УДК 519.682:336.763

В. Б. Говоруха, О. Ю. Лебідь

Академія митної служби України

ПРОБЛЕМА ФОРМАЛЬНОГО ПРЕДСТАВЛЕННЯ СМИСЛОВОГО ЗМІСТУ ТЕКСТУ

У праці розглядається основна проблема представлення смислового 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, смислове представлення тексту.*

В работе рассматривается основная 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.*

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

Keywords: *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 цієї проблеми недостатньо створити великі сховища словників для перекладу з однієї мови на іншу та відповідні правила перекладу. Тому згодом було створено мову-посередник, яка в свою чергу перетворилася у семантичну модель представлення сенсу текстів, що перекладаються.

Як відомо, природна мова є універсальною знаковою системою, що служить для обміну інформацією між людьми. Оскільки документи на вході документально-пошукових інформаційних систем, записані на природній мові, справедливо було б задатися питанням, а чи не можна використовувати природну мову як основний засіб представлення інформації під час всього циклу функціонування документально-інформаційних пошукових систем? Відповідь буде позитивною, якщо мова йде про ті інформаційно-пошукові системи, в яких відповідність між запитом і документом встановлює людина. Проте в сучасних документально-пошукових інформаційних системах ця операція виконується комп'ютером, що практично виключає застосування природної мови як основного засобу представлення інформації. Це пояснюється істотними недоліками природної мови з погляду машинної технології обробки інформації, таких 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 є семантичний аналіз.

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

Результати дослідження. Семантичний аналіз – процес виявлення смислового змісту слів і словосполучень в реченні. Семантичний аналіз забезпечує нормалізацію синтаксичної структури речень, розпізнавання термінів, класифікацію термінів по семантичних ознаках, з урахуванням синонімічних і гіпонемічних (загальне – приватне) класів, виявлення визначень термінів.

Тематику будь-якого терміну або тексту можна представити комбінацією базових семантичних категорій, що асоціюються з ним, число яких вже значно менше, ніж число слів. Таким чином, семантичний опис використовує замість слів укрупнені поняття – категорії, кожна з яких характеризується своїм набором термінів. Тому семантичне представлення змісту текстів супроводжується істотним стисненням інформації. Стиснення інформації при переході від лексичного до семантичного опису документів відбувається за рахунок використання деякого знання про структуру мови.

Терміни семантичний аналіз і машинне розуміння тексту приймаються еквівалентними. За основу в даній роботі узяті методи текстології отримання знань, що використовуються при розробці і ручному наповненні баз знань експертних систем. При такому підході процедури «розуміння» і «витягання 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 виділене лінгвосемантичне і програмне забезпечення. Перше використовується для опису моделі наочної області і представлене лінгвістичним і семантичним словниками, в термінах яких СРТ формує образ тексту. Програмне забезпечення реалізує відповідні методи аналізу. Роботу СРТ можна розділити 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 пропозицій вхідного тексту. На етапі графематичного аналізу виділяються текстові одиниці, такі як слова, речення і абзаци. Крім того, на цьому етапі виконується виключення незначущих слів і складних конструкцій, таких як вступні речення. На етапі морфологічного аналізу визначаються граматичні значення слів, такі як частина мови, рід, число і так далі На етапі синтаксичного аналізу визначається синтаксична структура речення. Найчастіше для опису синтаксису використовується V-мова.

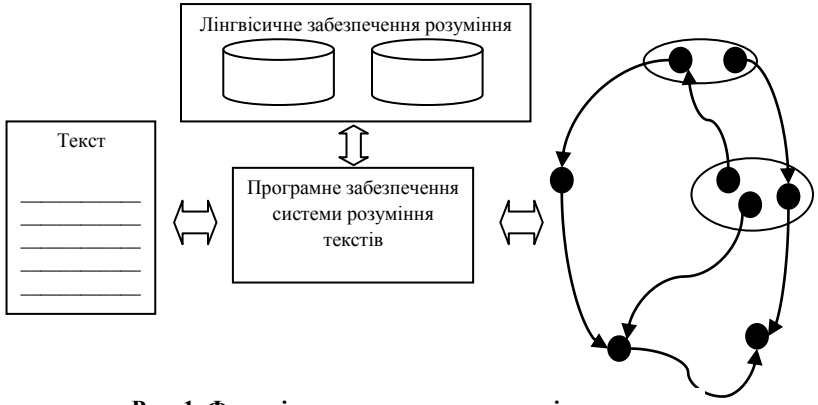

Рис. 1. Функціонування системи розуміння тексту

Семантичний модуль виконує смислову обробку тексту вхідних даних, представлених V–формулами, отриманими лінгвістичним модулем. Даний вид обробки називається інтерпретацією, оскільки згідно закладеної в словниках семантичного забезпечення моделлю наочної області виконується визначення формального сенсу окремих формул V–мови. Ця процедура виконується на етапі семантичного аналізу. На етапі міжфразового семантичного аналізу проводиться об'єднання семантичних представлень окремих речень в єдину семантичну мережу, що описує сенс всього тексту.

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

Розглянемо докладніше моделі пошуку. Перший підхід в них базується на теорії множин, другій на векторній алгебрі, а третій на теорії вірогідності. Всі ці підходи досить ефективні на практиці, проте в канонічному вигляді у всіх них є загальний недолік, який витікає з припущення, що контент документа, його основний зміст визначається безліччю ключових слів - термінів і понять, які входять в нього. Звичайно ж, такий підхід частково веде до втрати змістовних відтінків документів, проте дозволяє виконувати швидкий пошук і групування документів по формальних ознаках. Сьогодні ці підходи – найпоширеніші. Слід зазначити, що існують також 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 і їх групами. Очевидно, що такі підходи вимагають істотних витрат на підтримку баз знань і тезаурусів для кожної мови, тематики і виду документів, область їх застосування – професійні аналітичні системи.

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

У булевій моделі запитом користувача є логічний вираз, в якому ключові слова (терми запиту) зв’язуються операторами з теорії множин і відповідними їм 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 указуючи в явному виді логічних операцій. Найчастіше при цьому передбачається, що всі ключові слова з'єднуються логічною операцією AND – в цих випадках в результати пошуку включаються тільки ті документи, які містять одночасно всі ключові слова запиту. У системах, в яких пропуск між словами прирівнюється до оператора OR, в результати пошуку включаються документи, в які 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 із словника бази даних міститься список документів, в яких цей терм зустрічається. У індексі можуть зберігатися також частоти даного терма в кожному документі, що дозволяє сортувати список по убуванню частоти. Класична база даних, що відповідає булевій моделі, організована так, щоб по кожному терму можна 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, відповідних булевій моделі і що знайшли своє втілення в реальних системах. Одній з найбільш вдалих реалізацій структури бази даних інформаційно-пошукової системи на мейнфреймах фірми IBM була визнана модель даних системи STAIRS (Storage and Information Retrieval System), яка, завдяки початково вдалим архітектурним рішенням до цих пір продовжує розвиватися. База даних інформаційно-пошукових систем цієї традиційної архітектури складається з наступних основних таблиць [3]:

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

Процеси, що відбувалися при пошуку інформації в базі даних STAIRS, сьогодні реалізуються засобами сучасних СУБД і інформаційно-пошукових систем документального типу. Пошук терміну в базі даних здійснюється таким чином. Відбувається звернення до словникової таблиці, по якій визначається, чи входить слово до складу словника бази даних, і якщо входить, то визначається посилання на ланцюжок появ цього слова в документах. Відбувається звернення до масиву термів, по якому визначаються координати всіх входжень терма в текстову таблицю бази даних. По номеру документа відбувається звернення до запису таблиці покажчиків текстів. Кожен запис цього файлу відповідає одному документу в базі даних. По номеру документа відбувається пряме звернення до фрагмента текстової таблиці - документа і подальший його вивід. У разі, коли обробляється не один термін, а деяка їх комбінація, в результаті відпрацювання пошуку по кожному терміну запиту формується масив записів, що відповідає входженню цього терміну в базу даних.

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

У роботі проаналізовано деякі з методів семантичного аналізу текстів для подальшої класифікації та пошуку інформації. Надалі планується розробка методів та алгоритмів для розв’язання зазначених актуальних питань на основі методів семантичного аналізу та подальше їх впровадження у інформаційних системах.

Бібліографічні посилання

1. **Нильсон Н.** Принципы искуственного интеллектатьи – М.: Мир, 1985. – 374 с.
2. **Рубашкин В. Ш.** Представление и анализ смысла в интеллектуальных информационных системах – М.: Наука, 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 технологи – Харьков: Нац. аэрокосмический ун-т «ХАИ», 2005. – Вып. 29. – С. 169–174.

Надійшла до редколегії 12.05.2012