ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

УДК 004.93:159.95

О. В. БІСІКАЛО, О. В. ЯХИМОВИЧ

Вінницький національний технічний університет, м. Вінниця

ЗНАХОДЖЕННЯ КЛЮЧОВИХ СЛІВ АНГЛОМОВНОГО ТЕКСТУ ЗА ДОПОМОГОЮ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ ПАКЕТУ DKPRO CORE

Анотація. Запропоновано новий підхід до визначення ключових слів, який базується на знаходженні зв'язків між словоформами англомовного тексту за допомогою інструментальних можливостей пакету DKPro Core. Проілюстрований аналізом прикладів застосування підхід спрямовано на розв'язання задач ефективної обробки текстових документів – індексування, реферування, кластеризації та класифікації. В результаті теоретичного та експериментального дослідження встановлено, що розроблений підхід забезпечує кращу релевантність знаходження ключових слів у порівнянні з аналогами за критеріями точності та повноти.

Ключові слова: ключові слова, англійська мова, релевантність, лінгвістичний пакет, DKPro Core, синтаксичний аналіз

Аннотация. Предложен новый поход к определению ключевых слов, основанный на нахождении связей между словоформами англоязычного текста с помощью инструментальных возможностей пакета DKPro Core.Проиллюстрированный анализом примеров применения подход направлен на решение задач эффективной обработки текстовых документов – индексирования, реферирования, кластеризации и классификации. В результате теоретического и экспериментального исследования установлено, что разработанный подход обеспечивает более высокую по сравнению с аналогами релевантность по критериям точности и полноты.

Ключевые слова: ключевые слова, английский язык, релевантность, лингвистический пакет, DKPro Core, синтаксический анализ

Abstract. It is proposed a new approach to determining the keywords based on finding connections between word forms of the English text with the instrumental capabilities of package DKPro Core. The approach, which illustrated with examples of analysis, aimed at solving problems of efficient processing of text documents – indexing, abstracting, clustering and classification. As a result of theoretical and experimental studies it is found that the developed approach ensures better relevance of the keywords compared to similar criteria for accuracy and completeness.

Key words: keywords, English, method, relevance, linguistic package, DKPro Core, syntactic analysis.

Вступ

Якість виділення ключових слів з тексту грає важливу роль у бібліотечній справі, лексикографії та термінознавстві, а також в задачах інформаційного пошуку. В даний час обсяги і динаміка інформації, яка підлягає обробці в цих областях, роблять особливо актуальною задачу автоматичного визначення ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів: індексування, реферування, кластеризації та класифікації.

Актуальність

У переважній більшості доступних лінгвістичних систем, що орієнтовані на обробку природномовних текстів, присутні функції автоматичного виділення ключових слів. В основу реалізації таких функцій покладено відомі методи визначення ключових слів, які діляться на лінгвістичні та статистичні. Лінгвістичні методи ґрунтуються на значеннях слів, зокрема використовують онтології та семантичні дані про слово. На жаль, методи такого класу ресурсоємні на ранніх етапах: розробка онтологій, наприклад, вельми трудомісткий процес [1]. З іншого боку, статистичні методи супроводжуються значними обсягами «вербального шуму», який суттєво впливає на якість визначення ключових слів. Тому найбільш перспективними для дослідження, на думку авторів, є гібридні методи, для яких швидкість статистичної обробки тексту підсилюється можливостями сучасних лінгвістичних пакетів.

Мета

Мета роботи полягає у підвищенні релевантності визначення ключових слів англомовного тексту за рахунок врахування інформації щодо синтаксичних зв'язків між словами у реченнях тексту.

Задачі

1. Обгрунтування підходу до визначення ключових слів тексту на основі інформації про синтаксичні зв'язки.
2. Програмна реалізація підходу та експериментальна оцінка складових релевантності.

Інформаційна оцінка синтаксичного аналізу тексту для задачі визначення ключових слів

Розглянемо задачу визначення ключових слів тексту як певну інформаційну технологію, що має на вході текст, а на виході – множину з *l* ключових слів $W^k = \{w_1^k, ..., w_l^k\}$. Без применшення загальності будемо вважати, що текст *T* складається з *m* різних слів, а в окреме його *j*-те речення з *k* налічує *n* слів з *m* можливих, причому $m >> n$ та $m >> l$. Більшість відомих методів визначення ключових слів тексту беруть за основу частотний словник тексту, який фактично є списком або упорядкованою множиною пар $D = \{< w_i,\ f_i >\},\ i = \overline{1,m}$, де w_i – одне слово з *m*, а f_i – його частота ($f_i \geq f_{i+1}$,

$i = \overline{1, m-1}$), що визначена для *T*. За певною фільтрацією окремих незначущих категорій слів ключовими вважають перші *l* слів зі списку *D*, тобто, дещо спрощено маємо $W^k = \{w_1, ..., w_l\}$.

Проте результати парсерингу природних мов за допомогою сучасних лінгвістичних пакетів дозволяють на доступному програмному рівні [2] оперувати синтаксичними зв’язками між словами окремого речення. Окрім того, можливості цих пакетів дозволяють суттєво зменшити значення *m* шляхом об’єднання слів у словоформи, а останні – у леми та стеми. Отже, необхідно з’ясувати, які формальні переваги для визначення $W^k = \{w_1^k, ..., w_l^k\}$ надасть нам програмно-лінгвістичне забезпечення процедури синтаксичного аналізу всіх речень тексту *T*.

З інформаційної точки зору розуміння сенсу речення окремим суб’єктом супроводжується розпізнаванням а) окремих слів, з яких воно складається та б) зв’язків між парами цих слів з відповідною побудовою дерева таких зв’язків [3]. Вважатимемо, що всі ці процеси відбуваються шляхом порівняльного аналізу та залучення інформації з деякої загальнолінгвістичної бази знань суб’єкта розуміння. Якщо кожен з цих етапів супроводжується збільшенням інформації, то приймаємо робочу гіпотезу:

- Рівень загального розуміння тексту *T* може змінюватися від мінімально можливого до максимального в залежності від обсягу та інших параметрів загальнолінгвістичної бази знань суб’єкта;
- Якість визначення $W^k = \{w_1^k, ..., w_l^k\}$ пропорційна рівню загального розуміння тексту, що має підтверджуватися формальними ознаками.

Нехай будь-яке *j*-те речення з *k* складається з *n* різних слів, що не є досить жорстким обмеженням. Тоді зв’язне дерево парних залежностей такого речення налічує або *n–1* гілок, якщо не брати до уваги зворотну залежність між підметом та присудком, або *n* – якщо брати. Відповідно загальна кількість слів цього речення для подальшого поглибленого аналізу збільшується або до *2×n–2* або до *2×n*. Проте таке збільшення відбувається нерівномірно – для всіх не термінальних (кінцевих) вузлів дерева частоти відповідних слів не змінюються, а для термінальних (проміжних) можуть зрости суттєво. В таблиці 1 показані випадки зміни частот слів, які позначаються літерами a, b, …, f за порядком застосування, з урахуванням парних залежностей для різних типів речення.

Таблиця 1 – Аналіз збільшення частоти значимих слів унаслідок урахування парних залежностей для різних типів речення

№ з/п	Склад речення / кількість слів	Тип речення та граф його дерева залежностей	Частотна формула	Кінцева частота
1.	Ab / 2	Словосполучення (Коріння дерева)	A+b	2
2.	Abc / 3	Лінійна трійка (Бережи скарби природи)	A+2b+c	4
3.	Abcd / 4	Лінійна четвірка (Отримав переклад слова дивного)	A+2b+2c+d	6
4.	Abcde / 5	Розгалудження (Густий ліс нізвідки завершився провалля)	A+2b+c+3d+e	8
5.	Abcdef / 6	Група підмета (Сині примружені очі коханого говорили багато)	A+b+4c+d+2e+f	10
6.	Abcdef / 6	Група присудка (Досвідчений кінь борозну швидко відчує нюхом)	A+2b+c+d+4e+f	10
7.	Abcdef / 6	Обидві групи (Старий дід Еол зобрав всіх вітрів)	A+3b+c+2d+e+2f	10

Проведений аналіз на рівні одного речення показує, що збільшуються частоти тих слів (підкреслені у 3-му стовпчику таблиці), які потенційно можуть належати до множини ключових. Проведемо формальну оцінку такого збільшення для накладених обмежень щодо наявності виключно різних слів у реченні та не врахуванням зворотної залежності між підметом та присудком:

1. Мінімальне збільшення відсутнє за умови знаходження *i*-го слова з *m* серед не термінальних (кінцевих) вузлів дерева кожного речення, де це слово зустрічається, тобто $f_i^{\min} = 0,\ f_i^{new} = f_i$ $i = \overline{1, m}$.

2. Якщо *i*-те слово знаходиться у кожному з *k* речень тексту та, окрім того, відповідає у кожному реченні найбільш розгалуженому термінальному вузлу, то максимальне збільшення частоти складає $f_i^{\max} = \sum_{j=1}^{k}(n_j - 2), i = \overline{1,m}$. Відповідно $f_i^{new} = f_i + f_i^{\max} = k + \sum_{j=1}^{k}(n_j - 2) = \sum_{j=1}^{k}(n_j - 1)$.

3. В загальному та більш реальному випадку $f_i = z \mid z \le k$, тобто *i*-те слово знаходиться у *z* реченнях з *k* маємо $f_i^{new} = z + \sum_{j=1}^{z}(n_j - 2) = \sum_{j=1}^{z}(n_j - 1)$ як оцінку зверху збільшення частоти *i*-го слова.

Очікується, що експериментальні дослідження мають підтвердити справедливість отриманих формальних оцінок процесу визначення ключових слів тексту у встановлених межах.

Програмна реалізація запропонованого підходу в DKPro Core та проведення експерименту

Для експериментальної перевірки результатів теоретичного аналізу було розроблене програмне забезпечення на основі DKPro Core.

DKPro Core – це набір програмних компонентів для обробки природної мови, що базується на Apache UIMA framework. Він був побудований з метою підвищення продуктивності дослідників, які працюють з автоматичним аналізом мови. Підхід DKPro Core полягає в тому, що дослідники повинні мати можливість зосередитися на своїх реальних наукових питаннях, а не на розробці технологій [4, 5].

Визначення ключових слів відбувається за кількома етапами:

а) створення багаторівневої розмітки тексту;

б) синтаксична розмітка, що враховує складні залежності між парами лем;

в) заміна займенників в отриманих парах на відповідні до них іменники;

г) розбиття пар на окремі слова і визначення кількості зв'язків;

д) вибір перших n слів з найбільшою кількістю зв'язків, де n – кількість потрібних ключових слів.

Аналогами розробленої програми можуть бути сайти SEO оптимізації, де є функція визначення ключових слів. Для даного експерименту вибрані сервіси: advego.ru/text/seo/, rise-top.com/keywordstext.php та seotool.by/analiz/seo/key-wordstext.php.

Для проведення експерименту було взято текст статті з 1460 слів «A new pattern for historical geography: working with enthusiast communities and public history» [6].

Упорядкований список ключових слів з їх позиціями x_i, що були задані автором: Participation (1), Public (2) history (3), Enthusiast (4) communities (5), Museums (6), Heritage (7). Результати знаходження ключових слів для власної розробки і аналогів наведено в таблиці 2. Перед правильно знайденими ключовими словами вказана позиція, на якій знаходиться це слово у авторському списку.

Таблиця 2 – Результати пошуку ключових слів

Слова задані автором		власна розробка		rise-top		advego		seotool	
1	Participation		work		historical		historical		historical
2	Public	5	community	4	enthusiast	4	enthusiast	4	enthusiast
3	history		geography	5	communities		for	5	communities
4	Enthusiast	1	participation	1	participation	5	community	1	participation
5	communities	4	enthusiast		geography		this		work
6	Museums		geographer		work	6	museum		geography
7	Heritage	6	museum		research		geography		new

Кількісними характеристиками релевантності отриманих результатів обрано повноту (за Жаккаром і абсолютну) і точність (за евклідовою і манхеттенською відстанями). Проведено інтерпретацію обраних критеріїв до умов задачі визначення ключових слів.

Повноту за Жаккаром визначено як частку від ділення кількості знайдених ключових слів на різницю кількості можливих ключових слів заданих автором і знайдених програмно (в даному випадку по 7) і кількості знайдених ключових слів. Абсолютна повнота знаходиться як відношення кількості правильно знайдених ключових слів до загальної кількості ключових слів. На рисунку 1 наведена гістограма повноти за Жаккаром і абсолютної для власної розробки і аналогів.

Точність за евклідовою відстанню визначається за формулою:

$$d_{\text{e}} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad (1)$$

де n – кількість правильно визначених ключових слів;

x_i – позиція і-го ключового слова з n в авторському списку;

y_i – позиція і-го ключового слова з n в альтернативному списку, що знайдено програмно.

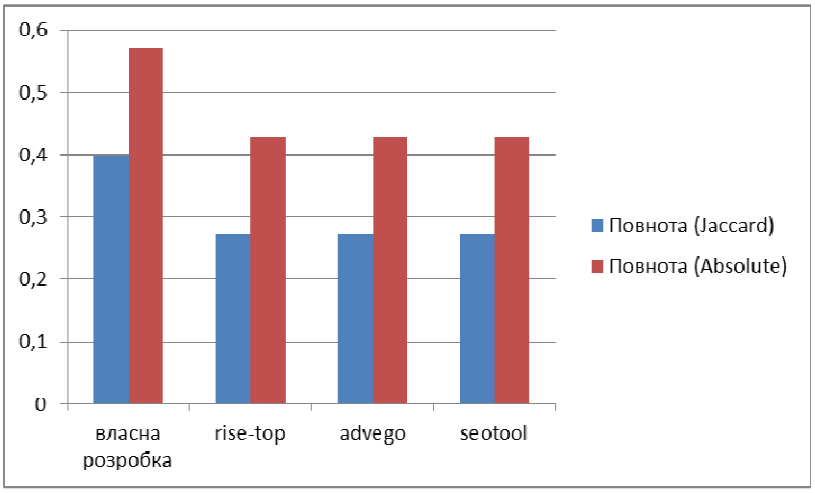

Рисунок 1 – Гістограма повноти за Жаккаром і абсолютної

Манхеттенська відстань визначається за формулою:

$$d_{\text{m}} = \sum_{i=1}^{n}\left|x_i - y_i\right|. \quad (2)$$

На рис. 2 наведена гістограма евклідової та манхеттенської відстані для власної розробки і аналогів.

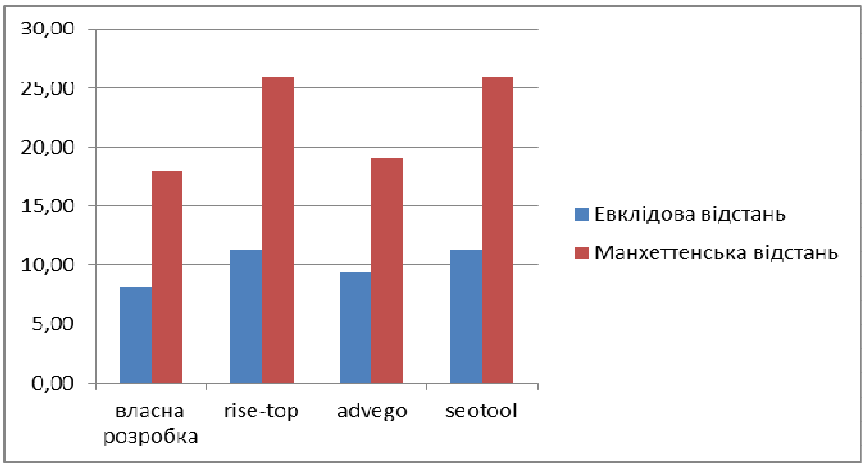

Рисунок 2 – Гістограма евклідової та манхеттенської відстані

Повнота знаходження ключових слів має бути якомога більшою, а відстань між позиціями ключових слів заданих автором і визначених програмно – якомога меншою. Як видно з гістограм власна розробка має кращі кількісні характеристики за різними метриками у порівнянні з аналогами – від 25% до 31,8% за повнотою, а також від 5,3% до 14% за точністю для тестового прикладу [6].

Висновки

1. Оскільки краща якість обробки тексту досягається лінгвістичними методами або ж при їх комбінації зі статистичними, систему автоматичного визначення ключових фраз з тексту природною мовою слід розробляти з використанням морфологічного словника (лексикону) і синтаксичних правил. Ці дані визначаються попередньо і зберігаються в базі даних. Текст підлягає обробці аналізатором, який виробляє інформацію про розділення тексту на абзаци, речення та окремі слова, що необхідно для подальшого оброблення. Кожне слово, виділене аналізатором, піддається морфологічному аналізу з метою побудови морфологічної інтерпретації, визначення основи слова і формування леми. На основі наявної інтерпретації тексту виконується побудова та наповнення синтаксичних груп і виявлення відношень між ними.

2. В роботі запропоновано підхід до визначення ключових слів, що базується на використанні додаткової інформації про складні залежності між членами англомовного речення. Для функціональної реалізації аналізатора тексту обрано популярний лінгвістичний пакет DKPro Core. Проведені експериментальні дослідження теоретичного обґрунтування підходу підтвердили його якісні та кількісні переваги у порівнянні з відомими аналогами. Для англомовного тексту обсягом 1460 слів отримано збільшення повноти визначення ключових слів (на 31,8% за Жаккаром та на 25% за абсолютним значенням) і покращення точності (на 14% за евклідовою і на 5,3% манхеттенською відстанями) у порівнянні з аналогами.

3. Якість отриманих результатів потенційно можна підвищити через окремий аналіз частин мови, оскільки ймовірність релевантності ключового слова, наприклад, іменника і прислівника буде відрізнятися. Окрім цього, варто оцінити збільшення частотних показників для ключових слів шляхом реалізації наявних в DKPro Core компонентів для визначення кореференційних зв'язків.

Список літератури

1. Ershov, Yu. S. (2014). Vydelenie kliuchevyh slov v russkoiazychnyh tekstah. Molodezhnyi nauchno-tehnicheskii vestnik. M.: FGBOU VPO "MGTU im. N. E. Baumana". Available: http://sntbul.bmstu.ru/file/out/730754. Last accessed 21.01.2015.

2. Bisikalo, O. V. (2013). Kontseptualna model systemy obraznoho analizu i syntezu pryrodno-movnykh konstruktsii. Matematychni mashyny i systemy, № 2, 184–187. ISSN 1028-9763.

3. Bisikalo, O. V. (2013). Formalni metody obraznoho analizu ta syntezu pryrodno-movnykh konstruktsii. Vinnytsia: VNTU, 316. ISBN 978-966-641-528-1.

4. Natural Language Processing: Integration of Automatic and Manual Analysis. (2014). Technischen Universität Darmstadt. Available: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf. Last accessed 21.01.2015.

5. Gurevych, I., Muhlhauser, M., Muller, Ch., Steimle, J., Weimer, M., Zesch, T. (2007, February 9). Darmstadt Knowledge Processing Repository Based on UIMA. Available: https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf. . Last accessed 21.01.2015.

6. Geoghegan, H. (2014). A new pattern for historical geography: working with enthusiast communities and public history. Journal of Historical Geography, № 46. Available: http://ac.els-cdn.com/S0305748814001029/1-s2.0-S0305748814001029-main.pdf?_tid=d45ec9e6-ba7b-11e4-b562-00000aab0f01&acdnat=1424600353_48bb4ef54fffbc3b800698d175c3c052. Last accessed 21.01.2015.

Відомості про авторів

Бісікало Олег Володимирович – доктор технічних наук, професор, декан ФКСА, кафедра автоматики та інформаційно-вимірювальної техніки, Вінницький національний технічний університет, Хмельницьке шосе 95, м. Вінниця, Україна, 21000.

Яхимович Олександр Вікторович – кафедра автоматики та інформаційно-вимірювальної техніки, Вінницький національний технічний університет, Хмельницьке шосе 95, м. Вінниця, Україна, 21000.