УДК 004.93:159.95
DOI: 10.15587/2312-8372.2018.149962

РОЗРОБКА МЕТОДУ ФІЛЬТРАЦІЇ ВЕРБАЛЬНОГО ШУМУ В ПРОЦЕСІ ПОШУКУ КЛЮЧОВИХ СЛІВ АНГЛОМОВНОГО ТЕКСТУ

Бісікало О. В., Яхимович О. В., Яхимович Я. В.

1. Вступ

У даний час обсяги і динаміка інформації, яка підлягає обробці в лексикографії та термінознавстві, а також в задачах інформаційного пошуку, роблять особливо актуальною задачу автоматичного визначення ключових слів. Дуже активно у сучасних інформаційних технологіях (ІТ) використовують ключові слова для створення і розвитку термінологічних ресурсів, для ефективної обробки документів, зокрема, індексування, реферування, кластеризації та класифікації [1].

Існує велика кількість доступних систем автоматичного виділення ключових слів, розроблених і орієнтованих на обробку природних мов. Ці системи засновані на певних методах визначення ключових слів, які діляться на лінгвістичні та статистичні. Лінгвістичні методи ґрунтуються на значеннях слів, зокрема, використовують онтології та семантичні дані про слово. Ці методи 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 є гібридні методи, для яких швидкість статистичної обробки тексту підсилюється можливостями сучасних лінгвістичних пакетів.

Актуальність та практична цінність напряму досліджень полягає у тому, що знайдені ключові слова можна використати для підвищення точності аналізу контенту сайту і підняття позиції сайту в результатах пошуку.

Ключове слово – слово в тексті, здатне в сукупності з іншими ключовими словами представляти текст. Набір ключових слів близький до анотації, плану і конспекту, які теж представляють документ з меншою деталізацією, але, на відміну від ключових слів, пов’язані у синтаксичні структури.

Вербальний шум або шумові слова – термін з теорії пошуку інформації за ключовими словами. Це такі слова, які не несуть смислового навантаження, тому їх користь та роль для пошуку не суттєва [2].

В процесі обробки проводиться виключення з досліджуваного тексту слів, які за визначенням не можуть бути значущими тому, що складають «шум». На відміну від ключових ці слова називаються нейтральними або стоповими (стоп словами). Такими є слова, що відносяться до службових частин мови, а також займенники [3].

2. Об'єкт дослідження та його технологічний аудит

Об’єкт дослідження – процес обробки вербальної інформації для визначення ключових слів в тексті.

Предмет дослідження – методи знаходження ключових слів в тексті, а також підходи до зменшення вербального шуму в процесі пошуку ключових слів.

Ключові слова мають ряд суттєвих ознак:

- високий ступінь повторюваності даних слів у тексті, частотність їх вживання;
- здатність знака (слова як вербальни ознаки певного поняття) конденсувати, згортати інформацію, виражену цілим текстом, об'єднувати «його основний зміст». Ця ознака особливо яскраво проявляється у ключових словах у позиції заголовку.

Наявність правильно підібраного набору ключових слів дозволить:

а) швидше знайти статтю користувачеві при пошуку по базі даних;

б) побачити статтю при перегляді інших схожих статей;

в) швидше зрозуміти тематичну і термінологічну область як однієї статті, так і журналу в цілому.

Все це служить одній меті: привернути увагу читачів до статті, яка є основним завданням будь-якого засобу масової інформації [4].

Однак вибір ключових слів є дуже непростою операцією і вимагає зваженого підходу. Слід вибирати ті ключові слова, які найбільш точно відображають специфіку розглянутої теми. При цьому необхідно уникати випадкових і загальних фраз, не рекомендується повторювати кілька разів одні й ті ж ключові слова. Отже, процес пошуку ключових слів є аналітичним [5].

3. Мета та задачі дослідження

Мета роботи полягає у підвищенні точності визначення ключових слів з англомовного тексту на основі розробки методу зменшення впливу вербального шуму.

Для досягнення поставленої мети необхідно вирішити такі задачі:

1. Розглянути підходи до зменшення вербального шуму при знаходженні ключових слів.
2. Обчислити чисельні показники зв’язків між словами та проаналізувати отримані результати як основу методу.
3. Формалізувати операції для кожного етапу методу та визначити кількісні характеристиками релевантності отриманих результатів в порівнянні з аналогами.

4. Дослідження існуючих рішень проблеми

Серед основних напрямів вирішення задачі пошуку ключових слів в тексті, виявлених в ресурсах світової наукової періодики, можуть бути виділені [6, 7]. Для відділення одиночних ключових слів використовується методи на основі закону Ципфа. Такі методи залежать від установки діапазону частот, в яких знаходяться значущі для тексту слова. Так як слова, які трапляються дуже часто, в основному виявляються вербальним шумом, а слова, що зустрічаються рідко, в більшості випадків, не мають вирішального смислового значення. Тому в кожному окремому випадку необхідно використовувати ряд евристик для визначення ширини діапазону, а також методик, що зменшують вплив цієї ширини. Одним із способів, як зазначено в роботі [8], є виключення, з кандедатів у ключові слова, слів, які не можуть бути значущими тому, що складають шум. Але у цій роботі не розглянуто зменшення шуму на основі синтаксичної інформації.

Робота [9] присвячена покращенню результатів розрахунку ваг термінів на основі алгоритму TF-IDF. Однак загальною рисою таких систем є те, що вони вимагають наявності інформації, отриманої з усієї колекції документів. Іншими словами, якщо метод, заснований на TF-IDF, використовується для створення уявлення про документ, то надходження нового документа в колекцію вимагає перерахунку ваг термінів у всіх документах. Отже, будь-які додатки, засновані на значеннях ваг термінів у документі, також будуть зачеплені. Це значною мірою перешкоджає використанню методів вилучення ключових термінів, що вимагають навчання, в системах, де динамічні потоки даних повинні оброблятися в режимі реального часу [10].

Для вирішення цієї проблеми було запропоновано алгоритм TF-ICF, зазначений в роботі [11]. Як розвиток цієї ідеї в [12, 13] запропоновано використовувати в якості навчального тезауруса Вікіпедію. Для розрахунків застосовується інформація, що міститься в анотованих статтях енциклопедії з вручну виділеними ключовими термінами. Проте, не враховується порядок проходження термінів у документі та їх синтаксична роль.

Альтернативний варіант вирішення проблеми, викладений в [14], передбачає використання лінгвістичних онтологій, які є більш-менш наближеними моделями існуючого набору слів заданої мови. Однак ці методи ресурсоємні на ранніх етапах: розробка онтологій вельми трудомісткий процес.

Метод, що служить для автоматичного формування тематичного корпусу з WEB показано в [15]. Однак відбором управляє порогове значення відносин частот термінів.

Автори роботи [16] підкреслюють важливість використання в якості кандидатів в ключові слова іменних груп, виділених за допомогою синтаксичного аналізатора. Хоча це твердження може бути розглянуто зі сторони інших синтаксичних одиниць, що використовуються при визначенні ключових слів.

Seotool – безкоштовний онлайн сервіс, що допоможе перевірити, чи релевантний написаний текст ключовим словам (згенерувавши автоматично ключі за вказаним текстом). Це допоможе отримати більш високий рейтинг в пошукових системах Яндекс і Google, так як сторінка матиме ключові слова, відповідні змісту сторінки, на якій розміщені. Також даний сервіс допоможе у генерації семантичного ядра сайту (при включеному режимі прибирати HTML код). Проте у генерації ключових слів і фраз використовуються тільки перші тисячу слів введеного тексту.

Є можливість відсоткового порівняння слів з шаблоном. Слова аналізованого тексту (контенту) будуть в процентному співвідношенні порівнюватися зі списком слів всього шаблону (тексту) шляхом морфологічного аналізу. При відповідності відсоткової рівності з будь-яким зі слів шаблону слово враховується, інакше – не враховується. Максимальна кількість слів шаблону не повинно перевищувати 250 слів [17].

Rise-Top допоможе скласти «начерки» ключових слів для сайту на основі використання для аналізу вказаного тексту. В якості відбору ключових слів застосовуються слова з найбільш високою щільністю в порядку зменшення їх

щільності до всього тексту [18]. Але у генерації ключових слів так само використовуються тільки перші 1000 слів обробленого тексту.

Advego (Адвего) – найбільший в Рунеті постачальник контенту і супутніх послуг для інтернет-сайтів. Для оптимізаторів і власників сайтів пропонуються унікальні статті, відгуки, публікації. Забезпечується просування в пошукових системах та розкрутка в соцмережах. Ресурс також має можливість визначати ключові слова [19].

Таким чином, результати аналізу дозволяють зробити висновок про те, що питання щодо розробки методу фільтрації вербального шуму в процесі пошуку ключових слів є перспективним та потребує подальшого вивчення.

5. Методи дослідження

Для підвищення точності визначення ключових слів задіяні статистичні методи обробки тексту, швидкість роботи яких підсилюється можливостями сучасних лінгвістичних пакетів.

Одним з таких пакетів є DKPro Core – це набір програмних компонентів для обробки природної мови, заснований на Apache UIMA framework.

Пакет DKPro Core – це більше, ніж деяка множина компонентів аналізу, які взаємодіють між собою. Він був побудований з метою підвищення продуктивності дослідників, що 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 відповідних технологій [20].

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

Повнота за Жаккаром, в даному випадку, визначається для двох множин ключових слів – заданої автором (еталонної) та визначеної програмно, дорівнює відношенню кількості елементів перетину цих множин до кількості елементів їх об'єднання. Тобто, це частка від ділення, де в чисельнику знаходиться кількість правильно знайдених програмою ключових слів, а в знаменнику – різниця суми елементів в двох множинах і кількості знайдених правильно ключових слів.

Абсолютна повнота знаходиться як відношення кількості правильно знайдених програмою ключових слів до кількості ключових слів.

Евклідова відстань визначається за формулою:

$$d_e = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

де n – кількість ключових слів;

x_i – позиція i-го ключового слова, визначеного автором;

y_i – позиція i-го ключового слова, визначеного програмно.

Манхеттенська відстань визначається за формулою:

$$d_m = \sum_{i=1}^{n} |x_i - y_i|.$$

Застосування пари формальних критеріїв і для повноти, і для точності, дозволить більш об'єктивно оцінити релевантність отриманих результатів пошуку ключових слів.

6. Результати дослідження

Згідно з [21] пропонується такий підхід до визначення ключових слів, що відбувається за три основних етапи:

1) створення багаторівневої розмітки тексту;

2) застосування синтаксичної розмітки, що враховує складні залежності між парами лем;

3) зменшення вербального шуму.

Сутність підходу, на відміну від відомих аналогів, полягає у визначенні кількості зв'язків для окремих слів і подальшим вибором перших *n* слів з найбільшою кількістю зв'язків, де *n* – кількість потрібних ключових слів.

Створення багаторівневої розмітки тексту і синтаксична розмітка, що враховує складні залежності між парами лем досягається засобами DKPro Core [20].

Фільтрацію вербального шуму пропонується забезпечити за допомогою таких операцій:

- заміна займенників на відповідні до них іменники;
- вилучення шумових зв'язків;
- вилучення шумових слів;
- вилучення стоп слів.

Заміна займенників на відповідні до них іменники (replace pronouns) дозволяє зменшити кількість займенників, а також збільшити кількість іменників, які можуть бути ключовими словами. Для методу зменшення вербального шуму при визначенні ключових слів англомовного тексту, що пропонується, заміна займенників здійснюється засобами DKPro Core [20].

Розглянемо вилучення словосполучень із типами зв'язків, які не несуть суттєвого смислового навантаження. Внаслідок дослідження виявлено, що такими зв'язками є DET, EXPL, FIXED, PUNCT, REF, ROOT.

DET – зв'язок визначника, що існує між номінально головним словом та його визначником. Найчастіше, слово, яке має тег частини мови DET, буде мати такий же зв'язок визначника DET і навпаки. Відомим винятком є те, що у деяких з наборів даних присвійний визначник (наприклад, такий як «my») у певний момент отримує тег частини мови DET, але зв'язок NMOD, що є паралеллю до інших присвійних конструкцій. Але це не повністю однаково для різних мов, у деяких мовах набагато чіткіше, ніж на англійській, виражено те, як присвійні визначники відносяться до прикметників, тому відношення NMOD не підлягає сумніву [22].

Приклади DET зв'язків наведено на рис. 1.

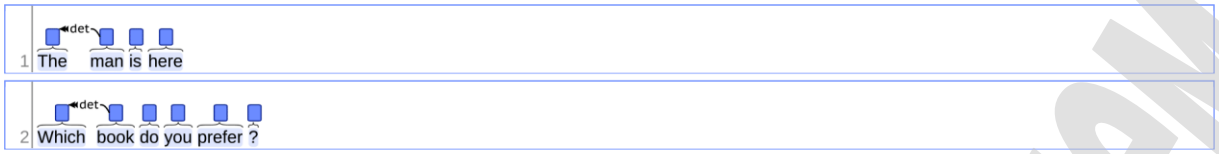

Рис. 1. Приклади шумових зв’язків DET

EXPL – це відношення, що фіксує вставні або плеонастичні номінали. Такі номінали з'являються в аргументній позиції предиката, але не виконують ніякої з семантичних ролей предиката. Основний предикат речення (дієслово або предикатний прикметник або іменник) є головним словом. В англійській мові це стосується деяких способів використання it і there: екзистенціальне there, а також it при використанні в експозиційних конструкціях [23].

Деякі мови не мають таких, подібних англійському, висловів, це стосується більшості мов pro-drop (мова, в якій певні класи займенників можуть бути опущені, коли вони прагматично або граматично інерційні). Також це явище часто називають нуль або нульовою анафорою [24]. У мовах з подібними висловами вони можуть бути розташовані там, де зазвичай з'являється основний аргумент: підмет та прямий (і, навіть, непрямий) додаток [25].

Приклади EXPL зв'язків наведено на рис. 2.

Рис. 2. Приклади шумових зв’язків EXPL

FIXED – використовується для певних сталих граматичних виразів, які ведуть себе як функціональні слова або короткі прислівники.

Сталі багатослівні вирази анотовано у рівній структурі, де всі наступні слова у виразі прикріплені до першого з використанням сталої мітки. Припущення полягає в тому, що ці вирази не мають внутрішньої синтаксичної структури (окрім з історичної точки зору) і що структурна анотація в принципі є довільною. Однак, на практиці, дуже важливо використовувати послідовну анотацію всіх сталих багатослівних виразів на всіх мовах [26].

Приклади FIXED зв'язків наведено на рис. 3.

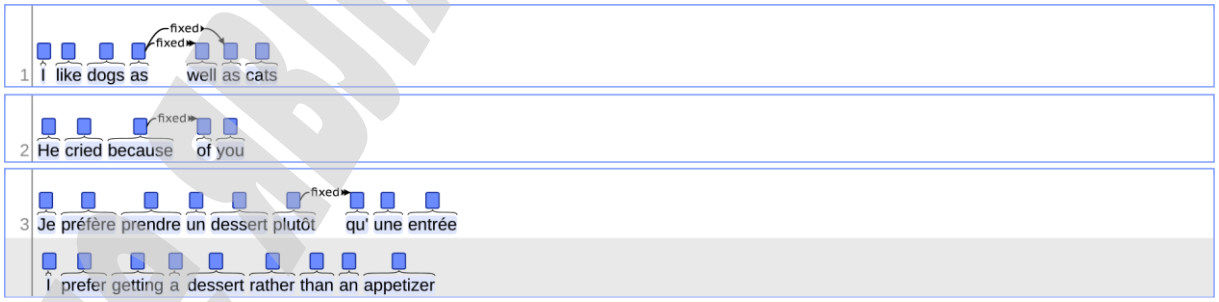

Рис. 3. Приклади шумових зв’язків FIXED

PUNCT – використовується для позначення будь-якої частини пунктуації в реченні чи частині тексту, якщо пунктуація зберігається в типізованих залежностях.

Токени з співвідношенням PUNCT завжди прикріплюються до змісту слів і ніколи не можуть мати залежностей. Оскільки PUNCT не є нормальним відношенням залежностей, звичайні критерії визначення головного слова не застосовуються. Натомість використовуються такі принципи:

1. Знак пунктуації, що розділяє скоординовані одиниці, додається до наступного зв'язку.
2. Знак пунктуації, що передує або слідує за незалежною одиницею, додається до цієї одиниці.
3. У межах відповідного підрозділу знак пунктуації прикріплюється до найвищого можливого вузла, який зберігає перспективу.
4. Парні знаки пунктуації (наприклад, цитати та дужки, іноді також дефіси, коми тощо) мають бути додані до одного слова, якщо це не порушує перспективу [27].

Приклади PUNCT зв'язків наведено на рис. 4.

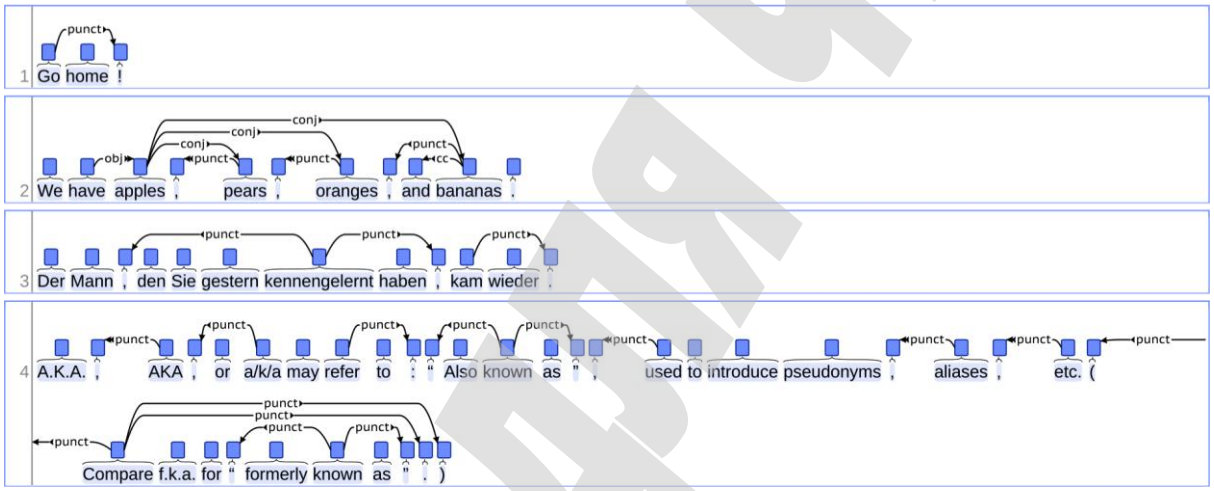

Рис. 4. Приклади шумових зв’язків PUNCT

REF – референт головного слова іменникового словосполучення, що є відносним словом, яке вводить відносне положення шляхом модифікації іменникового словосполучення. Наприклад для речення: «I saw the book which you bought», зв'язок REF буде між словами book і which [25].

ROOT – корневе граматичне відношення, що вказує на корінь речення. Фейковий вузол ROOT використовується як головний вузол. Вузол ROOT має індекс 0, оскільки індексація реальних слів у реченні починається з 1. У кожному дереві повинен бути тільки один кореневий вузол. Якщо основний предикат відсутній, але є багато одиничних залежностей, то одне з них підвищується до положення головного (кореневого), а до нього приєднуються інші одинаки [28].

Приклад ROOT зв'язку наведено на рис. 5.

Рис. 5. Приклади шумових зв’язків ROOT

Розглянемо вилучення шумових слів, які відносяться до неінформативних частин мови, що мають теги: CC, CD, DT, EX, IN, LS, MD, PDT, POS, PRP, PRP$, RP, SYM, TO, UH, WDT, WP, WP$, WRB, -LRB-, -RRB-.

CC – координуючі сполучення: and, but, nor, or, yet, plus, minus, less, times (multiplication), over (division), also for (because), so (i. e., so that), &, 'n, both, either, et, neither, therefore, v., versus, vs., whether.

CD – номер, число, кількість: one, two, 2, mid-1890, nine-thirty, forty-two, one-tenth, ten, million, 0.5, forty-seven, 1987, twenty, '79, zero, 78-degrees, eighty-four, IX, '60s, .025, fifteen, 271, 124, dozen, quintillion, DM2,000.

DT – визначник: a, an, every, no, the, another, any, some, all, both, del, each, either, half, la, many, much, nary, neither, such, that, them, these, this, those.

EX – екзистенціальне there: ненаголошений there, що викликає інверсію дієслова у відповідній формі та логічного суб'єкта. Наприклад: «There was a party in progress».

IN – прийменники або сполучники підпорядкування: among, around, astride, atop, behind, below, by, despite, for, if beside, if like, inside, into, near, next, on, out, pro, throughout, towards, until, upon, whether, within.

LS – список, елемент, маркер, цифри та літери, що використовуються як ідентифікатори елементів у списку: A, A., B, B., C, C., D, E, F, First, G, H, I, J, K, One, SP-44001, SP-44002, SP-44005, SP-44007, Second, Third, Three, Two, *, a, b, c, d, first, five, four, one, six, three, two.

MD – модальні допоміжні дієслова. Всі дієслова, які не приймають закінчення -s у формі третьої особи однини: can, could, dare, may, might, must, ought, shall, should, will, would, cannot, couldn't, need, ought, shouldn't.

PDT – префіксний визначник. Визначники, як елементи, що передують статті або присвійним займенникам: all, both, half, many, quite, such, sure, this. Наприклад: «all his marbles», «quite a mess».

POS – присвійне закінчення: іменники, що закінчуються маркером ' або 's.

PRP – особовий займенник: he, her, hers, herself, him, him, himself, hisself, I, it, itself, me, myself, one, oneself, ours, ourselves, ownself, self, she, she, thee, theirs, them, themselves, they, thou, thy, us, you.

PRP$ – присвійний займенник: her, his, its, mine, my, one’s, our, ours, their, thy, your.

RP – частка. В основному односкладові слова, що також двоскладові в якості прислівників напрямку: aboard, about, across, along, apart, around, aside, at, away, back, before, behind, by, crop, down, ever, fast, for, forth, from, go, high, i. e., in, into, just, later, low, more, off, on, open, out, over, per, pie, raising, start, teeth, that, through, under, unto, up, up-pp, upon, whole, with you.

SYM – символ. Технічні символи або вирази, які не є словами (% & ' " * + , . < = > @ A[fj] U.S U.S.S.R * ** ***).

TO – літерал to, як прийменник або інфінітивний маркер.

UH – вигук: amen, anyways, baby, dammit, diddle, Goodbye, Goody, Gosh, heck, Hey, honey, howdy, Hubba, huh, hush, Jee-sus, Jeepers, Kee-reist, man, my, oh, Oops, please, shucks, sonuvabitch, uh, well, whammo, whodunnit, Wow, yes.

WDT – wh-визначник: that, what, whatever, which, whichever.

WP – wh-займенник: that, what, whatever, whatsoever, which, who, whom, whosoever.

WP$ – присвійний wh-займенник: whose.

WRB – wh-прислівник, включаючи when, коли використовується в переносному значенні: how, however, whence, whenever, where, whereby, wherever, wherein, whereof, why.

-LRB- – відкрита дужка.

-RRB- – закрита дужка [29–31].

Щодо вилучення слів, які відносяться до списку стоп слів – це питання вже було досліджено. Список таких слів для англомовних текстів обгрунтовано і наведено в [32].

Проілюструємо результати визначення ключових слів на кожному кроці роботи методу, що пропонується на невеликому тексті, що складається з двох речень: «Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree».

Знайдені словосполучення і частини мови відповідних слів першого речення наведено в табл. 1, а для другого – в табл. 2.

Таблиця 1

Словосполучення і частини мови відповідних слів першого речення

Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review			
Головне слово	Тег частини мови головного слова	Залежне слово	Тег частини мови залежного слова
graduate	NN	born	VBN
born	VBN	honolulu	NNP
honolulu	NNP	hawaii	NNP
graduate	NN	obama	NNP
graduate	NN	is	VBZ
graduate	NN	a	DT
university	NNP	columbia	NNP
graduate	NN	university	NNP
school	NNP	harvard	NNP
school	NNP	law	NNP
university	NNP	school	NNP
graduate	NN	school	NNP
president	NN	where	WRB
president	NN	he	PRP
president	NN	was	VBD
university	NNP	president	NN
review	NNP	the	DT
review	NNP	harvard	NNP
review	NNP	law	NNP
president	NN	review	NNP

Таблиця 2

Словосполучення і частини мови відповідних слів другого речення

He was a community organizer in Chicago before earning his law degree			
Головне слово	Тег частини мови головного слова (Governor POS)	Залежне слово	Тег частини мови залежного слова (Dependent POS)
organizer	NN	he	PRP
organizer	NN	was	VBD
organizer	NN	a	DT
organizer	NN	community	NN
organizer	NN	chicago	NNP
organizer	NN	earning	VBG
degree	NN	his	PRP$
degree	NN	law	NN
earning	VBG	degree	NN

Типи зв'язків між головними і залежними словами у словосполученнях, приведеними до незмінної, основної форми слова, наведено для першого та другого речення в табл. 3, 4.

Таблиця 3

Зв'язки у словосполученнях першого речення

Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review					
Головне слово (Governor)	Залежне слово (Dependent)	Тип зв'язку (Dependency Type)	Головне слово (Governor)	Залежне слово (Dependent)	Тип зв'язку (Dependency Type)
graduate	bear	vmod	university	school	conj_and
bear	honolulu	prep_in	graduate	school	prep_of
honolulu	hawaius	appos	president	where	advmod
graduate	obama	nsubj	president	he	nsubj
graduate	be	cop	president	be	cop
graduate	a	det	university	president	rcmod
university	columbium	nn	review	the	det
graduate	university	prep_of	review	harvard	nn
school	harvard	nn	review	law	nn
school	law	nn	president	review	prep_of

Таблиця 4

Зв'язки у словосполученнях другого речення

He was a community organizer in Chicago before earning his law degree		
Головне слово (Governor)	Залежне слово (Dependent)	Тип зв'язку (Dependency Type)
organizer	he	nsubj
organizer	be	cop
organizer	a	det
organizer	community	nn
organizer	chicago	prep_in
organizer	earn	prepc_before
degree	his	poss
degree	law	nn
earn	degree	dobj

Розіб'ємо словосполучення на окремі слова і підрахуємо кількість зв'язків для кожного слова, тобто в скількох словосполученнях слово зустрічається. Відсортувавши слова за кількістю зв’язків отримаємо результати, які наведено в табл. 5.

Таблиця 5

Кандидати в ключові слова після розбиття словосполучень

Слово	Кількість зв'язків	Слово	Кількість зв'язків	Слово	Кількість зв'язків
graduate	6	degree	3	hawaius	1
organizer	6	a	2	community	1
president	5	honolulu	2	the	1
university	4	earn	2	his	1
school	4	bear	2	columbium	1
review	4	harvard	2	where	1
be	3	he	2	chicago	1
law	3	obama	1	–	–

Умовно словосполучення можна позначити:

G-[T]->D,

де G – головне слово (Governor); T – тип зв'язку (Dependency Type); D – залежне слово (Dependent).

На етапі заміни займенників на відповідні до них іменники (replace pronouns):

− словосполучення president-[nsubj]->he замінюється на president-[nsubj]->obama;

− словосполучення organizer-[nsubj]->he замінюється на organizer-[nsubj]->obama;

– словосполучення degree-[poss]->his замінюється на degree-[poss]->obama.

Кандидати в ключові слова, після заміни займенників на відповідні до них іменники, наведено в табл. 6.

Таблиця 6

Кандидати в ключові слова після заміни займенників

Слово	Кількість зв'язків	Слово	Кількість зв'язків	Слово	Кількість зв'язків
graduate	6	be	3	harvard	2
organizer	6	law	3	hawaius	1
president	5	degree	3	community	1
university	4	a	2	the	1
obama	4	honolulu	2	columbium	1
school	4	earn	2	where	1
review	4	bear	2	chicago	1

Після заміни займенників кількість кандидатів в ключові слова зменшилася з 23 до 21. До заміни займенників слово obama мало 1 зв'язок, а після – 4 зв'язки. І навпаки слова he з 2 зв'язками і his з одним зв'язком після заміни займенників мають нуль зв'язків, тому що словосполучення з ними були замінені на еквіваленти з іменниками.

Вилучення словосполучень із типами зв’язків, які не несуть суттєвого смислового навантаження (deleting noise relationship). Для даного тексту, видаляються словосполучення: graduate-[det]->a, review-[det]->the, organizer-[det]->a.

У результаті кількість кандидатів в ключові слова зменшиться до 19, що відображено в табл. 7.

Таблиця 7

Кандидати в ключові слова після видалення шумових зв'язків

Слово	Кількість зв'язків	Слово	Кількість зв'язків
graduate	5	honolulu	2
organizer	5	earn	2
president	5	bear	2
university	4	harvard	2
obama	4	hawaius	1
school	4	community	1
be	3	columbium	1
law	3	where	1
degree	3	chicago	1
review	3	–	–

Вилучення слів, що відносяться до шумових частин мови (deleting noise POS keywords). На даному кроці видаляється слово where з тегом частини мови WRB. Кандидатів в ключові слова будуть мати вигляд, наведений в табл. 8.

Таблиця 8

Кандидати в ключові слова після видалення шумових частин мови

Слово	Кількість зв'язків	Слово	Кількість зв'язків	Слово	Кількість зв'язків
graduate	5	be	3	bear	2
organizer	5	law	3	harvard	2
president	5	degree	3	hawaius	1
university	4	review	3	community	1
obama	4	honolulu	2	columbium	1
school	4	earn	2	chicago	1

На етапі видалення стор слів(deleting stop words) – видаляється стоп слово be, а табл. 9 містить 17 кандидатів в ключові слова.

Таблиця 9

Кандидати в ключові слова після видалення стоп слів

Слово	Кількість зв'язків	Слово	Кількість зв'язків	Слово	Кількість зв'язків
graduate	5	law	3	harvard	2
organizer	5	degree	3	hawaius	1
president	5	review	3	community	1
university	4	honolulu	2	columbium	1
obama	4	earn	2	chicago	1
school	4	bear	2	–	–

У результаті, після усіх запропонованих кроків методу, вдалося зменшити кількість кандидатів в ключові слова з 23 до 17, а також видалити шумові слова.

Розглянемо тепер відносно великий текст з метою визначення кількісних характеристик релевантності отриманих результатів у порівнянні з аналогами. Для цього було обрано текст «A Workingman's Poet», який складається з 3299 слів, та відомі ключові слова, що задані автором: american, literature, books, chicago, poetry, publishing, twentieth century, united states. За результатами експерименту маємо перші десять кандидатів в ключові слова, знайдені розробленим методом: sandburg, poem, write, poet, poetry, book, life, lincoln, learn, speak. Пошук ключових слів у цьому ж тексті було реалізовано за допомогою програм-аналогів.

Результати знаходження ключових слів розробленим методом і аналогами наведено в табл. 10.

Таблиця 10

Результати знаходження ключових слів розробленим методом і аналогами

Etalon keywords		Advego		Rise-top		Seotool		Our development	
1	American	–	sandburg	–	sandburg	–	his	–	sandburg
2	Literature	–	that	–	his	–	sandburg	–	poem
3	Books	–	for	–	lincoln	–	lincoln	–	write
4	Chicago	–	poem	5	poetry	–	poems	–	poet
5	Poetry	–	lincoln	–	poems	5	poetry	5	poetry
6	Publishing	5	poetry	–	who	–	who	3	book
7	Twentieth	–	work	1	american	1	american	–	life
8	Century	–	write	–	where	–	where	–	lincoln
9	United	1	american	–	had	–	years	–	learn
10	States	–	where	–	years	–	had	–	speak

Результати повноти і точності отриманих ключових слів наведено в табл. 11, 12 і на рис. 6, 7.

Таблиця 11

Результати повноти отриманих ключових слів

Name	Advego	Rise-top	Seotool	Our development
Completeness (Jaccard)	0,111111111	0,111111111	0,111111111	0,111111111
Completeness (Absolute)	0,2	0,2	0,2	0,2

Таблиця 12

Результати точності отриманих ключових слів

Name	Advego	Rise-top	Seotool	Our development
Euclidean distance	0,577061522	0,59749477	0,589067059	0,577061522
Manhattan distance	0,49	0,51	0,49	0,47

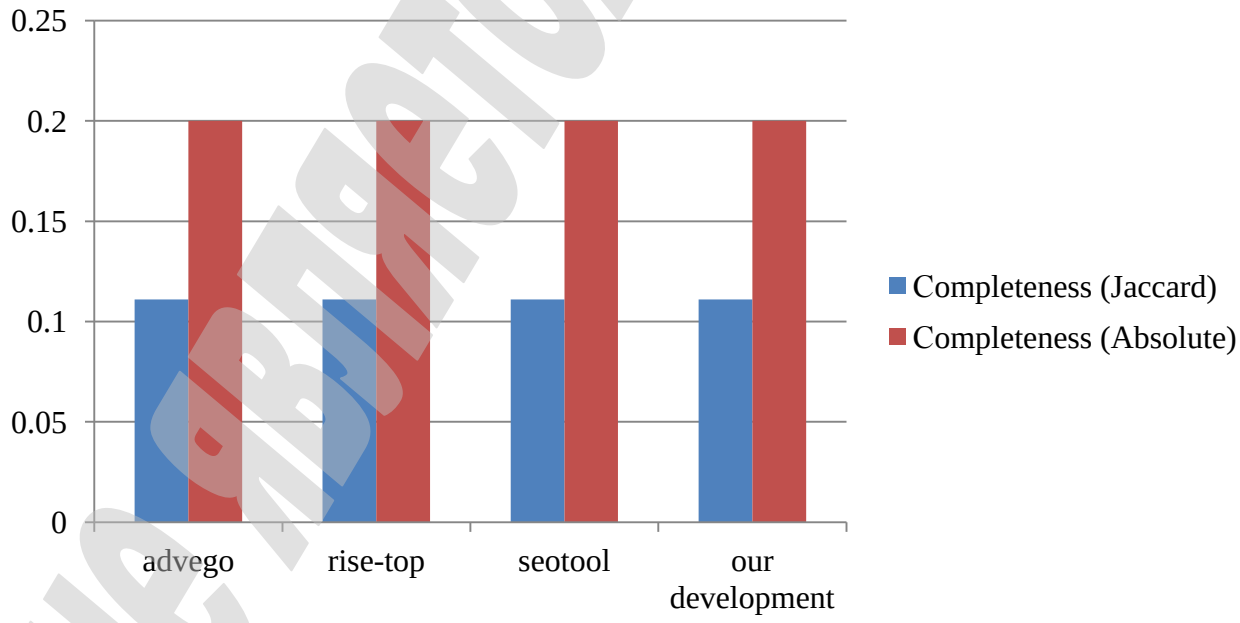

Рис. 6. Гістограми повноти за Жаккаром і абсолютної

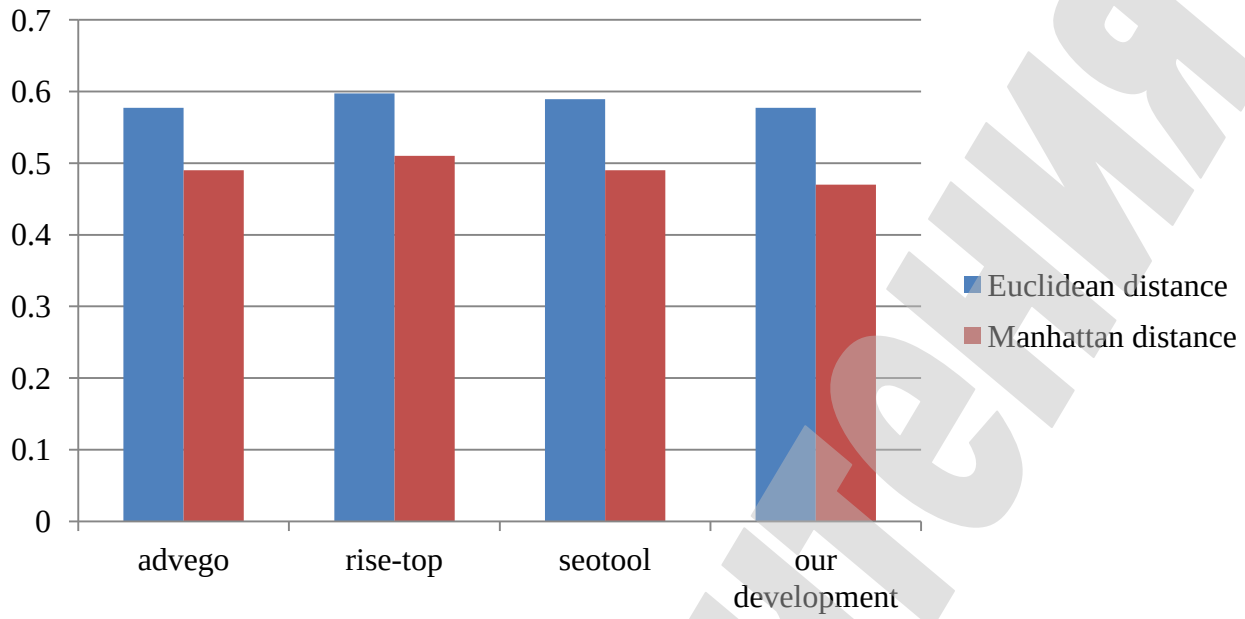

Рис. 7. Гістограми точності за евклідовою і манхеттенською відстанями

Повнота знаходження ключових слів повинна бути якомога більшою, а відстань між позиціями ключових слів заданих автором і визначених програмно якомога меншою.

Як видно з гістограм на рис. 6, 7 та табл. 11, 12, власна розробка для даного тексту має таку саму повноту, як і аналоги – 11 % та 20 %, проте кращі кількісні характеристики за точністю – 57,71 % та 47 %, ніж аналоги rise-top (59,75 %; 51 %) і seotool (58,91 %; 49 %). Також власна розробка має однакову точність за евклідовою відстанню, як і аналог advego, але, на відміну від нього, кращі характеристики за манхеттенською відстанню.

7. SWOT-аналіз результатів досліджень

Strengths. У порівнянні з аналогами представлена розробка, за результатами проведеного експерименту з текстом обсягом 3299 слів, має таку саму повноту, як і аналоги, проте, кращі кількісні характеристики за точністю, ніж аналоги rise-top і seotool. Також представлена розробка має однакову точність за евклідовою відстанню, як і аналог advego, але, на відміну від нього, кращі характеристики за манхеттенською відстанню. Ще однією перевагою в порівнянні з аналогами є те, що представлена розробка дозволяє повністю виключити шумові слова.

Weaknesses. До слабких сторін роботи методу можна віднести швидкодію його практичної реалізації засобами DKPro Core, зокрема, відносно задовгим для онлайн режиму є час створення багаторівневої розмітки тексту. Але це, в свою чергу, може бути виправлено за рахунок використання більш потужного апаратного забезпечення або платформ хмарних обчислень, що дозволяють мати у своєму розпорядженні віртуальний кластер комп'ютерів. Цього не важко досягти, оскільки додатки визначення ключових слів і зменшення вербального шуму написані на Java і можуть бути легко розгорнуті на таких платформах.

Opportunities. Перспективою подальших досліджень визначення ключових слів є проведення більш масштабних експериментів для текстів різних катего-

рій з метою визначення додаткових шляхів підвищення релевантності методу. Доцільно також використання нових лінгвістичних пакетів, що підтримують більше мов, в тому числі і українську.

Threats. Процес визначення ключових слів запропонованим методом є незалежним від процесів визначення ключових слів іншими методами, тому загроза негативної дії на об’єкт дослідження зовнішніх чинників відсутня.

Впровадження запропонованої методології не потребує додаткових витрат для компанії.

Аналогом розробленого методу можуть бути сайти SEO оптимізації з можливістю визначення ключових слів.

8. Висновки

1. Запропоновано метод, фільтрація вербального шуму у якому забезпечується такими формалізованими операціями:
   - заміна займенників на відповідні до них іменники;
   - вилучення шумових зв'язків;
   - вилучення шумових слів;
   - вилучення стоп слів.

Описані операції можна використовувати як додаткові модулі, що покращують результати знаходження ключових слів для методу визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core. А також для інших алгоритмів знаходження ключових слів.

2. Обчислення чисельних показників зв'язків між словами та аналіз отриманих результатів роботи на кожному етапі методу, що пропонується, проілюстровано на прикладі тексту з двох речень. За розглянутими у прикладі результатами вдалося зменшити кількість кандидатів в ключові слова з 23 до 17, а також повністю виключити шумові слова.

3. За результатами проведеного експерименту представлена розробка для тексту з 3299 слів, яка має таку саму повноту, як і аналоги – 11 % та 20 %, проте, кращі кількісні характеристики за точністю – 57,71 % та 47 %, ніж аналоги rise-top (59,75 %; 51 %) і seotool (58,91 %; 49 %). Представлена розробка також має однакову точність за евклідовою відстанню, як і аналог advego, але, на відміну від нього, кращі характеристики за манхеттенською відстанню.

Література

1. Ершов Ю. С. Выделение ключевых слов в русскоязычных текстах // Молодежный научно-технический вестник. 2014. № ФС77-51038. С. 70–79.

2. Гращенко Л. А. О модельном стоп-словаре // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. 2013. № 1 (150). С. 40–46.

3. Модели и методы автоматической классификации текстовых документов / Андреев А. М., Березкин Д. В., Сюзев В. В., Шабанов В. И. // Вестник МГТУ. Сер. Приборостроение. 2003. № 3. С. 64–94.

4. Абрамов Е. Г. Подбор ключевых слов для научной статьи // Научная периодика: проблемы и решения. 2011. № 1 (2). С. 35–40.

5. Даркулова К. Н., Ергешова Г. Необходимость выделения ключевых слов для свёртывания текста // Лингвистический анализ научного текста. VI Международная студенческая электронная научная конференція. Южно-Казахстанский государственный университет им. Мухтара Ауэзова Шымкент, 2014. С. 30–35.

6. Halkidi M., Batistakis Y., Vazirgiannis M. On clustering validation techniques // Journal of intelligent information systems. 2001. Vol. 17, Issue 2-3. P. 107–145. doi: http://doi.org/10.1023/a:1012801612483

7. Barahnin V. B., Tkachev D. A. Clustering of text documents based on composite key terms // Vestnik NSU. Series: Information Technology. 2010. Vol. 8, Issue 2. P. 5–14.

8. Гращенко Л. А. О модельном стоп-словаре // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. 2013. № 1 (150). С. 40–46.

9. Guo A., Tao Y. Research and Improvement of Feature Words Weight Based on TFIDF Algorithm // 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference. Chongqing, 2016. doi: http://doi.org/10.1109/itnec.2016.7560393

10. Sifting Micro-blogging Stream for Events of User Interest / Grineva M. et. al. // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. Boston, 2009. P. 327–333. doi: http://doi.org/10.1145/1571941.1572157

11. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams / Reed J. et. al. // 2006 5th International Conference on Machine Learning and Applications. Orlando, 2006. P. 258–263. doi: http://doi.org/10.1109/icmla.2006.50

12. Mihalcea R., Csomai A. Wikify!: linking documents to encyclopedic knowledge // Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. Lisbon, 2007. P. 233–242. doi: http://doi.org/10.1145/1321440.1321475

13. Astrakhantsev N. Automatic term acquisition from domain-specific text collection by using Wikipedia // Proceedings of the Institute for System Programming of RAS. 2014. Vol. 26, Issue 4. P. 7–20. doi: http://doi.org/10.15514/ispras-2014-26(4)-1

14. Özgür A., Hur J., He Y. The Interaction Network Ontology-supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature // BioData Mining. 2016. Vol. 9, Issue 1. doi: http://doi.org/10.1186/s13040-016-0118-0

15. Wong W., Liu W., Bennamoun M. Ontology learning from text // ACM Computing Surveys. 2012. Vol. 44, Issue 4. P. 1–36. doi: http://doi.org/10.1145/2333112.2333115

16. Korobkin D. M., Fomenkov S. A., Kolesnikov S. G. Method of ontology-based extraction of physical effect description // Vestnik Komp'iuternykh i Informatsionnykh Tekhnologii. 2015. P. 28–35. doi: http://doi.org/10.14489/vkit.2015.02.pp.028-035

17. Бесплатный онлайн-генератор ключевых слов с текста. URL: http://seotool.by/analiz/seo/keywordstext.php

18. Генератор ключевых слов с текста. URL: http://www.rise-top.com/keywordstext.php

19. Адвего. URL: http://wiki.advego.ru/index.php/Адвего

20. Natural Language Processing: Integration of Automatic and Manual Analysis. 2014. URL: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf

21. Method of determining of keywords in English texts based on the DKPro Core / Bisikalo O. V. et. al. // Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016. 2016. doi: http://doi.org/10.1117/12.2249225

22. Determiner. URL: http://universaldependencies.org/u/dep/det.html
23. Expletive and Reflexives. URL: http://universaldependencies.org/u/dep/expl.html
24. Welo E. Null Anaphora // Encyclopedia of Ancient Greek Language and Linguistics. 2013. doi: http://doi.org/10.1163/2214-448x_eagll_com_00000254
25. Manning C., de Marneffe M. Stanford typed dependencies manual. 2016. URL: https://nlp.stanford.edu/software/dependencies_manual.pdf
26. Fixed multiword. URL: http://universaldependencies.org/u/dep/fixed.html
27. Punctuation. URL: http://universaldependencies.org/u/dep/punct.html
28. Root. URL: http://universaldependencies.org/u/dep/root.html
29. Taylor A., Marcus M., Santorini B. The Penn Treebank: An Overview // Text, Speech and Language Technology. 2003. P. 5–22. doi: http://doi.org/10.1007/978-94-010-0201-1_1
30. Penn Treebank II Constituent Tags: Word level. URL: http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html#Word
31. Alphabetical list of part-of-speech tags used in the Penn Treebank Project. URL: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
32. Bougé K. Lists of stop words. URL: https://sites.google.com/site/kevinbouge/stopwords-lists