УДК 519.765: 004 (045)

МОДЕЛІ СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТІВ

Литвиненко О. Є., Бурко Д. А.
Національний авіаційний університет
litvinen@nau.edu.ua

Розглянуто основні технології обробки текстів у комп'ютерних системах. Виділено ефективний метод семантичного аналізу текстів, в основі якого лежить виділення семантичних рівнів.

The paper deals with considering of the main technologies of text analysis in computer systems. The efficient methods of the semantic analysis and semantic processing of texts was distinguished. The approach is based on the defining of semantic levels.

Постановка завдання дослідження

Необхідність створення помічників для людини в її науковій діяльності ставить завдання розробки систем або машин, здатних виконувати такі дії, для яких зазвичай потрібен людський розум. Один з найважливіших напрямів теоретичних розробок зі створення подібних систем пов'язаний з моделюванням інтелектуальної діяльності людини. При цьому виникають різного роду проблеми, зумовлені тим, що розумова діяльність людини різноманітна і включає в себе розпізнання образів, розв'язання задач, прогнозування, аналіз текстової інформації, доведення теорем тощо.

У сучасному науковому просторі приділяється значна увага такому феномену, як електронний текст. Саме текст розглядається як основне джерело інформації. Існують декілька підходів до його аналізу. Можна, наприклад, визначати тему й ідею текстів, аналізувати, оцінювати смислове навантаження або виділяти сферу, з якою вони пов'язані (математика, комп'ютерні науки, література, соціологія). Проте комп'ютерні системи обробки даних, такі як пошукові або порівняльні системи, таких умінь не мають. Вони аналізують інформацію інакше. Тому актуальною є проблема розробки комп'ютерних методів і алгоритмів моделювання діяльності людського мозку. Для аналізу текстової інформації використовуються алгоритми семантичної, морфологічної та синтаксичної обробки текстів.

Аналіз стану проблеми

Сьогодні проводяться активні дослідження в галузі комп'ютерного семантичного аналізу текстів. Спроба моделювання розуміння людиною семантичних зв'язків у тексті приводить до постановки питання про семантичну структуру мови і про рівні, на яких описуються значення слів. Фахівці в галузі семантики виокремлюють принаймні два рівні: семантичний і поверхнево-синтаксичний. Синтаксичні конструкції відображають будову речення у формі граматичних елементів мови (дієслово, іменник, прикметник). А семантичні — виражають змістове співвідношення між словами. Перші семантичні дослідження і спроби розробити семантичну мову-посередника між машиною і людиною були зроблені наприкінці 50-х років XX ст. Було побудовано серію, яка містила в собі 58 класифікаторів (назви елементарних сенсів) і рекурсивних правил для побудови формул для них. Класифікатори були зведені до кількасот шаблонів. Проте словники таких шаблонів робили реальне сприйняття змісту дуже спрощеним і особливого подальшого розвитку такі методи не набули.

Наступним кроком для розробки семантичного аналізу став метод, запропонований Ч. Філмором. Він ґрунтувався на тому, що виділяв змістові ролі елементів тексту:

- агент — «одухотворений» ініціатор подій;
- контрагент — сила, проти якої спрямована дія;
- об'єкт — предмет або сутність, яка рухається або змінюється, положення якої є об'єктом уваги;
- місце — фізичне тіло, яке відчуває вплив з боку ініціатора;
- адресат — особа, відносно якої відбувається дія;
- пацієнт — річ, на яку впливає ефект дії;
- результат — річ, яка виникає в результаті дії;
- інструмент — стимул або безпосередня причина події;
- джерело — місце, з якого щось направлене.

Семантичні ролі Філмора допомагають враховувати під час аналізу тексту семантичну структуру речення завдяки попередньому опису моделі світу в термінах «ролей».

Інший цікавий підхід було запропоновано Р. Шенком. Він полягає у представленні ідентичних речень, які мають різну структуру, за допомогою єдиної «концептуальної конструкції». Метод ґрунтується на тому, що однією з відмінностей структур при однаковому змісті є різноманітність граматичної ролі слів, які описують одну і ту саму ситуацію.

Заслуговує на увагу також метод, оснований на використанні семантичних графів. Один з таких графів може складатися з вершин та ребер зі знаками аг (бути аргументом), об (бути об'єктом), чс (бути часом). Наприклад, «Ми аналізуємо текст» (рис. 1). Проте, в цьому випадку побудова графу може бути дещо ускладнена,

оскільки речення, що розглядається, може бути об'єктом або причиною іншого речення.

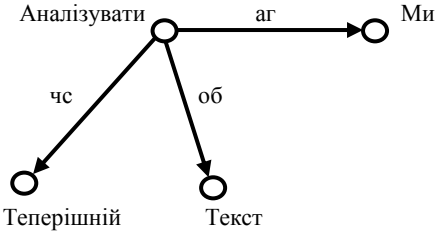

Рис. 1. Семантичний граф 1

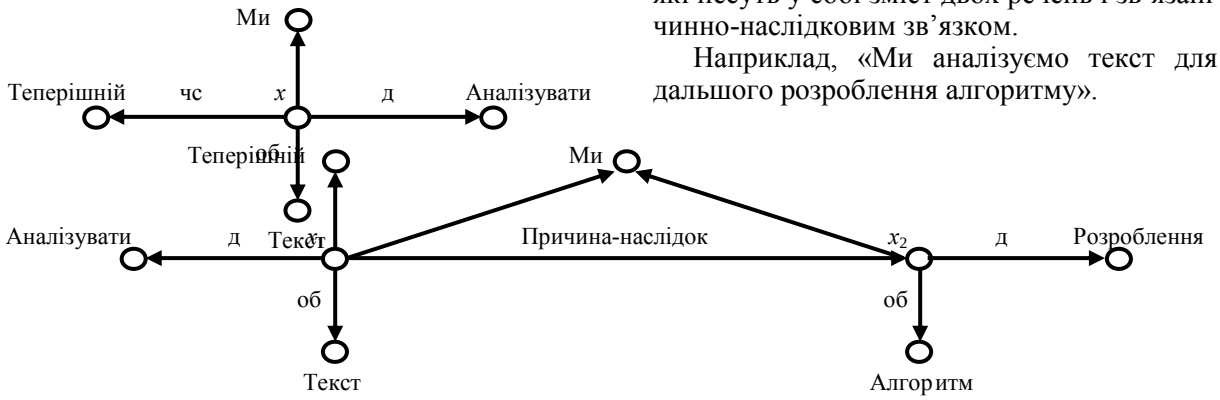

Рис. 3. Семантичний граф 3

Рис. 2. Семантичний граф 2

У графі на рис. 2 введено вершину x, яка відповідає всьому значенню речення, і ребро д — буде дія.

На графі рис. 3 введено дві вершини x_1 і x_2, які несуть у собі зміст двох речень і зв'язані причинно-наслідковим зв'язком.

Наприклад, «Ми аналізуємо текст для подальшого розроблення алгоритму».

Мета дослідження

Метою цього дослідження є:

- аналіз стану проблеми в сфері методів обробки текстів, що використовуються для побудови систем пошуку, класифікації та порівняння електронних документів;
- з'ясування механізму реалізації основних принципів змістового аналізу текстів;
- формування сучасного уявлення про текстовий контент-аналіз;
- пропозиція оптимального методу семантичного аналізу.

Моделі і алгоритми контент-аналізу текстів

Передача змісту в процесі обробки тексту складається з двох компонентів — змістового і граматичного. Сенс речення виникає як результат поєднання елементарних семантичних одиниць відповідно до визначених правил. Інакше кажучи, речення представляє код окресленого змісту, а синтаксис — умовну форму послідовності змістових одиниць, що дає змогу структурувати цей зміст. Головною умовою правильної семантичної інтерпретації тексту є контекст. Будь-яка галузь діяльності відображається конкретними термінами і взаємозв'язками між ними. Таку множину можна розбити на певну кількість типів термінів і типів взаємозв'язків (семантичних одиниць). Кожне речення можна перевести в текст, який складається з ряду термінів і типів зв'язків, без урахування граматичних особливостей, відображаючи кожен термін або зв'язок у певний тип. Цей процес має назву канонізації тексту, а зміст, який при цьому виникає, — канонічним сенсом тексту. Проте повна відмова від граматики не завжди виправдана. Іноді зміст речень визначається прийменниками, відмінками слів, і тому їхнє врахування може значно полегшити семантичний аналіз. Тому вводиться додаткова семантична одиниця — граматична роль лексем або їх частин для зазначення відповідних граматичних ознак мови.

Під час процесу аналізу тексту слід враховувати, що зміст також залежить від специфіки сфери, до якої він належить. Цим викликано впровадження третього типу семантичних одиниць — спеціальних ролей лексем. У загальному випадку канонічний зміст тексту визначається за допомогою значень семантичних одиниць усіх вказаних типів.

Останнім часом набув поширення контент-аналіз (аналіз змісту) текстів. Він передбачає пошук у тексті мовних індикаторів (одиниць аналізу, символів), певних змістових понять (категорій аналізу, термінів), визначення частоти їх уживання, оцінювання співвідношення з іншими одиницями і зі змістом усього твору.

Основними процедурами контент-аналізу текстів є (рис. 4):

- процедура символьного аналізу;
- процедура морфологічного аналізу;

- процедура термінологічного аналізу.

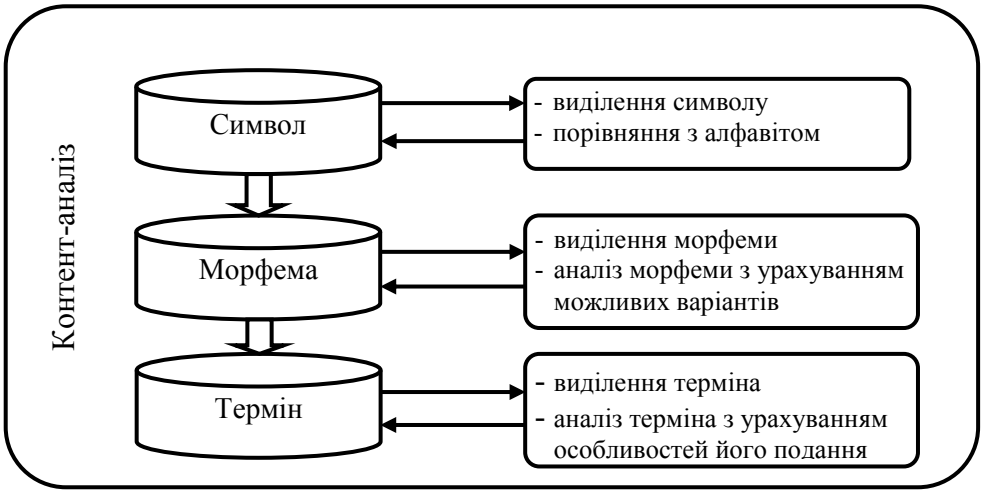

Рис. 4. Структурна схема алгоритму контент-аналізу

Процедура символьного аналізу полягає у порівнянні послідовності символів текстового фрагменту з символами алфавіту. Результат такої обробки переходить на етап морфологічного аналізу. Він проводиться у декілька кроків. На першому компонується підмножина морфем, а на другому — виділяються їх основні характеристики:

- один з символів у морфемі може бути замінений;
- символ може випадати;
- може бути доданий додатковий або зайвий символ;
- символи можуть бути змінені місцями.

Під час етапу термінологічного аналізу необхідно врахувати дві умови. Перша — особливості подання терміну в тексті:

- слова та словосполучення можуть мати різні рід, відмінок, множину;
- між словами у терміні, який складається з декількох слів, можуть стояти інші слова;
- відсутність строгого порядку слів у термінологічному словосполученні.

Другою умовою слід відзначити той факт, що в мовознавстві часто не визнається наявність у мові багатоосновних термінів, які вчені ідентифікують як концептуальні об'єднання. За кількістю компонентів виділяють такі типи термінологічних словосполучень: двокомпонентні, трикомпонентні, чотири-, п'яти- і шестикомпонентні. Аналіз словесних конструкцій термінів дає підставу вважати, що більшість із них двокомпонентні (рис. 5).

Привертає увагу той факт, що нині з'явилася тенденція до збільшення багатокомпонентних структур термінології.

Трикомпонентні словесні конструкції мають структуру, відображену на рис. 6.

Іменник
Прикметник + **іменник**
Дієслово

Рис. 5. Типові лексико-граматичні моделі двокомпонентних термінологічних словосполучень

Іменник + *іменник*
Прикметник + *іменник* **+ іменник**
Дієслово + *іменник*

Рис. 6. Типові лексико-граматичні моделі трикомпонентних термінологічних словосполучень

На практиці в багатьох текстах формуються багатослівні словосполучення, тобто ті, які складаються з чотирьох і більше слів. Вони вживаються для вираження складних понять, кожному з яких відповідає свій термін. Сьогодні спостерігається тенденція до їх збільшення (рис. 7).

Іменник + *іменник* + іменник

Іменник + *іменник* + прикметник **+ іменник**

Прикметник + *іменник* + прикметник

Рис. 7. Типові лексико-граматичні моделі мультикомпонентних термінологічних словосполучень

На основі цих факторів можна розробити характеристику рівня термінів, яка несе в собі інформацію про те, у якій позиції і з якими словами наявні у тексті терміни.

А. Я. Шайкевич висунув гіпотезу, що слова, які зв'язані між собою за змістом, у тексті повинні траплятися недалеко один від одного. І нав-паки, слова, які часто трапляються разом у тексті, — пов'язані між собою за змістом. Цю гіпотезу він використав для виділення семантичних рівнів при аналізі віршованих текстів. Але, якщо під час аналізу віршованих текстів за «одиницю аналізу» можна взяти один рядок, то виділення семантичного поля у звичайних прозових текстах є проблемою. Зрозуміло, що чим більша частина тексту вибирається як «одиниця», тим більшою може виявитися відстань між термінами. Слід зауважити, що семантична подібність слів буде залежати не лише від їх відстані відносно один одного, а й від їхнього граматичного розташування. Для того, щоб спростити процедуру аналізу взаємозв'язку термінів, пропонується виділяти один термін як «домінанту», а інший, який буде траплятися разом з ним, умовно називати «супровідним» (рис. 8). При цьому не слід забувати, що один і той самий термін може виступати як у ролі «домінанти», так і «супровідним». Після того, як зв'язки між окремими парами термінів будуть установлені, слова, які тісно пов'язані один з одним за сенсом, можна об'єднати у семантичні підгрупи тощо доки у підгрупі не з'являться цілі змістові фрази.

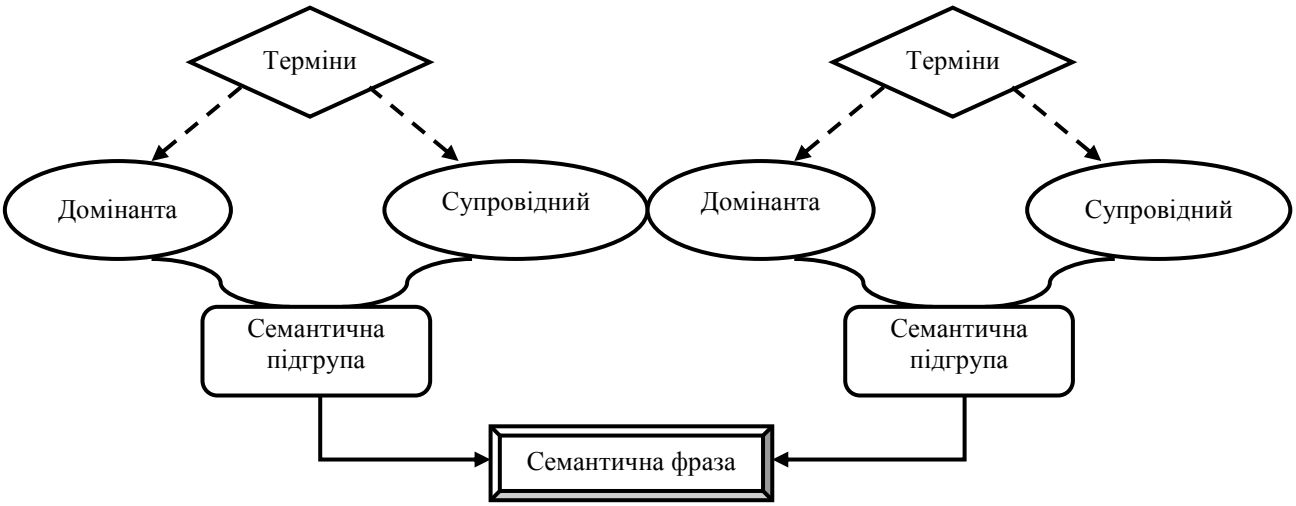

Рис. 8. Схематична модель термінологічного процесу аналізу тексту

Завдяки контент-аналізу можна отримати достатньо об'єктивний результат та зробити зміст тексту вимірюваним і придатним для точного обчислення.

Висновки

Кожен з розглянутих наукових методів має свої технології і технічні засоби. Їх розвиток дає змогу реалізувати у відповідних системах досконаліші функції та наділяє систему елементами інтелектуальної діяльності людини. Можна припустити, що найперспективнішим серед них є метод контент-аналізу, який дає змогу представити та проаналізувати семантичну інформацію тексту.

У статті подано модель семантичного аналізу текстів, яка може бути використана у комп'ютерно-лінгвістичній системі обробки текстової інформації. Запропонована модель дає змогу врахувати семантичні зміни (зміни порядку слів, зміни множини/роду/відмінку, вставляння слів у середину фрази), виразити характеристики тексту на рівні символів, морфем, термінів у вигляді послідовності множин.

Розглянута модель подання тексту ґрунтується на різних семантичних рівнях, що дає змогу зменшити розмір інформації, яка зберігається. Це збільшує швидкість подальшого аналізу за рахунок зменшення інформаційного навантаження тексту і кількості елементів, що обробляються.

Узагальнюючи проведені дослідження, можна зробити висновок, що лише чітке співвідношення контенту з окресленими особливостями мови дасть змогу перейти семантичному аналізу тексту на вищий рівень, більш наближений до елементів людського інтелекту.

ЛІТЕРАТУРА

1. *Анисимов А. В.* Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык / В. А. Анисимов. — К.: Наук. думка, 1991. — С. 208.

2. *Белоногов Г. Г., Быстров И. И., Козачук М.В. Новоселов А. П., Хорошилов А. А* Автоматический концептуальный анализ текстов: Сб. «Научно-техническая информация», серия 2, № 10, ВИНИТИ, 2002.

3. *Люгер Д. Ф.* Искусственный интеллект. Стратегии и методы решения сложных проблем / Н.И. Галаган (пер. с англ.). — 4-е изд. — М.; С.Пб.; К. : Издательский дом «Вильямс», 2005. — 863 с.

4. *Марченко О. О.* Моделювання семантичного контексту при аналізі текстів на природній мові. Вісник Київського університету. Сер. фіз.-мат. науки. — 2006. — № 3. — С. 230—235.

5. *Шемакин Ю. И., Романов А. А*. Компьютерная семантика. — М. : Научно-образовательный центр «Школа Китайгородской», 1995. — 344 с.

Стаття надійшла до редакції 12.01.10.