

АНАЛИЗ МЕТОДОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ ИЗ ТЕКСТОВЫХ ДАННЫХ

Ш. И. Мутаиров^а, К. И. Бушмелева^б*Сургутский государственный университет, Сургут, Российская Федерация*^а *usermage@gmail.com*^б *bkiya@yandex.ru*

Аннотация: статья посвящена сравнительному анализу методов извлечения ключевых слов из текстовых данных с использованием статистических и нейросетевых подходов, таких как YAKE, и языковых моделей на основе трансформеров (например, BERT). В работе исследуются их точность, производительность и применимость к различным типам текстов. Основное внимание уделено разработке рекомендаций по оптимизации этих методов для повышения эффективности обработки текстов в условиях роста объемов данных. Результаты основаны на экспериментах с реальными текстами из научных статей.

Ключевые слова: обработка естественного языка, извлечение ключевых слов, YAKE, BERT, точность, производительность.

Для цитирования: Мутаиров Ш. И., Бушмелева К. И. Анализ методов извлечения ключевых слов из текстовых данных. *Успехи кибернетики*. 2025;6(2):86–91.

Поступила в редакцию: 14.03.2025.

В окончательном варианте: 02.04.2025.

COMPARATIVE ANALYSIS OF KEYWORD EXTRACTION STRATEGIES

Sh. I. Mutairov^а, K. I. Bushmeleva^б*Surgut State University, Surgut, Russian Federation*^а *usermage@gmail.com*^б *bkiya@yandex.ru*

Abstract: we conducted a comparative analysis of keyword extraction methods based on statistical and neural network approaches, including YAKE and transformer-based language models such as BERT. We evaluated their accuracy, performance, and suitability for different text types. Our primary goal was to give recommendations for optimizing these methods to enhance text processing efficiency as data volumes increase. We performed experiments with real scientific texts.

Keywords: natural language processing, keyword extraction, YAKE, BERT, accuracy, performance.

Cite this article: Mutairov Sh. I., Bushmeleva K. I. Comparative Analysis of Keyword Extraction Strategies. *Russian Journal of Cybernetics*. 2025;6(2):86–91.

Original article submitted: 14.03.2025.

Revision submitted: 02.04.2025.

Введение

Рост объемов данных в цифровой среде сопровождается увеличением потребности в автоматизированных методах анализа текстов [1]. Извлечение ключевых слов находит широкое применение в различных областях:

- научные исследования: упрощение индексации статей и поиска релевантной информации;
- информационные системы: повышение качества поиска и классификации документов;
- маркетинг и аналитика: анализ текстовых данных для выявления трендов и предпочтений пользователей.

Традиционные подходы к извлечению ключевых слов, такие как статистические методы (TF-IDF, YAKE), успешно справляются с задачей в условиях ограниченных ресурсов, однако их возможности ограничены при анализе сложных текстов с глубоким контекстом [2], [3]. С развитием языковых моделей (ЯМ) на основе трансформеров (например, BERT) появилась возможность учитывать семантические связи и контекст, что значительно повышает качество извлечения. Однако такие модели требуют больших вычислительных ресурсов и времени, что ограничивает их применение в реальных задачах с большими объемами данных [4].

Целью данного исследования является сравнительный анализ методов извлечения ключевых слов — статистического метода YAKE и языковой модели BERT — с точки зрения их точности, скорости и применимости. В статье также рассматриваются пути оптимизации этих методов для

достижения баланса между качеством и производительностью.

Сравнительный анализ методов извлечения ключевых слов

Существующие методы различаются по подходам: статистические алгоритмы (например, YAKE, RAKE) быстры и эффективны для коротких текстов, тогда как модели глубокого обучения (BERT) лучше справляются с длинными и сложными текстами, требующими учета контекста. Выбор подходящего метода зависит от задачи, объема данных и доступных ресурсов, что делает сравнительный анализ особенно актуальным.

В таблице 1 представлено сравнение характеристик трех подходов к извлечению ключевых слов.

Таблица 1

Сравнение подходов к извлечению ключевых слов

Характеристика	RAKE	YAKE	Языковая модель
Подход	Статистический	Статистический + контекст	Нейронные сети, семантический анализ
Точность	Низкая	Средняя	Высокая
Контекст	Не учитывается	Учитывается частично	Глубокий анализ контекста
Скорость	Очень высокая	Высокая	Низкая
Ресурсы	Минимальные	Низкие	Высокие
Мультиязычность	Ограничена	Частично	Полная (при поддержке модели)
Простота настройки	Простая	Простая	Сложная
Применение	Для коротких текстов	Для коротких и средних текстов	Для сложных и длинных текстов

В текущем исследовании рассмотрены варианты использования YAKE и языковой модели BERT. RAKE не подходит, так как его низкая точность ограничивает его применимость.

Метод YAKE

YAKE (Yet Another Keyword Extractor) — это статистический метод, использующий локальные особенности текста для ранжирования ключевых слов [5, 6]. Он основан на комбинации нескольких метрик, объединенных в итоговую оценку:

$$\hat{f}_{score}(k_i) = TF(k_i) \cdot Pos(k_i) \cdot Spread(k_i) \cdot PC(k_i) \cdot Case(k_i), \quad (1)$$

где $TF(k_i)$ — частота термина в тексте;

$Pos(k_i)$ — взвешенная позиция слова (слова в начале текста считаются более значимыми);

$Spread(k_i)$ — разброс вхождений слова по тексту;

$PC(k_i)$ — плотность соседних слов (учитывает контекст);

$Case(k_i)$ — чувствительность к регистру.

Чем ниже значение $\hat{f}_{score}$, тем выше значимость фразы k_i .

Метод на основе BERT

BERT (Bidirectional Encoder Representations from Transformers) — это языковая модель, использующая трансформеры для создания контекстных эмбеддингов [4, 7]. Алгоритм извлечения ключевых слов с помощью BERT включает следующие шаги:

1. Получение векторов: создание эмбеддингов для текста и отдельных слов/фраз.
2. Выбор кандидатов: извлечение потенциальных ключевых слов (n-грамм).
3. Сравнение: вычисление косинусного сходства между вектором текста и векторами кандидатов.
4. Ранжирование: выбор слов с наибольшим сходством.

В данном исследовании применялась модель `deeppavlov/rubert-base-cased` (180М параметров), адаптированная для русского языка.

Экспериментальное исследование

Для анализа было обработано 18 научных статей. В таблицах 2 и 3 представлены результаты извлечения данных по 6 статьям. Каждая статья сопровождалась списком ключевых слов, заданных авторами, что позволило оценить точность методов. Тестирование проводилось на следующих параметрах:

- процент совпадений извлеченных слов с авторскими ключевыми словами;
- среднее время обработки одного файла в секундах.

Таблица 2

Результаты извлечения ключевых слов из статей с помощью языковой модели

№	Тема статьи	Ключевые слова (автор статьи)	Языковая модель (deeppavlov/rubert-based – 180M)
1.	RFM-анализ работы автозаправочных станций в системе Loginom	RFM-анализ, автозаправочные станции, система Loginom, low-code платформа	решения автоматизации, технологическое решение, student scientific, основными драйверами, rfm analysis
2.	Применение искусственных нейронных сетей в идентификации значимых электрокардиографических параметров операторов технологических установок	искусственные нейронные сети, ЭКГ-параметры, значимые диагностические признаки	операторов технологических, способами статистическая, ai обработка, биофизики, electrocardiographic parameters
3.	Метод расчета коэффициента достоверности на уровне заголовков сегментов и пакетов	коэффициент достоверности данных, заголовки сегмента и пакета, расчет, безопасность данных	исследования сетевого, модификацию трафика, зашифрованного, использование коэффициента, detect data
4.	Многоразрядные вычисления как область приложения модулярной арифметики	многоразрядные вычисления, модулярная система счисления, модулярная система счисления квадратичного диапазона, модульные операции, математическое моделирование, программирование на языке Python, оценка сложности алгоритмов	решения автоматизации, вычислительной теории, уровне машинных, компьютере выполнить, проблем прикладной
5.	Прототип роботизированной системы транспортировки деталей	роботизированная система, Robotino, транспортировка, компьютерное зрение	робототехники основанный, prototype of, wiki openrobotino, scientific supervisor, transportation computer
6.	Проектирование информационно-аналитической системы «Достижения студентов СурГУ»	информационно-аналитическая система, функциональная модель, информологическая и физическая модели	информационно-аналитическая, возможностей системы, диаграммы функциональной, образовательным учреждениям, типов данных

Таблица 3

Результаты извлечения ключевых слов с помощью YAKE

№	Тема статьи	Ключевые слова (автор статьи)	YAKE
1.	RFM-анализ работы автозаправочных станций в системе Loginom	RFM-анализ, автозаправочные станции, система Loginom, low-code платформа	RFM, LOGINOM, клиентов, RFM ANALYSIS, анализа, АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ УДК
2.	Применение искусственных нейронных сетей в идентификации значимых электрокардиографических параметров операторов технологических установок	искусственные нейронные сети, ЭКГ-параметры, значимые диагностические признаки	ОПЕРАТОРОВ ТЕХНОЛОГИЧЕСКИХ УСТАНОВОК, ЭКГ-параметров операторов технологических, ЭКГ-параметров операторов, обработки ЭКГ-параметров операторов, ИДЕНТИФИКАЦИИ ЗНАЧИМЫХ ЭЛЕКТРОКАРДИОГРАФИЧЕСКИХ
3.	Метод расчета коэффициента достоверности на уровне заголовков сегментов и пакетов	коэффициент достоверности данных, заголовки сегмента и пакета, расчет, безопасность данных	коэффициент достоверности данных, достоверности данных сетевого трафика, компьютерной сети предприятия, РАСЧЕТА КОЭФФИЦИЕНТА ДОСТОВЕРНОСТИ
4.	Многоразрядные вычисления как область приложения модулярной арифметики	многоразрядные вычисления, модулярная система счисления одинарного диапазона, модулярная система счисления квадратичного диапазона, модульные операции, математическое моделирование, программирование на языке Python, оценка сложности алгоритмов	МСС квадратичного диапазона, МСС квадратичного, квадратичного диапазона, МСС, квадратичного, диапазона
5.	Прототип роботизированной системы транспортировки деталей	роботизированная система, Robotino, транспортировка, компьютерное зрение	Иван Сергеевич Курочкин, ДЕТАЛЕЙ Иван Сергеевич, Сургутский государственный университет, ПРОТОТИП РОБОТИЗИРОВАННОЙ СИСТЕМЫ, ПРОТОТИП РОБОТИЗИРОВАННОЙ
6.	Проектирование информационно-аналитической системы «Достижения студентов СурГУ»	информационно-аналитическая система, функциональная модель, инфологическая и физическая модели	Сургутский государственный университет, ДОСТИЖЕНИЯ СТУДЕНТОВ СУРГУ, ДОСТИЖЕНИЯ СТУДЕНТОВ, кафедры АСОИУ, СТУДЕНТОВ СУРГУ

В таблице 4 приведена оценка экспериментальных данных по ранее представленным методам.

Таблица 4

Оценка экспериментальных данных

Характеристика	YAKE	Языковая модель
Точность	40%. Демонстрирует более высокую точность благодаря статистическому подходу, который лучше выявляет часто встречающиеся термины	30%. Показывает меньшую точность, так как склонен извлекать слова с глубоким контекстным значением, которые не всегда совпадают с авторскими ожиданиями
Производительность	0.30 сек.	1.52 сек.

Анализ экспериментальных данных, представленный в таблице 4, показал, что YAKE лучше справляется с текстами, где ключевые слова представляют собой часто повторяющиеся термины или фразы (например, «RFM», «коэффициента достоверности данных», «прототип роботизированной системы»). Это делает его более подходящим для статей с технической направленностью, где терминология играет ключевую роль.

BERT показывает потенциал в извлечении контекстно-зависимых слов, что может быть полезно для текстов с более абстрактной тематикой или сложной структурой, однако в данном эксперименте его результаты оказались менее точными из-за несоответствия с авторскими ожиданиями.

Оптимизация методов

Для получения преимущества обоих методов и достижения баланса между точностью и производительностью предлагается 2 варианта гибридного алгоритма (таблица 5).

Таблица 5

Описание вариантов комбинации методов для извлечения ключевых слов

№	Вариант комбинации	Преимущества	Недостатки	Вывод
1.	YAKE формирует кандидатов, ЯМ пересортирует и берет топ-N	быстрый отбор кандидатов (YAKE); семантическая точность от ЯМ; снижение шума	зависит от качества кандидатов YAKE; возможна потеря редких, но важных слов	хороший баланс скорости и точности, но ограничен первым шагом
2.	Средний вес от YAKE и KeyBERT, затем ранжирование	баланс статистики и семантики; учет сильных сторон обоих методов; высокая точность	сложность нормализации весов; возможны искажения при усреднении	лучший подход по точности и гибкости, но чуть сложнее в реализации

В первом варианте метод YAKE как статистический алгоритм быстро анализирует текст и выделяет потенциальные ключевые слова на основе частоты, позиции и других локальных характеристик. Это позволяет минимизировать начальные вычисления и сосредоточиться на наиболее вероятных терминах. Далее с помощью языковой модели пересчитываются веса кандидатов с использованием косинусного сходства между векторами текста и отдельных слов или фраз. Это улучшает точность за счет учета семантического контекста.

Во втором варианте метод объединяет статистические веса от YAKE с контекстными эмбедами от языковой модели. Средний вес позволяет сбалансировать точность и релевантность. Ранжирование на основе среднего веса снижает влияние ошибок одного из методов. Например, если YAKE переоценит частотный термин, а языковая модель скорректирует его семантическую значимость, итоговый результат будет более надежным.

Вариант 1 предпочтителен для сценариев с ограниченными ресурсами или необходимостью быстрой обработки (например, обработка потоков данных в реальном времени).

Вариант 2 лучше подходит для задач, где приоритет отдается максимальной точности и устойчивости, таких как анализ научной литературы или написание аннотаций.

Заключение

Экспериментальные данные показывают, что YAKE является более эффективным методом для извлечения ключевых слов из научных текстов с технической направленностью, обеспечивая точность 40% и высокую скорость обработки (0.30 сек.). BERT, несмотря на меньшую точность (30%) и большее время обработки (1.52 сек.), демонстрирует потенциал для задач, требующих глубокого контекстного анализа.

Ограничения эксперимента связаны с выборкой текстов: все статьи имеют техническую направленность, что могло повлиять на результаты в пользу YAKE. В дальнейшем целесообразно протестировать методы на текстах с более абстрактной тематикой, таких как гуманитарные или художественные произведения, чтобы оценить эффективность BERT в условиях сложного контекста.

Комбинированные подходы, объединяющие преимущества YAKE и BERT, позволяют достичь оптимального баланса между точностью и производительностью, что делает их перспективными для дальнейшего развития и применения в системах обработки текстов.

Результаты исследования могут быть использованы в системах автоматической индексации, научных поисковиках и аналитических платформах. В дальнейшем планируется расширить эксперименты и интегрировать методы с полнотекстовым и семантическим поиском.

ЛИТЕРАТУРА

1. 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;509:257–289.
2. Ванюшкин А. С., Гращенко Л. А. Оценка алгоритмов извлечения ключевых слов: инструментарий и ресурсы. *Новые информационные технологии в автоматизированных системах*. 2017;20:95–102. Режим доступа: <https://cyberleninka.ru/article/n/otsenka-algoritmov-izvlecheniya-klyuchevyh-slov-instrumentariy-i-resursy>.
3. Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов. *Новые информационные технологии в автоматизированных системах*. 2016;19:85–93. Режим доступа: <https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov>.
4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.
5. Горлевич Д. Алгоритмы для выделения ключевых слов: Rake, YAKE!, TextRank. *NewTechAudit*. Режим доступа: <https://newtechaudit.ru/algoritmy-dlya-vydeleniya-klyuchevyh-slov-rake-yake-textrank>.
6. Мутаиров Ш. И., Бушмелева К. И. Алгоритмы обработки и вычисления сходства текстовых данных пользователей социальных сетей. *Успехи кибернетики*. 2023;4(1):33–38.
7. Sharma P., Li Y. Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labeling. *Preprints.org*. Режим доступа: <https://www.preprints.org/manuscript/201908.0073/v1>.