ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

ПРОГРЕССИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

PROGRESSIVE INFORMATION TECHNOLOGIES

УДК 004.9

Бісікало О. В.[1], Висоцька В. А.[2]

[1]Д-р техн. наук, професор, декан факультету комп'ютерних систем і автоматики Вінницького національного технічного університету, Вінниця, Україна

[2]Канд. техн. наук, доцент кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка», Львів, Україна

ЗАСТОСУВАННЯ МЕТОДУ СИНТАКСИЧНОГО АНАЛІЗУ РЕЧЕНЬ ДЛЯ ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ УКРАЇНОМОВНОГО ТЕКСТУ

У статті подано застосування породжувальних граматик у лінгвістичному моделюванні. Опис моделювання синтаксису речення застосовують для автоматизації процесів аналізу та синтезу природномовних текстів. У статті показано особливості процесу синтезу речень різних мов із застосуванням породжувальних граматик. В роботі розглянуто вплив норм та правил мови на хід побудови граматик. Застосування породжувальних граматик має широкі можливості у розробленні та створенні автоматизованих систем опрацювання текстового контенту, для лінгвістичного забезпечення комп'ютерних лінгвістичних систем тощо. В природних мовах є ситуації, коли явища, залежні від контексту, описані як незалежні від контексту, тобто в термінах контекстно-вільних граматик. При цьому опис ускладнений через утворення нових категорій і правил. В статті подано особливості процесу введення нових обмежень на класи даних граматик через введення нових правил. При кількості символів в правій частині правил не меншій за ліву отримали нескорочені граматики. Потім при заміні лише одного символу отримали контекстно-залежні граматики. При наявності в лівій частині правила лише одного символу отримали контекстно-вільні граматики. Жодних наступних природних обмежень на ліві частини правил накласти вже не можна. Виходячи із важливості забезпечення автоматичного опрацювання текстового контенту в сучасних інформаційних засобах (наприклад, інформаційно-пошукових системах, системах машинного перекладу, семантичного, статистичного, оптичного та акустичного аналізу і синтезу мови, автоматизованого редагування, екстракції знань з текстового контенту, реферування та анотування текстового контенту, індексування текстового контенту, навчально-дидактичних, менеджменту лінгвістичних корпусів, інструментальні засоби укладання словників різних типів тощо), фахівці інтенсивно шукають нові моделі, способи їх опису та методи автоматичного опрацювання текстового контенту. Одним із таких способів є розроблення загальних принципів побудови лексикографічних систем синтаксичного типу та побудови за цими принципами зазначених систем опрацювання текстового контенту для конкретних мов. Будь-які засоби синтаксичного аналізу складаються з двох частин: бази знань про конкретну природну мову і алгоритму синтаксичного аналізу, тобто набору стандартних операторів опрацювання текстового контенту на основі цих знань. Джерелом граматичних знань є дані з морфологічного аналізу та різні заповнені таблиці понять та лінгвістичних одиниць. Вони є результатом емпіричного опрацювання текстового контенту на природній мові експертами з метою виділення основних закономірностей для синтаксичного аналізу.

Ключові слова: текст, україномовний, алгоритм, контент-моніторінг, ключові слова, лінгвістичний аналіз, синтаксичний аналіз, породжувальні граматики, структурна схема речення, інформаційна лінгвістична система.

НОМЕНКЛАТУРА

АОПМК – автоматичного опрацювання природномовного контенту;

ІТ – інформаційні технології;

ІС – інформаційна система;

$\widetilde{N}$ – іменна група;

$\widetilde{R}$ – дієслівна група;

$C = \{c_1, c_2 \ldots, c_{n_C}\}$ – множина комерційного контенту $c_r \in C$ при $r = \overline{1, n_C}$;

$T = \{t_1, t_2 \ldots, t_{n_T}\}$ – час $t_p \in T$ транзакції формування комерційного контенту при $p = \overline{1, n_T}$;

$U_K = \{U_{K1}, U_{K2}, U_{K3}, U_{K4}\}$ – множина критеріїв визначення ключових слів в контенті;

C_1 – відфільтрований комерційний контент;

C_2 – відформатований комерційний контент;

C_3 – комерційний контент з визначеною множиною ключових слів;

DOI 10.15588/1607-3274-2016-3-7

$\alpha_0 : (X, U_C, T) \to C_0$ – оператор створення комерційного контенту;

$\alpha_1 : (X, U_G, T) \to C_0$ – оператор збирання комерційного контенту;

$\alpha_2 : (C_0, T, U_B) \to C_1$ – оператор виявлення дублювання комерційного контенту;

$\alpha_3 : (C_1, U_{FR}, T) \to C_2$ – оператор форматування комерційного контенту;

$\alpha_4 : (C_2, U_K, T) \to C_3$ – оператор виявлення ключових слів комерційного контенту – відображення комерційного контенту в новий стан, який відрізняється від попереднього стану наявністю множини ключових слів, що загально описують його зміст;

$Noun \in U_{K1}$ – терми – іменників, словосполучень іменників або прикметника з іменником серед множини слів текстового контенту;

$Unicity$ – унікальності для термів;

$NumbSymb \in U_{K3}$ – кількість знаків без пробілів для $Noun \in U_{K1}$ при $Unicity \geq 80$;

$UseFrequency \in U_{K2}$ – частота появи ключових слів комерційного контенту. Для термів з $NumbSymb \leq 2000$ частота $UseFrequency$ є в межах $(6;8]$%, з $NumbSymb \geq 3000$ – $[2;4)$%, з $2000 > NumbSymb < 3000$ – $[4;6]$%;

$BUseFrequency$ – частота появи ключових слів на початку тексту;

$IUseFrequency$ – частота появи ключових слів в середині тексту;

$EUseFrequency$ – частота появи ключових слів в кінці тексту комерційного контенту;

$KeyWords \in U_{K4}$ – ключові слова.

ВСТУП

Побудова систем АОПМК та формалізація відповідних процесів лінгвістичного аналізу/синтезу вважається основною проблемою інтелектуалізації ІТ [1–2]. Стрімкий та бурхливий розвиток Інтернет та ІТ різко прискорив створення різноманітних інформаційних лінгвістичних ресурсів і активізував сучасні дослідження, спрямовані на розроблення та впровадження інформаційних лінгвістичних систем, математичних методів та програмного забезпечення АОПМК. Для автоматизації етапів аналізу/синтезу природно-мовних текстів створюють різні моделі процесів АОПМК, обґрунтовують ефективні алгоритми та структури подання природно-мовних масивів даних. Традиційно лінгвістичний аналіз масивів природно-мовних текстів подають як послідовність процесів морфологічного, синтаксичного та семантичного аналізу/синтезу. Для кожного процесу створені відповідні моделі, методи та алгоритми: орієнтовані на конкретні групи мов (морфолексичний аналіз); системні граматики Холідея, граматики Хомскі (N. Chomsky) [3–15], дерева підпорядкування та системи складових Гладкого [2], розширенні мережі переходів (синтаксис речення); класичні семантичні мережі та фреймові моделі Мінського (семантика тексту). Необхідність в автоматизації процесів АОПМК сприяла появі відповідних формальних та математичних лінгвістичних моделей і методів їх аналізу/синтезу. Активним є розвиток мовознавчих дисциплін для потреб галузі комп'ютерних наук та ІТ. Інтеграційні процеси в цій галузі наук сприяють активному залученню науковців в сфері досліджень АОПМК для розроблення та створення автоматизованих ІС опрацювання багатомовної текстової інформації.

Найбільш складні проблеми АОПМК зумовлені явищами полісемії, омонімії, ононімії тощо, які характеризують неоднозначність мови і ускладнюють процес виявлення коректного відображення семантично-синтаксичної структури тексту в формальне подання через логічну інтерпретацію. Це вирішують в межах семантичного аналізу. Але застосування ресурсооб'ємних продукційних правил логічно-семантичного аналізу ускладнює та уповільнює програми АОПМК. Під час розуміння тексту не часто застосовують логіку, в основному ж здійснюється асоціативний пошук семантичного концепту, що відповідає шуканому слову та є контексто-наближеним до власного оточення. Тому асоціативний пошук є перспективним методом інтерпретації природно-мовних масивів даних.

1 ПОСТАНОВКА ЗАДАЧІ

Для реалізації синтаксичного аналізу текстового контенту з метою знаходження ключових слів та зменшення етапів опрацювання тексту необхідно:

I) Відокремити в аналізованому термінальному ланцюжку (реченні українською мовою) дієслівну групу від іменної групи (ключовим словами можуть бути лише слова з іменної групи) – це відбувається за результатами стемінгу – аналіз закінчень та робота лише з тими словами, флексії яких відповідають прикметникам та іменникам (в українській мові дієслівну групу не входять прикметник та іменник);

II) В іменній групі після знаходження першої множини ключових слів (слів, які вживані в текст із певною частотою, в межах, заданої модератором, але ці слова можуть бути лише прикметник в називному відмінку чоловічого роду, іменник в називному відмінку або абревіатура) знаходять та аналізують сусідні слова знайдених ключовиків. При цьому шукаємо ключові словосполучення, тобто визначаємо терми $Noun \in U_{K1}$ як словосполучення іменників або прикметника з іменником серед множини слів текстового контенту, зокрема:

1. Якщо ключовим словом є прикметник (флексія слова ий – називний відмінок чоловічого роду). Тоді по тексту знаходяться всі слова, що вживані справа від цього прикметника в будь якому відмінку (пошук іде за основою цього прикметника) та будується для них частотний словник. Ті словосполучення, що вживанні більше за певний ліміт (але можуть бути вживані менше за самий прикметник) і є новими ключовими словами. Ліміт визначає модератор.

2. Якщо ключовим словом є іменник (флексія слова не ий), тоді аналізуються всі слова справа та злів від нього.

а. Спочатку перевіряються всі слова зліва від нього на наявність флексій ий. Будується також частотний слов-

ник. Визначається множина слів, які зустрічаються найчастіше за певний визначений модератором ліміт – це і є нові ключові сова.

b. Потім аналізуються всі слова справа – вони всі мають бути без флексії ий. Аналогічно будується частотний словник, за яким визначається множина ключових слів.

2 ОГЛЯД ЛІТЕРАТУРИ

Процес виведення термінального ланцюжка українською [1–2], в якій властивий вільний порядок слів у реченні, що, проте, не заперечує існування сталого порядку розміщення окремих мовних елементів [3–4]. Для простого повного речення з прямим порядком слів структурну схему вважатимемо фіксованою, основними синтаксичними категоріями такого речення будуть іменна та дієслівна групи [5–7]. Необмежена граматика, побудована на тих же засадах, що і у попередніх прикладах, не матиме застосування через свою складність [8–10]. Для утворення контекстно-залежної граматики введемо певні обмеження, перш за все, на структуру речення [11]. Спираючись на правила побудови речень української мови з прямим порядком слів (наприклад, прикметник стоїть у препозиції до іменника, елементи іменникової групи групуються навколо іменника тощо) [11–13], розглянемо іменну групу $\widetilde{N}$ такої структурної схеми $\widetilde{N} = \{AN\}$ або $\widetilde{N} = N^p$. Прикметник та іменник в іменній групі узгоджуються між собою за відмінком, числом та родом [14–15]. Ці граматичні категорії є також граматичними категоріями займенника. Розглядатимемо **дієслівну групу** $\widetilde{R}$ такої структурної схеми: $\widetilde{R} = R\widetilde{N}$ або $\widetilde{R} = \widetilde{N}R$. З огляду на граматичні характеристики дієслова в українській мові, узгодження між іменною та дієслівною групою відбувається за числом, родом та особою (табл. 1–2).

Розглядатимемо речення з іменною групою в третій особі і дієслівною групою в теперішньому часі. Скороченим позначеннями іменної групи є $\widetilde{N}_{РД,ЧЛ,ВД,ОС}$, а її складових – $A_{РД,ЧЛ,ВД}$, $N_{РД,ЧЛ,ВД,ОС}$, $N^{займ}_{РД,ЧЛ,ВД,ОС}$. За потреби наголосити на використанні різних значень граматичних категорій використаємо такі позначки: дві іменні групи з різними значеннями категорії, наприклад, роду, позначатимемо так: $\widetilde{N}_{РД',ЧЛ,ВД,ОС}$, $\widetilde{N}_{РД'',ЧЛ,ВД,ОС}$. Скороченим позначеннями дієслівної групи є $\widetilde{R}_{РД,ЧЛ,ЧС,ОС}$, дієслова – $R_{РД,ЧЛ,ЧС,ОС}$. Реалізація норм та правил української мови впливає на подання перетворень. Наприклад, відомо, що найбільш часто іменна група виражається іменником або займенником у називному відмінку, а форми дієслова у теперішньому часі для всіх родів однини співпадають (*він/вона/воно летить)*, і врахування таких закономірностей відповідно відображається у позначеннях іменної та дієслівної груп – $\widetilde{N}_{РД,Ч,н,ОС}$ і $\widetilde{R}_{ЧЛ,тп,ОС}$. Спосіб подання контекстно-залежної граматики, що виводить речення введеної структурної схеми (з урахуванням певних закономірностей української мови) приведемо на прикладі речення *У своїй найбільш важливій роботі він показує барвистий світ українського села в його неповторній привабливості*. Розглянемо граматику G_3=(V, T, S, P). Алфавіт (позначення синтаксичних категорій подамо без індексів – для зручності) V=(S, $\widetilde{N}$, $\widetilde{R}$, A, N, R, E, $N^{займ}$, #, *у*, *свій*, *найбільш*, *важливий*, *робота*, *він*, *показувати*, *барвистий*, *світ*, *український*, *село*, *в*, *неповторний*, *привабливість*), T=(#, *у*, *свій*, *найбільш*, *важливий*, *робота*, *він*, *показувати*, *барвистий*, *світ*, *український*, *село*, *в*, *неповторний*, *привабливість*), # – символ межі речення, S – початковий символ. Кожен крок виведення полягає в розгортанні одного з символів попереднього ланцюжка (так, при переході від ланцюжка 2 до ланцюжка 3 символ $\widetilde{R}_{од,тп,3}$ розгортається в три символи – $R_{од,тп,3\ ч,од,з,1\ с,од,о,3}$) або в заміні його іншим (наприклад, при переході від ланцюжка 10 до ланцюжка 11 символ $\widetilde{N}_{ч,од,з,1}$ замінюється на $N^{займ}_{ч,од,з,1}$), інші ж символи переписуються без зміни. Проміжний ланцюжок містить рівно один допоміжний символ на останньому місці, тобто речення породжується зліва направо. Регулярна граматика ніби передбачає, що може слідувати за вже виданою словоформою, причому глибина передбачення – один сусідній символ; кожен черговий вибір повністю обумовлюється лише одним попереднім вибором [12]. Із виведення речення в регулярній граматиці неможливо отримати природне подання структури безпосередніх складових цього речення порівняно в контекстно-залежній та контекстно-вільній граматиках. Регулярні граматики дають деяку структуру складових, як і взагалі всі граматики безпосередніх складових, однак, ці складові зазвичай носять формальний характер [14].

Таблиця 1 – Позначення граматичних категорій іменної групи в українській мові

Тип	Опис
Іменна група/ $\widetilde{N}$	прикметник/*A*, іменник/*N*, займенник/$N^{займ}$;
Число/*ЧЛ*	однина/*од*, множина/*мн*;
Рід/*РД*	чоловічий/*ч*, жіночий/*ж*, середній/*с*;
Відмінок/*ВД*	називний/*н*, родовий/*р*, давальний/*д*, знахідний/*з*, орудний/*о*, місцевий/*м*, кличний/к;
Особа/ОС	1-ша/*1*, 2-га/*2*, 3-тя/*3*

Таблиця 2 – Позначення граматичних категорій дієслівної групи в українській мові

Тип	Опис
Дієслівна група/ $\widetilde{R}$	дієслово/*R*, в межах іменної групи прикметник/*A*, іменник/*N*;
Число/*ЧЛ*	однина/*од*, множина/*мн*;
Рід/*РД*	чоловічий/*ч*, жіночий/*ж*, середній/*с*;
Особа/*ОС*	1-ша/*1*, 2-га/*2*, 3-тя/*3*;
Час/*ЧС*	теперішній/*тп*, минулий/*мн*, майбутній/*мб*

3 МАТЕРІАЛИ І МЕТОДИ

Виявлення ключових слів тематики контенту з фрагменту тексту забезпечимо за допомогою процесів, поданих на рис. 2.

Текст реалізує структурно подану діяльність, що передбачає суб'єкт і об'єкт, процес, мету, засоби і результат, які відображаються в змістовно-структурних, функціональних, комунікативних показниках. Одиницями внут-

рішньої організації структури тексту є алфавіт, лексика (парадигматика), граматика (синтагматика), парадигми, парадигматичні відношення, синтагматичні відношення, правила ідентифікації, висловлювання, між фразова єдність та фрагменти-блоки. На композиційному рівні виділяють речення, абзаци, параграфи, розділи, глави, підглави, сторінки тощо (речення, побічно пов'язані з внутрішньою структурою, не розглядаються – рис. 3). За допомогою бази даних (бази термінів/морфем і службових частин мови) та визначених правил аналізу тексту виконують пошук терміну (рис. 4а) на інформаційному ресурсі (рис. 4б).

Розглянемо синтаксичні аналізатори, що працюють у два етапи: ідентифікують змістовні лексеми та створюють дерево розбору (алг. 1).

Алгоритм 1. Синтаксичний аналізатор текстового контенту.

Етап 1. Ідентифікація змістовних лексем $U_{K1} \in U_K$ для комерційного контенту C_2.

Крок 1. Визначення ланцюжка термів у вигляді речення.

Крок 2. Ідентифікація іменної групи за допомогою словника основ.

Крок 3. Ідентифікація дієслівної групи за допомогою словника основ.

Етап 2. Створення дерева розбору зліва направо. Виведення дерева полягає в розгортанні одного з символів попереднього ланцюжка послідовності лінгвістичних змінних, або в заміні його іншим, інші ж символи переписуються без зміни. При розгортанні, замінювані/переписувані символи (предки) з'єднують безпосередньо з символами, які виходять в результаті розгортання, заміни або переписування (нащадками), та отримують дерево складових, або синтаксичну структуру для змісту комерційного контенту.

Крок 1. Розгортання іменної групи. Розгортання дієслівної групи.

Крок 2. Реалізація синтаксичних категорій словоформами.

Етап 3. Визначення множини ключових слів $\alpha_4 : (C_2, U_K, T) \to C_3$ для контенту C_2.

Крок 1. Визначення термів $Noun \in U_{K1}$ – іменників, словосполучень іменників або прикметника з іменником серед множини слів текстового контенту.

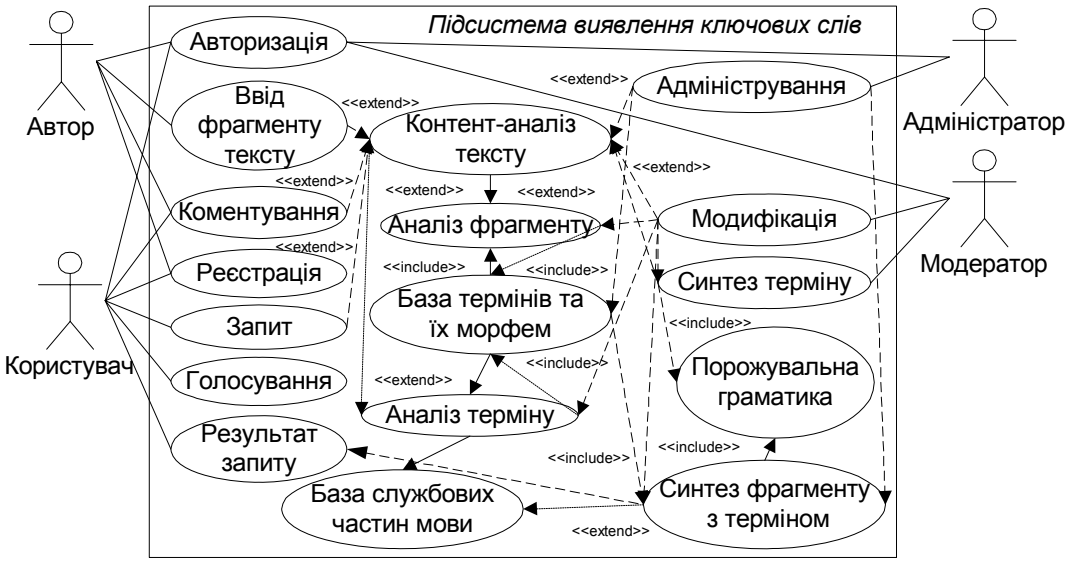


Рисунок 2 – Діаграма варіантів використання для виявлення ключових слів тематики контенту

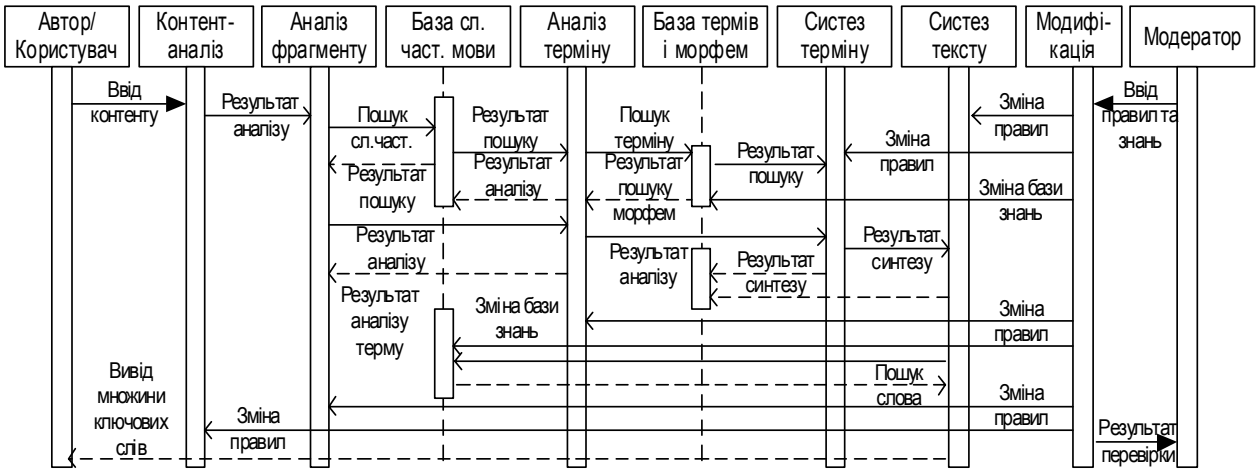
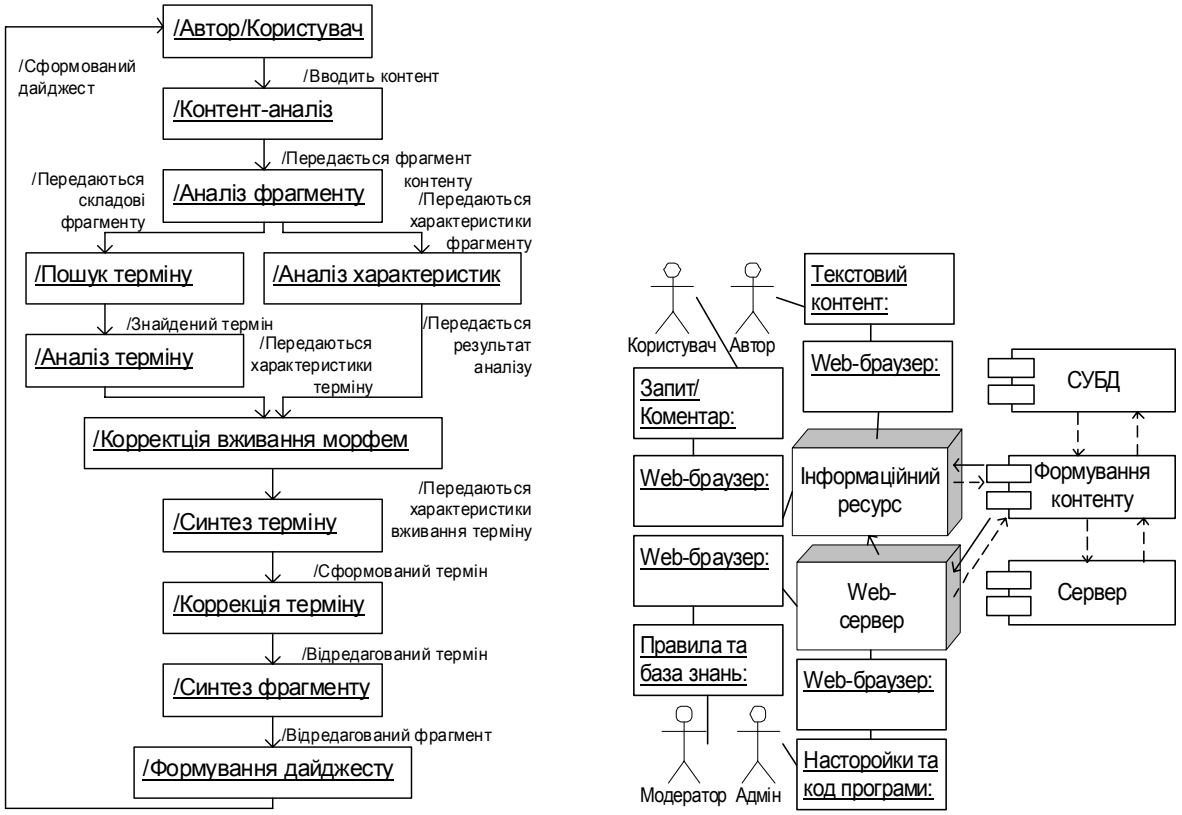


Рисунок 3 – Діаграма послідовності для процесу виявлення ключових слів тематики контенту



а б

Рисунок 4 – Діаграми: а – кооперації, б – компонентів виявлення ключових слів тематики контенту

Крок 2. Розрахунок унікальності *Unicity* для термів $Noun \in U_{K1}$.

Крок 3. Розрахунок $NumbSymb \in U_{K3}$ для $Noun \in U_{K1}$ при $Unicity \geq 80$.

Крок 4. Розрахунок $UseFrequency \in U_{K2}$ – частоти появи ключових слів контенту. Для термів з $NumbSymb \leq 2000$ частота *UseFrequency* є в межах $(6;8]$%, з $NumbSymb \geq 3000$ – $[2;4)$%, з $2000 > NumbSymb < 3000$ – $[4;6]$%.

Крок 5. Розрахунок *BUseFrequency* – частота появи ключових слів на початку тексту, *IUseFrequency* – частота появи ключових слів в середині тексту, *EUseFrequency* – частота появи ключових слів в кінці тексту контенту.

Крок 6. Порівняння значень *BUseFrequency*, *IUseFrequency* та *EUseFrequency* для розстановки пріоритетів. Ключові слова з більшими значеннями *BUseFrequency* мають більший пріоритет, ніж ключові слова з більшим значенням *EUseFrequency*.

Крок 7. Сортування ключових слів згідної їх пріоритетів.

Етап 4. Заповнення бази пошукових образів контенту C_3, тобто атрибутів $KeyWords \in U_{K4}$ – ключові слова, *Unicity* – унікальність ключових слів ≥ 80, *Noun* – терм, *NumbSymb* – кількість знаків без пробілів, *UseFrequency* – частота вживання ключових слів, *BUseFrequency* – частота вживання ключових слів на початку тексту, *IUseFrequency* – частота вживання ключових слів в середині тексту, *EUseFrequency* – частота вживання ключових слів в кінці тексту. Спираючись на правила породжувальної граматики виконується корекція терміну згідно правил його вживання у контексті (рис. 5).

Речення задають межі дії знаків пунктуації, анафоричних і катафоричних посилань. Семантика тексту зумовлена комунікативним завданням передавання інформації. Структура тексту визначається внутрішньою організацією одиниць тексту і закономірностями їх взаємозв'язку. Під час синтаксичного аналізу текст оформляють у структуру даних, наприклад, в дерево, яке відповідає синтаксичній структурі вхідної послідовності, і найкраще підходить для подальшого опрацювання. Після аналізу фрагменту тексту і терміну синтезують новий термін як ключове слово тематики контенту, використовуючи базу термінів та їх морфем (рис. 5).

Далі синтезуємо терміни для формування нового ключового слова, використовуючи базу службових частин мови. Принцип виявлення ключових слів за змістом (термами) базується на законі Зіпфа і зводиться до вибору слів із середньою частотою появи (найбільш вживанні слова ігнорують через «стоп-словники», а рідкісні слова тексту не враховують).

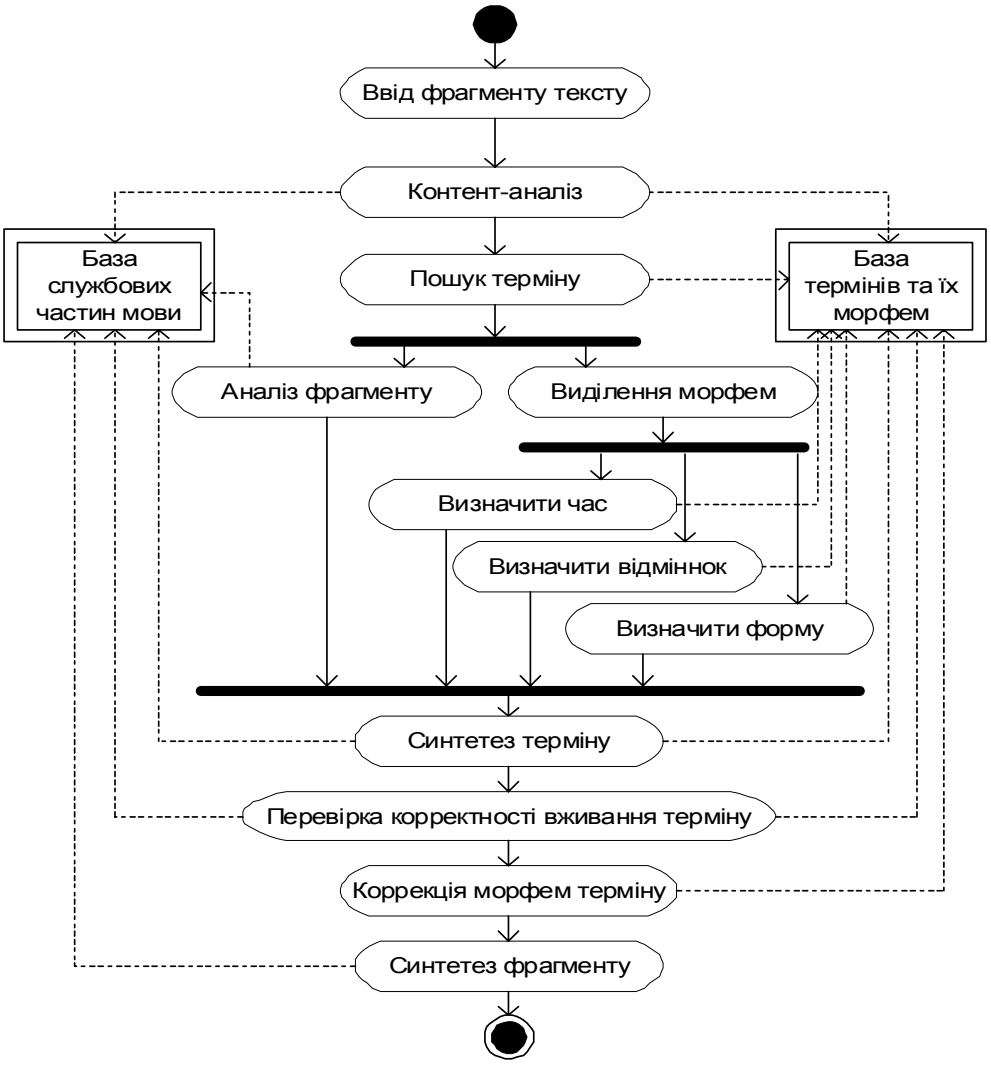


Рисунок 5 – Діаграма діяльності для процесу виявлення ключових слів тематики контенту

4 ЕКСПЕРИМЕНТИ

Лінгвістичною базою для експериментального дослідження обрано 100 наукових публікацій Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (http://science.lp.edu.ua/sisn), № 783 (http://science.lp.edu.ua/SISN/SISN-2014) та № 805 (http://science.lp.edu.ua/sisn/vol-cur-805-2014-2). Аналіз статистики функціонування системи виявлення множини ключових слів із 100 наукових статей було проведено у два етапи, зокрема:

1. Проаналізувати всі статті із перевіркою загальних заблокованих слів та тематичного словника.

2. Проаналізувати всі статті із перевіркою уточнених заблокованих слів та уточненого тематичного словника (з більшою кількістю запуску системи формується множина невідомих слів (відсутніх і в тематичному словнику і в множині заблокованих).

Окрім того на кожному етапі перевірка відбувалась в два кроки для кожної статті: аналіз всієї статті (http://victana.lviv.ua/index.php/kliuchovi-slova) та аналіз статті без початку (назва, автори, УДК, анотації двома мовами, авторські ключові слова двома мовами, місце роботи авторів) і без списку літератури для того, щоб визначити похибки точності формування множини ключових слів.

5 РЕЗУЛЬТАТИ

Аналіз статистики здійснювався за принципом порівняння множини авторських ключових слів (визначені та прописані в статті самими авторами цих робіт), множини ключових слів визначених за першим та другим етапами з різними вагами слів (але більше, за визначене в опції **Мін.вага слова, %* в межах [1,5]) з повними та скороченими текстами робіт (табл. 3) при середньому арифметичному значенні авторських ключових словосполучень / слів біля 5 (4,77), які в середньому утворені з 10 (9,82) слів. Вага слова розраховується як відносна частота появи основи цього слова у всьому тексті. В табл. 4 присутні такі позначення, як *A* (всього ключових слів, визначених системою при заданій вазі слова), *B* (змістовних слів зі списку утворених, тобто без невідомих абревіатур, дієслів, службових слів тощо), *C* (збіг слів з визначеними автором статті), *D* (точність збігу знайдених ключовиків з авторським ключовими словами), *E* (додаткові ключові слова, визначені системою, але не визначені автором статті).

Таблиця 3 – Статистичні дані досліджених обсягів текстів статей

Назва обсягу статті	Крок 1		Крок 2	
	Всього	Середнє арифметичне	Всього	Середнє арифметичне
Сторінок	956	9,56	828	8,28
Абзаців	16497	164,97	15263	152,63
Рядків	42553	425,53	36965	369,65
Слів	345580	3455,8	291247	2912,47
Знаків	2327209	23272,09	1974773	19747,73
Знаків та пробілів	2674889	26748,89	2265917	22659,17

Таблиця 4 – Статистичні дані досліджених змісту текстів статей

Назва	Вага слова	Етап 1					Етап 2				
		A	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
Крок 1	≥ 1	5,46	3,92	2,51	2,08	1,74	7,43	7,03	3,27	3	4,18
	≥ 2	1,08	0,88	0,63	0,59	0,26	2,67	2,64	1,65	1,54	1,12
	≥ 3	0,41	0,38	0,22	0,21	0,16	1,21	1,2	0,85	0,79	0,41
	≥ 4	0,15	0,13	0,09	0,09	0,04	0,46	0,45	0,33	0,31	0,15
	≥ 5	0	0	0	0	0	0	0	0	0	0
Крок 2	≥ 1	6,51	5,02	2,68	2,23	2,37	8,35	7,78	3,25	2,91	4,99
	≥ 2	1,34	1,11	0,74	0,72	0,39	3,12	3,07	1,81	1,67	1,43
	≥ 3	0,51	0,45	0,29	0,27	0,17	1,42	1,4	0,93	0,85	0,54
	≥ 4	0,19	0,17	0,12	0,12	0,05	0,73	0,72	0,45	0,42	0,31
	≥ 5	0,11	0,1	0,06	0,06	0,04	0,33	0,32	0,25	0,23	0,1

6 ОБГОВОРЕННЯ

В Інтернет-просторі зазвичай присутні інформаційні SEO-ресурси, які визначають ключові слова в межах $[100 \div 1000]$ слів в тексті, наприклад:

– http://msurf.ru/tools/keygeneratortext/;
– http://syn1.ru/tools/keygeneratortext/;
– http://webmasta.org/tools/keygeneratorurl/;
– http://labs.translated.net/terminology-extraction/;
– http://www.keywordstext.therealist.ru/.

Недолік таких SEO-ресурсів – неточність та некоректність опрацювання україномовних текстів при відсутності грамотно побудованих морфологічних словників, словників основ та заблокованих слів. Також основним недоліком більшості таких SEO-ресурсів обмеженість опрацювання обсягів текстових масивів даних. Для прикладу синтаксично проаналізована рядом SEO-ресурсів ця україномовна стаття, яка має понад 800 слів в тесті Частина перерахованих вище SEO-ресурсів не опрацьовує або некоректно опрацьовує такий великий обсяг інформації (рис. 6–7).

Одним із найкращих SEO-ресурсів є http://advego.ru/text/seo/, який найкраще працює з україномовних текстами (рис. 8). Провадить семантичний аналіз тексту онлайн та SEO-аналіз тексту. Результат найбільш наближений до отриманого розробленою системою.

Але є недоліки. Не визначає множину ключових слів, а лише частоту вживання слів, словосполучень та частин слів (які необов'язково є частинами слова як основа). Взагалі не працює з основами слова. Для цього SEO-ресурсу слова ключових та ключові є різними.

Розроблений SEO-ресурс http://victana.lviv.ua/kliuchovi-slova працює з основами слова, орієнтований на україномовні, російськомовні, англомовні тексти, а також змішаного типу (рис. 9). На прикладі цієї статті SEO-ресурс визначив наступну множину ключових слів {слово, ключових, контент, аналіз, chomsky, система}.

Повторюваність слів, раз: слово – 120; ключових – 49; контент – 46; аналіз – 39; chomsky – 37; система – 37. Автори визначили такі ключові слова: текст, україномовний, алгоритм, контент-моніторінг, ключові слова, лінгвістичний аналіз, синтаксичний аналіз, породжувальні граматики, структурна схема речення, інформаційна лінгвістична система.. Автори зазвичай більше визначають ключових слів порівняно з реальною ситуацією згідно закономірностей розподілу частоти слів за законом Зіпфа (George KingsleyZipf). Автор наукової статті зазвичай обирає за своїм розсудом кількість ключових слів в діапазоні від 2 до 10 слів (найчастіше – 3–5 ключовиків). Система ж визначає різну кількість слів, в залежності від стиля написання конкретного автора (існують такі статті, в яких система не знаходить за законом Зіпфа жодного ключового слова). Збіг списків виявлених ключовиків з авторськими без врахування зайвих слів, визначених авторами (повторюваність > 30 для обсягу тексту понад 4800 слів), складає відповідно для таких SEO-ресурсів:

– http://syn1.ru/tools/keygeneratortext/ – приблизно 35%;
– http://labs.translated.net/terminology-extraction/ – приблизно 57%;
– http://advego.ru/text/seo/ – приблизно 83%;
– http://victana.lviv.ua/kliuchovi-slova – приблизно 90%.

На рис. 10 приведено діаграму аналізу статистики формування системою множин всіх потенційних ключових слів порівняно з множиною, визначеною авторами статей.

Перший стовпчик – середньоарифметична кількість ключових слів, визначених автором (4,77), а другий – середньоарифметична кількість слів, які складають ці авторські ключові слова (9,82). Третій стовпчик – середньоарифметична кількість потенційних ключових слів, визначена системно на етапі 1, крок 1(5,46); четвертий – на етапі 1, крок 2 (6,51); п'ятий – на етапі 1, крок 1 (7,43);

Информация о тексте:

Всего слов в тексте:	5072
Обработано слов (без повторов):	1073

Результат

КЛЮЧОВИХ, контенту, АНАЛ, Chomsky, ться, сть, речення, групи, комерц, етап, Ключов, його, або, менник, появи, без, досл, Systems

Слова списком подробнее [Скрыть]

Слово	Вхождений	Частота (TF)
КЛЮЧОВИХ	43	0.008
контенту	40	0.008
АНАЛ	40	0.008
Chomsky	37	0.007
ться	22	0.004
сть	18	0.004
речення	17	0.003
групи	15	0.003
комерц	15	0.003
етап	13	0.003
Ключов	12	0.002
його	12	0.002
або	11	0.002
менник	11	0.002
появи	10	0.002
без	9	0.002
досл	9	0.002
Systems	9	0.002

Рисунок 6 – Результат аналізу цієї статті на SEO-ресурсі http://syn1.ru/tools/keygeneratortext/

#	Extracted term	Score
1	текстового контенту	65%
2	ключових слів	65%
3	комерційного контенту	62%
4	обработки текстового контента	62%
5	опрацювання текстового контенту	62%
6	для	61%
7	частота появи ключових слів	60%
8	аналізу	56%
9	слова	56%
10	систем	55%
11	при	55%
12	іменної групи	55%
13	синтаксичного аналізу	55%
14	правил	54%
15	систем опрацювання текстового контенту	53%
16	автоматического обработки текстового контента	53%
17	прикметника з іменником серед	53%
18	іменником серед множини слів	53%
19	лише одного символу отримали	53%
20	або прикметника з іменником	53%

Рисунок 7 – Результат аналізу цієї статті на SEO-ресурсі http://labs.translated.net/terminology-extraction/

Статистика текста

Наименование показателя	Значение
Количество символов	35927
Количество символов без пробелов	31118
Количество слов	4354
Количество уникальных слов	1589
Количество значимых слов	2873
Количество стоп-слов	1013
Вода	34.0 %
Количество грамматических ошибок	460
Классическая тошнота документа	8.12
Академическая тошнота документа	4.9 %

Семантическое ядро

Фраза/слово	Количество	Частота, %
слів	66	1.52
контент	54	1.24
ключових	45	1.03
ключових слів	42	0.96 / 1.93
chomsky	37	0.85
текст	36	0.83
система	29	0.67
текстового контенту	24	0.55 / 1.10
текстовой	24	0.55
граматика	22	0.51
аналізу	21	0.48
крок	21	0.48
речення	18	0.41

Слова

Слово	Количество	Частота, %
слів	66	1.52
контент	54	1.24
ключових	45	1.03
chomsky	37	0.85
текст	36	0.83
система	29	0.67
текстовой	24	0.55
граматика	22	0.51
аналізу	21	0.48
крок	21	0.48
речення	18	0.41
chomsky	16	0.37
частота	16	0.37

Стоп-слова

Слово	Количество	Частота, %
в	85	1.95
тот	68	1.56
of	60	1.38
п	56	1.29
з	48	1.10
на	45	1.03
слово	40	0.92
the	35	0.80
для	31	0.71
р	29	0.67
і	29	0.67
and	27	0.62
у	26	0.60

Рисунок 8 – Результат аналізу цієї статті на SEO-ресурсі http://advego.ru/text/seo/

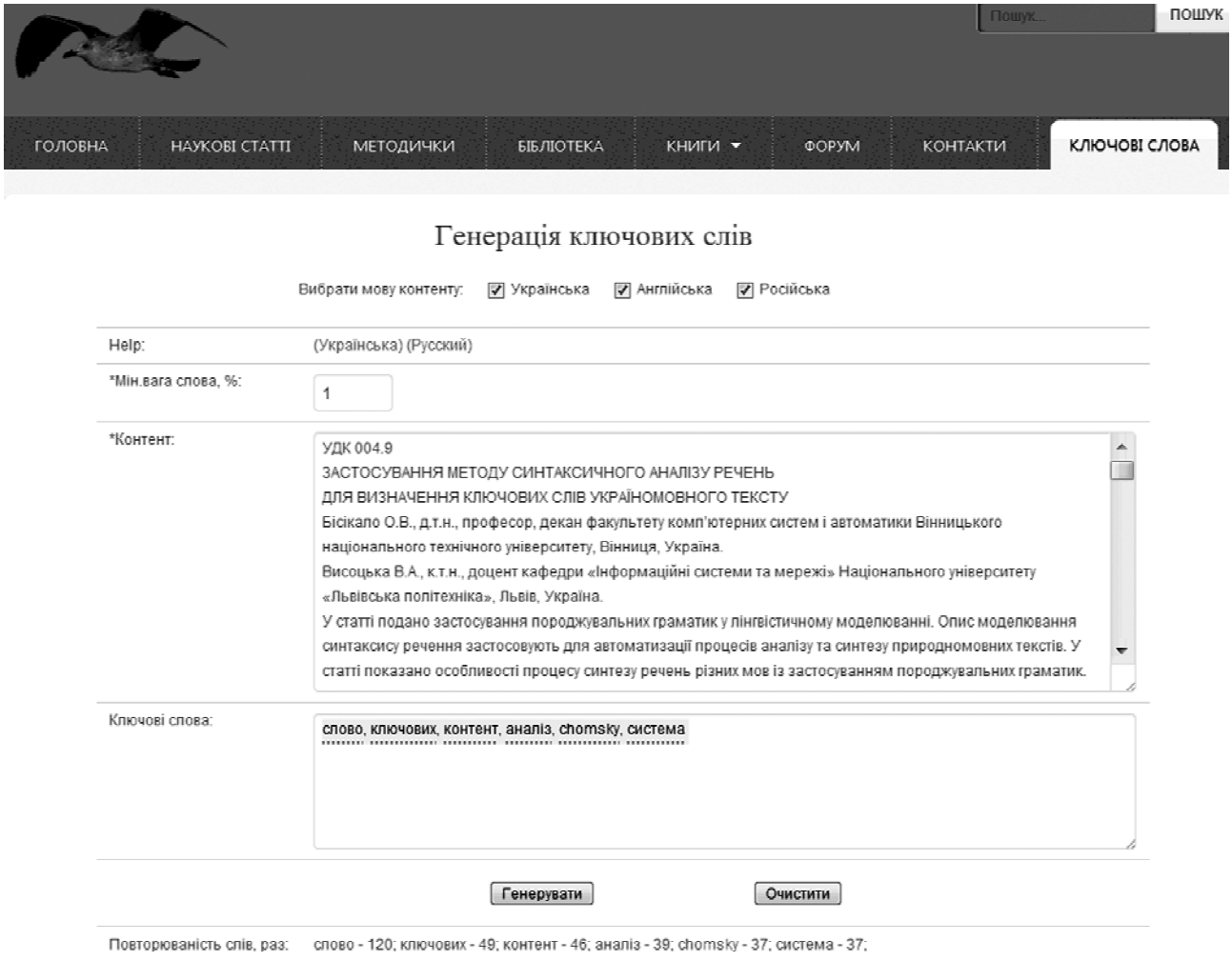


Рисунок 9 – Результат аналізу цієї статті на SEO-ресурсі http://victana.lviv.ua/kliuchovi-slova

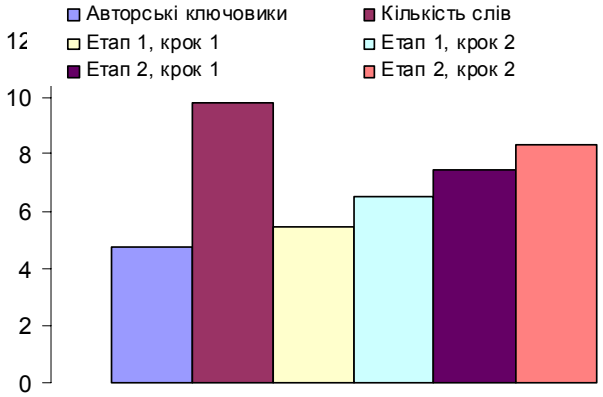


Рисунок 10 – Результати перевірки 100 статей

шостий – на етапі 2, крок 2 (8,35). Отже, автор статті в середньому зазвичай визначає більшу кількість слів (стовпчик 2) та меншу кількість ключових слів (стовпчик 1), ніж вона реально присутня в цій роботі.

ВИСНОВКИ

У статті розглянуто особливості методу синтаксичного аналізу україномовного текстового контенту, спрямованого на автоматичне виявлення значущих ключових слів вхідних текстів. Визначено роль і формальні ознаки синтаксичного аналізатора в процесі виявлення ключових слів тематики контенту, проведено декомпозицію процедур запропонованого методу на 4-х етапах. На відміну від відомих синтаксичних аналізаторів, запропонований метод забезпечує самовдосконалення та самонавчання автоматизованої системи визначення ключових слів за рахунок механізму ідентифікації значущих статистичних параметрів у визначених модератором межах. Експериментальне дослідження на матеріалах 100 наукових публікацій з двох номерів (783 та 805) Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (http://science.lp.edu.ua/sisn) підтвердило достовірність методу – для різних методик опрацювання первинного тексту середній збіг списків виявлених ключовиків з авторськими змінюється у проміжку 52,6–68,5%. Точність збігу ключових слів із авторськими коливається в проміжку 43,6–62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку 38,9–75,8% в залежності від етапів аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3–71,9% в залежності від етапів аналізу текстів статей. Потребує подальшого експериментального дослідження визначення ключових слів для інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо.

ПОДЯКИ

У статті розв'язана науково-практична задача автоматичного визначення ключових слів україномовного тексту в Інтернет-джерелах на основі синтаксичного аналізу речень відповідної текстової інформації. Роботу виконано в рамках спільних наукових досліджень кафедри інформаційних систем та мереж Національного університету «Львівська політехніка» на тему «Розроблення методів та засобів побудови інтелектуальних систем опрацювання інформаційних ресурсів з використанням онтологічного підходу», а також кафедри автоматики та інформаційно-вимірювальної техніки Вінницького національного технічного університету у межах діяльності науково-дослідного центру прикладної та комп'ютерної лінгвістики. Результати досліджень здійснювались у рамках держбюджетних науково-дослідних робіт за темами «Розробка методів, алгоритмів і програмних засобів моделювання, проектування та оптимізації інтелектуальних інформаційних систем на основі Web-технологій «ВЕБ» та «Інтелектуальна інформаційна технологія образного аналізу тексту та синтезу інтегрованої бази знань природно-мовного контенту».

СПИСОК ЛІТЕРАТУРИ

1. Берко, А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л. : НУЛП, 2009. – 612 с.
2. Математична лінгвістика / [В. Висоцька, В. Пасічник, Ю. Щербина, Т. Шестакевич]. – Л. : Новий Світ-2000, 2012. – 359 с.
3. Chomsky N. Three models for the description of language / N. Chomsky // I.R.E. Transactions on Information Theory. – 1956. – Vol. 2. – P. 113–124.
4. Chomsky N. On certain formal properties of grammars / N. Chomsky // Information and Control. – 1959. – Vol. 2. – P. 137–167.
5. Chomsky N. On the notion «Rule of Grammar» / N. Chomsky // Proceedings of the Twelfth Symposium in Applied Mathematics. – 1961. – P. 6–24.
6. Chomsky N. Context-free grammars and pushdown storage / N. Chomsky // Quarterly Progress Reports, Research Laboratory of Electronics, M.I.T. – 1962. – № 65. – P. 187–194.
7. Chomsky N. Formal properties of grammars / N. Chomsky // Handbook of Mathematical Psychology, New York : Wiley and Sons. – 1963. – Vol. 2. – P. 323–418.
8. Chomsky N. The logical basis for linguistic theory / N. Chomsky // Proc. IX-th Int. Cong. Linguists, 1962. – P. 91–111.
9. Chomsky N. Finite state languages / N. Chomsky, G. A. Miller // Information and Control. – 1958. – Vol. 1. – P. 91–112.
10. Chomsky N. Introduction to the formal analysis of natural languages / N. Chomsky, G. A. Miller // Handbook of Mathematical Psychology 2, Ch. 12, Wiley. – 1963. – Vol. 2. – P. 269–321.
11. Chomsky N. The algebraic theory of context-free languages / N. Chomsky, M. P. Schьtzenberger // Computer programming and formal systems, North-Holland. – 1963. – P. 118–162.
12. Chomsky N. Syntactic Structures / N. Chomsky. – Mouton, The Hague, 1957. – 117 p.
13. Chomsky N. Explanatory models in linguistics / N. Chomsky // Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress. Stanford University Press, Stanford, CA. – 1962. – P. 528–550.
14. Chomsky N. Aspects of the Theory of Syntax / N. Chomsky. – MIT Press, Cambridge, MA,1965. – 247 p.
15. Chomsky N. Conditions on transformations / N. Chomsky. – New York : Holt, Rinehart & Winston, 1973. – P. 232–286.

Стаття надійшла до редакції 17.02.2016.
Після доробки 29.02.2016.

Бисикало О. В.[1], Высоцкая В. А.[2]

[1]Д-р техн. наук, профессор, декан факультета компьютерных систем и автоматики Винницкого национального технического университета, Винница, Украина

[2]Канд. техн. наук, доцент кафедры «Информационные системы и сети» Национального университета «Львовская политехника», Львов, Украина

ПРИМЕНЕНИЕ МЕТОДА СИНТАКСИЧЕСКОГО АНАЛИЗА ПРЕДЛОЖЕНИЙ ДЛЯ ОПРЕДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ УКРАИНОЯЗЫЧНОГО ТЕКСТА

В статье представлены применения порождающих грамматик в лингвистическом моделировании. Описание моделирования синтаксиса предложения применяют для автоматизации процессов анализа и синтеза естественноязыковых текстов. В статье показаны особенности процесса синтеза предложений различных языков с применением порождающих грамматик. В работе рассмотрено влияние норм и правил языка на ход построения грамматик. Применение порождающих грамматик имеет широкие возможности в разработке и создании автоматизированных систем обработки текстового контента, для лингвистического обеспечения компьютерных лингвистических систем и тому подобное. В естественных языках есть ситуации, когда явления, зависящие от контекста, описаны как независимые от контекста, то есть в терминах контекстно-свободных грамматик. При этом описание затруднено из-за образования новых категорий и правил. В статье представлены особенности процесса введения новых ограничений на классы данных грамматик из-за введения новых правил. При количестве символов в правой части правил не меньшей левой получили несокращенные грамматики. Затем при замене только одного символа получили контекстно-зависимые грамматики. При наличии в левой части правила лишь одного символа получили контекстно-свободные грамматики. Никаких следующих природных ограничений на левые части правил наложить уже нельзя. Исходя из важности обеспечения автоматического обработки текстового контента в современных информационных средствах (например, информационно-поисковых системах, системах машинного перевода, семантического, статистического, оптического и акустического анализа и синтеза речи, автоматизированного редактирования, экстракции знаний текстового контента, реферирования и аннотирования текстового контента, индексирования текстового контента, учебно-дидактических, менеджмента лингвистических корпусов, инструментальные средства составления словарей различных типов и т.д.), специалисты интенсивно ищут новые модели, способы их описания и методы автоматического обработки текстового контента. Одним из таких способов является разработка общих принципов построения лексикографических систем синтаксического типа и построения по этим принципам указанных систем обработки текстового контента для конкретных языков. Любые средства синтаксического анализа состоят из двух частей: базы знаний о конкретном естественный язык и алгоритма синтаксического анализа, то есть набора стандартных операторов обработки текстового контента на основе этих знаний. Источником грамматических знаний сведения по морфологического анализа и различные заполнены таблицы понятий и лингвистических единиц. Они являются результатом эмпирического обработки текстового контента на естественном языке экспертами с целью выделения основных закономерностей для синтаксического анализа.

Ключевые слова: текст, украиноязычный, алгоритм, контент-мониторинг, ключевые слова, лингвистический анализ, синтаксический анализ, порождающих грамматики, структурная схема предложения, информационная лингвистическая система.

Bisikalo O. V.[1], Vysotska V. A.[2]
[1]Dr. Sc., Professor, Dean of Faculty for Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine
[2]PhD, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine

SENTENCE SYNTACTIC ANALYSIS APPLICATION TO KEYWORDS IDENTIFICATION UKRAINIAN TEXTS

This paper presents the generative grammar application in linguistic modelling. Description of syntax sentence modelling is applied to automate the processes of analysis and synthesis of texts in natural language. The article shows the features of the sentences synthesis indifferent languages of using generative grammars. The paper considers norms and rules influence in the language on the grammars constructing course. The use of generative grammars has great potential in the development and creation of automated systems for textual content processing, for linguistic providing linguistic computer systems, etc. The methods and tools development for automatic processing of text of commercial content in modern information technology are important and topical (for example, systems of information retrieval, machine translation, semantic, statistical, optical and acoustic analysis and synthesis of speech, automated editing, knowledge extracting from the text content, text content abstracting and annotation, textual content indexing, training and didactic, linguistic buildings management, instrumental means of dictionaries conclusion of various types, etc.). Specialists actively seeking new models of description and methods for automatic processing of text content. One of these methods is the development of general principles of lexicographic systems of syntactic type. It is important by these principles these systems construction of text content processing for specific languages. Any tools of syntactic analysis consists of two parts: a knowledge base about a particular natural language and algorithm of syntactic analysis (a set of standard operators of text content processing on this knowledge). The source of grammatical knowledge is data from morphological analysis and various filled tables of concepts and linguistic units. They are the result of the empirical processing of textual content in natural language of experts in order to highlight the basic laws for syntactic analysis.

Keywords: text, a Ukrainian, algorithm, content monitoring, keywords, linguistic analysis, parsing, generative grammar, structured scheme sentences, information linguistic system.

REFERENCES

1. Berko A., Vysotska V., Pasichnyk V. Systemy elektronnoyi kontent-komertsiyi. Leningrad, NULP, 2009, 612 p.
2. Vysotska V., Pasichnyk V., Scherbyna J., Shestakevych T. Matematychna linhvistyka. Leningrad, Novyy Svit-2000, 2012, 359 p.
3. Chomsky N. Three models for the description of language, *I.R.E. Transactions on Information Theory*, 1956, Vol. 2, pp. 113–124.
4. Chomsky N. On certain formal properties of grammars, *Information and Control*, 1959, Vol. 2, pp. 137–167.
5. Chomsky N. On the notion «Rule of Grammar», *Proceedings of the Twelfth Symposium in Applied Mathematics*, 1961, pp. 6–24.
6. Chomsky N. Context-free grammars and pushdown storage, *Quarterly Progress Reports, Research Laboratory of Electronics, M.I.T.*, 1962, No. 65, pp. 187–194.
7. Chomsky N. Formal properties of grammars, *Handbook of Mathematical Psychology, New York: Wiley and Sons*, 1963, Vol. 2, pp. 323–418.
8. Chomsky N. The logical basis for linguistic theory, *Proc. IX-th Int. Cong. Linguists*, 1962, pp. 91–111.
9. Chomsky N., Miller G. A. Finite state languages, *Information and Control*, 1958, Vol. 1, pp. 91–112.
10. Chomsky N., Miller G. A. Introduction to the formal analysis of natural languages, *Handbook of Mathematical Psychology 2, Ch. 12, Wiley*, 1963, Vol. 2, pp. 269–321.
11. Chomsky N., Schьtzenberger M. P. The algebraic theory of context-free languages, *Computer programming and formal systems, North-Holland*, 1963, pp. 118–162.
12. Chomsky N. Syntactic Structures. Mouton, The Hague, 1957, 117 p.
13. Chomsky N. Explanatory models in linguistics, *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress. Stanford University Press, Stanford, CA,* 1962, pp. 528–550.
14. Chomsky N. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA, 1965, 247 p.
15. Chomsky N. Conditions on transformations. New York, Holt, Rinehart & Winston, 1973, pp. 232–286.