УДК 004.9

О. В. Бісікало
Вінницький національний технічний університет,
кафедра автоматики та інформаційно-вимірювальної техніки

СТАТИСТИЧНИЙ АНАЛІЗ СКЛАДНИХ ЗАЛЕЖНОСТЕЙ У ТЕКСТІ

Розглянуто обґрунтування підходу до застосування складних залежностей між словоформами для розв'язання задач семантичного аналізу тексту. Сформульовані основні положення підходу та визначені у вигляді гіпотез основні його переваги. Запропоновано формальне поняття предметної області. Отримано статистичні та інформаційні оцінки зв'язків між лемами тексту, які технологічно можна визначити за допомогою сучасних лінгвістичних пакетів, зокрема DKPro Core.

Ключові слова: словоформа, лема, складна залежність, розподіл Парето, дерево зв'язків.

The approach to the application of complex dependencies between word-forms in resolving the semantic text analysis problems has been grounded in the article. General points and main advantages of the approach have been formulated. A formal notion of the subject area has been suggested. The statistical and information estimates of the relations between lemmas have been obtained.They can be determined technologically using modern language packs (DKPro Core).

Key words: word-form, lemma, difficult dependency, Pareto distribution, tree ties.

Вступ. Загальна постановка проблеми

Традиційно комп'ютерна лінгвістика пов'язана з математичною статистикою на рівні дослідження появи у тексті окремих символів та їх послідовностей (слів). Відомі з часів А. А. Маркова (1856–1922 рр.) та Дж. К. Ципфа (1902–1950 рр.) фундаментальні статистичні закономірності розподілу символів/слів природної мови отримані завдяки не тільки легкості фіксації окремого символу/слова у текстовому файлі, але й, більшою мірою, панівними лінгвістичними концепціями, що надають слову основоположне значення. Схожа ситуація і з етапами лінгвістичного аналізу тексту – найкращу ситуацію маємо з практично завершеним морфологічним аналізом, значно слабкіші результати демонструє синтаксичний аналіз речень, а в семантичному аналізі і до цього часу не вирішено низку проблемних питань. Багатозначність слова як головна проблема формального аналізу текстової інформації якраз і підвищує складність кожного наступного етапу лінгвістичного аналізу, отже, для її вирішення необхідно застосовувати нову інформацію та знання, отримати які потрібно з того самого корпусу текстів.

Сутність асоціативно-статистичного підходу до отримання знань з текстової інформації [1] полягає в моделюванні природних шляхів накопичення асоціацій між образами та закріпленні рефлексів шляхом повторень у людини. При цьому наголошується на дослідженні саме асоціативних зв'язків, яким у комп'ютерній лінгвістиці відповідають поняття синтагматичних та парадигматичних (загалом – складних) залежностей між словоформами/лемами/стемами, які, своєю чергою, близькі до поняття мовного образу. Важливе значення у такому моделюванні мають статистичні характеристики текстових параметрів, що використовуються в запропонованих моделях і методах, оскільки саме вони дозволяють визначити формальні обмеження і асоціативно-статистичного підходу загалом, і окремих його складових.

Аналіз останніх досліджень і публікацій

Відомо, що статистичні закономірності щодо повторень слів у тексті відповідають гіперболічному розподілу Парето згідно з законами Ципфа [2]. Численні експерименти показали – якщо виміряти кількість входжень кожного слова в тексті та взяти тільки одне значення з кожної

групи, що має однакову частоту, потім розташувати частоти по мірі їх спадання і пронумерувати, то порядковий номер частоти можна вважати рангом частоти (позначимо r_i ранг слова i). Тоді слова, що зустрічаються найчастіше, матимуть ранг 1, наступні за ними – 2 і так далі. Очевидно, що ймовірність зустріти довільне, заздалегідь вибране слово дорівнюватиме відношенню кількості входжень цього слова до загального числа слів у тексті $p = n_i / |M|$, де n_i – кількість входжень i-го слова, а $|M|$ – кількість слів у тексті. Перший закон Ципфа стверджує – добуток ймовірності виявлення слова в тексті до рангу його частоти є постійним числом

$$\frac{n_i \cdot r_i}{|M|} = C \,. \tag{1}$$

Розглянутий емпіричний закон показує, що поширеність слова в тексті змінюється за гіперболою залежно від кількості входжень. Наприклад, друге зі слів, що зустрічається найчастіше, зустрічається приблизно в два рази рідше, ніж перше, третє – у три рази рідше, ніж перше тощо. Значення константи C в різних мовах різна, але всередині однієї мовної групи залишається приблизно незмінною, незалежно від тексту. Для російськомовних текстів константа Ципфа приблизно дорівнює 0,08, а для англомовних – 0,1.

Перший закон не враховує той факт, що різні слова можуть входити в текст з однаковою частотою. Ципф та, у подальшому, Юл, встановили, що частота і кількість неоднакових слів, що входять у текст з цією частотою, також відповідають певній залежності. Якщо побудувати графік, відклавши за віссю абсцис частоту входження слова, а за віссю ординат – кількість різних слів у даній частоті, то отримана крива буде зберігати свій вид для всіх без винятку текстів. Як і для першого закону, це твердження правильне у межах однієї мови. Однак і міжмовні відмінності невеликі. Якою б мовою текст не був написаний, вигляд кривої за другим законом Ципфа залишиться незмінним, може незначно відрізнятися лише коефіцієнт гіперболи.

Формально залежності за 1-м (1) та 2-м законами Ципфа відповідають двопараметричному розподілу Парето для випадкової величини X

$$P(X < x) = 1 - \left(\frac{x_m}{x}\right)^{k^p} \,, \quad \forall x \geq x_m \,, \tag{2}$$

де параметри $x_m, k^p \geq 0$, причому x_m є коефіцієнтом масштабу, а k^p – початкове (найбільше) значення частоти для рангу 1. На рис. 1 показано приклад залежності частоти слова від рангу для $x_m = 1$ у відповідності до параметра k^p розподілу Парето.

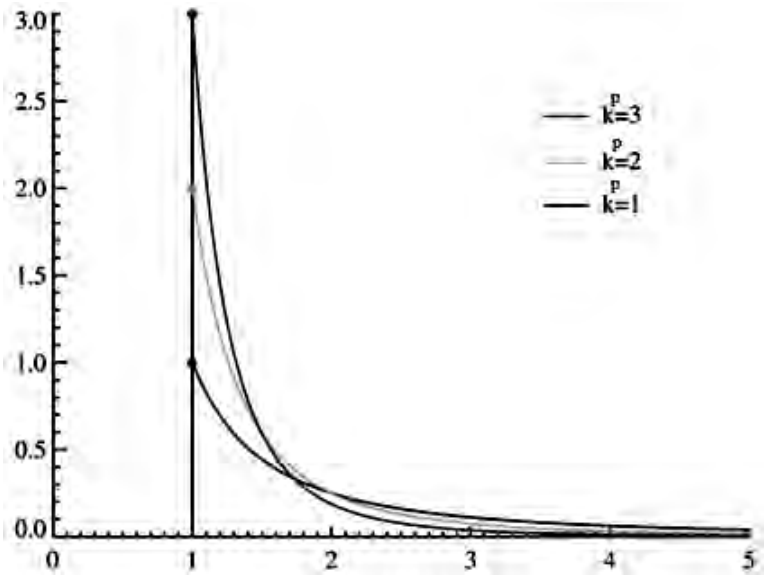

Рис. 1. Вигляд залежності для $x_m = 1$ за розподілом Парето

Прийнято вважати, що гіперболічна залежність (2) дозволяє при аналізі тексту виявити три групи слів [3]:

- початковий пік – найбільш частотні та короткі стоп-слова, що мають службове значення;
- ключові або найзначиміші з огляду на зміст тексту слова (середня частина гіперболи, оптимальний вибір ширини якої дозволяє досягти найвищої якості ключових слів);
- хвостова частина – найдовші, як правило, слова, що зустрічаються рідко та майже не характеризують зміст тексту.

Абсолютна більшість досліджень в окресленому напрямку стосувалася окремих слів та обґрунтування методик, що за рахунок застосування різного роду емпіричної та/або евристичної інформації мають за мету підвищити ефективність знаходження значимих (ключових) слів тексту. Проте результат розв'язання аналогічної задачі людиною базується зовсім не на статистиці окремих слів тексту, а на значенні цих слів, причому не загальному, а смисловому, у контексті змісту тих подій чи стану речей, які описуються обраним текстом.

Виділення проблеми

Актуальною проблемою дослідження вважатимемо отримання нової чисельної інформації щодо змістовних характеристик природномовного тексту та її застосування для ефективного розв'язання окремих задач комп'ютерної лінгвістики.

Формулювання мети

Мета дослідження полягає у теоретичному та експериментальному (на основі сучасних інструментальних засобів) обґрунтуванні підходу щодо інформативності статистичних ознак та параметрів для складних залежностей між словоформами / лемами тексту.

Статистична оцінка складних залежностей між словоформами (лемами)

Для розв’язання задачі дослідження послідовно розглянемо основні визначення та положення підходу, що пропонується, особливості статистичної та інформаційної оцінки ознак та параметрів для складних залежностей між такими лінгвістичними поняттями тексту, як словоформи та леми.

Основні визначення та положення підходу

У межах задекларованого підходу визначимо як основні поняття словоформи та леми. Без применшення загальності можна вважати систему понять слово-словоформа-лема довільної природної мови нестрогою ієрархічною системою. Верхнім рівнем ієрархії є кінцева множина лем I, кожна з яких складається тільки зі значимих словоформ, тобто з розгляду вилучено артиклі, прийменники та службові слова. Але, якщо леми являють собою множини словоформ, що не перетинаються між собою, то на нижчому рівні ієрархії маємо нестрогість – деякі слова одночасно можуть бути елементами множин різних словоформ, що відображає багатозначність слова.

Поняття складної залежності між лінгвістичними поняттями визначимо в загальному випадку як відношення $\Omega \subseteq I \times I$ між лемами тексту. Тоді лінгвістичною системою S будемо вважати таку, яка функціонально забезпечує:

- виокремлення з вхідного тексту всіх слів;
- формування припущень про відповідність кожного слова тексту певної словоформи та кожної словоформи до певної леми;
- синтаксичний аналіз всіх речень тексту;
- визначення та обробка складних залежностей між лемами тексту.

Врахуємо, що статистичний аналіз тексту на рівні окремих слів не дозволяє позбутися багатозначності слів, яка є однією з ключових проблем комп’ютерної лінгвістики. З метою залучення додаткових лінгвістичних засобів для зняття різних типів омонімії введемо формальне поняття предметної області як важливе обмеження запропонованого підходу до отримання знань з тексту. Припустимо, що всього у текстовій колекції виявлено $|M|$ різних слів, вилучення з яких стоп-слів привело до фіксації $|M'|$ значимих слів. У той же час зафіксовано n лем, кожна з яких є деякою множиною слів. Тоді предикат P визначає предметну область як

$$\forall w_i, \exists I_j \left| w_i \in I_j, w_i \notin I_k, k \neq j, i \in \overline{1,|M'|}, j \in \overline{1,n} \rightarrow P, \right. \quad (3)$$

де w_i – значимі слова, кожне з яких є елементом множини слів тільки однієї леми з I. Відзначимо, що визначення (3) окрім вилучення проблеми багатозначності слів, що, за різними джерелами, дає 10–15 % похибки статистичного аналізу тексту, також дозволяє зменшити статистичний вплив початкової частини кривої Парето за рахунок вилучення стоп-слів, що не входять до множини зафіксованих системою S (значимих) лем.

З урахуванням уведених вище понять сформулюємо тепер *гіпотезу 1* – розподіл Парето (2) справедливий не тільки для слів/словоформ/лем з певної предметної області, але й для множини виявлених зв'язків між ними $\Omega \subseteq I \times I$. Для експериментальної перевірки цієї гіпотези скористаємося трьома відомими англомовними текстами з відкритого джерела *Project Gutenberg* [4] та удосконаленими технологічними можливостями лінгвістичного пакета *DKPro Core* [5]. Для проведення експерименту було обрано англомовні (авторські) варіанти трьох текстів різного обсягу: «Аліса в країні див» (Л. Керол, уривок з 4204 слів), «Біле ікло» (Дж. Лондон, 48907 слів) та «Троє у човні без врахування собаки» (Дж. К. Джером, 67328 слів).

Розроблене в [5] програмне забезпечення на основі пакета *DKPro Core* визначило таку загальну кількість складних синтаксичних зв'язків між ідентифікованими у текстах лемами (вилучено артиклі, прийменники та службові слова): текст 1 – 2360, 2 – 25244, 3 – 33316. Експериментально отримані розподіли залежностей між лемами обраних трьох текстів наведені на рис. 2–4.

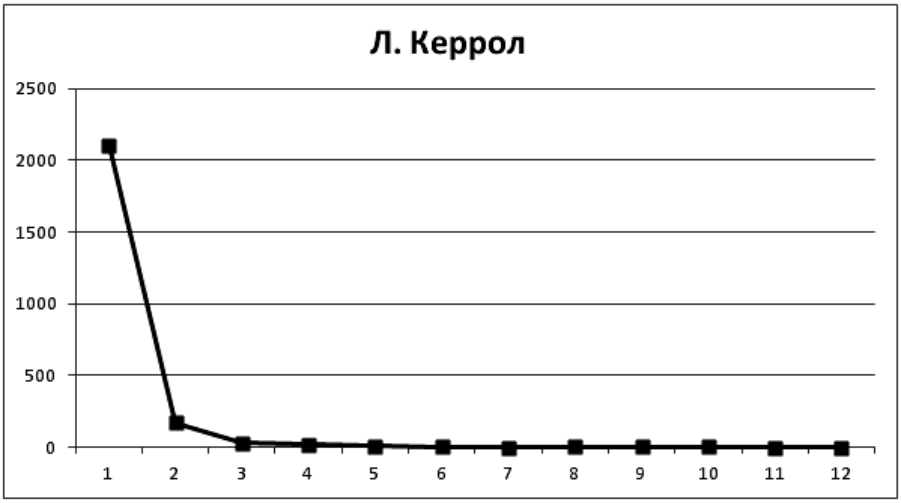

Рис. 2. Експериментальний розподіл 2360 залежностей між лемами для тексту 1

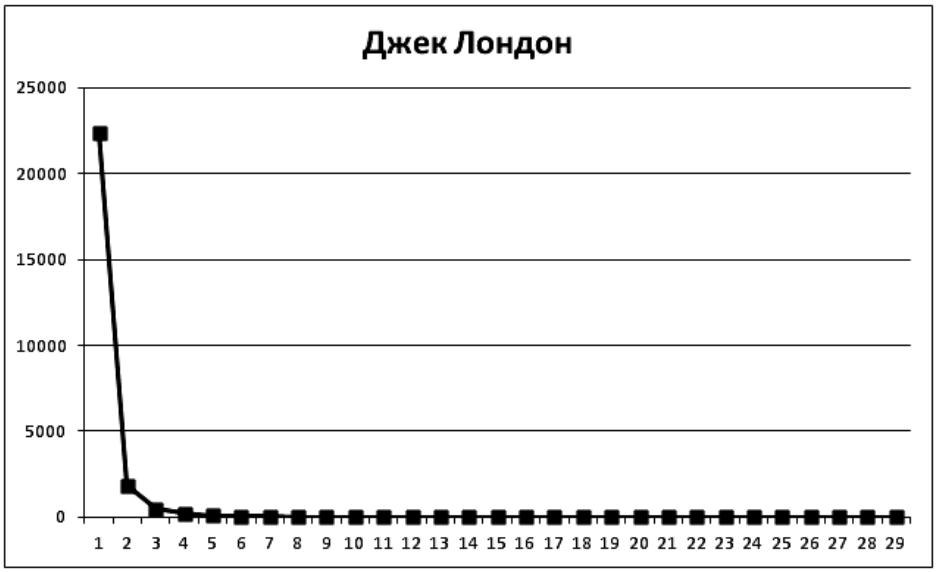

Рис. 3. Експериментальний розподіл 25244 залежностей між лемами для тексту 2

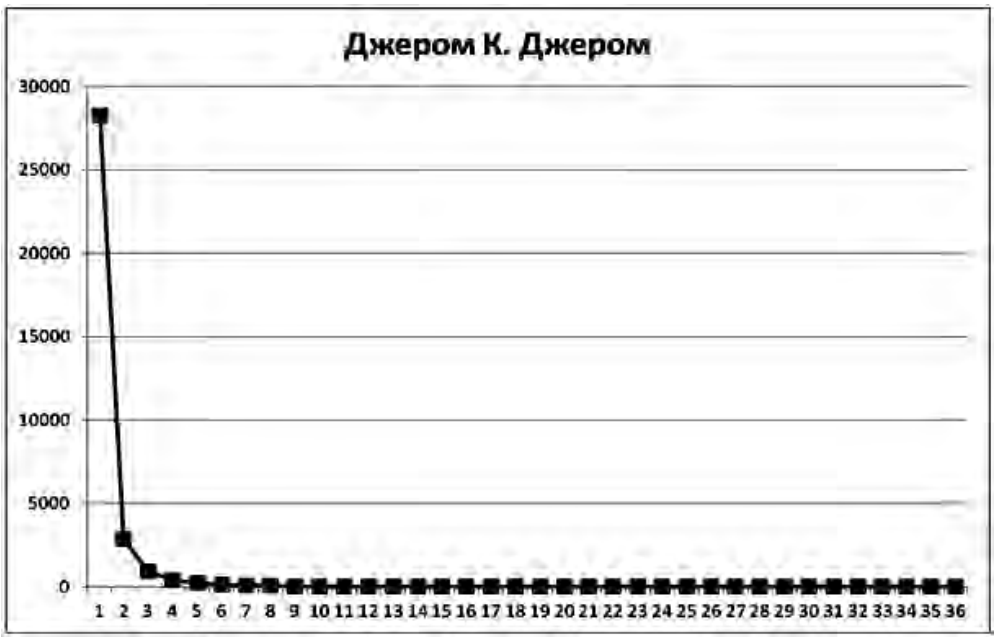

Рис. 4. Експериментальний розподіл 33316 залежностей між лемами для тексту 3

Візуальний аналіз всіх трьох рисунків свідчить про можливість прийняття у першому наближенні гіпотези 1 – вигляд гіперболічної кривої для різних за обсягом слів та залежностей між лемами текстів дуже схожий на розподіл Парето.

Статистична оцінка підходу

Задамо деяку множину складних залежностей між лемами $\Omega \subseteq I \times I$ на деякий момент часу такими параметрами: $k_{\lg}$ – кількість виявлених системою S зв'язків між *l*-ю та *g*-ю лемами з n відомих, m' – кількість ненульових елементів квадратної матриці, що відповідає відношенню Ω. Визначимо для дискретної випадкової величини $k_{\lg}$ статистичну оцінку математичного сподівання кількості повторень одного зв'язку. У цьому випадку статистична оцінка невідомого параметра $M[\omega]$, $\omega \in \Omega$ для теоретичного розподілу (генеральної сукупності) практично може бути визначена як

$$M[\omega] \to \lambda = k_{\Sigma} / m', \tag{4}$$

де

$$k_{\Sigma} = \sum_{l=1}^{n} \sum_{g=1}^{n} k_{\lg} .$$

Розглянемо оцінку достовірності запропонованого підходу з урахуванням *гіпотези 1* щодо справедливості розподілу Парето (2) не тільки для лінгвістичних понять (лем), але й для множини зв'язків між ними $\Omega \subseteq I \times I$. Відомі точкові статистичні оцінки параметрів генеральної сукупності за розподілом Парето показують, що математичне сподівання $M[\omega] = \frac{k^p x_m}{k-1}$ для великих значень k^p наближається до 1, отже, з урахуванням (4), $k_{\Sigma} \to m$. Такий висновок підсилює відома залежність для медіани $Me[\omega] = x_m \sqrt[k^p]{2}$, яка для великих значень k^p наближається до 0. Все це дозволяє висунути *гіпотезу 2* – досягнуте внаслідок запропонованого підходу переміщення найбільш вагомих лем та їх зв'язків у початковий пік кривої Парето істотно підвищує якість пошуку ключових слів для великих текстів.

Виконаємо формальну оцінку отриманого висновку через інтервальні статистичні оцінки параметрів. Визначимо та проаналізуємо довірчий інтервал для невідомого математичного

сподівання генеральної сукупності $M[\omega]$ при відомому середньоквадратичному відхиленні $\sigma = \frac{x_m}{k^p - 1} \cdot \sqrt[2]{\frac{k^p}{k^p - 2}}$ та статистичній оцінці λ (4). Будемо вимагати, щоб виконувалась рівність

$$P(|\lambda - M[\omega]| < \delta) = \gamma,$$

де γ – задана надійність, а δ – точність оцінки.

Як відомо $P(|X - m| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right)$, де Φ – функція Лапласа [6]. Підставивши відомі значення, отримуємо

$$P(|\lambda - m| < \delta) = 2\Phi\left(\frac{\delta}{x_m} \frac{k^p - 1}{\sqrt[2]{\frac{k^p}{k^p - 2}}} \cdot\right) = 2\Phi(t),$$

де $\Phi(t) = \frac{\gamma}{2}$, а $t \approx \delta \cdot (k^p - 1)$ для $x_m = 1$. Тоді $\delta \approx \frac{t}{k^p - 1}$ та

$$P(|\lambda - m| < \frac{t}{k^p - 1}) = 2\Phi(t). \quad (5)$$

За таблицею функції Лапласа маємо знайти аргумент *t*, якому відповідає значення функції Лапласа, дорівнює $\Phi(t) = \frac{\gamma}{2}$. При цьому залежність (5) показує, що зростання найбільшого значення частоти k^p призводить до практично лінійного збільшення точності оцінки, а тому порівняно малі обсяги виборки $|M'|$ не мають критичного значення.

Схожі висновки отримала Фрумкіна [7] на матеріалах експериментального дослідження індекса Хенлі та словника творів (мови) О. С. Пушкіна, а саме:

- для невеликих групп і малих рангових номерів закон Ципфа дає добре наближення, але для великих рангів і малих частот обсяг груп зростає до кількох тисяч слів;
- закон Ципфа не описує розподіл слів з малою ймовірністю.

Зі збільшенням довжини тексту збільшується кількість нових слів з малими ймовірностями, а апроксимація розподілу слів кривою Ципфа (Парето) погіршується.

Інформаційна оцінка підходу

Розглянемо інформаційні аспекти застосування складних залежностей між лінгвістичними поняттями у межах деякої інформаційної технології. Припустимо, що система *S* отримує на вході текст *T*, а на виході генерує множину з *l* ключових слів $W^k = \{w_1^k, ..., w_l^k\}$ цього тексту. Без применшення загальності будемо вважати, що текст *T* складається з *m* різних слів, а в окреме його *j*-те речення з *k* налічує *n* слів з *m* можливих, причому $m >> n$ та $m >> l$. Більшість відомих методів визначення ключових слів тексту беруть за основу частотний словник тексту, який фактично є списком або упорядкованою множиною пар $D = \{< w_i, \ f_i >\}, \ i = \overline{1, m}$, де w_i – одне слово з *m*, а f_i – його частота ($f_i \geq f_{i+1}, \ i = \overline{1, m-1}$), що визначена для *T*. За певною фільтрацією окремих незначущих категорій слів ключовими вважають перші *l* слів зі списку *D*, тобто, дещо спрощено маємо $W^k = \{w_1, ..., w_l\}$.

Результати парсерингу природних мов за допомогою сучасних лінгвістичних пакетів дозволяють на доступному програмному рівні [8] оперувати синтаксичними зв'язками між словами окремого речення. Окрім того, можливості цих пакетів дають змогу істотно зменшити значення *m* шляхом об'єднання слів у словоформи, а останні – у леми та стеми. Отже, необхідно з'ясувати, які

формальні переваги для визначення $W^k = \{w_1^k, ..., w_l^k\}$ надасть нам програмно-лінгвістичне забезпечення процедури синтаксичного аналізу всіх речень тексту *T*.

З інформаційного погляду розуміння сенсу речення окремим суб'єктом, зокрема системою S, супроводжується розпізнаванням а) окремих слів, з яких воно складається, та б) зв'язків між парами цих слів з відповідною побудовою дерева таких зв'язків [9]. Вважатимемо, що всі ці процеси відбуваються шляхом порівняльного аналізу та залучення інформації з деякої загальнолінгвістичної бази знань суб'єкта розуміння (S). Якщо кожен з цих етапів супроводжується збільшенням інформації, то приймаємо за робочу *гіпотезу 3*:

1) рівень загального розуміння тексту *T* може змінюватися від мінімально можливого до максимального залежно від обсягу та інших параметрів загальнолінгвістичної бази знань суб'єкта;

2) якість визначення $W^k = \{w_1^k, ..., w_l^k\}$ пропорційна до рівня загального розуміння тексту, що має підтверджуватися формальними ознаками.

Нехай будь-яке *j*-те речення з *k* складається з *n* різних слів, що не є доволі жорстким обмеженням. Тоді зв'язне дерево парних залежностей такого речення налічує або *n–1* гілок, якщо не брати до уваги зворотну залежність між підметом та присудком, або *n* – якщо брати. Оскільки кожному зв'язку (гілці дерева) відповідає два слова, 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 нетермінальних (кінцевих) вузлів дерева частоти відповідних слів не змінюються, а для термінальних (проміжних) можуть істотно зрости. У таблиці показані випадки зміни частот слів з урахуванням парних залежностей для різних типів речення.

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

№ з/п	Склад речення / кількість слів	Тип речення та граф його дерева залежностей	Кількість зв'язків / частотна формула	Кінцева частота
1	Ab / 2	Словосполучення (Бурштинова кімната)	1 / A+b	2
2	Abc / 3	Лінійна трійка (Місяць [на] небі сходить)	2 / A+b+2c	4
3	Abcd / 4	Лінійна четвірка (Отримав переклад слова дивного)	3 / A+2b+2c+d	6
4	Abcde / 5	Розгалуження (Густий ліс нізвідки завершився проваллям)	4 / A+2b+c+3d+e	8
5	Abcdef / 6	Група підмета (Сині примружені очі коханого говорили багато)	5 / A+b+4c+d+2e+f	10
6	Abcdef / 6	Група присудка (Голодний звір здобич миттєво відчує нюхом)	5 / A+2b+c+d+4e+f	10
7	Abcdef / 6	Обидві групи (Старий дід Еол зобрав всіх вітрів)	5 / A+3b+c+2d+e+2f	10

Навіть такий елементарний аналіз на рівні одного речення показує, що збільшуються частоти саме тих слів (позначені у таблиці підкресленням), які потенційно можуть належати до множини ключових. Виконаємо формальну оцінку такого збільшення для накладених обмежень щодо наявності винятково різних слів у реченні та неврахуванням зворотної залежності між підметом та присудком:

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

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

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

3) у загальному та реальнішому випадках $f_i = z \mid z \le k$, тобто коли *i*-те слово знаходиться у *z* реченнях з *k* маємо

$$f_i^{new} = z + \sum_{j=1}^{z}(n_j - 2) = \sum_{j=1}^{z}(n_j - 1) \quad (6)$$

як оцінку згори збільшення частоти *i*-го ключового слова. Отримана залежність (6) демонструє формальний вплив підходу до врахування складних залежностей між лінгвістичними поняттями на якість визначення ключових слів тексту.

Аналіз результатів наукового дослідження

Викладене обґрунтування інформативності статистичних ознак та параметрів для складних залежностей між лінгвістичними поняттями тексту має декілька обмежень. І якщо визначення для ієрархічної системи понять слово-словоформа-лема та, власне, лінгвістичної системи не суперечать традиційним поглядам у лінгвістиці, то введене поняття предметної області потрібно вважати істотним обмеженням запропонованого підходу. На практиці можна обрати два підходи до реалізації цього обмеження – корегування остаточних результатів людиною-експертом, що зазвичай використовується для доопрацювання розмічених текстів у корпусній лінгвістиці або застосування спеціалізованих програмних засобів, наприклад, у межах того самого DKPro Core. Потрібно розуміти, що перший випадок є ресурсоємним з погляду експертного часу, проте у другому досягти повної безпомилковості найближчим часом навряд чи вдасться.

Зауважимо, що експериментальне дослідження англомовних текстів для *гіпотези 1* було проведено через стандартні бібліотеки DKPro Core, а тому в результатах варто очікувати певний відсоток похибки від помилкового включення слова до омонімічної леми. Хоча така похибка, найшвидше, впливає не на сам закон розподілу Парето, а тільки на значення його параметрів, проте існує необхідність статистичного підтвердження *гіпотези 1*.

Окремих коментарів потребують виявлені пакетом DKPro Core чисельні параметри трьох обраних текстів. Якщо порівняно велика кількість артиклів та службових слів в англійській мові пояснює приблизно в два рази меншу кількість складних залежностей між лемами від кількості слів тексту, то міжтекстові відмінності такої частки наводять на роздуми. Експериментально отримані значення частки – текст 1 (1.78), текст 2 (1.94), текст 3 (2.02) – можливо, є ознакою авторського стилю, причому значні відмінності у статистці тексту 1 показують, що ми маємо справу з уривком тексту. Відсортований список залежностей між лемами показує значну кількість з них таких, що містять займенники, – це потребує додаткового аналізу та теж може бути ознакою стилю.

Висновки і перспективи подальших наукових розвідок

У роботі отримано формальні обмеження складових запропонованого підходу у вигляді статистичних оцінок колекції текстових документів певної предметної області – математичного сподівання кількості повторень зв'язку $\omega \in \Omega$, а також довірчих інтервалів для невідомого математичного сподівання генеральної сукупності $M[\omega]$ та оцінки згори збільшення частоти ключових слів. При цьому формально введене поняття предметної області застосовано як глобальне обмеження підходу, що дозволяє звести до практично неістотного рівня проблему багатозначності ключових слів тексту.

З огляду на важливість розв’язання задачі пошуку ключових слів потребують масштабного у розрізі обсягів тексту та предметних областей експериментального підтвердження *гіпотези 2* та *3*.

1. Квєтний Р.Н. Морфологічний аналіз слова на основі асоціативно-статистичного підходу / Р.Н. Квєтний, О.В. Бісікало, І.А. Кравчук // Вісник ЧДТУ. – 2010. – № 3. – С. 132–135. 2. Kechedzhy K. E. Rank distributions of words in additive many-step Markov chains and the Zipf law [Електронний ресурс] / K. E. Kechedzhy, O. V. Usatenko, V. A. Yampol’skii. – Режим доступу: http://arxiv.org/pdf/physics/0406099.pdf. 3. Каніщева О. В. Використання карт відношень (TRM) для автоматичного реферування / О. В. Каніщева // Вісник Нац. ун-ту "Львівська політехніка". – 2013. – № 770: Інформаційні системи та мережі. – С. 108–122. 4. Free ebooks – Project Gutenberg [Електронний ресурс] / Project Gutenberg Literary Archive Foundation. – Режим доступу: https://www.gutenberg.org/ 5. Бісікало О. В. Метод вилучення образних знань з англомовного тексту на основі інструментальних засобів пакету DKPro Core / О. В. Бісікало, І. Гуревич // Контроль і управління в складних системах: XII міжнар. конф., 14–16 жовтня 2014 р.: тези доповідей. – Вінниця, 2014. – С. 51. 6. Бочаров П. П. Теория вероятностей. Математическая статистика / П. П. Бочаров, А. В. Печинкин. – М.: Наука, 1998. – 325 с. 7. Фрумкина Р. М. Статистические методы изучения лексики / Р. М. Фрумкина. – М.: Наука, 1964. – 115 с. 8. Kotsyba N. Overview of the Ukrainian language resources within the multilingual European MULTEXT-East project, v. 4 / Kotsyba N. // Вісник Нац. ун-ту "Львівська політехніка". – 2013. – № 770: Інформаційні системи та мережі. – С. 122–129. 9. Бісікало О. В., Яхимович О. В. Метод визначення ключових слів англомовного тексту на основі DKPro Core // Технологический аудит и резервы производства: Информационные технологии. – 2015. – Т.1, № 2(21). – С. 26–30.