DOI: https://doi.org/10.15276/ict.02.2025.48
УДК 004.91/.77

Аналіз та вибір методів виявлення ключових слів у текстах: огляд існуючих підходів і практичне застосування

Діденко Тарас Володимирович[1)]
Магістр каф. програмної інженерії
ORCID: https://orcid.org/0009-0006-7880-2638; askuk2077@gmail.com
Кунгурцев Олексій Борисович[1)]
Канд. техніч. наук, професор каф. Програмної інженерії
ORCID: http:// orcid.org/ 0000-0002-3207-7315; kungurtsev@op.edu.ua
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У роботі розглядається проблема автоматичного виявлення ключових слів у текстах як важливого етапу обробки природної мови (NLP). Актуальність теми обумовлена стрімким зростанням обсягів текстових даних, що потребують систематизації та аналізу.

Проаналізовано основні підходи до виділення ключових слів: класичні статистичні методи (TF-IDF, RAKE, TextRank), сучасні семантичні алгоритми (BERT, KeyBERT, embeddings із кластеризацією), а також сторонні інструменти та API (ConceptNet, spaCy, HuggingFace Transformers). Показано, що статистичні методи відзначаються простотою реалізації, однак поступаються сучасним моделям за точністю, оскільки не враховують контекст і семантику.

Семантичні підходи забезпечують вищу якість результатів, проте є більш ресурсоємними. Особливу увагу приділено практичним експериментам з українськими текстами, які попередньо перекладалися англійською для використання англомовних моделей. Такий підхід дозволив отримати кращі результати, оскільки більшість бібліотек оптимізовані саме для англомовних корпусів. Однак спроби зворотного перекладу виявили проблеми зі збереженням змісту. Експериментальні дослідження показали, що KeyBERT продемонстрував найвищу ефективність серед розглянутих методів: він поєднує релевантність результатів, швидкодію та простоту інтеграції, що робить його придатним як для наукових досліджень, так і для прикладних інформаційних систем.

У висновках обґрунтовується доцільність використання KeyBERT у поєднанні з англомовними текстами як оптимального рішення для задачі виявлення ключових слів. Також окреслено перспективні напрями розвитку: підтримка мультимовних корпусів, адаптація під доменні тексти та оптимізація моделей для роботи з великими масивами даних.

Ключові слова: обробка природної мови; ключові слова» TF-IDF; RAKE; TextRank; BERT; KeyBERT; embeddings; spaCy; ConceptNet

Вступ. Автоматичне виявлення ключових слів у текстах є однією з базових і водночас найважливіших задач сучасної обробки природної мови (NLP). Воно дозволяє швидко ідентифікувати найбільш інформативні одиниці тексту, які відображають його основний зміст і можуть бути використані як компактні репрезентації для подальшого аналізу. Актуальність цієї проблеми безпосередньо пов’язана зі стрімким зростанням обсягів цифрової інформації, що продукується щодня у світі. Сотні тисяч нових статей, публікацій у соціальних мережах, наукових матеріалів та технічної документації потребують ефективної систематизації, і без автоматизованих інструментів опрацювати такі дані стає практично неможливо.

Завдання виявлення ключових слів має міждисциплінарний характер, адже поєднує в собі методи комп’ютерних наук, лінгвістики та штучного інтелекту.

Успішне його розв’язання відкриває можливості для розвитку більш складних інформаційних технологій, зокрема:

– пошукових систем, де ключові слова використовуються для формування релевантної видачі;

– рекомендаційних сервісів, що спираються на інтереси користувачів, відображені у текстових даних;

– систем моніторингу соціальних мереж, які дозволяють виявляти тренди, поширені теми та навіть прогнозувати суспільні настрої;

– наукової аналітики, де автоматизоване витягування ключових термінів сприяє швидшому опрацюванню великих обсягів академічних публікацій.

Незважаючи на широкий спектр існуючих методів, проблема вибору оптимального підходу залишається відкритою. Класичні статистичні алгоритми відзначаються простотою і швидкодією, однак часто втрачають точність через ігнорування контексту. Сучасні семантичні моделі, натомість, враховують значення слова у певному оточенні та демонструють значно кращі результати, проте є більш ресурсоємними і складними в реалізації.

Таким чином, дослідження і порівняння різних методів виділення ключових слів, а також визначення їх придатності для роботи з україномовними текстами, становлять не лише науковий інтерес, але й мають велике практичне значення для розробки ефективних програмних рішень у сфері аналізу даних.

Огляд існуючих підходів і бібліотек

Класичні статистичні методи

Одними з перших підходів до автоматичного виділення ключових слів були статистичні методи, які спираються на кількісні характеристики тексту.

1. TF-IDF (Term Frequency – Inverse Document Frequency) [1] – класичний підхід, що визначає вагу терміна на основі частоти його використання у документі та рідкості в усьому корпусі текстів. Цей метод був широко застосований у пошукових системах і завданнях інформаційного пошуку ще у 90-х роках. Його перевага полягає у простоті реалізації та швидкості роботи, особливо на великих колекціях документів. Проте обмеження очевидні: TF-IDF не враховує семантичних відносин між словами, чутливий до морфологічних варіацій і сильно залежить від розміру корпусу.

2. RAKE (Rapid Automatic Keyword Extraction) [2] – метод, що виділяє кандидати на ключові слова шляхом поділу тексту на фрази й подальшого статистичного зважування. Його часто застосовують для великих текстових колекцій, наприклад, при аналізі наукових статей або
новинних потоків. Основна перевага RAKE – швидкість і легкість у використанні без потреби додаткових лінгвістичних ресурсів. Водночас на коротких і насичених текстах (твіттер-повідомлення, відгуки) він часто дає слабкі результати.

3. TextRank [3] – графовий метод, що моделює текст як мережу зв’язків між словами чи фразами та застосовує алгоритм PageRank для визначення їхньої важливості. TextRank показує вищу якість результатів порівняно з TF-IDF та RAKE, особливо у випадку складних і структурованих документів. Він часто використовується для автоматичного підсумовування та тематичного аналізу. Недоліком є потреба у значних обчислювальних ресурсах і складніша реалізація.

Таким чином, класичні статистичні методи стали основою для розвитку подальших підходів, однак їхня ефективність обмежена через ігнорування контекстуальних і семантичних залежностей між словами.

Сучасні семантичні методи

З поширенням глибинного навчання і трансформерних моделей відбувся перехід від статистичних характеристик до семантичних представлень.

1. BERT [4] та його похідні (multilingual BERT, DistilBERT, RoBERTa) дали можливість враховувати значення слова в конкретному контексті. Це означає, що одна й та сама лексема може мати різне представлення залежно від сусідніх слів. Завдяки цьому такі моделі успішно застосовуються у завданнях пошуку інформації, аналізу тональності, чат-ботах та інтелектуальних системах підтримки. Основним недоліком залишається висока вимогливість до ресурсів – навіть на середніх корпусах обробка вимагає потужних обчислювальних засобів.

2. KeyBERT [5] – спеціалізована бібліотека, що поєднує вбудовування з BERT із простим алгоритмом пошуку найбільш релевантних ключових слів. Її головна перевага — баланс між якістю та швидкістю. KeyBERT має інтуїтивно зрозумілий API, що робить його придатним як для швидких дослідницьких прототипів, так і для інтеграції у виробничі

рішення. У багатьох випадках він демонструє результати, порівнянні з більш складними моделями, але при цьому значно простіший у використанні.

3. Embeddings + кластеризація – методологія, коли слова чи словосполучення переводяться у векторний простір, після чого групуються за допомогою алгоритмів кластеризації (наприклад, k-means, DBSCAN). Це дозволяє виявляти семантичні кластери та виділяти найбільш репрезентативні терміни. Підхід добре працює для великих корпусів, однак ефективність залежить від вибору параметрів кластеризації та векторних моделей. На малих текстах результат часто виходить нестабільним.

Сторонні API та бібліотеки

Крім локальних методів, існує низка бібліотек та API, що допомагають розширити можливості аналізу текстів.

1. ConceptNet API [6] – семантична база знань, яка відображає зв'язки між словами та поняттями (наприклад, «кіт» → «тварина», «пухнастий», «мяукає»). Це робить її корисною для розширення списку ключових слів і побудови онтологій. Однак точність витягу саме ключових слів залишається невисокою.

2. spaCy [7] – універсальна NLP-бібліотека з широким набором інструментів: токенізація, лематизація, визначення частин мови, NER (розпізнавання іменованих сутностей). Її часто застосовують як підготовчий етап для більш складних методів. Для довгих текстів spaCy працює стабільно, але на коротких корпусах точність виявлення ключових слів є невисокою.

3. HuggingFace Transformers [8] – найбільший репозиторій сучасних мовних моделей. Він надає доступ до десятків попередньо натренованих моделей (BERT, RoBERTa, DistilBERT тощо), які можна адаптувати під конкретні задачі. Головна перевага – гнучкість та величезна спільнота розробників, що забезпечує підтримку і швидкий розвиток інструментів.

Практичні експерименти

У ході експериментів спочатку було застосовано підхід із перекладом текстів. Українські тексти перекладалися англійською, після чого виконувалося витягування ключових слів за допомогою англомовних моделей. Це дало задовільний результат, адже більшість сучасних NLP-інструментів найкраще працюють саме з англійською мовою.

Однак при зворотному перекладі результатів на українську виникали проблеми: частина термінів спотворювалася або втрачала точність значення (наприклад, «artificial intelligence» перекладалося як «штучний інтелект», але словосполучення «intelligent system» могло перетворюватися на «розумна система» замість «інтелектуальна система»). Це ускладнювало використання методу в реальних умовах, тому було прийнято рішення залишати результати в англомовному форматі.

Додатково було протестовано кілька бібліотек:

– ConceptNet API надавав широкі семантичні зв'язки, але при безпосередньому виділенні ключових слів результати були малокорисними;

– Embeddings з кластеризацією дозволили сформувати семантичні групи термінів, проте виявилися надмірно ресурсоємними для невеликих текстів;

– spaCy дало посередні результати через невеликі обсяги аналізованих текстів;

– KeyBERT показав найкращий баланс між якістю та простотою використання: ключові слова були релевантними, а інтеграція у систему пройшла без складнощів.

Загалом класичні методи (TF-IDF, RAKE, TextRank) забезпечують базову ефективність, але поступаються сучасним семантичним підходам, особливо коли йдеться про короткі та варіативні тексти. Методи embeddings є перспективними, проте їхня складність і ресурсоємність обмежують практичне застосування. Найбільш збалансованим інструментом виявився KeyBERT.

Таблиця 1. **Порівняльний аналіз**

Метод/ Бібліотека	Переваги	Недоліки
TF-IDF	Простота, швидкість	Ігнорує контекст, низька точність
RAKE	Легкий у використанні, ефективний на довгих текстах	Слабкі результати на коротких текстах
TextRank	Враховує структуру тексту, краща точність	Витратний за ресурсами
BERT + Embeddings	Висока точність, контекстна обробка	Високі вимоги до ресурсів
KeyBERT	Баланс якості та швидкості, простота інтеграції	Залежність від моделей BERT
spaCy	Гнучка NLP-бібліотека, зручний інструментарій	Слабкі результати у витягу ключових слів
ConceptNet API	Семантичні зв’язки між словами	Низька точність для цієї задачі

Обґрунтування вибору остаточного рішення

У результаті порівняння обрано бібліотеку KeyBERT як оптимальне рішення. Вона забезпечує високу якість результатів навіть для невеликих обсягів даних, має зрозумілий API і не вимагає надмірних ресурсів. Використання англомовних моделей у поєднанні з KeyBERT дозволяє стабільно отримувати релевантні ключові слова, що робить цей інструмент придатним як для дослідницьких задач, так і для інтеграції у виробничі середовища.

Висновки. Автоматичне виявлення ключових слів у текстах залишається надзвичайно актуальною задачею сучасної обробки природної мови. Воно є базовим інструментом, що дозволяє систематизувати великі масиви даних, підвищувати ефективність пошукових систем, забезпечувати точність у рекомендаційних сервісах та слугує підґрунтям для аналізу тенденцій у соціальних мережах і наукових дослідженнях.

Проведений огляд та практичні експерименти показали, що класичні статистичні методи (TF-IDF, RAKE, TextRank) мають переваги у швидкості й простоті, проте їхня здатність відображати глибинний зміст тексту є обмеженою. Сучасні семантичні підходи, особливо ті, що базуються на BERT та його похідних, демонструють значно кращу точність і враховують контекст, але вимагають більших ресурсів і налаштувань.

KeyBERT, який поєднує в собі контекстуальні можливості моделей на основі BERT та простоту інтеграції, продемонстрував оптимальний баланс між якістю результатів і практичною зручністю використання. Досвід роботи з українськими текстами засвідчив, що їх переклад на англійську перед обробкою дозволяє досягти вищої якості витягу ключових слів, адже більшість сучасних бібліотек орієнтовані саме на англомовні корпуси.

Таким чином, доцільно рекомендувати застосування KeyBERT у поєднанні з англомовним представленням текстів як ефективне рішення для наукових і прикладних інформаційних систем. Водночас перспективним напрямом є розвиток мультимовних моделей, здатних працювати безпосередньо з українськими корпусами, а також адаптація алгоритмів під специфічні галузі знань – медицину, юриспруденцію, технічну документацію. Не менш важливим завданням майбутніх досліджень є оптимізація моделей для обробки великих потоків даних у реальному часі.

Отже, проведена робота має як теоретичне, так і прикладне значення, адже узагальнює існуючі підходи, демонструє їхні переваги та обмеження, а також визначає найефективніші інструменти для розв’язання конкретних практичних завдань у сфері обробки природної мови.

СПИСОК ЛІТЕРАТУРИ

1. Jones K. S. “A statistical interpretation of term specificity and its application in retrieval”. *Journal of Documentation*. 1972; 28 (1): 11–21.

2. Rose S., Engel D., Cramer N., Cowley W. “Automatic keyword extraction from individual documents”. *Text Mining: Applications and Theory*. *Wiley*, 2010. p. 1–20.

3. Mihalcea R., Tarau P. “TextRank: Bringing order into text”. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2004. p. 404–411.

4. Devlin J., Chang M.-W., Lee K., Toutanova K. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *Proceedings of NAACL-HLT*. 2019. p. 4171–4186.

5. Grootendorst M. “KeyBERT: Minimal keyword extraction with BERT”. *GitHub Repository*. 2020. – URL: https://github.com/MaartenGr/KeyBERT.

6. Speer R., Chin J., Havasi C. “ConceptNet 5.5: An open multilingual graph of general knowledge”. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2017. p. 4444–4451.

7. Honnibal M., Montani I., Van Landeghem S., Boyd A. “spaCy: Industrial-strength Natural Language Processing in Python”. *Journal of Open Source Software*. 2020; 5 (56): 1–5.

8. Wolf T., Debut L., Sanh V., Chaumond J., et al. “Transformers: State-of-the-art natural language processing”. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020. p. 38–45.

DOI: https://doi.org/10.15276/ict.02.2025.48
UDC 004.91/.77

Analysis and selection of keyword extraction methods in texts: review of existing approaches and practical application

Taras V. Didenko[1)]
Master’s Student of the Department of Software Engineering
ORCID: https://orcid.org/0009-0006-7880-2638; askuk2077@gmail.com
Oleksii B. Kungurtsev[1)]
PhD, Professor of the Department of Software Engineering
ORCID: http://orcid.org/0000-0002-3207-7315; kungurtsev@op.edu.ua
[1)] Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

The article addresses the problem of automatic keyword extraction from texts, which is an important stage in natural language processing (NLP). The relevance of the topic is driven by the rapid growth of textual data, which requires systematic organization and analysis. The main approaches to keyword extraction are analyzed: classical statistical methods (TF-IDF, RAKE, TextRank), modern semantic algorithms (BERT, KeyBERT, embeddings with clustering), as well as third-party tools and APIs (ConceptNet, spaCy, HuggingFace Transformers). It is shown that statistical methods are simple to implement but are less accurate than modern models because they do not account for context and semantics. Semantic approaches provide higher-quality results, although they are more resource-intensive. Particular attention is given to practical experiments with Ukrainian texts, which were pre-translated into English to use English-language models. This approach allowed better results, as most libraries are optimized for English corpora. However, attempts at back-translation revealed issues with preserving the original meaning. Experimental studies showed that KeyBERT demonstrated the highest effectiveness among the considered methods: it combines result relevance, speed, and ease of integration, making it suitable for both scientific research and applied information systems.

In conclusion, the use of KeyBERT in combination with English-language texts is justified as the optimal solution for keyword extraction tasks. Prospective directions for development are also outlined: support for multilingual corpora, adaptation to domain-specific texts, and optimization of models for large-scale data processing.

Keywords: Natural language processing; keywords; TF-IDF; RAKE; TextRank; BERT; KeyBERT; embeddings; spaCy; ConceptNet