УДК 519.682:336.763

В. Б. Говоруха, О. Ю. Лебідь

Академія митної служби України м. Дніпропетровськ

ПРОБЛЕМА ФОРМАЛЬНОГО ВІДОБРАЖЕННЯ ЗМІСТУ ТЕКСТУ

Розглянуто проблему відображення змісту тексту на основі семантичного аналізу. Проаналізовано засоби використання методів аналізу текстів для подальшого пошуку інформації та класифікації документів.

Ключові слова: *семантичний аналіз, система розуміння тексту, пошукова система, класифікація документів, змістове відображення тексту.*

Рассмотрена проблема представления смыслового содержания текста на основе семантического анализа. Проанализированы способы использования методов анализа текста для дальнейшего поиска информации и классификации документов.

Ключевые слова: *семантический анализ, система понимания текста, поисковая система, классификация документов, смысловое представление текста.*

The basic problem representation of the semantic content of text based on semantic analysis is considered. The different methods of text analysis methods for further information retrieval and document classification.

Key words: *semantic analysis, system understanding of the text, the search engine, classified documents, the semantic representation of text.*

Вступ. Розуміння текстів на природній мові та відповідний машинний переклад є одним з напрямків розвитку теорії штучного інтелекту. Останні дослідження в цьому напрямку показали, що для вирішення проблеми недостатньо створити великі сховища словників для перекладу з однієї мови на іншу та відповідні правила перекладу. Тому згодом було створено мову-посередник, яка, у свою чергу, перетворилася на семантичну модель відображення змісту текстів, що перекладаються.

Як відомо, природна мова – це універсальна знакова система, що служить для обміну інформацією між людьми. Оскільки документи на вході документально-пошукових інформаційних систем записані на природній мові, слушно було б поставити питання, а чи не можна

використовувати природну мову як основний засіб відображення інформації під час усього циклу функціонування документально-інформаційних пошукових систем? Відповідь буде позитивною, якщо йдеться про ті інформаційно-пошукові системи, в яких відповідність між запитом і документом встановлює людина. Проте в сучасних документально-пошукових інформаційних системах цю операцію виконує комп'ютер, тому виключається використання природної мови як основного засобу відображення інформації. Це пояснюється істотними особливостями природної мови з погляду машинної технології обробки інформації, такими як різноманіття засобів передачі змісту тексту, семантична неоднозначність, синонімія та багатозначність [1; 2].

Тому для організації пошуку та класифікації документів потрібен етап попереднього аналізу текстів документів. Одним із перспективних напрямів у системах пошуку і класифікації документів виступає семантичний аналіз.

Метою даної роботи є аналіз засобів використання методів семантичного аналізу текстів для класифікації та пошуку інформації.

Результати дослідження. Семантичний аналіз – процес виявлення змісту слів і словосполучень у реченні. Він забезпечує нормалізацію синтаксичної структури речень, розпізнавання термінів, їх класифікацію за семантичними ознаками, з урахуванням синонімічних і гіпонемічних (загальне – часткове) класів, виявлення визначень термінів.

Тематику будь-якого терміна або тексту можна відобразити комбінацією базових семантичних категорій, що асоціюються з ним, і кількість яких уже значно менша, ніж кількість слів. Таким чином, семантичний опис використовує замість слів укрупнені поняття – категорії, кожна з яких характеризується своїм набором термінів. Тому семантичне відображення змісту текстів супроводжується істотним стиском інформації. Стиск інформації при переході від лексичного до семантичного опису документів відбувається за рахунок використання певних знань про структуру мови.

Терміни «семантичний аналіз» і «машинне розуміння тексту» вважаються еквівалентними. За основу в даній роботі взято методи текстології отримання знань, що використовуються під час розробки і ручного наповнення баз знань експертних систем. У разі такого підходу процедури «розуміння» і «здобування знань» є ідентичними, а результат їх виконання формалізується у вигляді деякої семантичної структури. Аналогічно машинне розуміння розглядається у вигляді процесу формування семантичного образу для аналізованого тексту

на природній мові, що виконується системами розуміння тексту (СРТ) (рис. 1).

У СРТ виділено лінгвосемантичне і програмне забезпечення. Перше використовується для опису моделі наочної області і являє собою лінгвістичний і семантичний словники, в термінах яких СРТ формує образ тексту. Програмне забезпечення реалізує відповідні методи аналізу. Роботу СРТ можна поділити на два етапи: лінгвістична обробка і семантична інтерпретація, що виконуються відповідно лінгвістичним і семантичним модулями СРТ.

Лінгвістичний модуль об'єднує етапи безпосередньої обробки природної мови. На цих етапах з використанням словників лінгвістичного забезпечення відбувається первинна формалізація речень вхідного тексту. На етапі графоматичного аналізу виділяються текстові одиниці, такі як слова, речення і абзаци. Крім того, на цьому етапі виключаються незначущі слова і складні конструкції, такі як вступні речення. На етапі морфологічного аналізу визначаються граматичні значення слів, такі як частина мови, рід, число тощо. На етапі синтаксичного аналізу визначається синтаксична структура речення. Найчастіше для опису синтаксису використовується V-мова.

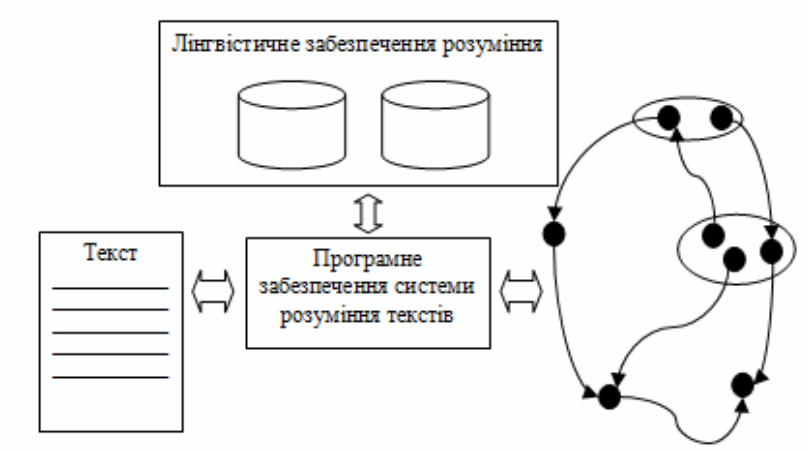

Рис. 1. Функціонування системи розуміння тексту

Семантичний модуль виконує змістовну обробку тексту вхідних даних, отриманими лінгвістичним модулем. Даний вид обробки називається інтерпретацією, оскільки згідно із закладеною в словниках семантичного забезпечення моделлю наочної області визначається формальний сенс окремих формул V-мови. Ця

процедура виконується на етапі семантичного аналізу. На етапі міжфразового семантичного аналізу об'єднуються семантичні зображення окремих речень в єдину семантичну мережу, що описує сенс усього тексту.

Моделі семантичного аналізу почали розвиватися у зв'язку з активним розвитком комп'ютерної обробки текстів. Особливо великі досягнення в цій галузі пов'язані з питаннями пошуку інформації в глобальній мережі Інтернет. Тому логічно розглядати моделі семантичного аналізу в контексті пошуку інформації, а також кластеризації та класифікації документів.

Розглянемо докладніше моделі пошуку. Перший підхід у них базується на теорії множин, другий – на векторній алгебрі, а третій – на теорії вірогідності. Усі ці підходи досить ефективні на практиці, проте в канонічному вигляді всі вони мають спільний недолік, який випливає з припущення, що контент документа, його основний зміст визначається множиною ключових слів – термінів і понять, які входять до нього. Звичайно ж, такий підхід частково веде до втрати змістовних відтінків документів, проте дозволяє виконувати швидкий пошук і групування документів за формальними ознаками. Нині ці підходи найпоширеніші. Слід також зазначити, що існують інші методи, наприклад семантичні, в рамках яких робляться спроби виявити зміст за рахунок аналізу граматики тексту, використання баз знань і тезаурусів, які відображають семантичні зв'язки між окремими словами та їх групами. Очевидно, що такі підходи вимагають істотних витрат на підтримку баз знань і тезаурусів для кожної мови, тематики і виду документів, галузь їх застосування – професійні аналітичні системи.

Булева модель, що базується на теорії множин, є класичною і найбільш поширеною моделлю зображення інформації. Її популярність пов'язана, перш за все, з простотою реалізації, що дозволяє індексувати і виконувати пошук у масивах документів великого обсягу. В даний час популярне об'єднання булевої моделі з векторно-просторовою моделлю алгебри зображення даних, що забезпечує, з одного боку, швидкий пошук з використанням операторів математичної логіки, а з іншого – якісне ранжування документів, що базується на вагах ключових слів, які входять до них. У рамках булевої моделі документи і запити зображуються у вигляді множини морфемних основ ключових слів (терми).

У булевій моделі запитом користувача слугує логічний вираз, в якому ключові слова (терми запиту) пов'язуються операторами з теорії множин і відповідними логічними операторами AND, OR та

NOT. У різних пошукових системах, що використовують булеву модель, зокрема в Інтернеті, користувачі при формуванні запитів можуть просто перераховувати ключові слова, не вказуючи в явному вигляді логічних операцій. Найчастіше при цьому передбачається, що всі ключові слова з'єднуються логічною операцією AND, у цих випадках до результатів пошуку включаються тільки ті документи, які містять одночасно всі ключові слова запиту. У системах, в яких пропуск між словами прирівнюється до оператора OR, у результати пошуку включаються документи, до яких входить хоч би одне з ключових слів запиту. При використанні булевої моделі база даних включає індекс, організований у вигляді масиву, в якому для кожного терма зі словника бази даних міститься список документів, в яких цей терм зустрічається. В індексі можуть зберігатися також частоти даного терма в кожному документі, що дозволяє сортувати список за убуванням частоти. Класична база даних, що відповідає булевій моделі, організована так, щоб за кожним термом можна було швидко дістати доступ до відповідного списку документів. Крім того, структура масиву забезпечує його швидку модифікацію при включенні в базу даних нових документів. У зв'язку з цими вимогами масив часто реалізується у вигляді B-дерева. Існує декілька підходів до формування архітектури пошукових систем, що відповідають булевій моделі і які знайшли своє втілення в реальних системах. Однією з найбільш вдалих реалізацій структури бази даних інформаційно- пошукової системи на мейнфреймах фірми IBM визнано модель даних системи STAIRS (Storage and Information Retrieval System), яка завдяки початковим вдалим архітектурним рішенням досі продовжує розвиватися. База даних інформаційно-пошукових систем цієї традиційної архітектури складається з таких основних таблиць [3]:

- текстова – містить текстову частину всіх документів;
- таблиці покажчиків текстів – включає покажчики місцезнаходження документів у текстовій таблиці, а також поля форматів усіх документів;
- словник – містить всі унікальні слова, що зустрічаються в полях документів, тобто ті слова, за якими може здійснюватися пошук. Слова можуть об'єднуватися в синонімічні ланцюжки;
- масив термів – містить списки номерів документів і координати окремих слів у полях документів.

Процеси, що відбувалися при пошуку інформації в базі даних STAIRS, сьогодні реалізуються засобами сучасних СУБД і

інформаційно-пошукових систем документального типу. Пошук терміна в базі даних здійснюється таким чином. Відбувається звернення до словникової таблиці, за якою визначається, чи входить слово до складу словника бази даних, і якщо входить, то визначається посилання на ланцюжок появи цього слова в документах. Відбувається звернення до масиву термів, за якими визначаються координати всіх входжень терма в текстову таблицю бази даних. За номером документа відбувається звернення до запису таблиці покажчиків текстів. Кожен запис цього файла відповідає одному документу в базі даних. За номером документа відбувається пряме звернення до фрагмента текстової таблиці – документа і подальше його виведення. У разі коли обробляється не один термін, а деяка їх комбінація, в результаті відпрацювання пошуку за кожним терміном запиту формується масив записів, що відповідає входженню цього терміна в базу даних.

Висновки. Актуальним напрямком розвитку систем машинного перекладу, систем пошуку інформації та класифікації документів є використання методів семантичного аналізу.

У статті проаналізовано деякі з методів семантичного аналізу текстів для подальшої класифікації та пошуку інформації. Надалі планується розробка методів та алгоритмів для вирішення зазначених актуальних питань на основі методів семантичного аналізу та подальше їх упровадження в інформаційних системах.

Бібліографічні посилання

1. Нильсон Н. Принципы искусственного интеллекта / Н. Нильсон – М., 1985. – 374 с.

2. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах / В.Ш. Рубашкин – М., 1989. – 258 с.

3. Ландэ Д. В. Определение тематической направленности запросов путем анализа набора рейтинговых источников / Д. В. Ландэ, С. М. Брайчевский // Открытые информационные и компьютерные интегрированные технологии. – Харьков, 2005. – Вып. 29. – С. 169–174.

Надійшла до редколегії 20.07.2012