ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2026. Т. 30, № 2 (112). С. 30-40 http://journal.uust.ru/index.php/Vestnik

УДК 004.912 doi 10.54708/19926502_2026_30211230
Код ГРНТИ 20.19.27

Гибридный метод извлечения ключевых слов из русскоязычных научно-технических текстов на основе синтаксического анализа и многофакторного ранжирования

А.А. Хорошилов[2], В.Н. Гокарев[1*], Д.В. Березкин[1], А.А. Хорошилов[2]

[1]Московский государственный технический университет имени Н.Э. Баумана, г. Москва, Россия
[2]Всероссийский институт научной и технической информации Российской академии наук, Московский авиационный институт, г. Москва, Россия

Аннотация. Статья посвящена разработке гибридного метода автоматического извлечения ключевых слов из русскоязычных научно-технических текстов. Предложенный метод объединяет морфосинтаксический анализ, генерацию кандидатов на основе обхода дерева синтаксических зависимостей и многофакторное ранжирование с использованием комбинации статистических (TF-IDF, C-value), графовых (TextRank) и позиционных признаков. Экспериментальная оценка на корпусе из 16818 русскоязычных научных статей из КиберЛенинки показала, что предложенный метод превосходит базовые алгоритмы (TF-IDF, RAKE, YAKE, TextRank) по метрикам F1@10 (улучшение на 2,3% – 0,268 против 0,262 у ближайшего конкурента TextRank), MAP (улучшение на 6,9% – 0,217 против 0,203) и семантическому покрытию текста (улучшение на 32,5% – 0,559 против 0,422). Особенно значительное преимущество достигается при извлечении многословных терминологических словосочетаний благодаря учету синтаксических связей между компонентами.

Ключевые слова: *извлечение ключевых слов, обработка естественного языка, деревья зависимостей, научно-технические тексты.*

***vgokarev@mail.ru**

Введение

Развитие науки и технологий в последние десятилетия привело к экспоненциальному росту объемов научно-технической литературы. По данным международных наукометрических баз данных, ежегодно публикуется более 2,5 миллионов научных статей, причем темпы роста продолжают увеличиваться [1]. Объемы информации достигли таких размеров, что исследователи и инженеры не способны самостоятельно ознакомиться со всеми релевантными публикациями даже в узкоспециализированных областях. Данный факт обуславливает активное развитие исследований в области автоматической обработки научно-технических текстов.

Автоматическое извлечение ключевых слов (КС) является одной из фундаментальных задач обработки естественного языка, результаты которой используются в информационном поиске, автоматическом реферировании, классификации документов и построении тематических онтологий [2]. Качественное извлечение ключевых терминов позволяет компактно представить содержание документа, облегчить навигацию по научным коллекциям и повысить релевантность поиска.

Существующие методы извлечения ключевых слов можно разделить на несколько категорий: статистические (TF-IDF [3], YAKE [4]), основанные на графах (TextRank [5], SingleRank [6]), использующие машинное обучение [7] и гибридные [8]. Однако большинство из них разрабатывались преимущественно для английского языка и не учитывают морфологическую и синтаксическую специфику русского языка. Статистические методы, такие как TF-IDF, эффективно выделяют частотно значимые термины, но не способны извлекать многословные терминологические словосочетания с учетом грамматического согласования. Алгоритм RAKE [9], ориентированный на разделители и стоп-слова английского языка, показывает неудовлетворительные результаты на русскоязычных текстах из-за отсутствия механизмов морфологической нормализации. Графовые методы, такие как TextRank, оперируют лексическими со-появлениями (co-occurrence) и не учитывают синтаксические связи между словами.

Цель данной работы – разработка гибридного метода извлечения ключевых слов из русскоязычных научно-технических текстов, объединяющего преимущества синтаксического анализа и многофакторного статистического ранжирования. Предлагаемый подход основан на генерации кандидатов с помощью обхода дерева синтаксических зависимостей и последующего ранжирования с использованием взвешенной комбинации признаков (TF-IDF, C-value, TextRank, позиционные характеристики).

Обзор существующих методов

Метод TF-IDF (Term Frequency – Inverse Document Frequency) является классической мерой важности термина в документе относительно коллекции [3]. Мера вычисляется как произведение частоты термина в документе (TF) и обратной документной частоты (IDF):

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times log\left(\frac{|D|}{DF(t,D)}\right), \quad (1)$$

где t – термин, d – документ, D – коллекция документов, $DF(t,D)$ – число документов, содержащих термин t.

Метод RAKE (Rapid Automatic Keyword Extraction) [9] – алгоритм, основанный на разделении текста на фразы с помощью стоп-слов и знаков препинания. Каждому слову-кандидату присваивается оценка, вычисляемая как отношение степени слова в графе со-появлений к его частоте:

$$Score(w) = \frac{deg(w)}{freq(w)}, \quad (2)$$

где $deg(w)$ – сумма со-появлений слова w с другими словами, $freq(w)$ – частота слова.

Метод YAKE (Yet Another Keyword Extractor) [4] – статистический метод без учителя, использующий пять признаков: позицию первого вхождения, частоту термина, контекстное разнообразие, связь с заглавием и количество вхождений в предложениях. Финальная оценка вычисляется как произведение нормализованных признаков.

Метод TextRank [5] – алгоритм извлечения ключевых слов, основанный на графовом представлении текста. Слова представляются как вершины графа, ребра соединяют слова, встречающиеся в пределах фиксированного окна. Важность вершины определяется алгоритмом PageRank:

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{\{ji\}}}{\sum_{V_k \in Out(V_j)} w_{\{jk\}}} WS(V_j), \quad (3)$$

где d – коэффициент затухания (обычно 0,85), $In(V_i)$ – множество вершин, указывающих на V_i, $Out(V_j)$ – множество вершин, на которые указывает V_j.

Характеристики каждого из методов кратко описаны в Табл. 1.

Таблица 1. Характеристики существующих методов извлечения ключевых слов.

Метод	Преимущества	Ограничения
TF-IDF	простота реализации, интерпретируемость, эффективность на больших коллекциях	не учитывает порядок слов и синтаксические связи; затруднено извлечение многословных терминов; не учитывает морфологию
RAKE	не требует обучения, быстрая работа	ориентирован на английский язык; качество сильно зависит от списка стоп-слов; не учитывает морфологическое согласование; генерирует грамматически некорректные фразы на русском языке

Продолжение Табл. 1

YAKE	не требует внешних ресурсов, учитывает позиционную информацию	не учитывает синтаксическую структуру; на русскоязычных текстах показывает низкую точность из-за отсутствия лемматизации
TextRank	учитывает глобальную структуру текста, не требует обучения	оперирует только со-появлениями в окне; не учитывает синтаксические зависимости; затруднено извлечение терминологических словосочетаний

Перечисленные методы разрабатывались преимущественно для английского языка и сталкиваются с рядом проблем при работе с русскоязычными текстами:

1. Русский язык характеризуется развитой системой словоизменения: одна лексема может иметь десятки словоформ. Без морфологической нормализации частотные характеристики термина распределяются между его формами, снижая точность ранжирования.

2. В отличие от аналитических языков, русский язык допускает вариативный порядок слов в словосочетаниях. Методы, основанные на последовательностях n-грамм, не учитывают эту особенность.

3. Научно-техническая терминология часто представлена словосочетаниями с согласованными определениями («глубокое обучение») и генитивными конструкциями («анализ данных»). Для их корректного извлечения необходим синтаксический анализ.

Предлагаемый гибридный метод

Предлагаемый метод состоит из трех основных этапов: предобработка текста, генерация кандидатов и ранжирование кандидатов. Общая архитектура представлена на Рис. 1.

На этапе предобработки текст подвергается следующим операциям:

1. Сегментация: текст разбивается на предложения с помощью библиотеки razdel [10], разработанной специально для русского языка и учитывающей особенности пунктуации научных текстов (сокращения, инициалы, нумерация).

2. Морфологический анализ и лемматизация: для каждого токена определяются лемма (начальная форма) и морфологические признаки (часть речи, падеж, число, род) с помощью библиотеки pymorphy3 [11]. Лемматизация обеспечивает объединение словоформ одной лексемы, что критично для корректного подсчета частотных характеристик.

3. Синтаксический анализ: для построения деревьев синтаксических зависимостей используется библиотека Stanza [12] с русскоязычной моделью. Дерево зависимостей представляет синтаксические связи между словами предложения, где каждая связь характеризуется типом (amod, nmod, conj и др.) и направлением.

Система генерирует кандидатов в ключевые слова с использованием трех стратегий: выделение одиночных кандидатов, N-граммы с POS-фильтрацией и обход дерева зависимостей.

При извлечении одиночных кандидатов все существительные и имена собственные из текста рассматриваются как потенциальные ключевые слова. Данная стратегия обеспечивает базовое покрытие однословных терминов.

Выделение с помощью N-грамм с POS-фильтрацией генерирует биграммы и триграммы, которые фильтруются по шаблонам частей речи. Основные шаблоны:

– ADJ + NOUN («нейронная сеть», «машинное обучение»);

– NOUN + NOUN_gen («анализ данных», «обработка текстов»);

– ADJ + ADJ + NOUN («глубокое нейронное обучение»).

Выделение фраз на основе синтаксических зависимостей – наиболее значимая стратегия, использующая дерево зависимостей для извлечения грамматически корректных

словосочетаний. Алгоритм обходит дерево и извлекает фразы, соответствующие следующим синтаксическим паттернам:

– amod (adjectival modifier): прилагательное + существительное. Пример: «глубокое обучение», «автоматическое реферирование»;

– nmod (nominal modifier): существительное + существительное в косвенном падеже (чаще всего, родительном). Пример: «анализ данных», «извлечение информации»;

– conj (conjunction): однородные члены, соединенные союзом. Пример: «методы и алгоритмы», «точность и полнота»;

– appos (appositional modifier): приложение. Пример: «язык Python», «метрика F1».

Алгоритм рекурсивно обходит зависимости, что позволяет извлекать многокомпонентные термины, например – «разработка новых методов анализа данных».

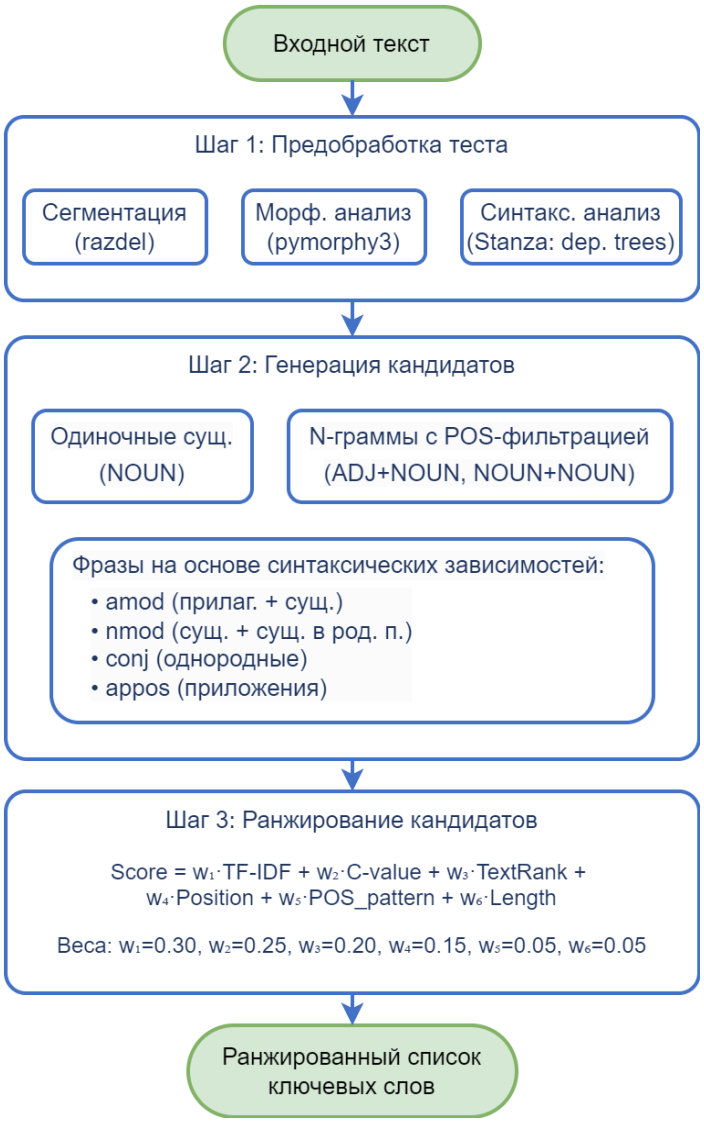

Рисунок 1. Архитектура гибридного метода извлечения ключевых слов.

После генерации кандидаты оцениваются и ранжируются на основе набора признаков. Финальная оценка вычисляется как взвешенная сумма нормализованных признаков:

$$Score(c) = \sum_{i=1}^{n} w_i \cdot f_i(c), \tag{4}$$

где c – кандидат, w_i – вес i-го признака, $f_i(c)$ – нормализованное значение признака.

Признаки ранжирования:

1. TF-IDF ($w_1 = 0{,}30$) – стандартная мера важности термина в контексте документа и коллекции. Вычисляется с помощью TfidfVectorizer из библиотеки scikit-learn [13].

2. C-value ($w_2 = 0{,}25$) – статистическая мера для выделения терминологических словосочетаний [14]. Учитывает частоту фразы и ее вложенность в более длинные фразы:

$$C\text{-}value(a) = \begin{cases} \log_2|a| \cdot f(a), & \text{если } a \text{ не вложен} \\ \log_2|a| \cdot \left(f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b) \right), & \text{иначе} \end{cases}, \quad (5)$$

где $|a|$ – длина кандидата в словах, $f(a)$ – частота появления кандидата a, T_a – множество более длинных фраз, включающих в себя a, $f(b)$ – частота появления кандидата b, включающего в себя a.

1. TextRank ($w_3 = 0{,}20$) – графовая мера центральности слов-кандидатов. Строится граф со-появления в пределах фиксированного окна, важность определяется алгоритмом PageRank.

2. Позиция первого вхождения ($w_4 = 0{,}15$). Основана на гипотезе, что ключевые слова чаще встречаются в начале документа (заголовок, аннотация, введение):

$$Position(c) = 1 - \frac{first\_pos(c)}{doc\_length}. \quad (6)$$

3. Соответствие POS-паттерну ($w_5 = 0{,}05$) – бинарный признак, равный 1, если кандидат соответствует одному из терминологических паттернов (ADJ+NOUN, NOUN+NOUN_gen, …).

4. Штраф за длину ($w_6 = 0{,}05$) учитывает количество слов в кандидате – более длинные специфичные термины получают больший вес, но применяется штраф для избыточно длинных фраз.

Веса подбирались эмпирически на валидационной выборке. Все признаки нормализуются к диапазону [0, 1] методом min-max нормализации.

Описание корпуса данных

Для экспериментальной оценки был сформирован корпус русскоязычных научных статей из открытой электронной библиотеки КиберЛенинка (cyberleninka.ru). Характеристики корпуса представлены в Табл. 2.

Таблица 2. Характеристики экспериментального корпуса.

Параметр	Значение
Количество статей	16818
Тематическая область	компьютерные и информационные науки
Источник	КиберЛенинка
Наличие авторских ключевых слов	100%
Среднее число авторских КС на статью	5–7
Язык	русский

Каждая статья содержит авторские ключевые слова, которые используются в качестве эталона для оценки качества извлечения. Следует отметить, что авторские ключевые слова представляют собой неидеальный эталон: авторы часто ограничены требованиями журналов (3–7 слов) и выбирают более общие термины для широкого охвата аудитории.

Метрики оценки

Для оценки качества извлечения используются стандартные метрики информационного поиска [15]:

– Precision@k – доля релевантных ключевых слов среди топ-k извлеченных: $P@k = \frac{|Extracted_k \cap Gold|}{k}$;

– Recall@k – доля найденных ключевых слов среди всех эталонных: $R@k = \frac{|Extracted»_k \cap Gold|}{|Gold|}$;

– F1@k – гармоническое среднее Precision и Recall: $F1@k = 2 \cdot \frac{P@k \cdot R@k}{P@k + R@k}$;

– MAP (Mean Average Precision) – средняя точность по всем релевантным документам;

– Hit Rate – доля документов, для которых хотя бы одно ключевое слово было найдено;

– типы сопоставления. Учитывая особенности задачи, применяются три режима сопоставления извлеченных и эталонных ключевых слов:

– Exact Matching – точное совпадение строк;

– Lemma Matching – совпадение после приведения к леммам;

– Partial Matching – частичное совпадение: пересечение множеств лемм непустое. Например, «анализ данных» и «анализ больших данных» считаются совпавшими.

Частичное совпадение (Partial Matching) более адекватно отражает семантическое соответствие, поскольку учитывает вариативность формулировок и расширенные версии терминов.

Дополнительные семантические метрики:

– Abstract Coverage – семантическое покрытие: насколько извлеченные ключевые слова покрывают содержание текста. Вычисляется как косинусное сходство между усредненными эмбеддингами ключевых слов и текста статьи (использовалась модель FastText для русского языка [16]);

– Keyword Diversity – разнообразие: среднее попарное косинусное расстояние между эмбеддингами извлеченных ключевых слов. Высокое разнообразие указывает на покрытие различных аспектов документа.

Базовые алгоритмы (бейзлайны)

Предложенный метод сравнивается со следующими базовыми алгоритмами:

– TF-IDF – реализация на основе TfidfVectorizer из scikit-learn с извлечением униграмм и биграмм;

– RAKE – реализация из библиотеки rake-nltk [17] с адаптированным списком стоп-слов для русского языка;

– YAKE – оригинальная реализация из библиотеки yake [4];

– TextRank – реализация из библиотеки summa [18] с настройкой окна со-появления равным 5.

Сравнительный анализ методов

Результаты экспериментальной оценки представлены в Табл. 2–3 и на Рис. 2. Основные метрики приводятся для режима частичного совпадения, наиболее адекватно отражающего семантическое соответствие.

Таблица 3. Результаты сравнения методов (частичное совпадение).

Метод	P@10	R@10	F1@10	MAP	Hit Rate
KWE	**0,277**	**0,355**	**0,295**	**0,294**	0,986
TextRank	0,274	0,345	0,289	0,237	**0,990**
TF-IDF	0,246	0,315	0,261	0,215	0,986
YAKE	0,173	0,220	0,183	0,134	0,945
RAKE	0,030	0,037	0,032	0,017	0,486

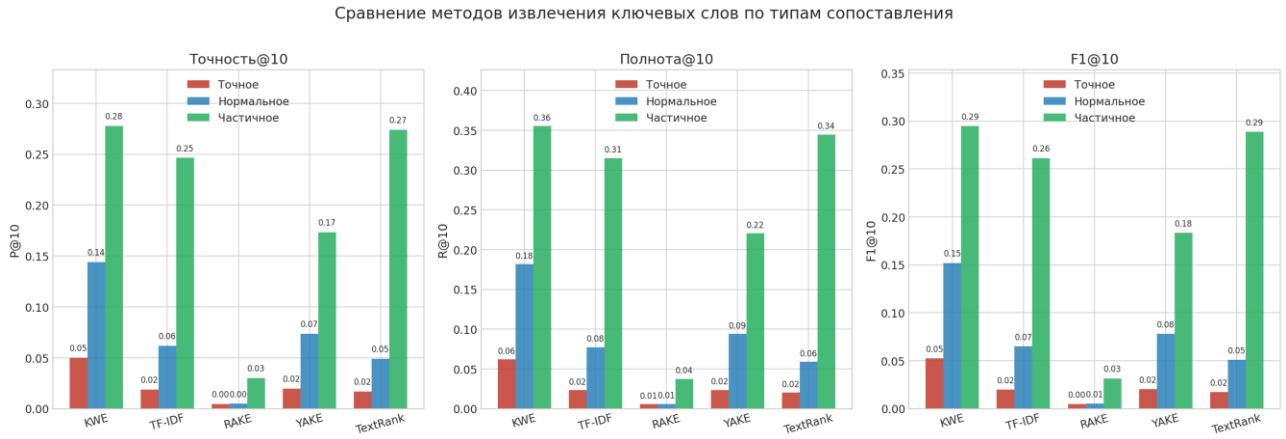

Рисунок 2. Сравнение методов по F1@10 для разных режимов сопоставления.

Предложенный гибридный метод KWE показывает лучшие результаты по основным метрикам качества: F1@10 = 0,268 (превышение над TextRank на 2,2%), MAP = 0,217 (превышение на 6,9%). При комплексной оценке с учетом всех метрик (нормализованный общий балл) KWE занимает первое место с баллом 6,58 против 5,87 у TextRank.

Особенно значительное превосходство наблюдается при сравнении режимов точного и нормального совпадений. Метод KWE показывает F1@10 = 0,146 при нормальном совпадении – почти втрое выше, чем у TextRank (0,052). Это объясняется способностью алгоритма извлекать грамматически корректные многословные термины благодаря использованию деревьев синтаксических зависимостей.

Предложенный метод обеспечивает наилучшее семантическое покрытие текста (Coverage = 0,559), превосходя ближайшего конкурента RAKE (0,489) на 14,3%. Это означает, что извлеченные ключевые слова полнее отражают содержание документа.

Алгоритм RAKE показывает неудовлетворительные результаты на русскоязычном корпусе (F1@10 = 0,042). Основная причина – ориентация на английский язык: алгоритм разделяет текст по стоп-словам без учета морфологического согласования, что приводит к генерации грамматически некорректных или неинформативных фраз.

Таблица 4. Семантические метрики и время обработки.

Метод	Покрытие	Разнообразие	Время обработки, мин
KWE	**0,528**	0,731	423,59
TextRank	0,402	**0,801**	37,55
TF-IDF	0,404	0,772	2,67
YAKE	0,424	0,621	393,64
RAKE	0,480	0,717	3,15

По данным Табл. 4 видно, что предложенный метод требует значительно большего времени обработки (423,59 минут на корпус) по сравнению с простыми методами (TF-IDF: 2,67 мин., RAKE: 3,15 мин.). Это обусловлено затратами на синтаксический анализ библиотекой Stanza. Однако для задач офлайн-индексирования научных коллекций данное время приемлемо.

Качественный анализ результатов

Для иллюстрации различий между методами приведем примеры извлеченных ключевых слов для статьи «Алгоритмы уточнения оси зеркальной симметрии, найденной методом сравнения подцепочек скелетных примитивов» [19] в Табл. 5.

Таблица 5. Пример извлеченных ключевых слов.

Метод	Топ-5 извлеченных ключевых слов
Авторские ключевые слова	зеркальная симметрия; бинарные растровые изображения; подцепочки скелетных примитивов; алгоритмы уточнения.
KWE	ось симметрии; скелетным методом; бинарного изображения; меры симметричности; сравнения подцепочек примитивов.
TextRank	симметрии; фигуры; оси; ось; the.
TF-IDF	симметрии; фигуры; оси; ось; точек.
YAKE	Mottl Vadim Vyacheslavovich; СИММЕТРИИ; Vadim Vyacheslavovich; оси симметрии; фигуры.
RAKE	этом процедура подвергается модификации; поддержке грантов рффи 14; объясняет столь неоднозначный результат; недостатком такого описания является; матрице находят оптимальное выравнивание.

Как видно из примеров в Табл. 5, метод KWE извлекает грамматически корректные многословные термины («ось симметрии», «скелетным методом», «бинарного изображения»), семантически близкие к авторским ключевым словам. Методы TextRank и TF-IDF извлекают только одиночные слова, теряя контекст. Кроме того, в топ попадают служебные слова («the») из-за двуязычных фрагментов в статье. Метод YAKE некорректно идентифицирует имена авторов как ключевые слова, а также извлекает слова в верхнем регистре из заголовка без нормализации. Метод RAKE генерирует длинные фрагменты, не являющиеся терминами («этом процедура подвергается модификации»), демонстрируя неприменимость алгоритма к русскому языку без существенной адаптации.

Заключение

В настоящей работе предложен гибридный метод автоматического извлечения ключевых слов из русскоязычных научно-технических текстов. Метод объединяет морфосинтаксический анализ с использованием библиотек razdel, pymorphy3 и Stanza, генерацию кандидатов на основе обхода деревьев синтаксических зависимостей с учетом паттернов русской научной терминологии (amod, nmod, conj) и многофакторное ранжирование с взвешенной комбинацией признаков (TF-IDF, C-value, TextRank, позиционные характеристики).

Экспериментальная оценка на корпусе из 16818 русскоязычных научных статей подтвердила эффективность предложенного подхода: достигнуты показатель F1@10 = 0,268 (на 2,3% выше лучшего базового метода TextRank), показатель MAP = 0,217 (превышение на 6,9%) и семантическое покрытие текста статьи 0,559, что на 32,5% выше по сравнению с TextRank. Детальный анализ профилей методов по частичным совпадениям, представленный на Рис. 3, наглядно демонстрирует эти конкурентные преимущества.

Рисунок 3. Радиальные профили методов.

В качестве дальнейших перспектив развития исследования планируется реализовать простое ранжирование кандидатов – автоматический подбор весов признаков с помощью методов машинного обучения (градиентный бустинг, нейронные сети), интегрировать контекстные эмбеддинги (BERT, RoBERTa) для улучшения семантического ранжирования, а также оптимизировать времени обработки за счет кэширования и распараллеливания синтаксического анализа.

Литература:

1. Bornmann L., Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Growth rates of modern science // Journal of the Association for Information Science and Technology. 2014. Vol. 66. No. 11. P. 2215–2222.
2. Turney P.D. Learning algorithms for keyphrase extraction // Information Retrieval. 2000. Vol. 2. No. 4. P. 303–336.
3. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24, No. 5. P. 513–523.
4. 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.
5. Mihalcea R., Tarau P. TextRank: Bringing order into texts // Proceedings of EMNLP. 2004. P. 404–411.
6. Wan X., Xiao J. Single document keyphrase extraction using neighborhood knowledge // Proceedings of AAAI. 2008. P. 855–860.

7. Meng R., Zhao S., Han S., He D., Brusilovsky P., Chi Y. Deep keyphrase generation // Proceedings of ACL. 2017. P. 582–592.
8. Boudin F. A Comparison of centrality measures for graph-based keyphrase extraction // Proceedings of IJCNLP. 2013. P. 834–838.
9. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents // Text Mining: Applications and Theory. 2010. P. 1–20.
10. Korobov M. razdel: Rule-based Russian sentence and word tokenization [webpage]. 2020. URL: https://github.com/natasha/razdel.
11. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts. 2015. P. 320–332.
12. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python natural language processing toolkit for many human languages // Proceedings of ACL: System Demonstrations. 2020. P. 101–108.
13. Pedregosa F., Varoqaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
14. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method // International Journal on Digital Libraries. 2000. Vol. 3. No. 2. P. 115–130.
15. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008. 544 p.
16. Grave E., Bojanowski P., Gupta P., Joulin A., Mikolov T. Learning word vectors for 157 languages // Proceedings of LREC. 2018. P. 3483–3487.
17. rake-nltk: Python implementation of RAKE algorithm [webpage]. URL: https://pypi.org/project/rake-nltk/.
18. Barrios F., López F., Argerich L., Wachenchauzer R. Variations of the similarity function of TextRank for automated summarization. arXiv: Computer Science. arXiv:1602.03606. arXiv, 2016.
19. Федотова С.А., Середин О.С., Кушнир О.А. Алгоритмы уточнения оси зеркальной симметрии, найденной методом сравнения подцепочек скелетных примитивов // Известия ТулГУ. Технические науки. 2016. № 11-1. С. 99–111. [Fedotova S.A., Seredin O.S., Kushnir O.A. The algorithms of adjustment of reflection symmetry axis found by the skeleton primitive sub-chains comparison method // Izvestiya Tula State University. 2016. No. 11-1. P. 99–111 (inRussian)].

Благодарности:
Исследование выполнена в рамках государственного задания: FFFU-2025-0010.
The research was conducted within the framework of the state assignment: FFFU-2025-0010.

Об авторах:
ХОРОШИЛОВ Александр Алексеевич, доктор технических наук, профессор, ведущий научный сотрудник Всероссийского института научно-технической информации Российской академии наук, khoroshilov@viniti.ru.
ГОКАРЕВ Вадим Николаевич, аспирант Московского государственного технического университета им. Н.Э. Баумана, vgokarev@mail.ru.
БЕРЕЗКИН Дмитрий Валерьевич, кандидат технических наук, доцент, Московский государственный технический университет им. Н.Э. Баумана, b17a52f@yandex.ru.
ХОРОШИЛОВ Алексей Алексеевич, кандидат технических наук, заведующий лабораторией Всероссийского института научно-технической информации Российской академии наук, alex_khor@viniti.ru.

Metadata:
Title: A hybrid method for extracting keywords from Russian-language scientific and technical texts based on syntactic analysis and multifactor ranking.

Author 1: Alexander Alexievich Khoroshilov, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, 20 Usievich st., Moscow 125190, Russia, khoroshilov@viniti.ru.
Author 2: Vadim Nikolaevich Gokarev, Bauman Moscow State Technical University, 5 2-ya Baumanskaya st., 105005 Moscow, Russia, vgokarev@mail.ru, ORCID: 0009-0002-4082-2925.
Author 3: Dmitriy Valeryevich Berezkin, Bauman Moscow State Technical University, 5 2-ya Baumanskaya st., Moscow, Russia, b17a52f@yandex.ru, ORCID: 0000-0001-7149-6072.
Author 4: Alexey Alexievich Khoroshilov, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, 20 Usievich st., Moscow 125190, Russia, alex_khor@viniti.ru.
Abstract: This article is devoted to the development of a hybrid method for automatically extracting keywords from Russian-language scientific and technical texts. The proposed method combines morphosyntactic analysis, candidate generation based on syntactic dependency tree traversal, and multivariate ranking using a combination of statistical (TF-IDF, C-value), graph (TextRank), and positional features. An experimental evaluation on a corpus of 16,818 Russian-language scientific articles from CyberLeninka showed that the proposed method outperforms baseline algorithms (TF-IDF, RAKE, YAKE, TextRank) in terms of F1@10 (2.3% improvement – 0.268 versus 0.262 for the closest competitor TextRank), MAP (6.9% improvement – 0.217 versus 0.203), and semantic text coverage (32.5% improvement – 0.559 versus 0.422). A particularly significant advantage is achieved when extracting multi-word terminological phrases due to the consideration of syntactic relationships between components.
Keywords: keyword extraction, natural language processing, dependency trees, scientific and technical texts.