УДК 004.89 DOI 10.18522/2311-3103-2024-5-143-151

Ж. Мохаммад

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ФРАЗ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Статья посвящена актуальной проблеме извлечения ключевых фраз из текстов на естественном языке, что является критически важной задачей в области обработки естественного языка и интеллектуального анализа текста. В ней подробно рассматриваются основные подходы к извлечению ключевых фраз (ключевых слов), включая как традиционные методы, так и современные подходы на основе искусственного интеллекта. В статье рассматривается набор широко используемых методов в этой области, таких как TF-IDF, RAKE, YAKE и методы, основанные на лингвистических анализаторах (парсерах). Эти методы опираются на статистические принципы и графовые структуры, но часто сталкиваются с проблемами, связанными с недостаточной способностью учитывать контекст текста. Большая языковая модель GPT-3 демонстрирует превосходное понимание контекста по сравнению с традиционными методами извлечения ключевых фраз. Эта продвинутая способность позволяет GPT-3 более точно идентифицировать и извлекать релевантные ключевые фразы из текста. Сравнительный анализ с использованием эталонного набора данных Inspec показывает значительно более высокую производительность GPT-3 с точки зрения средней точности (Mean Average Precision, MAP). Однако следует отметить, что, несмотря на высокую точность и качество извлечения, использование больших языковых моделей может быть ограничено в реальном времени из-за их более длительного времени отклика по сравнению с классическими статистическими методами. Таким образом, статья подчеркивает необходимость дальнейших исследований в этой области для оптимизации алгоритмов извлечения ключевых фраз с учетом требований реального времени и контекста текстов.

Извлечение ключевых слов; извлечение ключевых фраз; LLM; TF-IDF; большие языковые модели; GPT-3.

J. Mohammad

KEYPHRASE EXTRACTION BASED ON LARGE LANGUAGE MODELS

The article addresses the current problem of extracting key phrases from natural language texts, which is a critical task in the field of natural language processing and text mining. It examines in detail the main approaches to extracting key phrases (keywords), including both traditional methods and modern approaches based on artificial intelligence. The paper discusses a set of widely used methods in this field, such as TF-IDF, RAKE, YAKE, and linguistic parser-based methods. These methods are based on statistical principles and/or graph structures, but they often face problems related to their insufficient ability to take into account the context of the text. The GPT-3 large language model demonstrates superior contextual understanding compared to traditional methods for key phrase extraction. This advanced capability allows GPT-3 to more accurately identify and extract relevant key phrases from text. The comparative analysis using the Inspec benchmark dataset reveals GPT-3's significantly higher performance in terms of Mean Average Precision (MAP@K). However, it should be noted that despite high accuracy and extraction quality, the use of large language models may be limited in real-time applications due to their longer response time compared to classical statistical methods. Thus, the article emphasizes the need for further research in this area to optimize key phrase extraction algorithms, taking into account real-time requirements and text context.

Keyword extraction; key phrases extraction; LLMs; TF-IDF; Large language models; GPT-3.

Введение. В быстро развивающейся области обработки естественного языка (Natural Language Processing, NLP) извлечение ключевых фраз из текста является важной задачей с широким спектром применений, от классификации текстов до повышения эффективности поисковой оптимизации [1]. Извлечение ключевых фраз включает идентификацию наиболее значимых терминов или выражений в тексте, которые отражают его основные темы и концепции [2].

Традиционные подходы извлечения ключевых фраз включают как контролируемые, так и неконтролируемые методы. Контролируемые методы полагаются на аннотированные наборы данных для обучения моделей, способных идентифицировать ключевые фразы, в то время как неконтролируемые методы, такие как TF-IDF и тематическое моделирование, используют статистические меры для различения важных фраз без необходимости в размеченных данных. Эти подходы были инструментальными во многих приложениях, предлагая баланс простоты и эффективности [3].

Однако появление больших языковых моделей (англ. Large Language Models, LLM) привело к значительным изменениям в области обработки естественного языка. Модели, такие как GPT и BERT, продемонстрировали беспрецедентные возможности в понимании и генерации текста, похожего на человеческий, благодаря своим сложным архитектурам и обширной подготовке на разнообразных наборах данных [4].

LLM используют методы глубокого обучения и огромные объемы данных для захвата сложных моделей языка и контекстуальных нюансов, что делает их исключительно способными извлекать ключевые фразы с высокой точностью. Их способность понимать контекст и семантику превосходит традиционные методы, что позволяет проводить более точные и контекстуально релевантные извлечения.

В поддержку данной теоретической информации в статье представлен сравнительный анализ группы методов извлечения ключевых фраз. Кроме того, их производительность сопоставляется с эффективностью больших языковых моделей.

Аналитический обзор методов извлечения ключевых фраз. Методы извлечения ключевых фраз можно условно классифицировать на контролируемые и неконтролируемые, каждый из которых обладает уникальными приемами и областями применения [5, 6]. Ниже, представлено описание основных подходов, а также наиболее значимых методов и инструментов, используемых в данной области.

1. Неконтролируемые подходы. Методы данного подхода не требуют аннотированных обучающих данных и основываются на неотъемлемых свойствах текста. Ключевые методы включают:

А. Методы, основанные на частоте. Метод TF-IDF (Term Frequency-Inverse Document Frequency) вычисляет важность термина, учитывая его частоту в документе относительно его частоты в более крупном корпусе. Термины, которые часто встречаются в документе, но редко в других, считаются значимыми. Этот метод эффективен, но может упустить важные фразы, которые встречаются нечасто [7]. Другим примером этих методов являются *YAKE* и *RAKE* [8, 9]. Метод *YAKE* (Yet Another Keyword Extractor) основан на анализе текста без использования стоп-слов[1] (англ. stop-words), что позволяет выделять значимые слова и фразы. Для оценки важности слов *YAKE* применяет комбинацию пяти метрик: нормированная частота, местоположение в тексте, число предложений с выражением, число капитализированных употреблений и сходство со стоп-словами. Это делает его более гибким и точным по сравнению с другими методами, такими как *RAKE*, который оценивает ключевые слова на основе частоты, совместимости и других простых метрик.

Б. Методы на основе графов. Эти методы рассматривают документы как узлы в графе и используют алгоритмы для определения центральных или влиятельных узлов, которые соответствуют ключевым фразам. Методы, такие как PageRank и меры центральности, можно использовать для оценки и извлечения наиболее важных фраз. Такие методы, как TextRank, используют графовые структуры для представления связей между словами и фразами. Ранжируя фразы на основе их связности и важности в графе, данный метод может эффективно идентифицировать ключевые фразы [10, 11].

[1] Стоп-слова – это часто встречающиеся слова в языке, такие как "и/and", "в/in", "на/on", "с/with", которые не несут значительной смысловой нагрузки.

B. Методы на основе вероятности. Эти методы используют вероятностные модели и статистический анализ для определения ключевых фраз. Скрытое распределение Дирихле (англ. Latent Dirichlet allocation, LDA) – один из самых известных примеров этих методов. Это статистическая модель, которая предполагает, что каждый документ представляет собой смесь различных тем, а каждая тема характеризуется распределением слов [12, 13].

2. Контролируемые подходы. Методы этого подхода используют алгоритмы машинного обучения для классификации фраз на ключевые и неключевые. Эти методы, как правило, обеспечивают более высокую точность, однако требуют значительных затрат времени и ресурсов для подготовки обучающих данных [14].

3. Гибридные подходы. Объединение нескольких методов часто может привести к более точному извлечению ключевых фраз. Например, использование TF-IDF для создания начального списка фраз-кандидатов, а затем применение модели машинного обучения для уточнения и ранжирования этих фраз.

4. Лингвистические методы. Разметка частей речи (англ. Part-of-Speech Tagging, POS) является фундаментальным методом в обработке естественного языка (NLP). Он присваивает грамматические категории словам в тексте, таким как существительные, глаголы, прилагательные и наречия. Этот метод особенно полезен при извлечении ключевых фраз, где цель состоит в том, чтобы определить значимые фразы, которые отражают главные идеи документа [15, 16]. Ниже описаны некоторые приложения POS при извлечении ключевых фраз:

Идентификация фразы-кандидата. POS-теги помогают идентифицировать потенциальных кандидатов на роль ключевых фраз, фильтруя слова на основе их грамматических ролей. Например, именные фразы (комбинации прилагательных и существительных) часто являются целевыми, поскольку они обычно представляют ключевые концепции. Применяя определенные шаблоны POS, такие, как «прилагательное + существительное», процесс извлечения может сосредоточиться на более значимых фразах, уменьшая шум от нерелевантных слов.

Фильтрация нерелевантных слов. Используя POS-теги, процесс извлечения может исключить слова, которые с меньшей вероятностью будут способствовать значению текста, такие как стоп-слова (например, «and», «the»). Это повышает качество извлекаемых ключевых фраз, гарантируя, что рассматриваются только значимые термины, тем самым улучшая общую релевантность результатов [17].

Распознавание именованных сущностей (англ. Named Entity Recognition, NER). POS-тегирование может помочь в идентификации именованных сущностей, которые часто имеют решающее значение для извлечения ключевой фразы. Помечая слова как собственные имена, организации или местоположения, процесс извлечения может расставить приоритеты для этих сущностей [18].

Извлечение словосочетаний (англ. Collocation Extraction). POS-теги могут использоваться для идентификации словосочетаний – фраз, которые часто встречаются вместе. Анализируя POS-теги соседних слов, система извлечения может распознавать общие фразы, которые могут быть важны для понимания контекста и тем текста [19, 20].

Улучшение семантической осведомленности. Недавние подходы интегрировали POS-теги с семантической информацией для повышения производительности методов извлечения ключевых фраз. Учитывая типы слов, извлеченных из POS-тегов, а также контекстно-зависимые семантические критерии, процесс извлечения может создать более релевантные и контекстуально соответствующие ключевые слова.

5. Большие языковые модели (англ. Large Language Models, LLM). Большие языковые модели представляют собой один из самых передовых подходов к извлечению ключевых фраз. GPT-3, разработанная OpenAI, представляет собой современную большую языковую модель, которая использует методы глубокого обучения для понимания и создания текста, похожего на человеческий. GPT-3 использует архитектуру трансформатора, которая использует механизмы внутреннего внимания для обработки входного текста.

Это позволяет модели учитывать контекст слов по отношению друг к другу. Модель GPT-3 предварительно обучена на обширном корпусе текстовых данных, обучаясь предсказывать следующее слово в предложении. Хотя ее можно точно настроить для определенных задач, она также хорошо работает в условиях нулевого или небольшого количества попыток, генерируя выходные данные на основе минимальных примеров или подсказок [4, 21].

Преимущества использования GPT-3 для извлечения ключевых фраз:

♦ Контекстное понимание: способность GPT-3 интерпретировать контекст позволяет извлекать ключевые фразы, которые более релевантны и значимы по сравнению с традиционными методами, полагающимися исключительно на частоту или статистические показатели.

♦ Обучение с нуля: модель может выполнять извлечение ключевых фраз без необходимости использования аннотированных наборах данных. Это делает ее подходящей для приложений, где данных мало или их трудно маркировать.

♦ Гибкость: GPT-3 может адаптироваться к различным доменам и стилям текста, что делает ее универсальной для разных типов документов.

1. Постановка задачи. Задачу извлечения ключевых фраз можно сформулировать как задачу ранжирования ключевых фраз-кандидатов по степени их релевантности. Математически это можно описать следующим образом:

Пусть дан текст $\boldsymbol{T}$, состоящий из последовательности слов $\{\mathbf{w_1}, \mathbf{w_2}, \mathbf{..., w_n}\}$. $\boldsymbol{P}$ – множество ключевых фраз-кандидатов, извлеченных из $\boldsymbol{T}$. Введена $\boldsymbol{R(p)}$ – функция ранжирования, которая присваивает оценку каждой ключевой фразе-кандидате $\boldsymbol{p} \in \boldsymbol{P}$. Цель состоит в том, чтобы найти топ-$\boldsymbol{k}$ ключевых фраз из $\boldsymbol{P}$, имеющих наивысшие оценки ранжирования согласно $\boldsymbol{R(p)}$. Математически это можно выразить как задачу оптимизации:

$$\max_{p_1,p_2,\ldots,p_k \in P} \sum_{i=1}^{k} R(p_i), \tag{1}$$

где, $\boldsymbol{p_1, p_2, \ldots, p_k}$ – различные элементы множества $\boldsymbol{P}$, а $\boldsymbol{k}$ – желаемое количество извлекаемых ключевых фраз. Функция ранжирования $\boldsymbol{R(p)}$ может учитывать различные характеристики ключевой фразы-кандидата p, такие как: Частота появления $\boldsymbol{p}$ в $\boldsymbol{T}$; Позиция $\boldsymbol{p}$ в $\boldsymbol{T}$ (например, фразы из заголовков/введения могут иметь больший вес); Наличие стоп-слов в $\boldsymbol{p}$; Семантическая связность/релевантность слов в $\boldsymbol{p}$; Длина $\boldsymbol{p}$ (фразы умеренной длины могут быть предпочтительнее). Эти характеристики могут быть объединены в $\boldsymbol{R(p)}$ с использованием методов линейной регрессии, моделей обучения ранжированию или графовых методов, таких как PageRank.

Использование LLM в этом контексте помогает в вычислении более продвинутых характеристик для $\boldsymbol{R(p)}$, таких как семантическая связность и контекстуальная уместность ключевых фраз, что может привести к улучшению качества ранжирования по сравнению с традиционными статистическими методами.

2. Вычислительный эксперимент и анализ полученных результатов. В этом разделе проводится вычислительный эксперимент по оценке эффективности группы методов извлечения ключевых фраз. Данный эксперимент является расширением вычислительного эксперимента, проведенного в предыдущих работах [22, 23], в которых рассматривалась группа таких методов, как TF-IDF, *YAKE*, *RAKE*, $\mathrm{TF_{Spcay}}$, $\mathrm{TF_{stanza}}$ и $\mathrm{TF_{AllenNLP}}$. В данной работе производительность GPT-3 оценивается по сравнению с предыдущими методами при решении задачи извлечения ключевых фраз из документов набора данных *Inspec*.

Метрика оценки. Для оценки эффективности методов извлечения ключевых фраз используется модифицированная версия метрики MAP@K:

$$MAP@K = \frac{1}{Q}\sum_{q \in Q} \frac{1}{\min(m,K)} \sum_{i=1}^{K} P(i) \cdot rel(i), \tag{2}$$

где Q – множество ключевых фраз, определенных экспертами.

m – количество ключевых фраз в документе, которые соответствуют ключевым фразам Q.

P(i) – точность на позиции i в ранжированном списке ключевых фраз документа.

rel(i) – бинарная функция сопоставления, равная 1 если ключевая фраза на позиции i соответствует ключевой фразе q, и 0 иначе.

Таким образом, данная метрика позволяет оценить, насколько хорошо система упорядочивает ключевые фразы в документе в соответствии с ключевыми фразами, определенными экспертами.

3. Реализация и анализ результатов. Для реализации алгоритма на основе GPT-3 была использована бесплатная версия модели GPT-3, позволяющая отправлять 60 запросов каждый час[2].

Программное приложение было разработано с использованием языка программирования Python для взаимодействия с языковой моделью GPT-3 и получения необходимого ответа. Следующий фрагмент кода иллюстрирует часть реализации.

```python
import pandas as pd
from processing import *
from freeGPT import Client
model = "gpt3"
import json

my_prompt = """Extract the top 20 key phrases from the following text,
                ranking them according to their importance,
                and return the response in JSON format: {key phrases: list} """

def get_kws(my_prompt, doc):
    prompt = my_prompt + doc # doc: документ в обработке

    try:
        response = Client.create_completion(model, prompt) # Sending request
        return response
    except Exception as e:
        print(e)
        return [] # return empty list

```

Рис. 1. Листинг кода для извлечения ключевой фразы с помощью GPT-3

Для остальных методов использовалась та же реализация, что и в предыдущей работе автора [23].

В табл. 1 и рис. 2 представлены результаты сравнения различных методов извлечения ключевых фраз с использованием метрики MAP@K, где *K* варьируется от 1 до 20.

GPT-3 демонстрирует впечатляющие результаты в задаче извлечения ключевых фраз. Его производительность, особенно при небольших значениях K, превосходит другие методы, включая TF_{SpaCy} и TF_{Stanza}. Максимальное значение MAP@K равное 0.445 при k=1 указывает на то, что GPT3 очень эффективно в извлечении наиболее важных ключевых фраз. Даже при увеличении k до 10 и 20 результаты остаются относительно высокими, что свидетельствует о его способности захватывать важные фразы с более высокой точностью по сравнению с другими методами.

Высокая производительность GPT-3 может быть объяснена его архитектурой трансформеров и механизмом внимания. Трансформеры позволяют модели обрабатывать последовательность данных параллельно, что значительно увеличивает ее вычислительную мощность. Механизм внимания позволяет модели сосредоточиться на важных частях входных данных, игнорируя менее значимые детали. Это особенно полезно при извлечении ключевых фраз, так как модель может уделить внимание наиболее релевантным словам или фразам.

[2] freeGPT 1.3.5 https://pypi.org/project/freeGPT/.

Однако следует отметить, что использование GPT-3 может быть ресурсоемким и требовать значительных вычислительных мощностей, особенно при работе с большими наборами данных или в реальных приложениях с ограниченными ресурсами.

Таблица 1

Точность алгоритмов, измеренная с помощью MAP@K

	MAP@K			
Методы/ алгоритмы	@1	@5	@10	@20
TF	0,18	0,096	0,058	0,046
TF_{SpaCy}	**0,42**	**0,167**	**0,098**	0,076
TF_{Stanza}	0,24	0,136	0,082	0,063
$TF_{AllenNLP}$	0,25	0,13	0,08	0,06
YAKE	0,25	0,115	0,078	0,075
RAKE	0,23	0,107	0,087	**0,079**
GPT-3	**0,445**	**0,313**	**0,23**	**0,198**

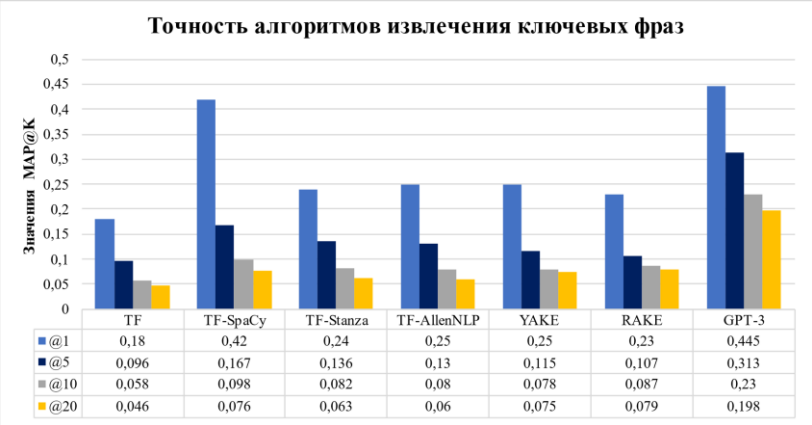

Рис. 2. Результаты извлечения ключевых фраз из набора данных Inspec, измеренные мерой MAP@K

Также отмечается, что методы, которые включают в себя дополнительную обработку с помощью лингвистического парсера (библиотек *SpaCy*, *Stanza* и AllenNLP), демонстрируют лучшие результаты по сравнению с чистым TF или другими методами, такими как *YAKE*, *RAKE*. Это указывает на то, что дополнительная семантическая обработка и анализ контекста значительно улучшают точность и качество извлечения ключевых фраз. Этот вариант может быть более предпочтительным по сравнению с GPT-3, при работе в режиме реального времени.

В результате можно сделать вывод, что использование GPT3 для извлечения ключевых фраз является нормальным, учитывая ее способность создавать связный текст и учитывать контекст. Однако, для задач не требующих таких расширенных возможностей, использование традиционных алгоритмов, таких как *YAKE* или *RAKE*, может быть приемлемым, учитывая вычислительные затраты GPT и время ответа на запрос. В таких случаях, традиционные алгоритмы могут быть более эффективными и экономными в ресурсах, несмотря на чуть ниже качество результатов.

Заключение. В данной статье рассмотрена важная задача извлечения ключевых фраз из текстов на естественном языке. Анализируются основные подходы к этой проблеме, в том числе традиционные и современные подходы, основанные на больших языковых моделях, в частности модели GPT-3.

Результаты данной работы показали, что традиционные методы, хотя и широко используемые, имеют свои ограничения, особенно в способности учитывать контекст текста. В отличие от них, GPT-3 продемонстрировала значительно лучшие результаты по критерию MAP@K, что подтверждает её эффективность в извлечении ключевых фраз. Тем не менее, важно отметить, что использование больших языковых моделей может быть затруднено в реальном времени из-за более длительного времени отклика по сравнению с классическими статистическими методами.

Таким образом, результаты данной работы подчеркивают необходимость дальнейших исследований и разработок в области оптимизации алгоритмов извлечения ключевых фраз. Это позволит не только повысить точность и качество извлечения, но и сделать эти алгоритмы более подходящими для применения в реальных условиях. В будущем стоит сосредоточиться на разработке гибридных подходов, которые смогут объединить преимущества как традиционных методов, так и современных технологий на основе искусственного интеллекта, чтобы обеспечить более эффективное решение задач извлечения ключевых фраз.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Hasan K.S., Ng V.* Automatic keyphrase extraction: A survey of the state of the art // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. – 2014. – Vol. 1 – P. 1262-1273.
2. *Schutz A.T.* Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods: M. App. Sc Thesis. – 2008.
3. *Mihalcea R., Tarau P.* Textrank: Bringing order into text // Proceedings of the 2004 conference on empirical methods in natural language processing. – 2004. – P. 404-411.
4. *Floridi L., Chiriatti M.* GPT-3: Its Nature, Scope, Limits, and Consequences // Minds and Machines. – 2020. – Vol. 30. GPT-3, No. 4. – P. 681-694.
5. *Kaur J., Gupta V.* Effective approaches for extraction of keywords // International Journal of Computer Science Issues. – 2010. – Vol. 7, No. 6. – P. 144.
6. *Giarelis N., Kanakaris N., Karacapilidis N.* A Comparative Assessment of State-Of-The-Art Methods for Multilingual Unsupervised Keyphrase Extraction // IFIP International Conference on Artificial Intelligence Applications and Innovations. – Springer, 2021. – P. 635-645.
7. *Ramos J.* Using tf-idf to determine word relevance in document queries // Proceedings of the first instructional conference on machine learning. – Citeseer, 2003. – Vol. 242. – P. 29-48.
8. *Rose S., Engel D., Cramer N., Cowley W.* Automatic keyword extraction from individual documents // Text mining: applications theory. – 2010. – Vol. 1. – P. 1-20.
9. *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. – Vol. 509. – P. 257-289.
10. *Alqaryouti O., Khwileh H., Farouk T., Nabhan A., Shaalan K*. Graph-Based Keyword Extraction // Intelligent Natural Language Processing: Trends and Applications: Studies in Computational Intelligence / eds. K. Shaalan, A.E. Hassanien, F. Tolba. – Cham: Springer International Publishing, 2018. – Vol. 740. – P. 159-172. – ISBN 978-3-319-67055-3.
11. *Beliga S., Meštrović A., Martinčić-Ipšić S.* An overview of graph-based keyword extraction methods and approaches // Journal of information and organizational sciences. – 2015. – Vol. 39, No. 1. – P. 1-20.
12. *Yijun G., Tian X.* Study on keyword extraction with LDA and TextRank combination // Data Analysis and Knowledge Discovery. – 2014. – Vol. 30, No. 7. – P. 41-47.
13. *Cho T., Lee J.-H.* Latent keyphrase extraction using LDA model // Journal of The Korean Institute of Intelligent Systems. – 2015. – Vol. 25, No. 2. – P. 180-185.
14. *Abulaish M., Anwar T.* A supervised learning approach for automatic keyphrase extraction // International Journal of Innovative Computing, Information and Control. – 2012. – Vol. 8, No. 11. – P. 7579-7601.
15. *Akhil K.K., Rajimol R., Anoop V.S.* Parts-of-Speech tagging for Malayalam using deep learning techniques // International Journal of Information Technology. – 2020. – Vol. 12, No. 3. – P. 741-748.
16. *Chiche A., Yitagesu B.* Part of speech tagging: a systematic review of deep learning and machine learning approaches // Journal of Big Data. – 2022. – Vol. 9. – Part of speech tagging. No. 1. – P. 10.
17. *Aro T.O., Dada F., Balogun A.O., Oluwasogo S.A.* Stop words removal on textual data classification. – 2019.
18. *Nadeau D., Sekine S.* A survey of named entity recognition and classification // Lingvisticae Investigationes. – 2007. – Vol. 30, No. 1. – P. 3-26.

19. *Das B., Pal S., Mondal S.K., Dalui D., Shome S.K.* Automatic keyword extraction from any text document using N-gram rigid collocation // Int. J. Soft Comput. Eng. (IJSCE). – 2013. – Vol. 3, No. 2. – P. 238-242.
20. *Evert S., Krenn B.* Exploratory collocation extraction // Phraseology 2005: The Many Faces of Phraseology. – 2005. – P. 113-115.
21. *Maragheh R.Y., Fang C., Irugu C.C., Parikh P., Cho J., Xu J., Sukumar S., Patel M., Korpeoglu E., Kumar S.* LLM-take: theme-aware keyword extraction using large language models // 2023 IEEE International Conference on Big Data (BigData). – IEEE, 2023. LLM-take. – P. 4318-4324.
22. *Мохаммад Ж.Х., Мансур А.М., Кравченко Ю.А., Кравченко Д.Ю.* Метод автоматического извлечения ключевых слов // Международный научно-технический конгресс «Интеллектуальные системы и информационные технологии – 2022». – 2022. – С. 91-97.
23. *Мохаммад Ж.Х., Мансур А.М., Кравченко Ю.А., Бова В.В.* Метод извлечения ключевых фраз на основе новой функции ранжирования // Информационные технологии. – 2022. – Т. 28, № 9. – С. 465-474.

REFERENCES

1. *Hasan K.S., Ng V.* Automatic keyphrase extraction: A survey of the state of the art, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* 2014, Vol. 1, pp. 1262-1273.
2. *Schutz A.T.* Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods: M. App. Sc Thesis, 2008.
3. *Mihalcea R., Tarau P.* Textrank: Bringing order into text, *Proceedings of the 2004 conference on empirical methods in natural language processing,* 2004, pp. 404-411.
4. *Floridi L., Chiriatti M.* GPT-3: Its Nature, Scope, Limits, and Consequences, *Minds and Machines,* 2020, Vol. 30. GPT-3, No. 4, pp. 681-694.
5. *Kaur J., Gupta V.* Effective approaches for extraction of keywords, *International Journal of Computer Science Issues,* 2010, Vol. 7, No. 6, pp. 144.
6. *Giarelis N., Kanakaris N., Karacapilidis N.* A Comparative Assessment of State-Of-The-Art Methods for Multilingual Unsupervised Keyphrase Extraction, *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2021, pp. 635-645.
7. *Ramos J.* Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning*. Citeseer, 2003, Vol. 242, pp. 29-48.
8. *Rose S., Engel D., Cramer N., Cowley W.* Automatic keyword extraction from individual document, *Text mining: applications theory,* 2010, Vol. 1, pp. 1-20.
9. *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, Vol. 509, pp. 257-289.
10. *Alqaryouti O., Khwileh H., Farouk T., Nabhan A., Shaalan K*. Graph-Based Keyword Extraction, *Intelligent Natural Language Processing: Trends and Applications: Studies in Computational Intelligence* / eds. K. Shaalan, A.E. Hassanien, F. Tolba. Cham: Springer International Publishing, 2018, Vol. 740, pp. 159-172. ISBN 978-3-319-67055-3.
11. *Beliga S., Meštrović A., Martinčić-Ipšić S.* An overview of graph-based keyword extraction methods and approaches, *Journal of information and organizational sciences,* 2015, Vol. 39, No. 1, pp. 1-20.
12. *Yijun G., Tian X.* Study on keyword extraction with LDA and TextRank combination, *Data Analysis and Knowledge Discovery,* 2014, Vol. 30, No. 7, pp. 41-47.
13. *Cho T., Lee J.-H.* Latent keyphrase extraction using LDA model, *Journal of The Korean Institute of Intelligent Systems,* 2015, Vol. 25, No. 2, pp. 180-185.
14. *Abulaish M., Anwar T.* A supervised learning approach for automatic keyphrase extraction, *International Journal of Innovative Computing, Information and Control,* 2012, Vol. 8, No. 11, pp. 7579-7601.
15. *Akhil K.K., Rajimol R., Anoop V.S.* Parts-of-Speech tagging for Malayalam using deep learning techniques, *International Journal of Information Technology,* 2020, Vol. 12, No. 3, pp. 741-748.
16. *Chiche A., Yitagesu B.* Part of speech tagging: a systematic review of deep learning and machine learning approaches, *Journal of Big Data,* 2022, Vol. 9. Part of speech tagging. No. 1, pp. 10.
17. *Aro T.O., Dada F., Balogun A.O., Oluwasogo S.A.* Stop words removal on textual data classification, 2019.
18. *Nadeau D., Sekine S.* A survey of named entity recognition and classification, *Lingvisticae Investigationes,* 2007, Vol. 30, No. 1, pp. 3-26.
19. *Das B., Pal S., Mondal S.K., Dalui D., Shome S.K.* Automatic keyword extraction from any text document using N-gram rigid collocation, *Int. J. Soft Comput. Eng. (IJSCE),* 2013, Vol. 3, No. 2, pp. 238-242.

20. *Evert S., Krenn B.* Exploratory collocation extraction, *Phraseology 2005: The Many Faces of Phraseology,* 2005, pp. 113-115.
21. *Maragheh R.Y., Fang C., Irugu C.C., Parikh P., Cho J., Xu J., Sukumar S., Patel M., Korpeoglu E., Kumar S.* LLM-take: theme-aware keyword extraction using large language models, *2023 IEEE International Conference on Big Data (BigData).* IEEE, 2023. LLM-take. pp. 4318-4324.
22. *Mokhammad Zh.Kh., Mansur A.M., Kravchenko Yu.A., Kravchenko D.Yu.* Metod avtomaticheskogo izvlecheniya klyuchevykh slov [Method of automatic keyword extraction], *Mezhdunarodnyy nauchno-tekhnicheskiy kongress «Intellektual'nye sistemy i informatsionnye tekhnologii – 2022»* [International scientific and technical congress "Intelligent systems and information technologies - 2022"], 2022, pp. 91-97.
23. *Mokhammad Zh.Kh., Mansur A.M., Kravchenko Yu.A., Bova V.V.* Metod izvlecheniya klyuchevykh fraz na osnove novoy funktsii ranzhirovaniya [Method of key phrase extraction based on a new ranking function], *Informatsionnye tekhnologii* [Information technologies], 2022, Vol. 28, No. 9, pp. 465-474.

Статью рекомендовал к опубликованию к.т.н. С.Г. Буланов.

Мохаммад Жуман Хуссайн – Южный федеральный университет; e-mail: zmohammad@sfedu.ru; г. Таганрог, Россия; тел.: 89185433526; кафедра систем автоматизированного проектирования им. В.М. Курейчика; соискатель.

Mohammad Juman Hussain – Southern Federal University; e-mail: zmohammad@sfedu.ru; Taganrog, Russia; phone: +79185433526; the Department of Computer-Aided Design Systems named after Viktor Mikhailovich Kureichik; applicant.