УДК 81.37
© О. Ляшко[1], І. Миклушка[1] , 2012

НАУКОВІ ТЕКСТИ ТА ОСОБЛИВОСТІ ВИДІЛЕННЯ В НИХ КЛЮЧОВИХ СЛІВ

Проаналізовано особливості наукових текстів та процедури виділення ключових слів як індикаторів предметної області.

The features of scientific texts and the procedures of the keywords (as subject indicators) separation methods have been analyzed.

1. ВСТУП

В умовах постійно зростаючої кількості інформації, що набувається людством, а зокрема інформації наукового характеру, та формування нових знань, з кожним днем все більш актуальною стає проблема опрацювання та систематизації текстових масивів, і, як результат, вдосконалення технологій пошуку. За таких умов важливим завданням є дослідження представлення даних в документах для оптимізації роботи пошукових систем за допомогою лексичних одиниць, що використовуються для представлення змісту наукового тексту.

Такими лексичними одиницями зазвичай вважаються ключові слова (КС)[6]. Відповідно, інструментом опису реального світу і наших знань про нього є модель предметної області, лінгвістичним аналогом якої є множина КС, які визначають максимум зв'язків з іншими словами даної наукової сфери.

2. ПОСТАНОВКА ПРОБЛЕМИ

Загальновідомим фактом є те, що в наукових текстах представлено певний предмет, натомість в художніх – його вид та опис. Це змушує читача по різному ставитись до можливості їх тлумачення. Проблема розуміння та тлумачення наукового тексту завжди буде актуальною, оскільки, з одного боку в науковому тексті завжди багато однозначних слів, він логічний, подача матеріалу аналітична, а з іншого – розуміння ускладнене із-за термінологічного насичення та інформаційної глибини. Ще однією причиною проблем з розумінням наукового тексту може бути неспівпадіння тезаурусів читача і автора. Зміст та манера викладу інформації в наукових текстах розділяють читачів на «підготовлених» та «випадкових». Щоб бути в числі «підготовлених», необхідно володіти не тільки частково співпадаючими професійними знаннями,

[1] Українська академія друкарства

але і близьким загальнокультурним фоном і приналежністю до одної наукової спільноти. Хоча варто відзначити, що навіть у вчених однієї і тієї ж галузі знань не завжди співпадають лінгвістична компетентність та знання про навколишній світ.

Для повноти розуміння тексту зазвичай недостатньо перечитати його лише один раз – повторне сприйняття інформації змінює її носій. Тобто при первинному сприйнятті інформації читачем у нього є певний тезаурус, який після прочитання тексту поповнюється новими знаннями та при повторному ознайомленні з даною інформацією читач сприймає її, відповідно до змін, що відбулись у його пізнанні. Також потрібно відзначити, що розуміння тексту не є сприйняттям його самого, натомість, розуміння тексту - це співвідношення його з предметним світом читача [1].

Розуміння тексту – це проникнення в його зміст. Зміст тексту формується на основі згорнутої, явно невираженої змістової структури, що виводиться у відповідності з об'ємом лінгвістичних, енциклопедичних, фонових знань читача та його досвіду. Ця структура - стиснуте змістове ядро тексту, видобуте із його КС, їх асоціативних зв'язків та імплікацій, що повинно бути збережено при всіх змінах змісту інтерпретаторами.

Основними складнощами в декодуванні інформації є незнання варіантів мовного коду. Наприклад, текст закону містить в собі елементи офіційно-ділового тексту, мова адвоката – ораторського, журнальна стаття - публіцистичного. А, відповідно, міняється використання емоційних можливостей слова, ускладнюється синтаксис, міняється логічна побудова тексту [6].

Розуміння тексту залежить від постійно зростаючої кількості факторів, що відбувається на фоні процесів інтеграції і диференціації знань. Центробіжні тенденції в науці призводять до того, що наступає момент, коли теоретики не можуть зрозуміти експериментаторів навіть в одній області знань. Загальна закономірність розуміння наукового тексту в тому, що чим більше теоретичного змісту, тим тяжчим є сприйняття тексту читачем [1].

Проблемою також є і те, які саме одиниці тексту є вирішальними для його розуміння. Значення окремих елементів для вираження загального змісту неоднакове, і разом з центральними елементами є другорядні елементи тексту. При зміні центральних елементів зміст тексту змінюється. Якщо ж змінити другорядні одиниці тексту, то загальний зміст може зберегтись [6].

В розумінні тексту різні слова виконуються різні ролі. Найважливіша роль належить КС які є сигналами області знань. Як семантичні

домінанти, що однаково розуміються членами одного наукового соціуму, вони формують однодумців та полегшують діалог між ними.

Набір КС – це каталог концептів, навколо яких виникають основні ідеї наукового тексту. Цей набір дає читачу уявлення про стан знань в даній області, він науково орієнтує читача. Значення КС пересікаються з семантикою континуума тексту, визначаючи його межі.

Чим краще ми знаємо предметну область, тим зрозуміліше текст – ми розуміємо значення семантичних домінант (ключових слів). Чим далі ми ознайомлюємось з текстом, тим ймовірніше зняття невизначеності розуміння, тобто кількість можливих інтерпретацій тексту зменшується, і він здобуває інформаційну визначеність.

КС пробуджують в читача лише релевантні для даного наукового тексту значення слів та словосполучень. Тобто однією з функцій КС є зняття семантичної неоднозначності елементів тексту [1].

Актуальним на даний момент є питання формування об'єктивної методики виділення КС в науковому тексті [6].

3. КЛЮЧОВІ СЛОВА, ЯК ІНДИКАТОРИ НАУКОВИХ ТЕКСТІВ

При створенні та експлуатації інформаційно-пошукових систем особливе значення має аналітико-синтетична переробка документів - реферування, редагування, індексування ключовими словами і рубриками і т. д. - для введення результатів цієї переробки в комп'ютерних системах з метою генерації баз даних (БД) і автоматизованої підготовки різних наукових видань [5].

В загальнонауковому плані КС розглядаються як найбільш важливий, вагомий для розуміння змістовий елемент. КС визначає зміст тексту та є носієм його смислу. Враховуючи вищесказане, можна стверджувати, що КС – це слова, найбільш значимі для розуміння змісту тексту. Висловлюючись лінгвістичними термінами, КС – це семантичні домінанти, що сигналізують про предметну область знань, що однаково розуміються членами одного соціуму, що полегшують діалог між ними та дозволяють проникнути в те, що знаходиться за текстом [6].

КС вважаються нормалізовані слова і стійкі словосполучення природної мови, які представляють собою терміни або поняття конкретної галузі науки і техніки і несуть в сукупності максимально повну і стислу інформацію про зміст документа (реферату). КС, що використовуються при індексуванні, можуть не міститися в явному вигляді в тексті документа. При формуванні предметних покажчиків в якості заголовків рубрик покажчиків використовуються тільки КС [5].

КС володіють декількома диференційними ознаками, що дозволяє впізнавати їх у науковому тексті: а) частотність, повторюваність в да-

ному тексті; б) семантична близькість слова до теми тексту; в) інформаційне насичення.

Для спрощення структури наукового тексту (для якого характерні складні розгорнуті синтаксичні конструкції), для надання йому стереотипізації, в ньому часто використовуються паралельні конструкції.

Семантична структура тексту є трирівневою структурою, що складається з рівня змісту, рівня сенсу і рівня зв'язку, що фактично відповідає поверхневому, глибинному і концептуальному розумінні тексту.

КС, якими індексують текст, повинні відображати всі основні аспекти його змісту, у тому числі й ті, які є "непрофільними" по відношенню до того розділу, де поміщений індексований документ, але які представляють інтерес для фахівців суміжних галузей знання.

Функціями КС в науковому тексті є: 1) формування змісту тексту і забезпечення збереження його в пам'яті; 2) створення структурно-семантичної єдності тексту, його цілісності; 3) задання функціонального стилю тексту, що розгортається; 4) відображення предметної області тексту, що розгортається, зняття семантичної неоднозначності; 5) маркування теми тексту; 6) визначення позиції слів чи словосполучень, що володіють максимальною передбачуваністю; 7) наближення до інваріанту стану; 8) виокремлення цілісного змісту; 9) передача узагальненого змісту тексту, вказання і обмеження направлення асоціацій, імплікацій читача; 10) комунікативна – засіб наукової комунікації [6].

Стандартний підхід в аналізі текстової інформації полягає у вивченні частотних характеристик мовних інваріантів: слів і словосполучень. Кожній предметній області відповідає свій словниковий частотний спектр, який при достатніх обчислювальних ресурсах можна вирахувати, маючи великий об'єм документів, що належать даній предметній області документів – вибірку [3].

Одним з основних аспектів аналітико-синтетичної переробки документів є так зване координатне індексування, тобто вираження основного смислового змісту документа. у вигляді певного набору ключових і пояснювальних слів.

Принцип одноразового реферування та індексування документів і багаторазового використання результатів для генерації БД та інформаційних видань пред'являє підвищені вимоги до якості обробки первинної інформації. Для цього необхідно розроблене лінгвістичне забезпечення, зокрема - списки основних ключових слів з різних областей науки і техніки, які є найважливішим інструментом координатного індексування документів. У технології одноразового реферування списки КС з різних галузей науки і техніки повинні бути єдиними і/або сумісними.

Списки підбираються шляхом "знімання" КС з пошукових образів документів та подальшою їх експертної оцінки, з урахуванням частоти входження терміна. Для визначення лексичного поля списків слід виявляти не тільки основні КС даної галузі науки або техніки (або так зване ядро термінології), а й терміни суміжних галузей, тобто виявляти ступінь перетину лексичних полів різних галузей науки і техніки.

Це особливо важливо в умовах одноразового реферування та необхідності підготовки єдиних і/або сумісних списків з різних галузей науки і техніки і, в перспективі, тезауруса, тому що в першу чергу необхідно узгоджувати термінологію, що зустрічається в суміжних галузях.

Списки призначені для: а) використання при розробці та експлуатації автоматизованих інформаційно-пошукових систем з різних галузей науки і техніки; б) уніфікації та стандартизації термінології в конкретній галузі науки і техніки; в) використання при координатному індексуванні документів; г) пошуку відомостей в БД і предметних покажчиках (оскільки ключові слова є заголовками рубрик у предметних покажчиках). В останньому випадку єдина форма подання КС особливо важлива для систематизації даних та зручності користування предметним покажчиком [5].

При умові, що є певний набір зразків спектрів різних предметних областей, можна ідентифікувати предметну область, порівнюючи отриманий спектр зі зразками. Проте, це потребує значних обчислювальних потужностей. Саме тому і виникла задача оптимізації розміру словникового спектру з можливістю ідентифікації його з предметною областю. Бажано обмежити словниковий спектр тільки значимими словам і словосполученнями, такими, що будуть визначати предметну область [3].

Очевидно, що значимість для слів, що дуже часто зустрічають, повинна знижуватися, тому що зазвичай це службові слова - прийменники і т. п. Для обліку цієї особливості алгоритм обчислення ваги модифікується наступним чином: а) вводиться список так званих "шумових" або "стоп"-слів; б) вводиться спеціальна обробка для термінів, що зустрічаються частіше певного граничного значення [2].

Варто зазначити, що виділена множина частотно-інваріантних слів і словосполучень ("стоп"-слів), зберігає свою частотну характеристику незалежно від предметної області. До таких слів можуть відноситись займенники, сполучники, частки, дієслова, невеликі слова. Недолік такого способу в тому, що "стоп"-словник може містити значимі для деякої предметної області слова і словосполучення.

Інший підхід – це співставлення частотного спектра вибірки документів зі спектром-зразком. Значимість даного слова для предметної

області тим більше, чим більша відносна насиченість вірогідності частоти вживання даного слова в даній предметній області порівняно з деякою суттєво більшою областю (над-областю), підмножиною якої є досліджувана область. При такому підході порівнюються дві величини: насиченість частоти вживання слова в області і надобласті. Це дозволяє виявити всі часто вживані слова, що виокремлюють дану область та підвищити точність виявлення КС у вибраному тексті [3].

Щоб відповісти на питання, чим і як забезпечується ідентифікація та пізнаваність КС, необхідно виділити їх диференціальні семантичні і формальні ознаки. В першу чергу це частотність слова в мові і частота його реального вжитку. В наукових текстах частота реального вжитку значно вища за частотність у мові. Статус КС забезпечує максимум зв'язків з іншими словами даної предметної області [1].

В якості КС використовуються перераховані нижче групи понять, якщо в рефераті міститься оригінальна інформація про позначуваний ними предмет: 1) об'єкти дослідження; 2) процеси і явища; 3) речовини та матеріали; 4) джерела речовин, що забруднюють навколишнє середовище; 5) характеристики процесів; 6) методи дослідження, аналізу; 7) прилади, апарати, пристрої, вироби; 8) закони, рівняння, функції; 9) загальнонаукові терміни та інші слова. КС, що представляють собою загальні поняття, контролюються за спеціальним словником - "Списком основних ключових слів".

Список основних ключових слів має єдину впорядковану формалізовану логічну структуру, в ньому усунені синонімія, полісемія а також омонімія між окремими термінами, всі слова наведені в стандартній словникової формі, тобто в називному відмінку однини чи множини, і в деяких випадках між ними встановлені смислові зв'язки.

Списки КС можуть стати основою тлумачних словників з різних галузей науки і техніки, а також основою двомовних словників, що використовуються при автоматизованому або звичайному перекладі. При систематичному контролі нормалізованої лексики за частотою появи нових термінів можна судити про зміну напрямків дослідження даної галузі науки або техніки, її точках росту, зміну термінологічного поля і т, д. Ці статистичні дані можуть бути використані при прийнятті рішення про актуалізацію лінгвістичного забезпечення системи підготовки інформаційних продуктів, рубрикатора галузі науки або техніки, про узгодження термінології в суміжних галузях знання, при вирішенні питань генерації проблемних фрагментів баз даних і/або підготовки випусків реферативних журналів і т. д. [5].

В сучасних дослідженнях проблема КС розглядається у зв'язку з розробкою інформаційно-пошукових систем (ІПС), в рамках теорії комунікації, психолінгвістики, когнітивної лінгвістики. Відповідно і

підходи до визначення даного поняття суттєво відрізняються. Дослідники звертають увагу на ієрархію мовних засобів в тексті, на існування в ньому домінантних і субдомінантних мовних засобів, виділяючи при цьому різні аспекти даного мовного явища [6].

Існує множина варіантів інформаційного пошуку, але найбільш поширене завдання - пошук текстового документа за запитом, який представляє собою деяку фразу або набір слів, введених користувачем. Це історично перший вид інформаційного пошуку, і найбільш часто застосовується в даний час. Існує ряд інших завдань таких, як класифікація документів, кластеризація, автоматизоване анотування і т. д.

Виходячи з цього, науковцями було сформовано декілька моделей представлення документів за допомогою їх лексичних одиниць. Наразі в ІПС використовуються такі моделі: модель множини слів, бінарна модель (з "вагою" слів) та інші нелексичні моделі.

Основною гіпотезою моделі "множини слів" інформаційного пошуку є те, що релевантний документ містить ті ж терміни, що і запит. Під терміном звичайно розуміють окреме слово природної мови, на якій створено документ. Всі відомі сучасні мови, що використовуються людиною, ґрунтуються на словах як мінімальній одиниці, яка має самостійне значення. Інформаційна система повинна виділити слова із збереженого в ній представлення документа [2]. Метою представлення документів в ІПС є виявлення та експлікація теми тексту чи його розділу, а репрезентантом теми наукового (спеціального) тексту є саме термін [4].

Вся множина символів розбивається на дві множини - символи, які можуть входити в слово (букви і цифри), і символи-розділювачі (пробіли, знаки пунктуації). Підпрограма виділення слів сканує текст, вибираючи послідовності символів слів, обмежених роздільниками. Однак такий алгоритм має ряд похибок при роботі з природними текстами: особливості написання, переноси слів, написання слів через дефіс, абревіатури, що пишуться через крапки, номери, дати й числа, містять крім цифр різні символи пунктуації і т. д. У системах інформаційного пошуку для обробки подібних ситуацій застосовується ускладнення описаного алгоритму, коли в якості слова формується множина можливих варіантів, як слова, з’єднання символами-роздільниками, так і слова цілком.

У випадку простої моделі пошукова система може розглядати тільки наявність/відсутність слів в документі. Така модель документа називається бінарною. Ускладненням цієї моделі є підхід, де для кожного слова вказується не тільки наявність, але й певна "вага". Таким чином, модель документа з моделі множини слів перетворюється в модель множини пар "слово - вага".

Існують відомі методи визначення "ваги" слова: а) статистичний підхід - "вага" слова визначається із статистичних характеристик його входження в текст документа; б) місце появи слова – слово по-різному "зважується" в залежності від того, зустрілося воно в заголовку, початку тексту і т. д.; в) оформлення слова. Слово може бути виділено шрифтом або положенням, що може бути враховано при визначенні "ваги". Як правило, реалізується деякий комбінований метод, коли "вага" обчислюється на підставі комбінації згаданих ознак.

Використовуючи статистичний підхід, можна взяти слова, що зустрічаються в деякій текстової колекції і відсортувати їх по частоті. При цьому позиція слова в цьому списку називатиметься рангом слова.

Найбільш поширеними методами "зважування" слів в документі є: а) кількість входжень слова в даному документі; б) частота входження слова в документі; в) логарифм частоти входження слова [2]

Частотний метод "зважування" слів став одним з перших, що були запропоновані науковцями і на даний момент досі використовується у великій кількості ІПС. Хоча даний метод і заслужено критикується із-за якості індексування, проте протягом останнього часу ведуться дослідження стосовно удосконалення частотного методу індексування наукових (а в ширшому розумінні – спеціальних) текстів [4].

Варто зазначити, що при формуванні списку КС бажано враховувати той очевидний факт, що більш важливі з точки зору сенсу слова автори текстів розміщують у заголовку документа чи його розділів, або з початку його тексту.

При дослідженні наукових текстів варто звертати увагу також на текстові виділення у документі, оскільки, зазвичай, всі значимі слова, терміни, поняття виділяються автором засобами форматування тексту (виділення курсивом, капітель, підкреслення, пів жирне виділення та ін.) [2].

Ігнорування останніх двох факторів в більшості випадків призводить до погіршення якості пошуку, тому їх варто використовувати на рівні з іншими методами виділення семантично значимих елементів тексту.

4. ВИСНОВОК

Отже, можна зробити висновок, що наразі є необхідність перегляду особливостей дослідження та стиснення наукових текстів методом автоматизованого, чи навіть автоматичного опрацювання системами індексування та стиснення даних. Враховуючи стрімке зростання об’ємів інформації та можливості її накопичення (стиснення даних відіграє достатньо значну роль у системах збереження інформації),

особливо гостро постає питання стосовно збільшення релевантності, точності та стислості пошукових видач в ІПС, оскільки, навіть якщо всі результати пошукової видачі будуть релевантними, то користувач може бути не здатний опрацювати величезний масив результатів цього пошуку. Саме за таких умов зростає роль ключових слів (та їх функцій) як особливих лексико-семантичних одиниць наукового тексту. Адже саме за допомогою ключової лексики є можливість відобразити як зміст наукового тексту, так і його семантику, а, відповідно, і в результатах роботи ІПС надавати користувачу інформацію, що співпадає з його запитами.

1. Н.А. Шехтман. Ключевые слова в научном тексте // НТИ, сер.2, Информ. процессы и системы № 10. – Москва, 2003. – С. 35-39.2. М.В. Губин. Модели и методы представления текстового документа в системах информационного поиска // НТИ, сер.1, Орг. и методика информ. работы № 12. – Москва, 2004. – С. 12-24. 3. А. Антонов. Контент-анализ: некоторые особенности решения ГАЛАКТИКА-Зум // НТИ, сер.2, Информ. процессы и системы № 4. – Москва, 2002. 4. Э.Ф. Скороходько. Роль системо- и текстообусловленых характеристик термина в частотном индексировании научных текстов // НТИ, сер.2, Информ. процессы и системы № 8. – Москва, 2002. – С. 1-6.
5. В.А. Быков, Е.Б. Дудин, М.М. Мельникова. Принципы составления списков основных ключевых слов по различным отраслям науки и техники // НТИ, сер.1, Орг. и методика информ. работы № 12. – Москва, 2002. – С. 20-22.
6. Т.Н. Москвитина. Ключевые слова и их функции в научном тексте // Вестник ЧГПУ №11. – Челябинск, 2009. – С. 270-283.