УДК 004.5

МОРФОЛОГІЧНИЙ АНАЛІЗ СЛОВА НА ОСНОВІ АСОЦІАТИВНО-СТАТИСТИЧНОГО ПІДХОДУ

Квєтний Р.Н., *д.т.н., професор,*
Бісікало О.В., *к.т.н., доцент,*
Кравчук І.А.
Вінницький національний технічний університет

Работа посвящена актуальной задаче компьютерной лингвистики. Подход к анализу морфем построен исходя из общей проблемы понимания смысла естественного языка. Решение задачи получено на основе введения грамматики, распознающей слова из тезауруса образов.

Ключові слова: *отримання знань з природно-мовних текстів, асоціативно-статистичний метод, семантична мережа, формальна граматика, тезаурус образів*

Work is devoted to an actual task in the field of computer linguistics. The approach to the analysis of morphemes constructed from a general problem of understanding of a natural language sense. The task decision is received on the basis of introduction of the grammar what identifies the words from the thesaurus of images.

Key words: *receipt of knowledges from naturally linguistic phototypographs associative statistical method, semantic network, formal grammar, thesaurus of appearances.*

Вирішальну роль в побудові сучасних систем автоматизованої обробки текстової інформації відіграють методи комп'ютерної лінгвістики. Однією з ключових наукових проблем комп'ютерної лінгвістики є проблема моделювання процесу розуміння змісту текстів (переходу від природно-мовних (ПМ) конструкцій до формалізованого подання їх сенсу). Побудова лінгвістичних процесорів, що забезпечують спілкування користувачів з інтелектуальними автоматизованими системами природною мовою або мовою, наближеною до природної, є на сьогодні надзвичайно важливим завданням [1, с.13-14].

Семантико-синтаксичний аналіз текстів в системах автоматизованої обробки інформації проводиться з метою формалізованого подання їх структури – виділення в тексті окремих лексем (змістовних одиниць) і встановлення зв'язків між ними [1, с. 97].

Створення автоматичного модуля синтаксичного аналізу є на сьогодні одним із найактуальніших завдань, вирішення якого дозволило б досягти високого рівня формалізації мовних структур у різноманітних прикладних цілях: від створення систем автоматичного розпізнавання мови до пошукових систем в мережі Інтернет.

Метою даного дослідження є розробка методики побудови бази знань з морфології на основі виявлення в ПМ конструкціях носіїв мови певних закономірностей.

Задачу отримання знань з природно-мовних (ПМ) текстів традиційно розв'язують за допомогою синтаксично-семантичного моделювання мови. Перспективні результати також демонструють моделі, які побудовані на основі запитів до бази даних, та досить популярні останнім часом онтології. Проте всі ці підходи вимагають значної попередньої роботи із залученням висококваліфікованих експертів [2].

Асоціативно-статистичний підхід до аналізу текстів відображає природний шлях отримання онтологій. Він призводить до побудови асоціативної семантичної мережі, яку складають пов'язані між собою значимі слова тексту, що найбільш часто зустрічаються. Узагальнено, ці слова відповідають поняттям об'єкту, методу, якості методу або об'єкту, які несуть основне семантичне навантаження [3]. Для кожного такого поняття формується набір асоціативних зв'язків до/від інших понять, у сполученні з якими воно зустрічалося в тексті.

Потужна фізіологічна основа асоціативно-статистичного підходу полягає у природному накопиченні асоціацій між образами та закріпленні рефлексів шляхом повторень. При цьому слова можна вважати вербальними ознаками мовних образів, тоді кожному вузлу словника мовних образів (інакше – тезауруса образів) відповідає список слів-ознак.

Розглянемо можливості подібного підходу з [2], який будемо вважати найближчим аналогом. Для кожного поняття з тексту формується

набір асоціативних (значеннєвих) зв'язків, тобто список інших понять, у сполученні з якими воно зустрічалося в реченнях тексту. При цьому вважається, що чим частіше зустрічаються разом два поняття в реченнях тексту, тим вище ймовірність того, що вони зв'язані за змістом.

Статистичні дані про зв'язки понять у тексті, їх імовірнісний розподіл, дозволяють оцінити внесок цих понять у загальний зміст тексту і, таким чином, ранжувати теми за інформативністю. У підсумку кожній темі семантичної мережі привласнюється так звана тематична вага. Максимальне значення тематичної ваги (рівне 100) відповідає ключовій (найважливішій) темі тексту. Близьке до нуля значення ваги теми показує, що вона лише мимохіть згадана в тексті, і в ньому мало відомостей, що відносяться до даної теми. Зв'язки між парами тем, у свою чергу, також мають характеристики – ваги зв'язків (від 0 до 100). Велике значення ваги зв'язку від однієї теми до іншої, близьке до 100, указує на те, що більша частина інформації в тексті, що стосується першої, стосується в той же час і другої теми – перша тема майже завжди викладається в контексті другої. Мале значення ваги відбиває той факт, що перша тема слабко зв'язана з другою (викладається незалежно від неї). Зв'язок між парою тем мережі завжди двосторонній, однак, зв'язок від першої теми до другої не завжди має таку ж саму вага, що й зворотна – від другої до першої. Таке розходження у вагах може вказувати на те, що одна тема є підтемою іншої.

Зрештою, семантична мережа являє собою тематичний індекс аналізованих текстів, що використовується для пошуку документів за темою. По кожній з тем мережі формується набір зв'язних фрагментів тексту – цитат, що відносяться до відповідної теми, які представляють тематичне резюме (реферат) тексту. Крім того, виконується ранжування цих фрагментів по вагам (від 0 до 100), які відбивають їх інформативність для відповідної теми. Загальне резюме тексту формується з найбільш інформативних фрагментів за ключовими темами документа [4].

При аналізі тексту можна скористатися семантичною мережею, побудованою на базі інших текстів, прийнятих за еталонні. Наприклад, якщо за текстами певної предметної області побудована семантична мережа, її можна використовувати для фільтрації інформації з інших текстів. У цьому випадку в текстах виявляються тільки ті теми, які містяться в еталонній мережі, і резюме будуються тільки по цих темах. Порівняння семантичних мереж різних текстів дозволяє встановити ступінь їхньої значеннєвої подібності, що може використовуватись для автоматичної класифікації документів за заданими рубриками чи пошуку документів за заданим текстом.

В основу схожого до аналогу запропонованого підходу до отримання морфологічних знань з тексту покладено ідеї Тьюринга та Хомського щодо формалізації природних шляхів опанування дитиною рідної мови [5]. Такий підхід є зворотнім до існуючих систем морфологічного аналізу, базу знань яких побудовано шляхом формалізації закладених експертами-філологами множини правил [6]. Головна відмінність запропонованого підходу від аналогу полягає у зміщенні акценту з частоти появи понять на частоту появи зв’язків між поняттями як головної інформаційної ознаки тексту.

Вхідну інформацію для розв’язку задачі морфологічного аналізу маємо у вигляді словника мовних образів *Thesaurus*. В основу словника покладено реляційну базу даних синтагматичних і парадигматичних зв’язків, що накопичуються в результаті синтаксично-статистичного аналізу текстів [3].

Формальним результатом аналізу тексту є орієнтований граф, що узагальнює зв’язки між окремими словами речення. У кожній вершині графу міститься статистично отриманий список спільнокореневих слів, які відповідають одному мовному образу. З метою морфологічного аналізу мовних образів застосуємо контекстно-вільну граматику, що розпізнає слова

$$G = (V_T, V_N, S, P), \qquad (1)$$

де V_T – алфавіт термінальних (основних) символів (морфем);

V_N – алфавіт нетермінальних символів (метазмінних), причому $V_T \cap V_N = \varnothing$;

S – стартовий символ;

P – скінчений набір правил $j \to y$, де $j \in V_N$, а $y \in F(V)$ – довільні слова вільної напівгрупи слів над алфавітом $F = V_T \cup V_N$.

Розглянута граматика (1) породжує мову L як множину $L(G)$ всіх правильних слів у граматиці G. Будемо вважати, що до складу множини V_N входять w *(СЛОВО)*, x *(ПРЕФІКС)*, y *(КОРІНЬ)* та z *(СУФІКС)*. Тоді маємо $V_N = \{w, x, y, z, v\}$, а $S = w$.

Постановка задачі: на основі аналізу словника образів *Thesaurus* необхідно визначити V_T – множину морфем, або таких частин слова, що мають певні словотворчі значення.

Аналізуючи морфологічну структуру слова флексійних мов на основі формально-структурного підходу [6], визначимо базові правила з P для слів w та y:

1. x є префіксом w ($x > w$), якщо $w = xy$,
2. z є суфіксом w ($z < w$), якщо $w = yz$.

В загальному випадку слово з одним префіксом, суфіксом та закінченням має вигляд $w = xyz$, хоча не виключаються варіанти слова з кількома префіксами та/або суфіксами. В нормальній формі Бекуса (БНФ) склад слова флексійної мови можна задати таким виразом: *<СЛОВО>::=<КОРІНЬ>|<ПРЕФІКС><СЛОВО>|<СЛОВО><СУФІКС>*.

У прийнятих раніше позначеннях формалізуємо правила визначення множин коренів, префіксів та суфіксів як

$$Root(L) = \{y \mid (\exists y \in L)\}, \quad (2)$$

$$Pref(L) = \{x \mid (\exists y \in L) x > w\}, \quad (3)$$

$$Suf(L) = \{z \mid (\exists y \in L) z < w\}, \quad (4)$$

де $Root, Pref, Suf \subseteq V_T$.

За допомогою правил (2)÷(4) забезпечується можливість перевірки належності довільного слова w до множини $L(G)$ всіх правильних слів у граматиці (1). Отже, задача зводиться до обґрунтування конструктивних алгоритмів попереднього відокремлення морфем з V_T на основі словника *Thesaurus*. Концептуальні ідеї побудови таких алгоритмів полягають в наступному:

1. Алгоритм $ROOT$ – визначає корінь як найбільш довгу спільну послідовність символів для всіх слів списку, що відповідають окремому мовному образу.
2. Алгоритм $PREF$ – генерує гіпотезу про префікс як спільну послідовність символів, що зустрічаються перед коренями декількох мовних образів.
3. Алгоритм SUF – генерує гіпотезу про суфікс як спільну послідовність символів, що зустрічаються після коренів декількох мовних образів.

Відсутність в системі експертних знань з філології та статистичний характер підходу вимагають обов'язкову зовнішню перевірку отриманих гіпотез щодо морфем з V_T. Таку перевірку пропонується проводити на основі порівняння результатів морфологічного аналізу, наприклад, російськомовного варіанту *Thesaurus* з відомим продуктом Диалинг [7] – морфологічним словником, розробленим колективом під керівництвом А.В. Сокирко для наукових експериментів.

Розглянемо на прикладі морфологічний аналіз кількох отриманих з словника *Thesaurus* множин слів (T_i), що відповідають різним мовним образам:

T_1 = *{модель, змоделювати, моделювання, модельний};*

T_2 = *{мова, безмовний, мовлення, мовець, мовити, мовний, перемовини, розмова, розмовний};*

T_3 = *{друк, друкарський, роздрукувати, друкар, друкарня, друкарка, передрукувати};*

T_4 = *{ключ, ключовий, включити, включення, переключення}.*

Аналізуючи множини, визначаємо для кожної з них спільну для всіх слів частину – корінь. Для наших множин це будуть відповідно *"модел", "мов", "друк", "ключ"*. На основі вищезгаданих означень, визначаємо інші значущі частини кожного слова (префікси, суфікси та закінчення).

Аналізуючи частини слів, що передують кореню, визначаємо множину префіксів {*"з", "без", "роз", "пере", "в"*}.

Аналізуючи і порівнюючи складові слів після кореня, виділяємо суфікси у вигляді множини {*"юв", "ат", "ен", "н", "ан", "ськ", "ув", "к", "ов", "ар"*}.

Зауважимо, що відсутність окремого алгоритму визначення закінчень слів може призвести до невірного тлумачення структури слова. Під введене означення суфікса як послідовності символів, що зустрічаються після коренів мовних образів, підпадають і закінчення. Так, для розглянутого прикладу помилковим є виділення в якості суфіксів множини закінчень {*"я", "и", "ий", "а"*}. Для виключення неправильних гіпотез необхідно проводити порівняння з наперед визначеними закінченнями або звертатися до вчителя з метою зовнішньої верифікації.

Визначені на основі словника образів *Thesaurus* морфеми вносимо до множини V_T граматики (1). Отже, поставлену задачу розв'язано та отримано формальні засоби для статистичного визначення V_T.

Висновки та перспективи. Формальне визначення V_T і частини правил з P на основі асоціативно-статистичного аналізу множини текстів є важливим компонентом лінгвістичного процесору, що має здатність до самонавчання. До перспективних задач розвитку даного напрямку досліджень можна віднести:

a) програмну реалізацію алгоритмів $ROOT$, $PREF$ та SUF на основі граматики (1);
b) побудову алгоритму END, який генерує гіпотезу про закінчення як спільну послідовність символів, що зустрічаються після

кореня або суфікса двох та більше мовних образів;

c) визначення базових правил з P на основі формально-семантичного підходу [8], який за думкою ряду відомих авторів переважає формально-структурний.

ЛІТЕРАТУРА

1. Белоногов Г. Г. Компьютерная лингвистика и перспективные информационные технологии. – М.: Русский мир, 2004. – 248 с.
2. Стюарт Рассел. Искусственный интеллект: современный подход, 2-е изд.: Пер. с англ. / Рассел Стюарт, Норвиг Питер. – М.: Вильямс, 2006. – 1408 с.
3. Бісікало О.В. Концептуальні основи моделювання образного мислення людини / О.В. Бісікало. – Вінниця: ПП Балюк І.Б., ВДАУ, 2009. – 163 с.
4. Обработка знаний: технологии анализа и поиска текстовой информации [Електронний ресурс] / Компания Гарант-Парк-Интернет. – Режим доступу: http://www.olap.ru/basic/knowledge.asp
5. Бісікало О.В. Асоціативно-статистичний метод морфологічного аналізу слів / О.В. Бісікало, І.А. Кравчук // Інформаційні технології та комп'ютерна інженерія: тези доповідей міжнар. наук.-практ. конф., м. Вінниця, Україна, 19-21 травня 2010 р. – Вінниця, 2010. – С. 110–111.
6. Морфология и словообразование [Електронний ресурс] / Слово. Филология. – Режим доступу: http://www.portal-slovo.ru/philology/37380.php
7. Русский морфологический словарь [Електронний ресурс] / Компания АОТ (Автоматическая Обработка Текста). – Режим доступу: http://www.aot.ru/docs/rusmorph.html
8. План морфемного анализа [Електронний ресурс] / Культура письменной речи. Русский язык. – Режим доступу: http://gramma.ru/RUS/?id=2.17

Квєтний Р.Н., д.т.н., професор, Вінницький національний технічний університет.

Бісікало О.В., к.т.н., доцент, Вінницький національний технічний університет.

Кравчук І.А., студент, Вінницький національний технічний університет.