3118.2022, поддержанного грантом Президента Российской Федерации для молодых ученых – кандидатов наук.

Для цитирования

Морозов Д. А., Глазкова А. В., Тютюльников М. А., Иомдин Б. Л. Генерация ключевых слов для аннотаций русскоязычных научных статей // Вестник НГУ, Серия: Лингвистика и межкультурная коммуникация. 2023. Т. 21, № 1. С. 54–66. DOI 10.25205/1818-7935-2023-21-1-54-66

Keyphrase Generation for Abstracts of the Russian-Language Scientific Articles

Dmitry A. Morozov[1], Anna V. Glazkova[2]
Mikhail A. Tyutyulnikov[3], Boris L. Iomdin[4]

[1,3]Novosibirsk State University
Novosibirsk, Russia

[2]Tyumen State University
Tyumen, Russia

[3]Vinogradov Russian Language Institute RAS
Moscow, Russia

[1]morozowdm@gmail.com, https://orcid.org/0000-0003-4464-1355
[2]a.v.glazkova@utmn.ru, https://orcid.org/0000-0001-8409-6457
[3]mishatyty@gmail.com, https://orcid.org/0000-0001-5367-3944
[4]iomdin@ruslang.ru, https://orcid.org/0000-0002-1767-5480

Abstract

In this paper, we attempted to adapt various well-known algorithms for keyword selection to a very specific text corpus containing abstracts of Russian academic papers from the mathematical and computer science domain. We faced several challenges including the lack of research in the field of keyword extraction for Russian, the absence of large text corpora of academic abstracts, and the insufficient length of the abstracts. Keywords are often found in the full text of the paper and can simply be highlighted, whereas abstracts may not include keywords in an explicit form. At the same time, it is abstracts that are usually in the public domain, so automatic selection of keywords from them would significantly facilitate the process of searching for papers. Moreover, an automatic keyword selection would be useful even for papers for which keywords were already specified by the authors. During the study, we found that authors often use unique keywords for their papers. This complicates their systematization on a given topic. For visualizing the results, we have created a web resource keyphrases.mca.nsu.ru, where young/beginning scholars can form an approximate list of keywords for their first research paper.

Keywords

article, lack of large corpora of abstracts, words, overall length of annotations, key

Acknowledgments

The work was carried out within the framework of project No. MK-3118.2022, supported by a grant from the President of the Russian Federation for young scientists - candidates of science.

For citation

Morozov D. A., Glazkova A. V., Tyutyulnikov M. A., Iomdin B. L. Keyphrase Generation for Abstracts of the Russian-Language Scientific Articles. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2023, vol. 21, no. 1, pp. 54–66. (in Russ.) DOI 10.25205/1818-7935-2023-21-1-54-66

Введение

Системы поиска для экспертов обладают рядом особенностей, отличающих их от, например, систем поиска в интернете. Такие системы зачастую предоставляют пользователю намного более широкие возможности настройки поиска. Так, в лингвистических корпусах текстов нередко присутствует возможность сформулировать запрос к нескольким аспектам искомого, например к грамматическим признакам и форме слова одновременно, или отобрать для поиска лишь часть документов в качестве подкорпуса. Один из способов обогатить поисковые возможности при использовании поискового движка в подобном узком экспертном поле – кластеризация документов по тематике текста с предоставлением в дальнейшем возможности поиска по кластерам. Примером такого механизма служат ключевые слова, повсеместно используемые в научных журналах. При этом они могут быть плохо выбраны (слишком специфичные или наоборот слишком абстрактные; вместо употребимого ключевого слова авторы используют менее распространенный синоним и т. д.) или отсутствовать вовсе.

Решением в такой ситуации могла бы стать автоматическая разметка ключевыми словами. Однако у подавляющего числа статей текст не является общедоступным, при этом почти всегда доступны тексты аннотаций. Они обладают рядом недостатков, таких как размер текста и доля ключевых слов, встречающихся в тексте в явном виде. Поэтому мы решили исследовать, насколько хорошо можно сгенерировать ключевые слова на таком материале. Поскольку нам не удалось найти крупных корпусов аннотаций, было принято решение собрать свой корпус и проанализировать на нем широко известные методы генерации ключевых слов.

1. Ключевые слова

Ключевые слова представляют собой одно- или многокомпонентные лексические группы, которые отражают основное содержание документа [Шереметьева, Осминин, 2015]. Автоматическое извлечение ключевых слов – важная задача в области обработки естественного языка, инструменты для решения которой могут выступать в роли необходимого компонента систем информационного поиска в различных предметных областях. В частности, список ключевых слов является обязательным элементом текста научной статьи, который позволяет упрощать ее поиск и систематизацию и, следовательно, оказывает влияние на видимость статьи научному сообществу и ее цитируемость [Тихонова, Косычева, 2021; Ghanbarpour, Naderi, 2019].

С точки зрения характера используемых алгоритмов подходы к извлечению ключевых слов делятся на методы обучения без учителя и с учителем (unsupervised и supervised learning соответственно). В ходе обучения без учителя тексты представляются в виде наборов признаков (например, словам могут быть сопоставлены их частотности), после чего производится ранжирование слов текста на основании выделенных признаков и выбор нужного количества слов, имеющих наивысший ранг. К этому классу методов относятся, в частности, статистические алгоритмы (например, TF-IDF, KPMiner [El-Beltagy, Rafea 2009], YAKE! [Campos et al., 2020]) и алгоритмы, основанные на построении графов (TextRank [Mihalcea, Tarau, 2004], TopicRank [Bougouin et al., 2013]). В качестве признаков могут быть использованы векторы, полученные из современных нейросетевых моделей (к примеру, метод KeyBERT [Grootendorst, 2020]).

Методы обучения с учителем предполагают наличие обучающей выборки для настройки алгоритма. Обучающая выборка состоит из текстов, из которых извлекаются ключевые слова, и эталонного списка ключевых слов, составленного, как правило, специалистом-экспертом. На основании этой выборки производится выбор параметров алгоритма, чтобы впоследствии он мог использоваться для извлечения ключевых слов из других текстов. Один из широко используемых методов обучения с учителем – KEA [Witten et al., 1999], использующий наивный байесовский классификатор для определения вероятности того, что слово является ключевым. Ряд методов основан на применении нейронных сетей [Meng et al., 2017; Chen et al., 2020]. Традиционные методы демонстрируют достаточно высокие результаты на англоязычных корпусах текстов, однако обладают рядом ограничений. В частности, большинство существующих методов способны извлечь только те ключевые слова, которые в явном виде присутствуют в исходном тексте. На практике же списки ключевых слов могут включать слова или фразы, напрямую не употребляющиеся в тексте (гиперонимы, синонимы и т. д.). Кроме того, методы обучения с учителем требуют наличия обучающей выборки, формирование которой может быть затруднено для малоресурсных языков. Тестирование методов обучения без учителя так же требует наличия размеченного текстового корпуса для оценки качества поиска ключевых слов.

Есть ряд исследований, посвященных адаптации существующих методов извлечения ключевых слов для русского языка [Sandul, Mikhailova, 2018; Sokolova et al., 2018; Wienecke, 2020; Koloski et al.,2021]. В перечисленных работах различаются используемые метрики и тематика статей, поэтому сравнивать полученные в них результаты затруднительно. Для получения объективных данных о результативности методов требуется проанализировать их результаты на материале русскоязычного корпуса научных текстов из различных источников.

2. Эксперимент

2.1. Данные

Для исследования был подготовлен датасет [Morozov, Glazkova, 2022], содержащий аннотации, заголовки, ключевые слова, год публикации и названия журналов. Данные были собраны из интернет-ресурсов «Киберленинка»[1] (7 044 статей из 421 журнала) и MathNet[2] (3 605 статей из пяти журналов). Для эксперимента мы исключили из рассмотрения те статьи, в которых приведено менее пяти ключевых слов. Краткое описание обеих частей собранного датасета приведено в таблице 1.

Таблица 1

Краткая характеристика использованных датасетов

Table 1

Brief Description of the Used Datasets

	Киберленинка	**MathNet**	**Совокупность**
Статей	5 433	3 091	8 524
Журналов	421	5	426
Среднее количество ключевых слов на статью	9,22	10,55	9,70
Суммарное количество ключевых слов	50 100	32 619	82 719
Количество уникальных ключевых слов	12 333	10396	18 565
Средняя доля ключевых слов, встречающихся в тексте аннотаций	42,41 %	43,64 %	42,86 %

2.2. Методы

Для проведения сравнения мы выбрали ряд алгоритмов, используемых в задаче выделения ключевых слов: TF-IDF, YAKE!, KEA, KeyBERT, TopicRank. KEA – это обучение с учителем (supervised), все остальные – без учителя (unsupervised). Среди алгоритмов без учителя можно выделить три группы: статистические (TfIdf, YAKE!), графовые (TopicRank) и нейросетевые (KeyBERT).

Алгоритм TF-IDF (от **T**erm **F**requency – **I**nverse **D**ocument **F**requency) часто применяется во многих задачах компьютерной лингвистики. Идея этого алгоритма состоит в выделении относительно редких слов, встречающихся в документе чаще, чем в среднем по корпусу. Для каждого слова текста вычисляется частота употребляемости в тексте, равная отношению числа вхождений слова в текст к общему числу слов текста (TF). Далее для этого слова вычисляется доля текстов корпуса, в которых встречается это слово, и от полученного значения вычисляется логарифм обратного к нему (IDF). Итоговый вес слова равен произведению TF и IDF. К примеру, если в тексте слово «ключевой» встретилось 15 раз, объем текста равен 1 000 слов, а всего это слово встретилось в 12 из 100 текстов корпуса, значение TF-IDF будет равным:

[1] https://cyberleninka.ru/.

[2] https://www.mathnet.ru/.

$$\frac{15}{1000} \times log\left(\frac{100}{12}\right) = 0{,}046$$

Кандидатами на роль ключевых слов становятся слова с наибольшим вычисленным значением среди всех слов текста.

Алгоритм YAKE! (Yet Another Keyword Extractor) [Campos et al., 2020] использует для определения значимости слова набор вычислимых эвристик: вероятность написания слова с большой буквы, наиболее вероятную позицию слова в предложении (предполагается, что ключевые слова чаще стоят в начале), частота употребления слова в тексте, долю предложений, содержащих слово, а также разнообразие слов, которые могут стоять до и после исследуемого. Более подробно описание формул, по которым эти эвристики вычисляются и агрегируются в одно значение, приведено в оригинальной статье.

TopicRank [Bougouin, 2013] в отличие от двух предыдущих алгоритмов является графовым. Все последовательности идущих подряд прилагательных и существительных в тексте обозначаются как потенциальные ключевые фразы, а затем объединяются в группы близких по семантике с использованием алгомеративной иерархической кластеризации. На следующем шаге среди получившихся кластеров выбираются наиболее значимые при помощи известного алгоритма PageRank [Page et al., 1998], лежавшего в основе поисковой системы Google.

Среди исследованных алгоритмов единственный, использующий так называемое обучение с учителем, – KEA (Keyphrase Extraction Algorithm) [Witten et al., 1999]. В ходе его работы обучается наивный байесовский классификатор, выделяющий ключевые слова на основании метрики TF-IDF и некоторых численных характеристик (например, предположительной позиции внутри предложения) среди всего текста.

Алгоритм KeyBERT [Grootendorst, 2020] представляет собой нейросетевой алгоритм на основе алгоритма BERT [Devlin et al., 2019], появившегося в 2019 году. BERT обучается, предсказывая пропущенные в предложении слова и прогнозируя, является ли второе предложение продолжением первого. Таким образом можно обучать модели на гигантских моноязычных корпусах без дополнительной разметки, а затем дообучать их на конкретных задачах. Благодаря очень высокому качеству и легко достигаемому результату BERT быстро стал одним из наиболее широко применяемых алгоритмов в обработке естественного языка. Принцип работы KeyBERT заключается в поиске слова или словосочетания, чей семантический вектор (эмбеддинг), полученный из модели BERT, больше всего схож с эмбеддингом текста в целом. Преимущество этого алгоритма заключается в возможности выбора из широкого списка предобученных BERT-моделей.

Для экспериментов мы использовали реализацию алгоритма KeyBERT из библиотеки KeyBERT[3], предобученную BERT-модель rubert_base_cased [Kuratov, Arkhipov, 2019] и реализацию алгоритмов YAKE!, KEA и TopicRank из библиотеки PKE[4] [Boudin, 2016], адаптировав их для русского языка и конкретного набора данных. Для корректной работы алгоритмов из библиотеки PKE используется spacy[5]-модель языка, в нашем случае – ru_core_news_lg. В качестве кандидатов на роль ключевых слов рассматривались 1-, 2- и 3-граммы.

2.3. Метрики

Для оценки качества мы использовали три метрики: F-мера, ROUGE-1 [Lin, 2004] и BERTScore [Zhang et al., 2019]. Для того чтобы исключить проблемы, связанные с согласованием, мы нормализовали все слова в каждой ключевой фразе при помощи библиотеки pymorphy2 [Korobov, 2015].

[3] https://github.com/MaartenGr/KeyBERT.

[4] https://github.com/boudinfl/pke.

[5] https://spacy.io/.

F-мера вычисляется с помощью показателей точности и полноты для двух списков ключевых слов: полученного с помощью алгоритма и составленного вручную автором текста. При этом точность определяется как доля ключевых слов (*n*-грамм) из списка ключевых слов, полученных с помощью алгоритма, которые присутствуют в списке, составленном автором текста вручную. Полнота представляет собой долю найденных алгоритмом ключевых слов относительно всех ключевых слов, подобранных вручную. Значения точности и полноты вычисляются на основе таблицы для каждой *n*-граммы текста:

Таблица 2

Методика вычисления точности и полноты

Table 2

Methodology for Evaluating Precision and Recall

N-грамма текста		**Авторский (экспертный) выбор ключевых слов**	
		Входит в список ключевых слов, составленный вручную	Не входит в список ключевых слов, составленный вручную
Автоматический выбор ключевых слов	Входит в список ключевых слов, полученный с помощью алгоритма	*TP*	*FP*
	Не входит в список ключевых слов, полученный с помощью алгоритма	*FN*	*TN*

TP в таблице 2 означает истинно положительное решение, то есть случай, когда *n*-грамма входит в список ключевых слов, составленный вручную, и определяется алгоритмом как ключевое слово. *TN* – истинно отрицательное решение (*n*-грамма не является ключевым словом и не определяется таковым с помощью алгоритма). *FP* – ложноположительное решение (*n*-грамма не входит в список ключевых слов, составленный вручную, однако распознается алгоритмом как ключевое слово). *FN* – ложноотрицательное решение (*n*-грамма входит в список ключевых слов, составленный вручную, но алгоритм не считает ее ключевым словом).

На основании составленной таблицы точность (precision), полнота (recall) и F-мера (F1-score) вычисляются по формулам:

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$

$$F1_{score} = \frac{2 \times Precision \times Recall}{Precision+Recall}.$$

Метрика ROUGE-1 показывает степень сходства униграмм (отдельных слов) авторского и полученного списков ключевых слов и вычисляется по принципу нахождения F-меры:

$$ROUGE_1 = \frac{2 \times ROUGE_1(Recall) \times ROUGE_1(Precision)}{ROUGE_1(Recall)+ROUGE_1(Precision)}.$$

Значение показателя полноты для данной метрики (ROUGE-1 (Recall)) показывает, какая доля отдельных слов, полученных алгоритмом, присутствует в списке ключевых слов, составленном вручную. Точность метрики ROUGE-1 (ROUGE-1 (Precision)) характеризует долю отдельных слов, входящих в список ключевых слов, составленный вручную, относительно общего количества слов, полученных алгоритмом. Для вычисления точности и полноты ис-

пользуются выражения, аналогичные формулам вычисления точности и полноты для F-меры, логика расчета значений *TP*, *TN*, *FP* и *FN* для каждой униграммы текста представлена в таблице 3.

Таблица 3

Методика вычисления метрики ROUGE

Table 3

Method for Evaluating the ROUGE Metric

Униграмма текста		**Авторский (экспертный) выбор ключевых слов**	
		Входит в список униграмм, содержащихся в ключевых словах, составленных вручную	Не входит в список униграмм, содержащихся в ключевых словах, составленных вручную
Автоматический выбор ключевых слов	Входит в список униграмм, содержащихся в ключевых словах, полученных с помощью алгоритма	*TP*	*FP*
	Не входит в список униграмм, содержащихся в ключевых словах, полученных с помощью алгоритма	*FN*	*TN*

В отличие от ранее описанных метрик BERTScore оценивает не наличие точных совпадений в списках ключевых слов, составленных вручную и машинным способом, а семантическую близость между ними. Метрика использует векторные представления документов (document embeddings), полученные из современных контекстуализированных лингвистических моделей [Devlin et al., 2019]. За счет предварительного обучения лингвистических моделей на больших текстовых корпусах полученные представления документов могут использоваться в широком спектре задач обработки естественного языка. BERTScore получает представление двух текстов (списков ключевых слов в нашем случае) из лингвистической модели, после чего попарно оценивает близость токенов текстов с помощью меры косинусного сходства. На основании дистрибутивной гипотезы [Harris, 1954] расстояние между близкими по смыслу токенами будет меньше, чем между более далекими по смыслу. Полученные показатели близости используются для расчета значения метрики с помощью точности и полноты совпадений токенов по принципу F-меры. Текст разбивается на токены с помощью токенизатора выбранной лингвистической модели. В данной работе мы использовали мультиязычную модель BERT[6] для получения векторных представлений текстов.

2.4. Результаты

В таблицах 4–6 приведены значения метрик, получившиеся при сравнении сгенерированных ключевых слов с авторскими. Для каждой пары «алгоритм – датасет» приведены три значения: метрики для 5, 10 и 15 сгенерированных слов. Для каждой метрики и каждого набора данных лучший результат выделен полужирным шрифтом.

[6] https://huggingface.co/bert-base-multilingual-cased.

Таблица 4

Вычисленное значение F-меры (%)

Table 4

F1-Value (%)

Алгоритм	**Киберленинка**			**MathNet**			**Совокупность**		
	Количество сгенерированных слов								
	5	10	15	5	10	15	5	10	15
TF-IDF	3,84	2,99	2,39	5,63	4,84	4,03	4,49	3,66	2,99
YAKE!	2,94	5,34	6,46	3,28	5,38	6,36	3,06	5,35	6,43
KEA	4,32	4,34	4,58	3,30	4,34	4,34	3,93	4,34	4,48
TopicRank	4,79	5,07	5,03	5,13	5,15	4,90	4,91	5,10	4,98
KeyBERT	1,78	2,51	3,08	1,38	1,96	2,24	1,64	2,31	2,78

Таблица 5

Вычисленное значение ROUGE-1, %

Table 5

ROUGE-1 Values, %

Алгоритм	**Киберленинка**			**MathNet**			**Совокупность**		
	Количество сгенерированных слов								
	5	10	15	5	10	15	5	10	15
TF-IDF	17,02	17,16	16,73	17,01	17,30	16,48	17,01	17,21	16,64
YAKE!	17,69	21,22	21,40	16,14	19,75	20,53	17,13	20,68	21,09
KEA	12,77	15,12	16,78	13,76	16,35	17,31	13,12	15,56	16,98
TopicRank	21,24	22,41	22,40	20,31	21,27	20,87	20,90	21,99	21,85
KeyBERT	14,81	16,31	16,79	13,18	14,58	14,79	14,22	15,69	16,06

Таблица 6

Вычисленное значение BERTScore, %

Table 6

BERTScore Values, %

Алгоритм	**Киберленинка**			**MathNet**			**Совокупность**		
	Количество сгенерированных слов								
	5	10	15	5	10	15	5	10	15
TF-IDF	68,14	66,55	64,03	68,45	67,23	65,74	68,25	66,80	64,65
YAKE!	69,23	68,24	67,04	68,82	67,89	66,93	69,08	68,11	67,00
KEA	68,73	66,23	65,59	69,24	67,62	66,62	68,91	66,75	65,97
TopicRank	73,60	73,65	73,61	73,13	73,03	72,77	73,43	73,43	73,30
KeyBERT	68,87	67,93	66,96	68,66	67,95	66,95	68,80	67,94	66,96

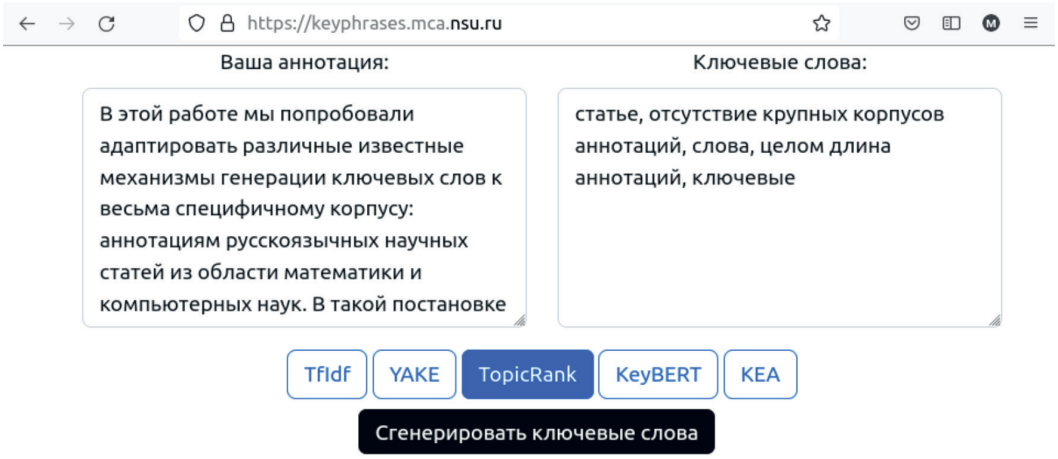
Использованные нами метрики оценивают качество различных аспектов подбора ключевых слов. С позиции применения F-меры лучшие результаты были получены с помощью алгоритма YAKE!. Это значит, что в наших экспериментах YAKE! лучше остальных методов спра-

вился с генерацией ключевых слов, точно совпадающих со словами, подобранными вручную. Наиболее высокий результат по метрике ROUGE-1 был достигнут алгоритмом TopicRank. Напомним, что ROUGE-1 оценивает количество совпадающих униграмм в составленном вручную списке ключевых слов и списке, полученном с помощью компьютерных методов. TopicRank также показал лучшее качество по полученным значениям метрики BERTScore, оценивающей семантическое сходство текстов. Алгоритм KeyBERT, использующий векторные представления документов, полученные из современных лингвистических моделей, продемонстрировал сравнительно высокое качество с позиции метрики BERTScore, однако показал самое низкое качество среди рассмотренных алгоритмов в большинстве случаев для других метрик. Кроме того, для разных метрик лучшие показатели были достигнуты при различном количестве ключевых слов (15 для F-меры, 10 для ROUGE-1, 5 и 10 для BERTScore).

Наши эксперименты показывают, что выбор алгоритма автоматического подбора ключевых слов зависит от специфики задачи, стоящей перед автором текста, и авторских предпочтений. Некоторые алгоритмы лучше справляются с извлечением ключевых слов, присутствующих в тексте в явном виде, а другие направлены на генерацию ключевых слов, семантически полно описывающих текст, но включающих в себя меньшее количество подстрок исходного текста. Кроме того, большинство существующих алгоритмов не способно самостоятельно определять необходимое количество ключевых слов. Выбор данного параметра также производится автором текста.

2.5. Сервис автоматического подбора ключевых слов

Для демонстрации того, как работают представленные в статье алгоритмы, мы разработали и выложили в публичный доступ сервис KeyPhrases[7] (интерфейс представлен на рис.), который по аннотации генерирует при помощи выбранного алгоритма ключевые слова. Несмотря на ряд очевидных недостатков, среди которых сразу бросается в глаза ненормализованность генерируемых фраз, сервис позволяет получить представление о том, насколько хорошо можно автоматизировать выбор ключевых слов. К примеру, ключевые слова для настоящей статьи были сгенерированы именно так. Сервис будет интересен и для начинающих исследователей, которые оформляют свою первую заявку на конференцию и не могут подобрать удачные ключевые слова.



Пример работы сервиса
Developed web service

[7] https://keyphrases.mca.nsu.ru/.

Заключение

Выбор метода автоматического подбора ключевых слов в первую очередь подразумевает определение требований к их специфике и параметрам алгоритма. Наши эксперименты показали, что качество и приоритет тех или иных алгоритмов зависит от того, какую метрику мы выбираем для их оценки. Так, некоторые методы лучше справляются с выделением словосочетаний, наиболее похожих на ключевые слова, подобранные вручную, а другие более успешно формируют список ключевых слов, семантически полно описывающих содержание текста. При этом большинство широко используемых методов обладает рядом ограничений. В частности, количество ключевых слов и максимальную длину n-граммы, которая может быть признана ключевым словом, требуется задавать вручную. Кроме того, методы, рассмотренные в данной статье, основаны на извлечении ключевых слов из исходного текста, однако на практике ключевые слова часто представляют собой обобщенные понятия или синонимы понятий, упоминаемых в тексте, что подтверждается данными статистического анализа собранного нами корпуса. Рассмотренные алгоритмы не способны справиться с задачей генерации ключевых слов, отсутствующих в тексте в явном виде. Реализация и адаптация алгоритмов генерации ключевых слов, напрямую не встречающихся в тексте, для русского языка является одним из перспективных направлений дальнейших исследований.

Другое наблюдение, сделанное в ходе анализа собранного корпуса, заключается в том, что ключевые слова, подобранные авторами текстов, часто являются уникальными. Так, каждое отдельное ключевое слово встречается в корпусе в среднем реже, чем три раза, то есть многие слова встречаются только в одном-двух текстах. Подобранные таким образом ключевые слова не могут быть непосредственно использованы для систематизации научных текстов в электронных библиотеках, поскольку дают представление скорее о методологии конкретной статьи, чем о ее предметной области. В таком случае использование ключевых слов для автоматического упорядочивания текстов требует дополнительных знаний о структуре предметной области или анализа других элементов текста статьи и требует дальнейшего изучения. Также стоит отметить, что в рамках данной работы мы рассматривали научные тексты по математике, информационным технологиям и смежным наукам. Вопрос о переносимости сделанных выводов на тексты других предметных областей нуждается в дополнительном исследовании.

Список литературы

Тихонова Е. В., Косычева М. А. Эффективные ключевые слова: стратегии формулирования // Health, Food & Biotechnology. 2021. № 4 (3). С. 7–15.

Шереметьева С. О., Осминин П. Г. Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2015. № 1 (12). С. 76–81.

Boudin F. PKE: an open source python-based keyphrase extraction toolkit // Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations / ed. by H. Watanabe. The COLING 2016 Organizing Committee. 2016. Pp. 69–73.

Bougouin A., Boudin F., Daille B. TopicRank: Graph-based topic ranking for keyphrase extraction // Proceedings of the Sixth International Joint Conference on Natural Language Processing / ed. by R. Mitkov and J. C. Park. Asian Federation of Natural Language Processing. 2013. Pp. 543–551.

Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. 2020. 509. Pp. 257–289.

Chen W., Chan H. P., Li P., King I. Exclusive Hierarchical Decoding for Deep Keyphrase Generation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics / ed. by D. Jurafsky, J. Chai, N. Schluter and J. Tetreault. Association for Computational Linguistics. 2020. Pp. 1095–1105.

Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT / ed. by J. Burstein, C. Doran, T. Solorio. Association for Computational Linguistics. 2019. Pp. 4171–4186.

El-Beltagy S. R., Rafea A. KP-Miner: A keyphrase extraction system for English and Arabic documents // Information Systems. 2009. № 1 (34). Pp. 132–144.

Ghanbarpour A., Naderi H. A model-based method to improve the quality of ranking in keyword search systems using pseudo-relevance feedback // Journal of Information Science. 2019. № 4 (45). Pp. 473–487.

Grootendorst M. KeyBERT: Minimal Keyword Extraction with BERT. 2020 [Электронный ресурс]. URL: http://doi.org/10.5281/zenodo.4461265 (дата обращения: 29.11.2022).

Harris Z. S. Distributional structure // Word. 1954. № 2-3 (10). Pp. 146–162.

Koloski B., Pollak S., Škrlj B., Martinc M. Extending Neural Keyword Extraction with TF-IDF tagset matching // Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation / ed. by H. Toivonen, M. Boggia. Association for Computational Linguistics. 2021. Pp. 22–29.

Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // International conference on analysis of images, social networks and texts / ed. by M. Yu. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V. G. Labunets. Springer, Cham. 2015. Pp. 320–332.

Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. 2019 [Электронный ресурс]. URL: https://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf (дата обращения: 29.11.2022).

Lin C. Y. ROUGE: A package for automatic evaluation of summaries // Text summarization branches out. Association for Computational Linguistics. 2004. Pp. 74–81.

Meng R., Zhao S., Han S., He D., Brusilovsky P., Chi Y. Deep Keyphrase Generation // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. by R. Barzilay, M.-Y. Kan. Association for Computational Linguistics. 2017. Pp. 582–592.

Mihalcea R., Tarau Pp. TextRank: Bringing order into text // Proceedings of the 2004 conference on empirical methods in natural language processing / ed. by D. Lin, D. Wu. Association for Computational Linguistics. 2004. Pp. 404–411.

Morozov D., Glazkova A. Keyphrases CS&Math Russian, Mendeley Data. 2022 [Электронный ресурс]. URL: http://doi.org/10.17632/dv3j9wc59v.1 (дата обращения: 29.11.2022).

Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab. 1998 [Электронный ресурс]. URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf (дата обращения: 02.12.2022).

Sandul M., Mikhailova E. Keyword extraction from single Russian document // Proceedings of the Third Conference on Software Engineering and Information Management (full papers) / ed. by Y. Litvinov, M. Akhin, B. Novikov ,V. Itsykson. CEUR Workshop Proceedings, 2018. Pp. 30–36.

Sokolova E., Moskvina A., Mitrofanova O. Keyphrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithms // Data analytics and management in data intensive domains: Proceedings of the XX International Conference – DAMDID/RCDL’2018 / ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. FRC CSC RAS, 2018. Pp. 369–372.

Wienecke Y. Automatic Keyphrase Extraction From Russian-Language Scholarly Papers in Computational Linguistics: University Honors Theses. Portland State University, 2020. 36 p.

Witten I. H., Paynter G. W., Frank E., Gutwin C., Nevill-Manning C. G. KEA: Practical automatic keyphrase extraction // Proceedings of the fourth ACM conference on Digital libraries / ed. by N. Rowe, E. A. Fox. Association for Computing Machinery, 1999. Pp. 254–255.

Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT // International Conference on Learning Representations. 2019 [Электронный ресурс]. URL: https://openreview.net/pdf?id=SkeHuCVFDr (дата обращения: 29.11.2022).

References

Boudin, F. PKE: an open source python-based keyphrase extraction toolkit. Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations. Osaka, Japan, 2016, pp. 69–73.

Bougouin, A., Boudin, F., Daille, B. TopicRank: Graph-based topic ranking for keyphrase extraction. Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan, 2013, pp. 543–551.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 2020, 509, pp. 257–289.

Chen, W., Chan, H. P., Li, P., King, I. Exclusive Hierarchical Decoding for Deep Keyphrase Generation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020, pp. 1095–1105.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. Minneapolis, USA, 2019, pp. 4171–4186.

El-Beltagy, S. R., Rafea, A. KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information Systems*, 2009, no. 1 (34), pp. 132–144.

Ghanbarpour, A., Naderi, H. A model-based method to improve the quality of ranking in keyword search systems using pseudo-relevance feedback. *Journal of Information Science*, 2019, no. 4 (45), pp. 473–487.

Grootendorst, M. KeyBERT: Minimal Keyword Extraction with BERT, 2020. Available at: http://doi.org/10.5281/zenodo.4461265 (accessed 29.11.2022).

Harris, Z. S. Distributional structure. Word, 1954. no. 2-3 (10), pp. 146–162.

Koloski, B., Pollak, S., Škrlj, B., Martinc, M. Extending Neural Keyword Extraction with TF-IDF tagset matching. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Online, 2021, pp. 22–29.

Korobov, M. Morphological analyzer and generator for Russian and Ukrainian languages. International conference on analysis of images, social networks and texts. Yekaterinburg, 2015, pp. 320–332.

Kuratov, Y., Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". Moscow, 2019. Available at: https://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf (accessed 29.11.2022).

Lin C. Y. ROUGE: A package for automatic evaluation of summaries. Text summarization branches out. Osaka, Japan, 2004, pp. 74–81.

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., Chi, Y. Deep Keyphrase Generation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017, pp. 582–592.

Mihalcea, R., Tarau, P. TextRank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing. Barcelona, Spain, 2004, pp. 404–411.

Morozov, D., Glazkova, A. Keyphrases CS&Math Russian, Mendeley Data, 2022. Available at: http://doi.org/10.17632/dv3j9wc59v.1 (accessed 29.11.2022).

Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the web, Stanford InfoLab, 1998. Available at: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf (accessed 02.12.2022).

Sandul, M., Mikhailova, E. Keyword extraction from single Russian document. Proceedings of the Third Conference on Software Engineering and Information Management (full papers). Saint Petersburg, 2018, pp. 30–36.

Sheremetyeva, S. O., Osminin, P. G. [On Methods and Models of Keywords Automatic Extraction]. *Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Lingvistika* [Bulletin of South Ural State University, Series «Linguistics»], 2015, no. 1 (12), pp. 76–81. (In Russ.)

Sokolova, E., Moskvina, A., Mitrofanova, O. Keyphrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithms. Data analytics and management in data intensive domains: Proceedings of the XX International Conference – DAMDID/RCDL'2018. Moscow, 2018, pp. 369–372.

Tikhonova, E. V., Kosycheva, M. A. Effective Keywords: Strategies for Their Formulation. *Health, Food & Biotechnology*, 2021, no. 4 (3), pp. 7–15. (In Russ.)

Wienecke, Y. Automatic Keyphrase Extraction From Russian-Language Scholarly Papers in Computational Linguistics: University Honors Theses. Portland State University, 2020. 36 p.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G. KEA: Practical automatic keyphrase extraction. Proceedings of the fourth ACM conference on Digital libraries. Berkeley, USA, 1999, pp. 254–255.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations. Online, 2019 Available at: https://openreview.net/pdf?id=SkeHuCVFDr (accessed 29.11.2022).

Информация об авторах

Морозов Дмитрий Алексеевич, младший научный сотрудник Лаборатории прикладных цифровых технологий Международного математического центра Новосибирского национального исследовательского государственного университета

Глазкова Анна Валерьевна, канд. тех. наук, доцент кафедры программного обеспечения Института математики и компьютерных наук Тюменского государственного университета

Тютюльников Михаил Андреевич, инженер, Лаборатория прикладных цифровых технологий Международного математического центра Новосибирского национального исследовательского государственного университета

Иомдин Борис Леонидович, канд. филол. наук, ведущий научный сотрудник Института русского языка им. В. В. Виноградова РАН

Information about the Authors

Dmitry A. Morozov, junior researcher, Laboratory of Applied Digital Technologies, Mathematical Center in Akademgorodok, Novosibirsk State University

Anna V. Glazkova, Cand. Sc. (Technology), Associate Professor, Department of Software, Institute of Mathematics and Computer Science of the University of Tyumen

Mikhail A. Tyutyulnikov, engineer, Laboratory of Applied Digital Technologies, Mathematical Center in Akademgorodok, Novosibirsk State University

Boris L. Iomdin, Cand. Sc. (Philology), Leading Researcher at the Vinogradov Russian Language Institute of the Russian Academy of Sciences

Статья поступила в редакцию 12.12.2022;
одобрена после рецензирования 11.01.2023; принята к публикации 13.01.2023

The article was submitted 12.12.2022;
approved after reviewing 11.01.2023; accepted for publication 13.01.2023