УДК 004.5

О. В. Бісікало, канд. техн. наук, доц.;
І. А. Кравчук, студ.

АНАЛІЗ МОРФОЛОГІЧНОЇ СТРУКТУРИ СЛОВА НА ОСНОВІ АСОЦІАТИВНО-СТАТИСТИЧНОГО ПІДХОДУ

Розглянуто актуальну задачу комп'ютерної лінгвістики — автоматизацію морфологічного аналізу слова. Описано підхід, що базується на природному накопиченні асоціацій між образами та закріпленні рефлексів шляхом повторень. Подано алгоритм роботи програмного забезпечення, створеного на основі запропонованого підходу.

Найпоширенішою формою подання знань є природно-мовні (ПМ) тексти. Тому задача ефективної обробки таких текстів є актуальною. ПМ засоби спілкування людини з ЕОМ постійно розвиваються, залишаючись одним із найперспективніших способів побудови користувацького інтерфейсу до складних інформаційних систем [1].

Обробка ПМ текстів має широке прикладне застосування: розробка лінгвістичного процесору, що забезпечує спілкування з користувачами природною мовою або мовою, наближеною до природної, а також інтелектуальні пошукові системи, створення словників, автоматизоване реферування тексту тощо [2].

Задача отримання корисної інформації з ПМ тексту являє собою ідентифікацію в ньому певних елементів. На вході системи отримання інформації маємо текст на природній мові, на виході — певні заповнені структури даних, що дозволяють проводити подальшу автоматизовану або ручну обробку інформації [3].

Лінгвістичні процесори складаються з компонентів, що один за одним обробляють вхідний текст. Вихід одного процесора є входом іншого. Виділяються такі компоненти:

— графематичний аналіз (виділення слів, цифрових комплексів, формул тощо),
— морфологічний аналіз (побудова морфологічної інтерпретації слів вхідного тексту),
— синтаксичний аналіз (побудова синтаксичної інтерпретації речень вхідного тексту),
— семантичний аналіз (побудова семантичного графу тексту) [4].

Метою морфологічного аналізу є визначення морфологічної інформації словоформ для використання її на подальших етапах обробки ПМ текстів.

Постає задача розробки методики побудови бази знань з морфології на основі виявлення в ПМ конструкціях певних закономірностей.

Для розв'язання поставленої задачі запропоновано асоціативно-статистичний підхід до отримання знань з ПМ тексту, що базується на природному накопиченні асоціацій між образами та закріпленні рефлексів шляхом повторень.

Перевагою запропонованого підходу перед іншими методами розв'язання поставленої задачі (синтактико-семантичний підхід, онтології, запити до бази даних) є відсутність необхідності попередньої роботи із залученням висококваліфікованих експертів.

Вхідною інформацією для розв'язання поставленої задачі є словник мовних образів Thesaurus.

Задано контекстно-вільну граматику для розпізнавання слів:

$$G = (V_T, V_N, S, P), \tag{1}$$

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 $\phi \to \psi$, де $\phi \in V_N$, а $\psi \in F(V)$ — довільні слова вільної напівгрупи слів над алфавітом $F = V_T \cup V_N$.

До складу множини *VN* входять *w* (слово), *x* (префікс), *y* (корінь) та *z* (суфікс) [5].

Аналізуючи словник мовних образів, необхідно статистично визначити V_T — множину морфем (частини слова, що мають певне значення).

Морфеми поділяються на два основних типи — кореневі (корені) та афіксальні (афікси).

Корінь є основною значущою частиною слова. Кореневі морфеми можуть утворювати слово як разом з афіксами, так і без них.

Афікс — допоміжна частина слова, що приєднується до кореня і служить для словотвору та вираження граматичних значень. Найбільш поширені два типи афіксів — префікси, що стоять перед коренем, та постфікси, що стоять після кореня. В залежності від значення, що виражається постфіксами, вони розділяються на суфікси (мають деривацийне або словотвірне значення) і флексії або закінчення (мають реляційне значення, тобто вказують на зв'язок з іншими членами речення) [6]. В українській мові, яка належить до флексійних мов, використовуються як префікси, так і постфікси (суфікси та закінчення).

Проаналізувавши структуру слова української мови, визначаємо такі етапи морфологічного аналізу:

— визначення коренів;

— визначення префіксів;

— визначення суфіксів;

— визначення закінчень.

Формалізоване представлення префіксів та суфіксів є таким:

— x є префіксом w $(x \triangleright w)$, якщо $w = xy$;

— z є суфіксом w $(z \triangleleft w)$, якщо $w = yz$ [5].

Робота програмного забезпечення, розробленого для реалізації запропонованого підходу, складається з двох етапів:

— виокремлення морфем з мовних образів;

— морфологічний аналіз для заданого користувачем слова.

Схема роботи першого етапу наведена на рис. 1.

Вхідні дані для роботи програмного забезпечення (словник мовних образів) подаються у вигляді реляційної бази даних. В розробці було використано базу *SQLite*. Окрім мовних образів, база даних містить окремі таблиці для коренів, префіксів, суфіксів та закінчень.

На першому кроці завантажується словник мовних образів з бази даних в робоче середовище програми.

Слід відмітити, що відсутність ефективного алгоритму виділення закінчень за невеликих обсягів словника мовних образів *Thesaurus* може привести до виокремлення їх як суфіксів. Щоб цього уникнути, закінчення можуть задаватися до початку роботи програми експертом (зовнішнім вчителем) та завантажуються одразу ж після завантаження мовних образів. Після цього у кожному слові виокремлюються закінчення шляхом співставлення з заданими морфемами та відсікаються, залишаючи тільки основу слова.

Наступний крок — визначення коренів, аналізуючи окремо кожний мовний образ. Корінь визначається як найдовша спільна послідовність символів для всіх слів списку, що відповідають окремому мовному образу.

Рис. 1. Виокремлення морфем з мовних образів

Після визначення коренів розпочинається виокремлення префіксів та суфіксів. На цих кроках проводиться аналіз всіх мовних образів. Префікси та суфікси визначаються як спільна послідовність символів, що зустрічається перед (для префіксів) коренем та після (для суфіксів) кореня декількох мовних образів.

Визначені морфеми заносяться до відповідних таблиць бази даних.

Схему роботи другого етапу роботи програмного забезпечення — морфологічний аналіз для

заданого користувачем слова — показано на рис. 2.

На другому етапі роботи програми здійснюється аналіз слова, введеного користувачем за допомогою розробленого інтерфейсу. Результат аналізу відображається користувачу – виділені частини слова відділяються прямою лінією, а в подальшому планується додатково позначати всі типи морфем власними кольорами. Інтерфейс користувача показаний на рис. 3.

Рис. 2. Морфологічний аналіз слова, заданого користувачем

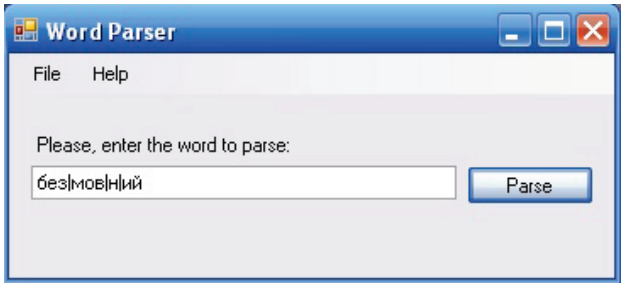

Рис. 3. Інтерфейс користувача

Таким чином, на першому етапі роботи програмного забезпечення статистично визначаються морфеми. Так ми формуємо множину V_T граматики (1). Отже, поставлену задачу розв'язано та отримано формальні засоби для статистичного визначення V_T.

Висновки

Запропоновано метод автоматизації морфологічного аналізу слова, що базується на асоціативно-статистичному підході до отримання знань з тексту. Розв'язано задачу введенням граматики, що розпізнає слова з тезауруса образів. Наведено алгоритм роботи програми, що виконує морфологічний аналіз слів на основі асоціативно-статистичного підходу. У розвиток запропонованого підходу планується розробити модулі верифікації отриманих результатів морфологічного аналізу за допомогою зовнішнього вчителя та на основі порівняння з відкритими ресурсами корпусної лінгвістики.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Компьютерный синтаксический анализ: описание моделей и направлений разработок / [Г. Д. Карпова, Ю. К. Пирогова, Т. Ю. Кобзарева, Е. В. Микаэлян] // Итоги науки и техники: серия «Вычислительные науки». — 1991. —Т. 6. — М. : ВИНИТИ.

2. Белоногов Г. Г. Компьютерная лингвистика и перспективные информационные технологии / Г. Г. Белоногов. — М. : Русский мир, 2004. — 248 с.

3. Анализа текста [Електронний ресурс] / История компьютера. — Режим доступа : http://chernykh.net/content/view/1038/1121/.

4. Автоматическая обработка текста [Електронний ресурс] / Компания АОТ (автоматическая обработка текста). — Режим доступа : http://www.aot.ru/.

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. Реформатский А. А. Введение в языковедение / А. А. Реформатский. — М. : Аспект Пресс, 1996. — 536 с.

Рекомендована кафедрою автоматики та інформаційно-вимірювальної техніки

Стаття надійшла до редакції 11.03.11
Рекомендована до друку 12.04.11

Бісікало Олег Володимирович — доцент кафедри автоматики та інформаційно-вимірювальної техніки;
Кравчук Ірина Анатоліївна — студентка Інституту магістратури, аспірантури та докторантури.

Вінницький національний технічний університет, Вінниця