Міністерство освіти і науки України
Вінницький національний технічний університет

ЯХИМОВИЧ ОЛЕКСАНДР ВІКТОРОВИЧ

УДК 004.891

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОШУКУ КЛЮЧОВИХ СЛІВ НА ОСНОВІ ПАРСИНГУ АНГЛОМОВНИХ ТЕКСТІВ

05.13.06 – інформаційні технології

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
кандидата технічних наук

Вінниця – 2021

Дисертацією є кваліфікаційна наукова праця на правах рукопису.

Робота виконана на кафедрі автоматизації та інтелектуальних інформаційних технологій у Вінницькому національному технічному університеті, Міністерство освіти і науки України.

Науковий керівник: доктор технічних наук, професор
Бісікало Олег Володимирович,
Вінницький національний технічний університет,
декан факультету комп'ютерних систем і автоматики

Офіційні опоненти: доктор технічних наук, професор
Хайрова Ніна Феліксівна,
Національний технічний університет «Харківський політехнічний інститут», професор кафедри інтелектуальних комп'ютерних систем

доктор технічних наук, професор
Шевченко Ігор Васильович,
Кременчуцький національний університет ім. Михайла Остроградського, професор кафедри автоматизації та інформаційних систем

Захист відбудеться «08» квітня 2021 р. о 17:00 годині на засіданні спеціалізованої вченої ради Д 05.052.01 у Вінницькому національному технічному університеті за адресою: 21021, Україна, м. Вінниця, вул. Хмельницьке шосе, 95, ГНК, ауд. 210.

З дисертацією можна ознайомитись у бібліотеці Вінницького національного технічного університету за адресою: 21021, Україна, м. Вінниця, вул. Хмельницьке шосе, 95, ГНК.

Автореферат розісланий «04» березня 2021 р.

Вчений секретар
спеціалізованої вченої ради С. М. Захарченко

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Обґрунтування вибору теми дослідження. Завдання пошуку ключових слів тексту виникає у бібліотечній справі, лексикографії та термінознавстві, а також в усіх задачах інформаційного пошуку. В даний час обсяги і динаміка інформації, яка підлягає обробці в цих областях, роблять особливо актуальною задачу автоматичного пошуку ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів – індексування, реферування, кластеризації та класифікації. Популярність та затребуваність пошукових машин для кожного користувача Інтернету висуває високі вимоги до релевантності результатів пошуку, яка значно залежить від якості пошуку ключових слів у текстовій інформації.

Натепер існує значна кількість доступних систем автоматичного пошуку ключових слів, розроблених і орієнтованих на обробку природних мов. В основу роботи таких систем покладено відомі методи виявлення ключових слів тексту, які умовно можна розділити на дві категорії − експертно-лінгвістичні та статистичні. Перша група методів ґрунтується на значеннях слів, отриманих експертним шляхом, зокрема використовують словники з зібраними семантичними даними про кожне слово або онтології предметних областей. Такі методи ресурсоємні на ранніх етапах – розробка онтологій, наприклад, вельми трудомісткий процес. Тому найбільша кількість відомих напрацювань у напряму експертно-лінгвістичної обробки текстів відома для визнаної мови міжнародного спілкування – англійської. З іншого боку, традиційні статистичні методи, які є фактично мово-незалежними, супроводжуються значними обсягами «вербального шуму», що суттєво впливає на якість пошуку ключових слів. Внаслідок цього статистичні методи зазвичай супроводжуються емпіричними процедурами, налаштованими на конкретний клас задач, що суттєво звужує їхню область застосування.

Підвищення якості процесу отримання ключових слів вимагає залучення додаткової інформації, бажано універсального, а не специфічного характеру. Зокрема, відсутня формальна постановка та розв'язок задачі пошуку ключових слів як згортки інформації у тексті. Не враховуються у відомих методах пошуку ключових слів результати аналізу зв’язків між лексичними одиницями тексту, а інформаційна оцінка результатів парсингу тексту не береться до уваги як складова відповідного критерія якості.

Потрібно звернути увагу на гібридні методи пошуку ключових слів, для яких швидкість статистичної обробки тексту підсилюється можливостями сучасних лінгвістичних пакетів, що мають найбільш розвинутий функціонал, у першу чергу, для англійської мови. Тому актуальним науковим завданням є підвищення якості пошуку ключових слів у англомовному тексті шляхом розробки інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів.

Зв’язок роботи з науковими програмами, планами, темами. Дослідження, результати яких представлено в дисертації, проводились відповідно до пріоритетних тематичних напрямів науково-технічних розробок на період до 2020 року «Технології та засоби розробки програмних продуктів і

систем», затверджених постановою Кабінету Міністрів України №556 від 23.08.2016 р. Зокрема, дослідження проводились згідно планів науково-дослідної роботи кафедри автоматизації та інтелектуальних інформаційних технологій Вінницького національного технічного університету. Автор брав участь у виконанні науково-дослідних робіт «Інтелектуальна інформаційна технологія образного аналізу тексту та синтезу інтегрованої бази знань природно-мовного контенту» (№ ДР 0114U003462) та «Ідентифікація прихованих залежностей в онлайнових соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики» (№ ДР 0117U000575).

Мета і завдання дослідження. Мета дослідження полягає у підвищенні якості пошуку ключових слів у англомовному тексті.

Для досягнення поставленої мети необхідно розв’язати наступні задачі:

1. Аналіз й порівняльна характеристика відомих підходів, методів та засобів пошуку ключових слів.
2. Побудова математичної моделі процесу пошуку ключових слів на основі інформаційної оцінки результатів парсингу тексту.
3. Розробка методу пошуку ключових слів на основі визначення зв'язків між словоформами.
4. Розробка методу зменшення впливу вербального шуму на пошук ключових слів.
5. Експериментальне дослідження результатів запропонованих методів у порівнянні з аналогами.
6. Розробка та апробація інформаційної технології пошуку ключових слів англомовного тексту.

Об’єкт дослідження − процес обробки вербальної інформації для пошуку ключових слів у англомовному тексті.

Предмет дослідження – моделі, методи та технологічні засоби пошуку ключових слів у англомовному тексті.

Методи досліджень. Для побудови моделі пошуку ключових слів використано методи теорії множин та теорії інформації. Розробка методів пошуку ключових слів та зменшення впливу вербального шуму базувалася на основі поєднання методів лінгвістичного аналізу англомовного тексту та статистичних методів. Під час оцінки достовірності запропонованої інформаційної технології застосовано методи вимірювання у метричних просторах.

Наукова новизна отриманих результатів. В ході розв’язання поставлених задач були отримані наукові результати.

Удосконалено модель пошуку ключових слів, яка, на відміну від існуючих, побудована на основі інформаційної оцінки результатів парсингу тексту та враховує результати аналізу зв’язків між лексичними одиницями тексту, що дозволило формалізувати критерій якості процесу пошуку ключових слів.

Уперше розроблено метод пошуку ключових слів, який, на відміну від існуючих, базується на знаходженні синтаксичних зв’язків між словоформами у реченнях англомовного тексту за допомогою технологічних можливостей

парсингу сучасних лінгвістичних пакетів. Запропонований метод дає змогу підвищити чисельні характеристики якості пошуку ключових слів, а саме повноту (за Жаккаром) і точність.

Удосконалено метод зменшення впливу вербального шуму на пошук ключових слів, який, на відміну від існуючих, побудовано на основі стенфордської класифікації зв'язків між лексичними одиницями речення, що дозволило підвищити якість результатів пошуку ключових слів у порівнянні з основним методом.

Набула подальшого розвитку інформаційна технологія пошуку ключових слів, яка, на відміну від існуючих, враховує додаткову інформацію процесів парсингу речень у межах послідовного застосування двох запропонованих методів, що дозволило уточнити чисельні оцінки змістовних параметрів тексту та підвищити якість пошуку його ключових слів.

Практичне значення отриманих результатів. Прикладні результати дисертаційного дослідження полягають у формальному описі методики пошуку ключових слів англомовного тексту, створенні алгоритму її реалізації та розробці програмного забезпечення, що знаходить ключові слова на основі врахування значимих зв’язків між словоформами у реченнях англомовного тексту та подальшої фільтрації вербального шуму.

Створені моделі, алгоритми та програмні засоби можуть бути використані при вирішенні практичних задач комп’ютерної лінгвістики, які потребують знаходження ключових слів, наприклад, для підвищення точності аналізу контенту сайту і підняття позиції сайту в результатах пошуку. Використання мово-незалежних засобів запропонованої інформаційної технології у поєднанні з необхідними, згідно з отриманою специфікацією, технологічними ресурсами лінгвістичного аналізу інших природних мов дозволить розширити область застосування інформаційної технології, зокрема на українську мову.

Робота впроваджена на ТОВ НВП «СПІЛЬНА СПРАВА» (акт про результати впровадження від 10.01.2020), а також в навчальний процес кафедри АІІТ Вінницького національного технічного університету, що підтверджено виданням навчального посібника “A lexical relationships-based keywords selection in an English text”. Результати експериментів показали, що запропонована інформаційна технологія одночасно збільшує у межах від 8,1% до 12,7% повноту за метрикою Жаккара та від 9,1% до 14,3% абсолютну точність пошуку ключових слів для англомовних текстів обсягом 140-1400 слів у порівнянні з аналогами.

Особистий внесок здобувача. Усі результати, які складають основний зміст дисертаційної роботи, отримані здобувачем самостійно. У роботах [9], [12], [13], [14], [15], [16] здобувачеві належать усі теоретичні та практичні результати. У роботах, опублікованих у співавторстві, здобувачу належать: [1] – експериментально підтверджено кращу релевантність результатів розробленого методу пошуку ключових слів у порівнянні з аналогами; [2], [7], [19] – розробка формальних складових методу зменшення впливу вербального шуму на пошук ключових слів; [3], [10], [11] – розробка програмного забезпечення для методу пошуку ключових слів на основі обробки

синтаксичних зв’язків засобами пакету DKPro Core; [4] – експериментально підтверджено, що результати пошуку ключових слів з додатковою заміною займенників містять менше стоп-слів у порівнянні з частотним словником; [5] – запропоновано підхід до інформаційної оцінки процесів парсингу тексту у межах задачі пошуку ключових слів; [6] – запропоновано методику побудови онтологій на основі знаходження відношень між термінами як складової інформаційного пошуку; [8] – застосування запропонованої інформаційної технології пошуку ключових слів для покращення ідентифікації токсичних коментарів в соціальних мережах; [17], [18] – розробка модулів програмного забезпечення на основі пакету DKPro Core для запропонованої інформаційної технології; [20], [21] – реалізація програмного експерименту з використанням лінгвістичного пакету DKPro Core на англомовних текстах різної довжини; оцінка чисельних характеристик релевантності отриманих результатів за критеріями повноти та точності.

Апробація матеріалів дисертації. Основні результати та дисертаційна робота в цілому апробовані на восьми науково-практичних конференціях:

- міжнародній Інтернет-конференції «Молодь в технічних науках: дослідження, проблеми, перспективи» (МТН-2015), м. Вінниця, 23-26 квітня 2015 р.;
- першій міжнародній конференції «Адаптивні технології управління навчанням», м. Одеса, 23-25 вересня 2015 р.;
- третій міжнародній науковій конференції «Вимірювання, контроль та діагностика в технічних системах» (ВКДТС-2015), м. Вінниця, 27-29 жовтня 2015 р.;
- XLIV, XLV, XLVI, XLVII та XLVIII науково-технічних конференціях підрозділів ВНТУ факультету комп'ютерних систем та автоматики, м. Вінниця, щорічно у 2015 − 2019 рр.

Публікації. За темою дисертації з викладенням її основних результатів опубліковано 21 науковій праці, серед яких 1 стаття у закордонному фаховому періодичному виданні, що входить до SCOPUS, 1 стаття у вітчизняному періодичному виданні, що індексується у SCOPUS, 6 статей у спеціалізованих фахових виданнях України, що індексуються міжнародними бібліометричними та наукометричними базами даних, 8 публікацій в матеріалах та тезах доповідей, 2 звіти про науково-дослідну роботу. Також, отримано 1 патент України на корисну модель (Пат. 135223 UA), опубліковано закордонну монографію та електронний навчальний посібник.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків. Основний зміст викладено на 138 сторінках друкованого тексту, містить 67 рисунки, 17 таблиць. Список використаних джерел містить 157 найменувань. Загальний обсяг роботи – 193 сторінки.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі наведено загальну характеристику роботи, обґрунтовано актуальність теми, зазначено її зв’язок з науковими програмами, планами та

темами, сформульовано мету та напрям досліджень. Визначено основні задачі досліджень, наукову новизну та практичне значення основних результатів, а також відомості про їх впровадження, апробацію та публікації.

У першому розділі на основі огляду та аналізу сучасної літератури в галузі пошуку ключових слів показано, що ключові слова є основним засобом упорядкування важливого масиву інформації, яким є науковий журнал або база наукових статей певного наукового напряму. Зокрема, ключові слова потрібні для систематизації множини статей, оскільки дозволяють швидше знайти статтю, групувати схожі статті, класифікувати статті у межах інших структурних групувань. З іншого боку, на основі застосування ключових слів реалізовано велику кількість методів пошуку, класифікації та оцінки інформації.

Показано, що існує велика кількість доступних систем автоматичного пошуку ключових слів, розроблених і орієнтованих на обробку природних мов. Ці системи засновані на значній кількості відомих методів пошуку ключових слів, які діляться на експертно-лінгвістичні та статистичні. При цьому експертно-лінгвістичні методи ресурсоємні на ранніх етапах: розробка онтологій, наприклад, вельми трудомісткий процес. З іншого боку, статистичні методи супроводжуються значними обсягами «вербального шуму», який суттєво впливає на якість пошуку ключових слів. Тому найбільш перспективними для дослідження, на думку автора, є гібридні методи, для яких швидкість статистичної обробки тексту підсилюється можливостями баз знань та функціоналом сучасних лінгвістичних пакетів.

Найбільшою суттєвою проблемою відомих методів знаходження ключових слів лишається недостатня якість процесу автоматизованої обробки вербальної інформації для пошуку ключових слів в тексті. Дуже важливим, з огляду на це, є вибір найбільш інформативної ознаки (критерія) при пошуку ключових слів як в традиційних умовах обробки колекції текстів, так і в умовах аналізу блогів та мікроблогів для WEB. Підвищення якості процесу пошуку ключових слів вимагає залучення додаткової інформації, бажано універсального, а не специфічного характеру. Зокрема, не враховуються у відомих методах пошуку ключових слів результати аналізу зв’язків між лексичними одиницями тексту, а інформаційна оцінка результатів парсингу тексту не береться до уваги як складова відповідного критерія якості.

Інтенсивний розвиток методів та засобів підтримки пошукових машин в Інтернеті став причиною появи досить популярних на даний час систем та ресурсів, що є інструментами розкрутки та підвищення рейтингу сайтів. Тому можна вважати, що вдосконалення математичних моделей і методів з метою підвищення якості пошуку ключових слів має необхідну інструментальну підтримку і може бути швидко та ефективно впровадженим у SEO-індустрію.

На основі проведеного аналізу визначено задачі дослідження.

У другому розділі розглянуто довільний текст як множину синтаксично зв'язаних упорядкованих слів, які, в свою чергу, є підмножиною слів мови: $w \in \{W\} \subset \{\omega\}$, де w – слово, W – множина слів в тексті, ω – множина слів мови.

Кожне слово має основну форму (канонічна форма, лема), до якої можна звести слово шляхом його морфологічного аналізу. Формально це представлено застосуванням до слова функції нормалізації m морфологічного аналізу: $m(w) = b,\ w \in \{W^F\}$, де $\{W^F\}$ – множина словоформ слова w, b – основна форма слова w. Основні властивості функції нормалізації: в результаті нормалізації будь-якого слова отримуємо основну форму слова $m(b) = b$; для будь-якого слова нормалізація дає таку основну форму слова, яка належить множині словоформ $\forall w(w \in \{W\}) \Rightarrow m(w) = b,\ b \in \{W\}$.

Для задачі пошуку ключових слів представимо довільний текст T як впорядкований набір слів w_i і символів c_i:

$$T = \{w_i, c_i\},\ w_i \in \{W_T\} \subset \{\omega\},\ c_i \in \{C\} \subset \{\varsigma\}, \qquad (1)$$

де $\{W_T\}$ – множина слів в тексті T, що є підмножиною множини слів мови; c_i – знаки пунктуації, цифри, пробіли, переходи на новий рядок та інші символи, що не є буквами; $\{C\}$ – множина символів в тексті, що є підмножиною $\{\varsigma\}$ – множини всіх символів; i – порядковий номер слова або символу в тексті.

Між словами w і символами c існують зв'язки R в тексті, які можуть бути трьох видів: зв'язок між i-м словом та l-м словом: $(w_i, w_l) \in R_j$, де j – довільний порядковий номер зв'язку в тексті; зв'язок між i-м словом та l-м символом: $(w_i, c_l) \in R_j$; зв'язок між i-м символом та l-м символом: $(c_i, c_l) \in R_j$.

Будемо вважати ключовими словами $K = W^k$ такі, які стисло представляють сутність тексту T:

$$W^k(T) = \{w^k\},\ w^k \in \{\omega\}, \qquad (2)$$

де w^k – ключове слово, що належить множині слів мови $\{\omega\}$.

Розглянуто процес знаходження ключових слів як згортку текстової інформації за певними критеріями. У цьому випадку ключові слова не завжди можливо формально знайти в тексті. Також, в якості ключових, можуть використовуватися: синоніми, терміни з якими текст може бути пов'язаний логічно, власні назви, з якими асоціюється текст. В роботі розглядається варіант, коли всі ключові слова знаходяться в тексті K_T.

Задачу пошуку ключових слів K_T для тексту T описано як знаходження таких слів $W^k(T)$, які належать цьому тексту і входять до складу множини слів мови:

$$W^k(T) = \{w^k\},\ w^k \in \{W_T\} \subset \{\omega\}. \qquad (3)$$

Ключове слово w^k є словом w з тексту T, яке приведено до основної форми b

$$w^k = b\ = m(w),\ w \in \{W^F\}, \qquad (4)$$

де $\{W^F\}$ – множина словоформ одного слова w; $m(w)$ – функція нормалізації морфологічної форми слова w. З огляду на задачу дослідження, позначемо словосполучення як:

$$G \rightarrow [DT] \rightarrow D, \tag{5}$$

де G – головне слово (Governor); $[DT]$ – тип зв'язку (Dependency Type); D – залежне слово (Dependent), при чому для зручності розгляду зв'язків з (5), слова краще представляти у формі (3), але для пошуку ключових слів – у формі (4).

Формалізовано задачу пошуку ключових слів тексту як параметричну ідентифікацію функції згортки вербальної інформації за критерієм максимума інформації у зв'язках $I(G \rightarrow [DT] \rightarrow D)$, які поєднують n обраних ключових слів тексту T між собою та з усіма m' значущими словами цього тексту:

$$W^k(T) = argmax\, f(W_T), \tag{6}$$

де $W^k(T) = \{w_i^k \mid i = \overline{1,n}\}$ – множина n обраних ключових слів тексту з усіх m' значущих слів цього тексту $W_T = \{w_j \mid j = \overline{1,m'}\}$, а

$$f(W_T) = \sum_{i=1}^{n}\sum_{j=1}^{m'} I(G_i \rightarrow [DT] \rightarrow D_j) + \sum_{i=1}^{n}\sum_{j=1}^{m'} I(G_j \rightarrow [DT] \rightarrow D_i) \tag{7}$$

– функція згортки вербальної інформації в процесі пошуку ключових слів з T.

Розглянуто задачу пошуку ключових слів тексту як певну інформаційну технологію, що має на вході текст, а на виході – множину з l ключових слів $W^k = \{w_1^k, ..., w_l^k\}$. Текст T складається з m' різних слів, а окреме його j-те речення з k, в загальному випадку, налічує n слів з m можливих, причому $m' >> n$ та $m' >> l$. Більшість відомих методів пошуку ключових слів тексту беруть за основу частотний словник тексту, який фактично є списком або впорядкованою 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, у першому наближенні маємо маємо $W^k = \{w_1, ..., w_l\}$.

Зауважимо, що технології парсингу природних мов, реалізовані в сучасних лінгвістичних пакетах дозволяють на доступному програмному рівні 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.

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. Всі ці процеси відбуваються шляхом порівняльного аналізу та залучення інформації з деякої загальнолінгвістичної бази знань окремого (обраного)

суб'єкта розуміння. При цьому, якщо кожен з цих двох етапів супроводжується збільшенням інформації, то:

- 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 пошуку $W^k = \{w_1^k, ..., w_l^k\}$ пропорційна рівню загального розуміння тексту, що має підтверджуватися формальними ознаками.

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 її врахувати. 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.

Характерно, що збільшуються частоти саме тих слів, які потенційно можуть належати до множини ключових. Проведемо формальну оцінку такого збільшення з урахуванням обмежень щодо наявності виключно різних слів у реченні та не взявши до уваги зворотну залежність між підметом та присудком:

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

б) 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 $f_i^{new} = f_i + f_i^{max} = k + \sum_{j=1}^{k}(n_j - 2) = \sum_{j=1}^{k}(n_j - 1)$;

в) 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.

На основі проведеної інформаційної оцінки результатів парсингу тексту отримано формальні межі лінійного збільшення кількості інформації для значущих слів тексту, з яких обираються ключові. На відміну від існуючих моделей, така інформація враховується запропонованим підходом для автоматизації пошуку ключових слів тексту.

У межах запропонованої математичної моделі обгрунтовано обов'язкове врахування 31 значущого типу зв'язків Стенфордської класифікації (*ACL, ADVCL, ADVMOD, AMOD, APPOS, AUX, CASE, CCOMP, CLF, COMPOUND, CONJ, COP, CSUBJ, DEP, DISCOURSE, DISLOCATED, FLAT, GOESWITH, IOBJ, LIST, MARK, NMOD, NSUBJ, NUMMOD, OBJ, OBL, ORPHAN,*

PARATAXIS, REPARANDUM, VOCATIVE, XCOMP), в процесі пошуку ключових слів та виключення з процесу аналізу тексту семи неінформативних типів зв'язків (*CC, DET, EXPL, FIXED, PUNCT, REF, ROOT*), а також 21-го тегу, якими позначаються неінформативні частин мови (*CC, CD, DT, EX, IN, LS, MD, PDT, POS, PRP, PRP$, RP, SYM, TO, UH, WDT, WP, WP$, WRB, -LRB-, -RRB-*). Це і є головним обмеженням моделі, що пропонується.

Отже, удосконалена модель пошуку ключових слів (1)-(5) побудована на основі інформаційної оцінки результатів парсингу тексту та враховує результати аналізу зв’язків між лексичними одиницями тексту, що дозволило формалізувати критерій якості (6)-(7) процесу пошуку ключових слів.

У третьому розділі розроблено загальний підхід до пошуку ключових слів з урахуванням зв'язків між членами речення. Підхід реалізовано на базі двох методів, що забезпечують такі технологічні етапи: створення багаторівневої розмітки тексту; застосування синтаксичної розмітки, що враховує складні залежності між парами лем; зменшення вербального шуму; вибір перших *n* слів з найбільшою кількістю зв’язків, де *n* – кількість потрібних ключових слів.

Розглянуто суттєві особливості методів, що пропонуються, зокрема колізію при знаходженні ключових слів – це рівність значень частоти для двох чи більше кандидатів у ключові слова, причому вибрати в якості ключових з них потрібно тільки частину. Пропонується використовувати комбінований підхід для зменшення колізії, зокрема спочатку перевіряти ключові слова з однаковою частотою на зв'язність. На другому етапі, якщо у блоці потенційних ще залишилися ключові слова з однаковою частотою, вибираються спочатку іменники, потім дієслова, а потім інші частини мови.

Зменшення кількості шумових слів пропонується досягти за допомогою наступних кроків: заміна займенників на відповідні до них іменники; вилучення словосполучень із типами зв’язків, які не несуть суттєвого смислового навантаження; вилучення слів, які відносяться до неінформативних частин мови; вилучення слів, які відносяться до списку стоп-слів.

Розроблено алгоритм пошуку ключових слів, що забезпечує:

1. Створення багаторівневої розмітки тексту.

2. Отримання синтаксичної розмітки, що враховує складні залежності між парами лем.

3. Виключення неінформативних для аналізу типів зв’язків – вилучаються слова, які не відносяться до самостійних частин мови.

4. Заміну займенників в отриманих парах на відповідні до них іменники.

5. Отримання пар ключових слів.

6. Розбиття пар на окремі слова і визначення кількості зв’язків; даний крок дозволяє збільшити шанси отримання більш вагомих ключових слів, оскільки багато з них замінюються в наступних реченнях займенниками.

7. Вибір перших n слів з найбільшою кількістю зв’язків, де n – кількість потрібних ключових слів.

На рис. 1 представлено схему алгоритму аналізу тексту, який є базовим для двох взаємопов'язаних методів пошуку ключових слів, що пропонуються.

На вхід алгоритму надходить текст, для якого створюється багаторівнева розмітка, що включає набір речень з яких він складається. Далі, на основі синтаксичного розбору кожного речення знаходяться зв'язки між членами цього речення. Наступний крок - для кожного зв'язку визначається словосполучення, що містить головне і залежне слово та тип зв’язку між ними. Останній етап - накопичення словосполучень в масиві(currentPhrases).

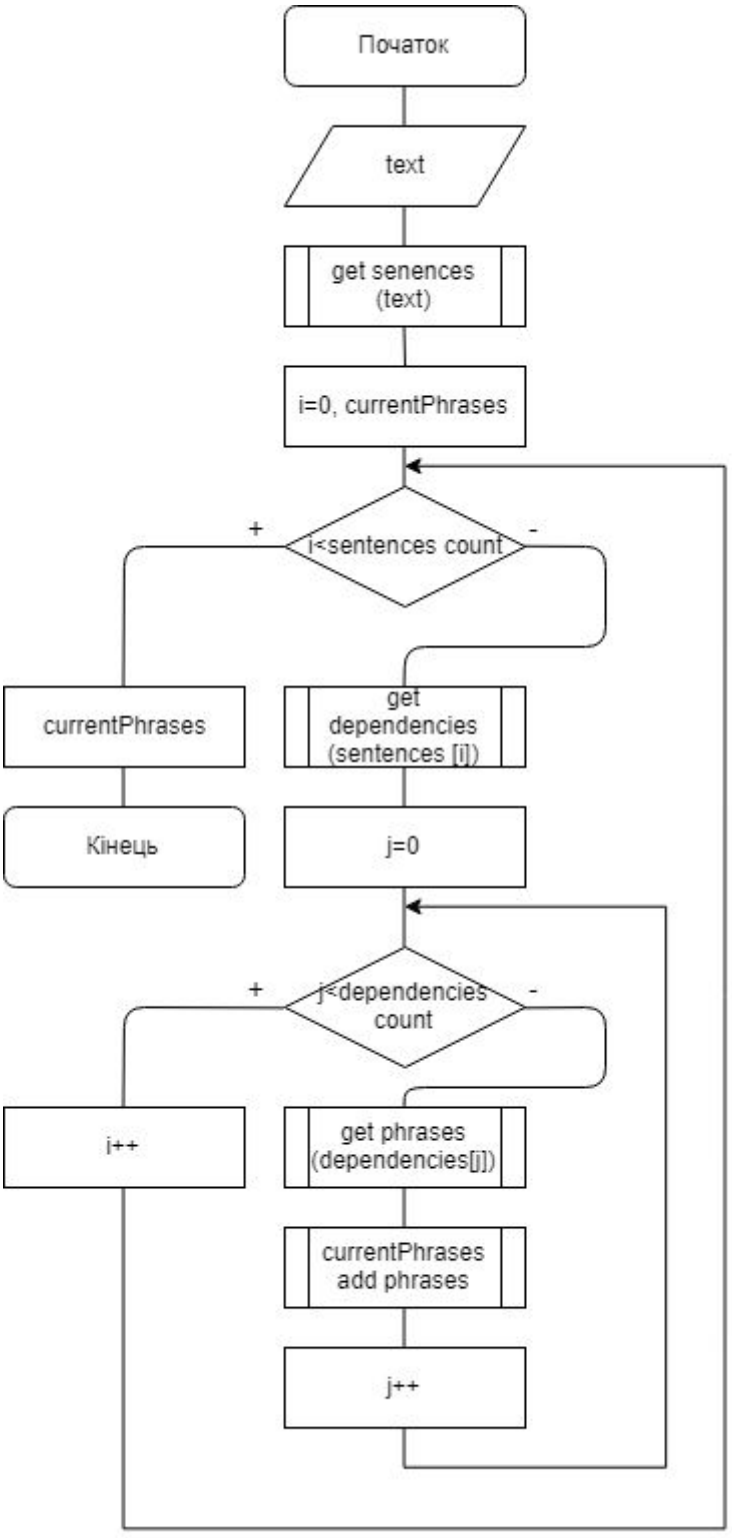

Рисунок 1 – Схема алгоритму аналізу тексту

Алгоритм фільтрації вербального шуму отримує словосполучення (currentPhrases), що отримано з тексту в результаті його розбору. Далі словосполучення розбиваються на окремі слова і підраховується кількість зв'язків для кожного з них.

Створення багаторівневої розмітки тексту і побудову синтаксичної розмітки, що враховує складні залежності між парами лем доцільно здійснювати за допомогою сучасних лінгвістичних пакетів, що мають відповідний функціонал для парсингу текстів.

Наразі такі можливості для англійської мови присутні у цілого ряду відомих лінгвістичних пакетів, що відрізняються, у першу чергу, базовою мовою програмування. Найбільш популярними з них вважаються такі: NLTK для мови Python, DKPro Core та Stanford NLP для мови Java, SharpNLP для .NET, MeTA для C++, Apache OpenNLP для Java або C++ тощо. Для реалізації методів пошуку ключових слів і зменшення вербального шуму обрано пакети DKPro Core та Stanford NLP.

На рис. 2 представлено UML діаграму класів. Зокрема, клас NLPKRootScreen відповідає за відображення інтерфейсу користувача і зв’язаний з головним класом запропонованого підходу NLPKTextProcessing, що відповідає за логіку пошуку ключових слів. Як видно з діаграми, клас NLPKTextProcessing зберігає масив словосполучень (currentPhrases) для того, щоб після розбору тексту можна було багаторазово робити пошук ключових слів, при цьому застосовуючи всі модулі фільтрації вербального шуму або тільки деякі з них. Елементами масиву словосполучень є об’єкти типу NLPKPhrase, в яких зберігається інформація про головне і залежне слова, а також тип зв'язку між ними.

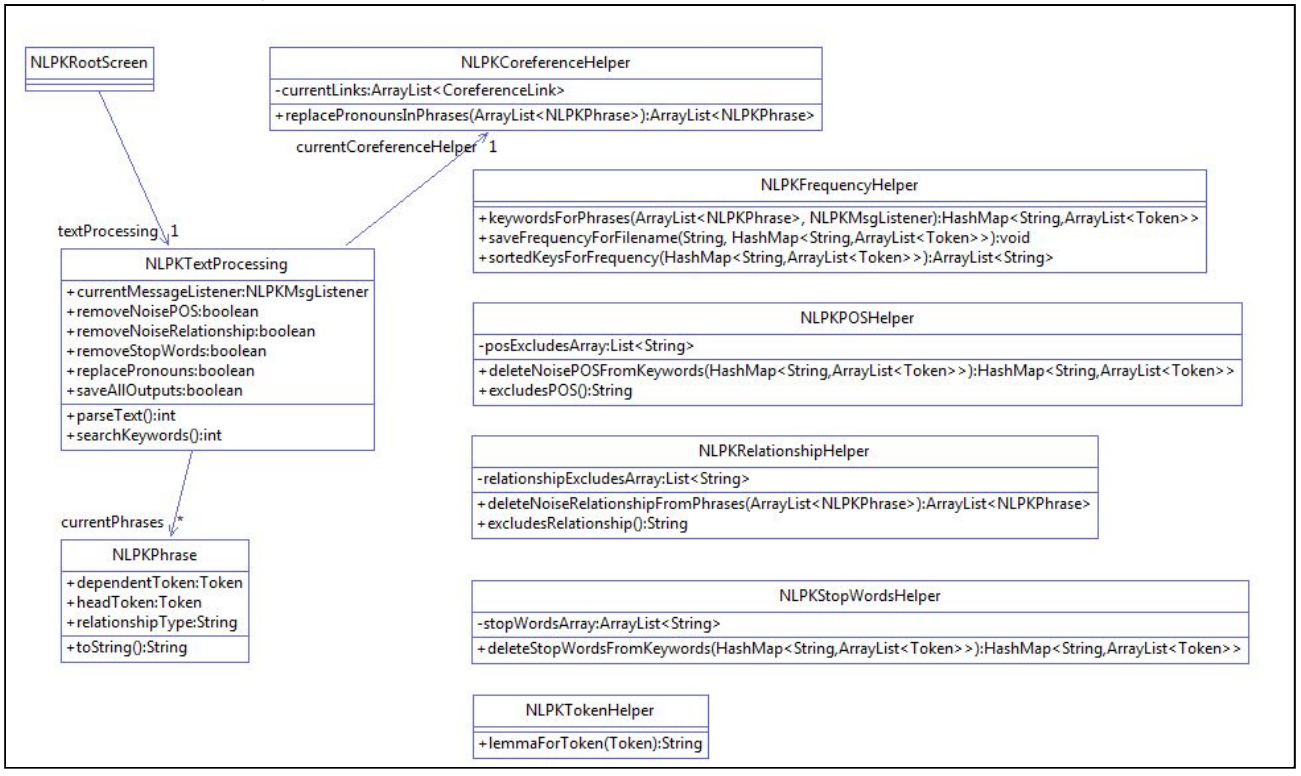

Рисунок 2 – UML діаграма класів

Головне і залежне слово є об'єктами типу Token, що зберігає інформацію про слово в тексті. Такою інформацією є: порядковий номер в тексті першого і останнього символу слова; лінк на попереднє і наступне слово; основна форма слова; тег частини мови тощо. Також, клас NLPKTextProcessing створює об'єкт класу NLPKCoreferenceHelper на етапі розбору тексту. У цьому об'єкті зберігається інформація про займенники, які зустрічаються в тексті та слова, з якими вони пов'язані.

На рис. 3 зображена діаграма зв'язків між класами. Основним класом в ній також є NLPKTextProcessing, який відповідає за оброблення тексту і пошуку ключових слів та використовує клас NLPKFrequencyHelper для підрахунку кількості зв'язків для певного слова з тексту.

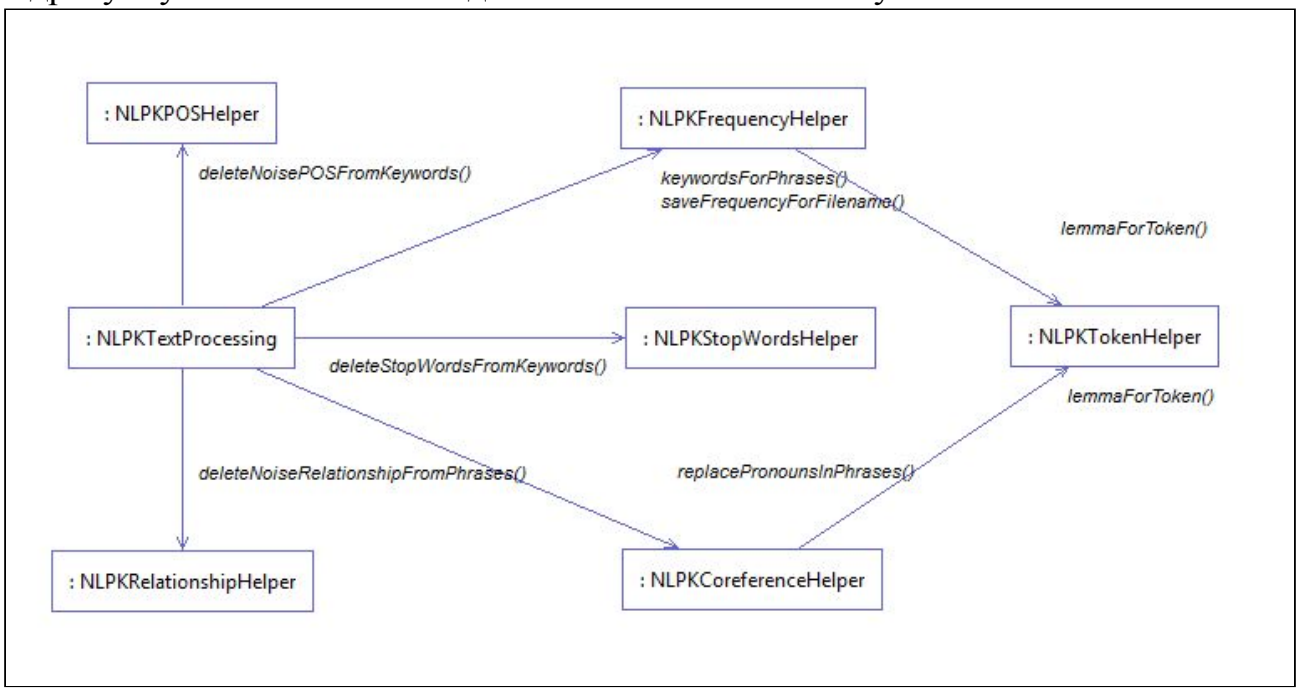

Рисунок 3 – Діаграма зв'язків

NLPKFrequencyHelper, в свою чергу, пов'язаний з класом NLPKTokenHelper, робота якого полягає в знаходженні основної форми відповідного слова. Це потрібно для того, щоб зв'язки накопичувалися для унікальних слів, без повторів. Також головний клас NLPKTextProcessing запускає, по черзі, модулі фільтрації вербального шуму:

- модуль заміни займенників (NLPKCoreferenceHelper);
- модуль вилучення словосполучень, які не несуть суттєвого смислового навантаження (NLPKRelationshipHelper);
- модуль вилучення слів, які відносяться до неінформативних частин мови (NLPKPOSHelper);
- модуль вилучення стоп-слів (NLPKStopWordsHelper), який пов'язаний з модулем заміни займенників NLPKCoreferenceHelper і класом NLPKTokenHelper, оскільки займенники у словосполученнях мають замінюватися на іменники в основній формі слова.

Отже, розроблений метод пошуку ключових слів базується на знаходженні синтаксичних зв’язків між словоформами у реченнях

англомовного тексту за допомогою технологічних можливостей парсингу сучасних лінгвістичних пакетів. Разом з ним удосконалено метод зменшення впливу вербального шуму на пошук ключових слів, який побудовано на основі стенфордської класифікації зв'язків між лексичними одиницями речення згідно моделі з другого розділу.

У четвертому розділі розроблено інформаційну технологію пошуку ключових слів та її структуру, яка забезпечує послідовне застосування основного та додаткового методів, обгрунтованих у розділі 3. Структуру інформаційної технології (рис. 4) розроблено за модульним принципом з метою подальшого розвитку програмного забезпечення шляхом інтеграції нових модулів в існуючу архітектуру, розширюючи таким чином функціонал системи.

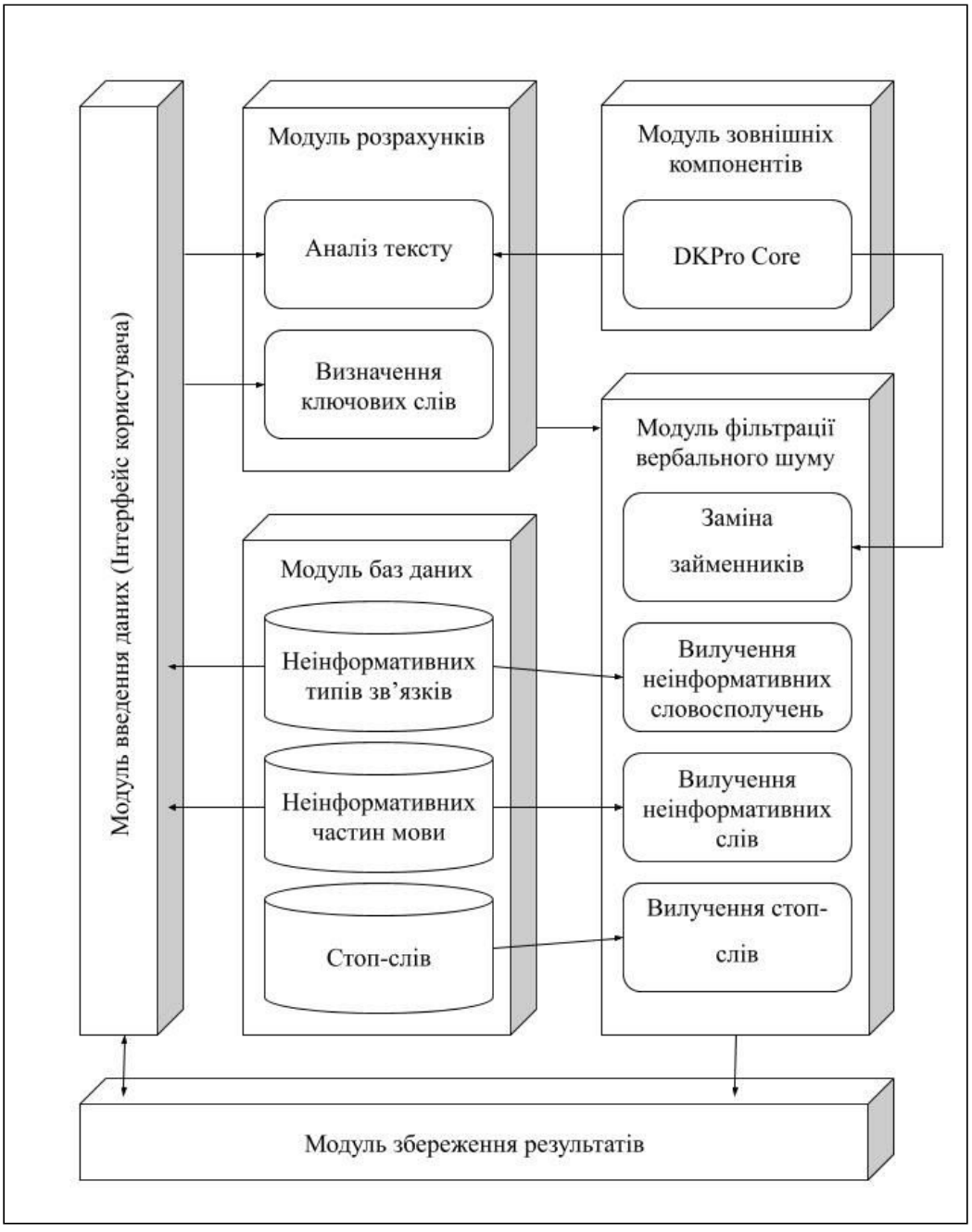

Рисунок 4 – Структура інформаційної технології

Модуль введення даних (Інтерфейс користувача) відповідає за: вибір з носіїв інформації тексту, для якого необхідно знайти ключові слова; відображення типів зв’язків, які не несуть суттєвого смислового навантаження; відображення списку неінформативних частин мови; запуск аналізу тексту та пошуку ключових слів; вибір модулів фільтрації вербального шуму – всіх або тільки деяких з доступних; відображення прогресу аналізу тексту та логів виконання.

Модуль розрахунків здійснює аналіз тексту та пошук ключових слів. Аналіз тексту передбачає створення багаторівневої розмітки тексту та отримання синтаксичної розмітки, що враховує складні залежності між парами лем. Модуль зовнішніх компонентів відповідає за роботу з DKPro Core і завантаження необхідних для його роботи бібліотек.

Модуль баз даних відповідає за роботу з інформацію про неінформативні типи зв’язків, а також слова, які відносяться до неінформативних частин мови та список стоп-слів. Модуль збереження результатів зберігає ключові слова знайдені після аналізу зв'язків, і потім, після кожного кроку фільтрації вербального шуму, на носії інформації.

За допомогою розробленого програмного забезпечення, в якому реалізовано дружній до користувача інтерфейс, було проведено значну кількість експериментів. Зокрема були опрацьовані тексти з офіційними англомовними перекладами Послань президента РФ до Федеральних зборів 2013 та 2014 років, де метод пошуку ключових слів порівнювався з частотним статистичним методом. Результати порівняння отриманих ключових слів для 2013 року свідчать, що при знаходження однакової кількості (10) значущих ключових слів власна розробка додає додатково 2 стоп-слова, а частотний словник – 11. Аналогічне порівняння для 2014 року демонструє схожу пропорцію надлишковості стоп-слів при пошуку 10 значущих ключових слів – власна розробка додає 4 стоп-слова, а частотний словник – 20. Це означає, що запропонований підхід забезпечує збільшення питомої ваги значущих ключових слів у межах від 0,33-0,48 до 0,71-0,83 (від 148% до 251%), при цьому приблизно на 80% стабільно відфільтровуються стоп-слова у кожному експерименті. Порівняльний аналіз складу значущих ключових слів для цих двох текстів, отриманих відомим та запропонованим підходами, додатково продемонстрував досить показові результати, які певним чином виходять за межі формального лексичного аналізу. Наприкінці 2013 року президент РФ говорив: we, I, work, need, system. А рівно через рік, у 2014, він уже казав: I, Russia, people, Russian, work – застосування формальних засобів інформаційної технології недвозначно свідчить про суттєву зміну у публічній політичній риториці сусідньої країни.

Кількісними характеристиками релевантності отриманих результатів обрано повноту (за Жаккаром) і точність (абсолютну). Проведено інтерпретацію обраних критеріїв до умов задачі пошуку ключових слів, зокрема повнота за Жаккаром визначається за формулою:

$$J = \frac{n(A \cap B)}{n(A)+n(B)-n(A \cap B)} = \frac{n(A \cap B)}{n(A \cup B)}, \qquad (8)$$

де A – множина ключових слів, заданих автором;

B – множина ключових слів, знайдених програмно;

$n(A)$ – кількість елементів множини ключових слів, заданих автором;

$n(B)$ – кількість елементів множини ключових слів, знайдених програмно;

$n(A \cap B)$ – кількість елементів множини-результату перетину двох множин;

$n(A \cup B)$ – кількість елементів множини-результату об'єднання двох множин.

Абсолютна точність, у свою чергу, визначається, як відношення кількості правильно знайдених програмою ключових слів до кількості еталонних ключових слів (заданих автором):

$$a = \frac{n(A \cap B)}{n(A)} \tag{9}$$

Проведено порівняльний аналіз запропонованої інформаційної технології з системами-аналогами, якими є сайти SEO оптимізації, де є функція пошуку ключових слів: seotool, keywordstext, advego. За результатами пошуку ключових слів для короткої анотації з 141 слова «Variability management in software product lines using adaptive object and reflection» власна розробка знаходить найбільше слів, заданих автором – 6 слів з 11. При цьому програми-аналоги знаходять для цього тексту по 5 слів. Отже, запропонована інформаційна технологія одночасно збільшує на 8,1% повноту за метрикою Жаккара (до 38%) та на 9,1% абсолютну точність (до 55%) пошуку ключових слів у порівнянні з програмами-аналогами.

Аналогічний експеримент проведено для тексту статті, що має на порядок більше слів – 1460 «A new pattern for historical geography: working with enthusiast communities and public history». Власна розробка стабільно демонструє кращі кількісні характеристики повноти і точності знаходження ключових слів, порівняно з аналогами – 40% та 57%. Для цього експерименту також одночасно підвищено на 12,7% повноту та на 14,3% точність отриманих ключових слів.

У більш масштабній апробації двох запропонованих методів, проведено експеримент з текстом «Participation, archival activism and learning to learn», що складається з 1588 слів. Послідовне використання удосконаленого методу до зменшення шуму після основного методу пошуку ключових слів також дозволило покращити загальні результати – збільшено на 7,5% повноту за метрикою Жаккара (до 20%) та на 11,1% абсолютну точність (до 33,3%) пошуку ключових слів. Ще однією суттєвою перевагою в порівнянні з аналогами є те, що запропонована інформаційна технологія дозволила повністю (до 100%) виключити шумові слова: проведено заміну чотирьох шумових слів з дев'яти першого списку на чотири значущих слова-кандидата у ключові слова.

Проведено масштабні експериментальні дослідження для оцінки перспективного напрямку впровадження запропонованої інформаційної технології, що полягає у пошуку ключових слів для дуже коротких англомовних текстів. Експериментальну базу склали близько 30 000 відібраних

коротких анотацій до фільмів, що супроводжувалися заданими ключовими словами та були розбиті на категорії за жанрами.

Потрібно звернути увагу на явну “незручність” експериментального лексичного матеріалу, наприклад, у багатьох випадках еталонним є тільки одне ключове слово або еталонні ключові слова жодного разу не зустрічаються в тексті, текст анотації складається лише з кількох слів, тощо. Не зважаючи на такі складнощі, запропонована інформаційна технологія знаходить одне і більше ключових слів, заданих автором для 42.3% опрацьованих анотацій. При цьому 51.2% усіх позитивних результатів отримано з немалою точністю від 0.2 до 0.55, а 41.1% позитивних результатів отримано з повнотою в діапазоні від 0.12 до 0.37.

Проведено аналіз значення точності та повноти результатів пошуку ключових слів для п’яти найбільших категорій фільмів: Drama, Comedy, Thriller, Romance, Action. У середньому результати, наведені у таблиці 1, приблизно однакові, що дає підстави стверджувати про лексичну незалежність запропонованої інформаційної технології для англомовних текстів.

Таблиця 1 – Результати середнього значення точності та повноти для найбільших категорій фільмів

Назва	Action	Romance	Thriller	Comedy	Drama
Absolute	0,22142	0,22766	0,220114	0,231997	0,227988
Jaccard	0,133538	0,138171	0,131441	0,14134	0,137378

До обмежень у застосуванні запропонованої інформаційної технології можна віднести швидкодію її 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.

Впровадження запропонованої інформаційної технології не потребує додаткових витрат для компанії. Аналогами та перспективними об'єктами її впровадження можуть бути сайти SEO оптимізації з можливістю пошуку ключових слів, а також автоматизовані системи підтримки англомовного контент-аналізу. Запропонована інформаційна технологія пошуку ключових слів враховує додаткову інформацію процесів парсингу речень у межах послідовного застосування двох запропонованих методів, що дозволило уточнити чисельні оцінки змістовних параметрів тексту та підвищити якість пошуку його ключових слів. Визначено обмеження та перспективні напрями впровадження розробленої інформаційної технології.

У додатках подано документи щодо впровадження результатів роботи; список публікацій за темою дисертації та відомості про апробацію результатів; практичні поради при складанні списку ключових слів; схематичне зображення робочого процесу в DKPro Lab; граф зв'язків між словами.

ВИСНОВКИ

Внаслідок проведеного дослідження було вирішене актуальне наукове завдання підвищення якості пошуку ключових слів у англомовному тексті шляхом розробки інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів. Мета і задачі дослідження формувалися на гіпотезі про те, що підвищення якості процесу та результатів отримання ключових слів для природно-мовного тексту вимагає залучення додаткової інформації про цей текст, причому універсального, а не специфічного характеру.

В роботі запропоновано підхід до пошуку ключових слів, що базується на використанні додаткової інформації універсального характеру про складні залежності між членами англомовного речення. За результатами математичного моделювання формалізовано задачу пошуку ключових слів тексту як параметричну ідентифікацію функції згортки вербальної інформації за критерієм максимума інформації у зв'язках, які поєднують n обраних ключових слів тексту T між собою та з усіма m' значущими словами цього тексту.

У межах запропонованої математичної моделі обгрунтовано обов'язкове врахування 38 значущих типів зв'язків в процесі пошуку ключових слів та виключення з процесу аналізу тексту 7-ми неінформативних типів зв'язків, а також 21 тегу, якими позначаються неінформативні частин мови. Модель дозволила формалізувати критерій якості процесу пошуку ключових слів.

Уперше розроблено метод пошуку ключових слів, який, на відміну від існуючих, базується на знаходженні синтаксичних зв’язків між словоформами у реченнях англомовного тексту за допомогою технологічних можливостей парсингу сучасних лінгвістичних пакетів. Для зменшення колізії при знаходженні ключових слів спочатку перевіряються ключові слова з однаковою частотою на зв'язність, а далі, якщо у блоці потенційних ще залишилися ключові слова з однаковою частотою, вибираються спочатку іменники, потім дієслова, а потім інші частини мови.

Зменшення кількості шумових слів пропонується досягти за допомогою удосконаленого методу, що передбачає заміну займенників на відповідні до них іменники; вилучення словосполучень із типами зв’язків, які не несуть суттєвого смислового навантаження; вилучення слів, які відносяться до неінформативних частин мови; вилучення слів, які відносяться до списку стоп-слів.

Для розробки програмного забезпечення було обрано мову програмування Java та лінгвістичний пакет DKPro Core – набір програмних компонентів для обробки природної мови, заснований на Apache UIMA framework. Розроблені структура інформаційної технології та відповідне програмне забезпечення.

Експериментально підтверджено теоретичні оцінки формальних меж лінійного збільшення кількості інформації для значущих слів тексту, з яких обираються ключові, внаслідок проведення парсингу та врахування результатів аналізу зв’язків між лексичними одиницями тексту. Досягнуто збільшення питомої ваги значущих ключових слів у межах від 0,33-0,48 до 0,71-0,83 (від 148% до 251%) у порівнянні з частотним словником.

Експериментальна апробація запропонованої інформаційної технології показала її переваги за кількісними характеристиками. Так, у порівнянні з 3-ма програмами-аналогами запропонована інформаційна технологія одночасно збільшує у межах від 8,1% до 12,7% повноту за метрикою Жаккара та від 9,1% до 14,3% абсолютну точність пошуку ключових слів для англомовних текстів обсягом 140-1400 слів. Послідовне використання удосконаленого методу до зменшення шуму після основного методу пошуку ключових слів дозволило для тексту з приблизно 1600 слів підвищити на 7,5% повноту за метрикою Жаккара (до 20%) та на 11,1% абсолютну точність (до 33,3%) пошуку ключових слів, при цьому повністю відфільтровано шумові слова.

Масштабний експеримент на основі 30 000 відібраних коротких анотацій до фільмів з відомими (заданими авторами) ключовими словами також продемонстрував перспективні результати. На максимально “незручному” для такої задачі природно-мовному матеріалі запропонована інформаційна технологія знайшла одне і більше ключових слів, заданих автором для 42.3% опрацьованих анотацій. Середнє значення повноти точності для знайдених розробленою технологією ключових слів відповідно складає 23.3% і 14.2%.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

[1] O.V. Bisikalo, W. Wójcik, O.V. Yahimovich, and S. Smailova, "Method of determining of keywords in English texts based on the DKPro Core", *Proceedings of SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016*, 100314T, 2016. DOI:10.1117/12.2249225.

[2] O. Bisikalo, A. Lisovenko, O. Jahumovuch, S. Trachenko, and M. Pradivliannyi, "System of computational linguistic on base of the figurative text comprehension", *Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing (DSMP 2016)*, pp. 69-74, 2016. DOI: 10.1109/DSMP.2016.7583510.

[3] О.В. Бісікало, та О.В. Яхимович, "Метод визначення ключових слів англомовного тексту на основі DKPro Core", *Технологический аудит и резервы производства: Информационные технологии*, Том 1, № 2(21), с. 26-30, 2015. ISSN 2226-3780.

[4] О.В. Бісікало, А.І. Лісовенко, О.В. Яхимович, та С.С. Траченко, "Визначення змістовних ознак тексту на основі аналізу зв’язків між лексичними одиницями", *Вісник НТУ «ХПІ». Серія "Механіко-технологічні системи та комплекси"*, № 21 (1130), с. 83-89, 2015. ISSN 2411-2798.

[5] О.В. Бісікало, та О.В. Яхимович, "Знаходження ключових слів англомовного тексту за допомогою інструментальних засобів пакету DKPro Core", *Інформаційні технології та комп’ютерна інженерія*, № 2(34), с. 10-14, 2015. ISSN 1999-9941.

[6] О.В. Бісікало, та О.В. Яхимович, "Автоматизоване визначення лексичних технологій з тезаурусу технічного спрямування", *Оптико-електронні інформаційно-енергетичні технології*, № 1 (31), с. 26-38, 2016. ISSN 1681-7893.

[7] О.В. Бісікало, О.В. Яхимович, та Я.В. Яхимович, "Розробка методу фільтрації вербального шуму в процесі пошуку ключових слів англомовного тексту", *Технологический аудит и резервы производства: Информационные технологии*, № 6(44), с. 33–41, 2018. ISSN 2226-3780.

[8] С.Д. Штовба, О.В. Штовба, О.В. Яхимович та М.В. Петричко, "Вплив синтаксичних зв’язків у реченнях на якість ідентифікації токсичних коментарів в соціальній мережі", *Наукові праці ВНТУ*, № 4, с. 1-8, 2019. DOI: https://doi.org/10.31649/2307-5376-2019-4-35-42.

[9] О.В. Яхимович, "Визначення ключових слів англомовного тексту з використанням технології DKPRO CORE", *Молодь в технічних науках: дослідження, проблеми, перспективи (МТН-2015) : Матеріали міжнародної Інтернет-конференції*, Вінниця, 2015, с. 72-74. ISBN 978-966-924-027-9.

[10] О.В. Бісікало, О.В. Яхимович, А.І. Лісовенко, та Траченко С.С., "Підтримка діалогу з навчальним контентом", *Адаптивні технології управління навчанням: матеріали першої міжнародної конференції*, Одеса, 2015, с. 97-100.

[11] О.В. Бісікало, О.В. Яхимович, А.І. Лісовенко, та Траченко С.С., "Моделювання процесів побудови парадигматичних зв’язків між словоформами на основі вимірювання текстової інформації", *Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)*, Вінниця, 2015, с. 119-121.

[12] О.В. Яхимович, "Застосування інструментальних засобів пакету DKPRO CORE для визначення ключових слів англомовного тексту", *XLIV науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2015, с. 7.

[13] О.В. Яхимович, "Визначення ключових слів з тексту повідомлень мікроблогів", *XLV науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2016, с. 1119-1121.

[14] О.В. Яхимович, "Колізія при знаходженні ключових слів", *XLVI науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2017, с. 1312-1314.

[15] О.В. Яхимович, "Зменшення вербального шуму при визначенні ключових слів", *XLVII науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2018, с. 1463-1465.

[16] О.В. Яхимович, "Формалізація задачі визначення ключових слів тексту", *XLVIII науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2019, с. 1136-1138.

[17] О.В. Бісікало, Р.Н. Квєтний, С.Г. Кривогубченко, Л.Є Азарова, та О.В. Яхимович, "Синтез інтегрованої бази знань природно-мовного контенту", ВНТУ, Вінниця, Д/б № 0114U003462, 2015.

[18] О.В.Бісікало, Р.Н. Квєтний, С.Г. Кривогубченко, А.І. Лісовенко, та О.В. Яхимович, "Розв’язання семантико-залежних задач обробки

природно-мовних об’єктів на основі бази знань", ВНТУ, Вінниця, Д/б № 0114U003462, 2016.

[19] О. В. Бісікало, А. І. Лісовенко, О. В. Яхимович, та В. В. Шолота, "Спосіб автоматичного пошуку ключових слів з використанням технології DKPro Core", *МПК G06F 17/21, G06F 17/27, G06F 17/28. № и 2019 00016*, Черв. 25, 2019.

[20] O. Bisikalo, and A. Yahimovich. *Keyword search based on lexical relationships in the text*. Beau Bassin-Rose Hill, Mauritius: Lap Lambert Academic Publishing, 2019. ISBN 978-620-0-00314-0.

[21] O. Bisikalo, and A. Yahimovich. *Lexical relationships-based keywords selection in english texts*. Vinnytsia, Ukraine: VNTU, 2020.

АНОТАЦІЯ

Яхимович О. В. **Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів. – Кваліфікаційна наукова праця на правах рукопису.**

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 «Інформаційні технології». – Вінницький національний технічний університет, Вінниця, 2021.

Робота присвячена розробці інформаційної технології пошуку ключових слів на основі автоматизації процесів парсингу англомовних текстів.

Удосконалено модель пошуку ключових слів, яка, на відміну від існуючих, побудована на основі інформаційної оцінки результатів парсингу тексту та враховує результати аналізу зв’язків між лексичними одиницями тексту, що дозволило формалізувати критерій якості процесу пошуку ключових слів.

Уперше розроблено метод пошуку ключових слів, який, на відміну від існуючих, базується на знаходженні синтаксичних зв’язків між словоформами у реченнях англомовного тексту за допомогою технологічних можливостей парсингу сучасних лінгвістичних пакетів.

Удосконалено метод зменшення впливу вербального шуму на пошук ключових слів, який, на відміну від існуючих, побудовано на основі зв'язків між лексичними одиницями речення, що дозволило підвищити якість результатів пошуку ключових слів у порівнянні з основним методом.

Набула подальшого розвитку інформаційна технологія пошуку ключових слів, яка, на відміну від існуючих, враховує інформацію процесів парсингу речень, що дозволило уточнити чисельні оцінки змістовних параметрів тексту та підвищити якість пошуку його ключових слів.

Ключові слова: інформаційна технологія, ключові слова, вербальний шум, лінгвістичний пакет, DKPro Core, синтаксичний аналіз, словосполучення.

АННОТАЦИЯ

Яхимович А. В. **Информационная технология поиска ключевых слов на основе парсинга англоязычных текстов. – Квалификационная научная работа на правах рукописи.**

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 «Информационные технологии». – Винницкий национальный технический университет, Винница, 2021.

Работа посвящена разработке информационной технологии поиска ключевых слов на основе автоматизации процессов парсинга англоязычных текстов.

Усовершенствована модель поиска ключевых слов, которая, в отличие от существующих, базируется на основе информационной оценки результатов парсинга текста и учитывает результаты анализа связей между лексическими единицами текста, что позволило формализовать критерий качества процесса поиска ключевых слов.

Впервые разработан метод поиска ключевых слов, который, в отличие от существующих, основан на обработке синтаксических связей между словоформами в предложениях англоязычного текста с помощью технологических возможностей парсинга современных лингвистических пакетов.

Усовершенствован метод уменьшения влияния вербального шума на поиск ключевых слов, который, в отличие от существующих, построен на основе связей между лексическими единицами предложения, что позволило повысить качество результатов поиска ключевых слов по сравнению с основным методом.

Получила дальнейшее развитие информационная технология поиска ключевых слов, которая, в отличие от существующих, учитывает информацию процессов парсинга предложений, что позволило уточнить численные оценки содержательных параметров текста и повысить качество поиска его ключевых слов.

Ключевые слова: информационная технология, ключевые слова, вербальный шум, лингвистический пакет, DKPro Core, синтаксический анализ, словосочетание.

ABSTRACT

Yahimovich O. V. **Information technology of searching keywords based on parsing English texts. – Qualification research paper, manuscript copyright.**

Thesis for the degree of a candidate of technical sciences in specialty 05.13.06 «Information technology». – Vinnytsia National Technical University, Vinnytsia, 2021.

The qualification research is dedicated to developing the informational technology of searching keywords based on the automation process of parsing English texts.

The purpose of the dissertation research is to increase the quality of searching keywords in English texts.

The scientific novelty of the qualification research paper is:

1. The model of searching keywords has been improved, which, unlike the existing ones, is based on the information evaluation of parsing text results and takes

into account the results of analysis of relationships between lexical units of text, which allowed to formalize the quality criterion of searching keywords process.

2. For the first time, searching keywords method has been developed, which, unlike the existing ones, is based on finding syntactic relationships between word forms in sentences of English text with the help of technological capabilities of parsing of modern linguistic packages. The proposed method allows to improve the numerical characteristics of searching keywords quality, namely completeness (according to Jacquard) and accuracy.

3. The method of reducing the impact of verbal noise for searching keywords has been improved, which, unlike the existing ones, is based on the Stanford classification of relationships between lexical units of a sentence, which has improved the quality of results of searching keywords compared to the main method.

4. The information technology of searching keywords has been further developed, which, unlike the existing ones, takes into account additional information of sentence parsing processes within the bounds of the consistent use of the two proposed methods, which allowed to refine numerical estimates of content parameters of the text and improve the quality of searching keywords.

The practical value of the results obtained in the qualification research paper is as follows: formal description of the method of searching keywords in the English text, creating an algorithm for its implementation and developing software that finds keywords based on significant relationships between word forms in sentences of the English text and subsequent filtering of verbal noise.

Created models, algorithms and software can be used in solving practical problems of computational linguistics, which require searching keywords, for example, to improve the accuracy of site content analysis and raise the position of the site in search results. The use of language-independent tools of the proposed information technology of searching keywords in combination with the needed, according to the obtained specification, technological resources of linguistic analysis of other natural languages will expand the scope of information technology, in particular to use in the Ukrainian language.

The results of the qualification research paper were implemented at LLC «SPILNA SPRAVA» (act on the results of implementation from 10.01.2020), as well as to the educational process of the Automation and Intelligent Information Technologies Department of Vinnytsia National Technical University, which is confirmed by the publication of the textbook “A lexical relationships-based keywords selection in an English text”. The results of the experiments showed that the proposed information technology simultaneously increases in the range from 8.1% to 12.7% the completeness according to the Jacquard metric and from 9.1% to 14.3% the absolute accuracy of searching keywords for English texts of 140-1400 words in comparison with analogues.

Keywords: information technology, keywords, verbal noise, linguistic package, DKPro Core, syntactic analysis, phrase.

Підписано до друку 02.03.2021 р. Формат 21х29.7 1/4
Наклад 100 прим. Зам. № 2021-019.
Віддруковано в інформаційному редакційно-видавничому
центрі Вінницького національного технічного університету
м. Вінниця, вул. Хмельницьке шосе, 95. Тел.: 65-18-06
Свідоцтво суб'єкта видавничої справи
серія ДК №3516 від 01.07.2009 р.