

УДК 378.4

doi: 10.15622/rcai.2025.024

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ НА КОЛЛЕКЦИЯХ НАУЧНЫХ ПУБЛИКАЦИЙ

Н.А. Назаров (*straider105@gmail.com*)

М.Р. Шарифуллин (*sharifullin2107@mail.ru*)

В.О. Толчеев (*tolcheevvo@mail.ru*)

Национальный исследовательский университет «МЭИ», Москва

Проанализированы прикладные задачи интеллектуального анализа текстовых данных, для решения которых важно проводить извлечение ключевых слов (КС). Отмечается, что наиболее часто и эффективно КС используются для обработки и анализа англоязычных научных (полнотекстовых и библиографических) документов. Проведено сравнение качества выявления КС ($F1@K$ -мера) на общеизвестных и свободно распространяемых коллекциях текстовых научных данных (Inspec, SemEval-2010, Krapivin), а также датасете из новостных сообщений (DUC2001). В ходе экспериментальных исследований рассмотрены различные технологии извлечения КС: статистический алгоритм YAKE, графовые алгоритмы TopicRank и MultipartiteRank, нейросетевые модели KeyBERT и PromptRank. Проанализирована зависимость параметров от длины документов, оценены показатели качества.

Ключевые слова: извлечение ключевых слов, интеллектуальный анализ текстов, научные публикации, графовые алгоритмы, нейросетевые модели, статистический анализ.

Введение

В интеллектуальном анализе текстовых данных (Text Mining) большое внимание уделяется правильному выбору информативных терминов, которые способны наиболее полно описать смысл документов. Для этого обычно используются отдельные слова или словосочетания. Такие ключевые слова (КС) широко применяются в информационном поиске, суммаризации текстов (аннотирование-реферирование); анализе тональности, при выявлении структуры и визуализации документальных коллекций, ведении диалога в вопросно-ответных системах, мониторинге электрон-

ных сообщений, обнаружении тематических сообществ в сети Интернет, поисковой оптимизации и продвижении сайтов [Song et. al., 2023], [Ajallouda et. al., 2022]. В последнее время КС активно применяются в промпт-инжиниринге (Prompt Engineering) для составления запросов к большим языковым моделям (Large Language Model, LLM), а также при автоматизированной разметке (кластеризации) текстовых данных [Sahoo et. al., 2025], [Liu et. al., 2018], [Ванюшкин и др., 2018].

Дадим определение КС – неслучайно встречающиеся в документах важные понятия и/или их комбинации, отражающие содержание документа и формирующие его смысловое ядро [Шереметьева, 2015]. Объединение отдельных информативных терминов в общие конструкции чаще всего способно обеспечить «приращение смысла» и упростить интерпретацию анализируемых текстов.

Эффективность использования КС для решения задач Text Mining существенно зависит от стиля документа (художественный, публицистический, официально-деловой, научный, разговорный и т.п.) и его размера. Большие художественные и публицистические тексты допускают множественные трактовки, отражая субъективное восприятие и понимание прочитанного. В этом случае КС будут индивидуальными для каждого из читателей, заметно различаясь между собой. При обработке более формальных и частично структурированных материалов (научные статьи, юридические документы, медицинские карты и рецепты, новостные сообщения) извлекаемый набор КС чаще всего одинаково интерпретируется специалистами и представляет «общий код», позволяющий обмениваться важной информацией. По мнению ряда специалистов, научная публикация после прочтения «сворачивается» в ограниченное количество КС (8-10 слов и словосочетаний), которые хранятся в памяти человека и ассоциируются с конкретным текстом [Шереметьева, 2015], [Москвитина, 2009], [Москвитина, 2018].

Отнесение отдельных терминов и их сочетаний к КС зависит от ряда факторов: частоты и места совместной встречаемости, контекста появления, принадлежности к определенным частям речи (обычно к существительным и прилагательным), специфики предметной области. Чем более концентрировано излагаются сведения в документе, тем информативнее получаются выделяемые из него КС. Так, в полнотекстовых научных работах наибольшая «концентрация смыслов» содержится в названии, аннотации, ключевых словах, введении и заключении статьи, в библиографических описаниях – в названии, аннотации и ключевых словах (иногда КС отсутствуют, так как не указаны авторами и не присвоены при рубрикации в электронных библиотеках).

В данной работе рассматриваются англоязычные полнотекстовые и библиографические научные документы. Проводится комплексная экспериментальная оценка точностных и временных характеристик десяти из-

вестных методов выявления КС, определяются те подходы, которые обладают самыми высокими показателями качества обнаружения КС и наилучшим образом подходят для решения прикладных задач обработки научных текстов на основе КС (прежде всего кластеризация, визуализация и классификация).

1. Методы автоматического извлечения КС

К настоящему времени, несмотря на интенсивные исследования, не разработано универсального SOTA (State of The Art) подхода для выделения КС [Song et. al., 2023], [Митрофанова и др., 2022], [Musunuru et al., 2024]. В данной работе рассматриваются известные статистические, графовые и нейросетевые (эмбединговые) методы извлечения ключевых слов (МИКС). Их объединяет общий подход к выявлению КС, которое проводится в автоматическом режиме (т.е. используется обучение без учителя). В данной работе реализуется комплексное исследование десяти МИКС, часто используемых на практике [Ajallouda et. al., 2022], [Campos et. al., 2018], [Bougouin et. al., 2013], [Boudin, 2018], [Grootendorst, 2020], [Kong et. al., 2023]:

- 1) статистические алгоритмы RAKE, YAKE;
- 2) графовые методы TextRank, SingleRank, EmbedRank, MDERank, TopicRank, MultipartiteRank;
- 3) нейросетевые модели KeyBERT (на основе модели BERT) и PromptRank (на основе генерации КС с помощью большой языковой модели T5).

Путем экспериментальных исследований с использованием коллекций документов Inspec, SemEval-2010, Krapivin, DUC2001 выделены пять наиболее высокоточных МИКС (YAKE, TopicRank, MultipartiteRank, KeyBERT, PromptRank), которые далее подробно рассматриваются в данной работе.

1.1. Статистический метод YAKE

YAKE [Campos et. al., 2018] является многоязычным подходом (применим для различных языков, включая русский) и основан на извлечении статистических характеристик текста. Выделяемые КС могут быть названиями или аббревиатурами (). При выборе КС учитывается их местоположение в тексте (), частота слова () количество предположений с кандидатом в КС (), сходство со стоп-словами (). Итоговый вес КС вычисляется по формуле:

$$\frac{\sum_{i=1}^n \left(\frac{w_i}{\sum_{j=1}^n w_j} \right)}{\sum_{i=1}^n \left(\frac{w_i}{\sum_{j=1}^n w_j} \right)}, \quad (1.1)$$

1.2. Графовый метод TopicRank

Этот метод [Bougouin et. al., 2013] является модификацией TextRank, однако вершинами (полного неориентированного) графа являются не слова (как в TextRank), а темы (кластер из похожих однословных и многословных выражений). Ключевые фразы-кандидаты выбираются из последовательности соседних существительных с одним или несколькими предшествующими прилагательными. Затем они группируются по темам с помощью иерархической агломеративной кластеризации [Boudin, 2018], чтобы выбрать КС, наилучшим образом «представляющие» получившиеся кластеры. Для каждого документа составляется ранжированный (по важности) список тем. Алгоритм TopicRank предусматривает выполнение следующих шагов: предобработка, выявление КС-кандидатов, составление тем (кластеризация КС-кандидатов), ранжирование тем, формирование наиболее информативных и релевантных КС. Полный граф позволяет учесть взаимозависимость тем, вес ребра между двумя темами (вершинами) и вычисляется на основе близости их ключевых фраз:

(1.2)

Расстояние между позициями ключевых фраз и рассчитывается по формуле:

$$\text{---}, \quad (1.3)$$

Здесь $\text{pos}(\quad)$, $\text{pos}(\quad)$ – множество позиций вхождений ключевых фраз и в документе, – расстояние между позициями (позиция КС) и (позиция КС). Таким образом, чем больше значение, тем сильнее связь между КС и. Рейтинг тем в методе определяется по алгоритмам ранжирования PageRank или TextRank. Формула расчета PageRank имеет вид:

$$\text{---} \quad (1.4)$$

Здесь – важность i -ой вершины, In множество вершин, входящих в i -ую вершину ребра, Out множество вершин, связанных с i -ой вершиной исходящими из неё ребрами, – коэффициент затухания, который задается пользователем (при значениях 0,1-0,3 выбираются средне связанные КС, при 0,8-1 – сильно связанные КС). В алгоритме ранжирования TextRank используется «улучшенная» формула PageRank:

$$\text{---} \quad (1.5)$$

где – число вершин связанных с темой, отношение под знаком суммы показывает, насколько сильно тема поддерживает.

После ранжирования тем выбираются ключевые фразы из самых важных тем. Для этого используется одна из стратегий: выбор фразы, которая первой встречается в документе; выбор наиболее частотной фразы в теме; выбор фразы – центроида, наиболее схожей с другими КС в кластере.

1.3. Графовый метод **MultipartiteRank**

Метод **MultipartiteRank** [Boudin, 2018] является модификацией **TopicRank**. Однако в нем используется многосторонний направленный граф, где вершины в отличие от **TopicRank** представляют КС-кандидаты, которые связаны только в том случае, если они относятся к разным темам. В алгоритме предусматривается корректировка весов кандидатов при вычислении их важности, чем выше значение, тем фразы из начала документа получают больший вес :

$$- , \quad (1.6)$$

1.4. Нейросетевой метод **KeyBERT**

Для выделения КС **KeyBERT** [Grootendorst, 2020] использует предварительно обученную модель **BERT**, что позволяет учитывать контекст появления КС в документе. В **KeyBERT** не вводятся ограничения на допустимые части речи и в качестве КС используются не только существительные и прилагательные, но и глаголы. Выбор репрезентативных КС осуществляется на основе расчета косинусного сходства между эмбедингом (вложением) КС и эмбедингом всего текста (отбираются КС, имеющие наибольшие значения косинусной меры).

1.5. Нейросетевой метод **Promptrank (T5)**

Этот метод [Kong A. et. al., 2023] извлечения ключевых фраз основан на использовании большой языковой модели **T5 (Text-To-TextTransfer Transformer)**, имеющей архитектуру кодера-декодера, и применяет шаблоны (prompts) для выбора наилучших КС. Выбор фраз-кандидатов осуществляется с помощью выделения частей речи (PoS, Part-of-Speech). Для ранжирования КС-кандидатов документ вводится в кодировщик и вычисляется вероятность того, что будет сгенерирована такая же ключевая фраза. Чем выше вероятность, тем более точно КС соответствует документу и получает соответствующий ранг.

2. Описание используемых датасетов

Выявление КС – слабоформализованный процесс, который зависит от ряда факторов, в частности, от размера документов и количества КС, которые наилучшим образом описывают тексты.

Для настройки параметров и сопоставления различных методов в данной работе применяется **F1@K**-мера (модификация метрики **F1-score**), которая рассчитывается для топ-К ключевых фраз. **F1@K**-мера выбрана исходя из того, что она интуитивно понятна и используется в большинстве про-

фильных публикаций для сопоставления различных МИКС. При определении $F1@K$ -меры необходимо вычислить значения точности и полноты для извлеченных КС (и), аналогично тому, как это делается в рекомендательных системах [Bjadon A. et. al., 2024]:

$$\frac{\sum_{k=1}^K \text{Precision}_k}{K}; \quad (2.1)$$

$$\frac{\sum_{k=1}^K \text{Recall}_k}{K}; \quad (2.2)$$

$$\frac{\sum_{k=1}^K \text{F1}_k}{K}. \quad (2.3)$$

Для корректного сравнения МИКС необходимо провести исследования на известных и общедоступных коллекциях научных документов, имеющих «золотой стандарт» – КС, сформированные экспертами (ассессорами). В данной работе экспериментальные исследования проводятся на датасетах, которые содержат англоязычные научные публикации разного размера: Inspec, SemEval-2010 и Krapivin.

Набор данных Inspec содержит 2000 библиографических описаний публикаций в области информационных технологий и компьютерных наук с временным охватом 1998–2002 гг. Его основное предназначение – поддержка исследований в области обработки естественного языка, в частности, разработка методов выявления ключевых слов. Средний размер документа составляет 137 слов.

Набор данных SemEval-2010, созданный для проведения конкурса Semantic Evaluation по оценке качества МИКС, состоит из 244 научных статей из цифровой библиотеки АСМ, временной охват 2000–2009 гг. Средний размер публикации составляет 230 слов.

Набор данных Krapivin включает 2000 полнотекстовых научных статей по тематикам Computer Science за 2009 год, которые были получены из цифровой библиотеки АСМ. Чаще всего датасет используется для исследований в области извлечения ключевых слов и кластеризации текстов. Средний размер документов составляет 8940 слов.

3. Экспериментальное исследование методов автоматического извлечения КС на общедоступных текстовых коллекциях

3.1. Настройка гиперпараметров МИКС (с учетом различий в размерах документов)

Обычно в научной статье автор указывает не менее 5 КС. Для больших публикаций требуется предоставить расширенный список от 10 до 15 КС. Поэтому при сопоставлении МИКС чаще всего анализируются показатели качества для 5, 10 и 15 КС.

Прежде всего настроим гиперпараметры методов и выберем значения, которые обеспечивают наибольшие значения метрики качества $F1@K$. Экспериментальные исследования показали, что наилучшие гиперпараметры методов несущественно зависят от длины документа и практически идентичны для трех выборок (Inspec, SemEval-2010, Krapivin). В методах Yake и TopicRank все параметры оказались одинаковыми. В PromptRank(T5) различия имеются в значениях весового коэффициента учета позиции слов в тексте (`position_factor` равен $1.2e8$ у Inspec и $1.2e9$ у Krapivin и SemEval-2010). В методе MultipartiteRank различаются способы объединения кандидатов `complete` и `average`, а также порог схожести для кластеризации кандидатов (`threshold`). В методе KeyBert имеются наиболее заметные различия: по числу выделяемых КС (Inspec, SemEval-2010 – 8 КС, Krapivin – 15 КС) и отбору КС (`use_mmr`). Большинство гиперпараметров методов не существенно зависят от размера документов и обладают высокой универсальностью, что позволяет их применять для различных датасетов.

Настройка гиперпараметров позволила в ряде случаев улучшить качество выявления КС по сравнению с результатами, которые приводятся в широкоизвестных профильных публикациях [Kong et. al., 2023]. Метод Yake показывает более высокие результаты, чем опубликованные ранее, на датасетах Krapivin и SemEval, TopicRank на датасетах Inspec и SemEval2010, MultipartiteRank на датасетах Inspec и SemEval2010, PromptRank на Inspec и Krapivin. KeyBert только на датасете SemEval2010 при показателях $F1@10$, $F1@15$.

В данной работе используются следующие настройки гиперпараметров:

- 1) Yake: `n = 3`, `dedupLim = 0,9`, `dedupFunc = seqm`, `windowsSize = 2`, `top = 15`.
- 2) TopicRank: `threshold = 0,2`, `method = average`, `heuristic = first`, `n = 15`, `edundancy_removal = False`, `stemming = False`.
- 3) MultipartiteRank: `threshold = 0,2`, `method = average`, `alpha = 1,1`, `n = 15`, `edundncy_removal = True`, `stemming = False`.
- 4) KeyBert: `keyphrase_ngram_range = (1,3)`, `use_mmr = False`, `diversity = False`, `use_maxsum = False`, `nr_candidates = False`.
- 5) PromptRank(T5): `max_len = 512`, `temp_en = Book`, `temp_de = This book mainly this about`, `mode = base`, `enable_pos = True`, `position_factor = 1.2e9`, `length_factor = 0,6`.

3.2. Сравнительный анализ показателей качества МИКС (в зависимости от размера документов)

Далее приводятся результаты исследований МИКС, полученные на разных наборах данных при указанных значениях гиперпараметров. Для оценки качества рассчитываются $F1@5$, $F1@10$, $F1@15$. Результаты представлены в табл. 1, а их визуализация на рис. 1.

Таблица 1

Inspec	Yake	TopicRank	MultipartiteRank	PromptRank	KeyBert
F1@5	6.16	16.69	22.69	32.02	5.94
F1@10	8.60	23.92	27.72	38.26	7.86
F1@15	9.65	27.79	29.64	38.93	9.03
Krapivin					
F1@5	10.50	5.44	7.72	16.13	4.32
F1@10	11.25	7.03	8.30	16.89	5.21
F1@15	10.65	7.46	7.90	17.45	6.13
SemEval					
F1@5	16.39	12.15	13.56	17.24	8.54
F1@10	19.40	15.12	16.43	20.66	13.05
F1@15	19.40	15.12	16.43	21.28	14.52

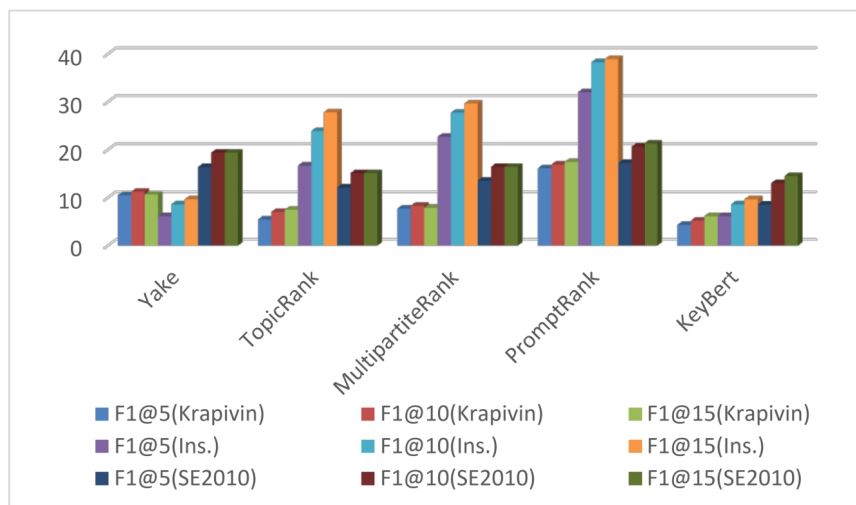


Рис. 1. Значения F1@K-меры на датасетах Inspec, SemEval-2010, Krapivin для различных методов

Анализ результатов исследований позволяет сделать выводы:

- 1) Показатели качества существенно зависят от размера документов. Все методы имеют низкие значения F1@5, F1@10, F1@15 на полнотекстовом датасете Krapivin.
- 2) Наилучшее качество продемонстрировал метод PromptRank(T5), который превзошёл остальные МИКС на рассмотренных коллекциях научных документов.
- 3) «Аутсайдером» практически на всех выборках оказался KeyBert.

3.3. Оценка точностных характеристик МИКС на коллекции новостных сообщений

В предыдущем эксперименте самые высокие показатели качества получены для частично структурированных коротких (библиографических) документов. Проанализируем, можно ли распространить сделанные выводы на другие небольшие тексты, в частности новостные сообщения. Для этого проведем экспериментальное изучение МИКС на коллекции DUC2001. Это набор данных, содержащий 308 новостных сообщений за 2001 год. Средний размер текстов – 845 слов.

Далее в табл. 2 приводятся значения $F1@5, 15$, полученные при использовании исследуемых методов для DUC2001.

Таблица 2

DUC2001	Yake	TopicRank	MultipartiteRank	PromptRank	KeyBert
$F1@5$	12.05	14.09	23.75	27.39	3.01
$F1@10$	14.44	19.31	25.83	31.59	3.10
$F1@15$	15.29	22.59	25.38	31.01	3.14

Исходя из результатов, представленных в табл. 1 и 2, можно сделать следующие выводы. На наборе новостных сообщений DUC2001 все методы достигают существенно более высоких показателей качества, чем в случае обработки коллекции полнотекстовых научных документов Krapivin. Однако PromptRank, который является «лидером» на всех датасетах, показывает меньшие значения $F1@5, F1@10, F1@15$ на DUC2001 по сравнению с Inspec, аналогичное поведение демонстрируют MultipartiteRank и TopicRank. Вместе с тем YAKE, в отличие от остальных МИКС, значительно улучшает показатели качества на DUC2001. Отметим также, что KeyBERT не способен обеспечить качественное выделение КС на DUC2001, показывая крайне низкие значения $F1@5, F1@10, F1@15$. В публикации [Rao S. et. al., 2022] причинами плохих результатов KeyBERT называют некорректное обрезание ключевых фраз и выделение большого числа КС, относящихся к глагольной группе. На рис. 2 (для данных из табл. 1 и 2) приведены показатели качества методов на датасетах Inspec (библиографические научные документы) и DUC2001 (короткие новостные сообщения).

Наряду с анализом точностных характеристик МИКС существенный интерес представляет их ресурсозатратность, прежде всего временная сложность. Теоретические оценки (с использованием О-нотации) указаны разработчиками в цитируемых нами публикациях, в данной работе приводятся значения процессного времени, что позволяет сравнить производительность методов на датасетах разного размера, состоящих из библиографических, полнотекстовых, новостных документов. Экспериментальные замеры (в секундах) на аппаратной платформе с моделью процессора E5-2640v3 и GPU - RTX 3060ti указаны в табл. 3.

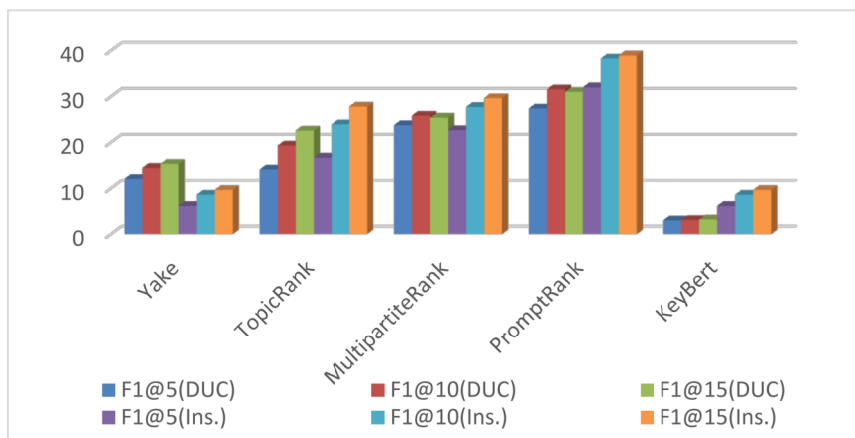


Рис. 2. Значения F1@K-меры на датасетах Inspec и DUC2001

Таблица 3

Метод \ Датасет	Inspec	SemEval-2010	Krapivin	DUC2001
YAKE	45	6	63	8
TopicRank	273	30	344	69
MultipartiteRank	338	47	409	81
KeyBERT	895	99	782	163
PromptRank	2051	264	2107	342

Из табл. 3 следует, что наиболее быстрым МИКС является статистический метод YAKE, затем идут графовые методы (TopicRank и MultipartiteRank), наиболее медленными и ресурсозатратными являются нейросетевые модели (KeyBERT и PromptRank). Таким образом, при выявлении КС справедлива общеизвестная закономерность – чем точнее метод, тем больше его алгоритмическая сложность и дольше ожидание результатов.

Заключение

Как и в большинстве нетривиальных слабоформализуемых задач, при выявлении КС из текстов не удастся найти единственное универсальное решение. В связи с этим для извлечения КС используются различные подходы, каждый из которых имеет как сильные стороны, так и определенные недостатки. В них по-разному выделяются наиболее существенные КС и даются не полностью совпадающие «приближения» к пониманию смысла научной публикации (или новостного сообщения).

В работе проведен сравнительный анализ наиболее известных и эффективных МИКС, оценена их точность и быстродействие на научных (полнотекстовых и библиографических) коллекциях документов и новостном датасете. Выявлены методы-«лидеры» (PromptRank) и методы-«аутсайдеры» (прежде всего KeyBERT), исследованы характеристики МИКС в зависимости от способа (правила) выявления КС, размера текста и вида документа (научный или новостной). Сформулированные при проведении исследований выводы и рекомендации позволяют сделать шаг в сторону большей формализации изучаемой проблемы, достичь лучшего понимания области применения каждого из методов, оценить степень их универсальности, а также разработать процедуры кластеризации, визуализации и классификации на основе использования извлеченных КС.

Список литературы

- [Ванюшкин и др., 2018] Ванюшкин А.С. Гращенко Л.А. Опыт автоматизированной разметки текстов ключевыми словами // Материалы IV Всероссийской научно-практической конференции «Современные проблемы физико-математических наук». – 2018. – С. 320-325.
- [Митрофанова и др., 2022] Митрофанова О.А., Гаврилюк Д.А. Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // TerraLinguistica. – 2022. – Т. 13, № 4. – С. 22-40.
- [Москвитина, 2009] Москвитина Т.Н. Ключевые слова и их функции в научном тексте // Вестник Южно-Уральского государственного гуманитарно-педагогического университета. – 2009. – № 11. – С. 270-283.
- [Москвитина, 2018] Москвитина Т.Н. Методы выделения ключевых слов при реферировании научного текста // Вестник Томского государственного педагогического университета. – 2018. – № 8. – С. 45-50.
- [Шереметьева, 2015] Шереметьева С.О. Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного университета. – 2015. – Т. 12, № 1. – С. 76-81.
- [Ajallouda et. al., 2022] Ajallouda L., Fagroud F.Z., Zellou A., Benlahmar E.A Systematic Literature Review of Keyphrases Extraction Approaches // International Journal of Interactive Mobile Technologies (iJIM). – August 2022. – 16(16). – P. 31-58.
- [BJadon et. al., 2024] BJadon A., Patil A. A Comprehensive Survey of Evaluation Techniques for Recommendation Systems // arXiv:2312.16015v2, 12 Jan 2024.
- [Bougouin et. al., 2013] Bougouin A., Boudin F., Daille B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction // International Joint Conference on Natural Language Processing (IJCNLP). – 2013. – P. 543-551.
- [Boudin, 2018] Boudin F. Unsupervised Keyphrase Extraction with Multipartite Graphs // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2018. – P. 667-672.
- [Campos et. al., 2018] Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Collection-independent automatic keyword extractor // Proceedings of 40th European Conference on Information Retrieval (ECIR). – 2018. – P. 806-810.

- [**Grootendorst, 2020**] Grootendorst M. KeyBERT: Minimal Keyword Extraction with BERT. – 2020. – URL: <http://doi.org/10.5281/zenodo.4461265> (дата обращения: 15.02.2025).
- [**Kong et. al., 2023**] Kong A., Zhao S., Chen H., Li Q., Qin Y., Sun R., Bai X. PromptRank: Unsupervised Keyphrase Extraction Using Prompt // In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. – 2023. – P. 9788-9801.
- [**Liu et. al., 2018**] Liu Q., Kawahara D., Li S. Scientific Keyphrase extraction: extracting candidates with semi-supervised data augmentation // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. – 2018. – P. 183-194.
- [**Musunuru et al., 2024**] Surveying Keyword Extractor: Classification, Applications, and Empirical Analysis // 2024 Parul International Conference on Engineering and Technology (PICET). – IEEE, 2024. – C. 1-8.
- [**Papagiannopoulou et. al., 2020**] Papagiannopoulou E., Tsoumakas G. A review of keyphrase extraction // Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery. – 2020. – 10(2).
- [**Rao et. al., 2022**] Rao S, Nasirian Sara, Ghoshal Parijat. Keyword Extraction in Scientific Documents // arXiv:2207.0188v2 7 Jul 2022.
- [**Song et. al., 2023**] Song M., Feng Y., Jing L. A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models // In Findings of the Association for Computational Linguistics: EACL. – 2023. – P. 2153-2164.
- [**Sahoo et. al., 2025**] Sahoo P., Singh A.K., Saha S., Jain V., Mondal S., Chadha A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications // arXiv:2402.07927v2 [cs.AI] 16 Mar 2025.