УДК 004.9

О.В. Бісікало, І.А. Кравчук

Вінницький національний технічний університет, Вінниця

КОНЦЕПТУАЛЬНІ АЛГОРИТМИ ВИОКРЕМЛЕННЯ МОРФЕМ ДЛЯ РЕАЛІЗАЦІЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ

У статті зроблено огляд існуючих напрямків морфологічного аналізу як складової частини лінгвістичної обробки текстової інформації та запропоновано алгоритми для виокремлення коренів, суфіксів, префіксів та закінчень, що реалізуються при виконання морфологічного аналізу слів з використанням асоціативно-статистичного підходу до отримання знань. Вхідними даними для запропонованих алгоритмів виокремлення морфем є тезаурус мовних образів. Введено показник "коефіцієнт входження" для статистичної оцінки визначення суфіксів та префіксів, що дозволило коректно визначати у слові кілька префіксів чи суфіксів.

Ключові слова: *обробка природномовних текстів, морфологічний аналіз, тезаурус мовних образів.*

Вступ

Постановка проблеми. Інформація відіграє все більшу роль в усіх сферах людської діяльності.

Виділяють два основних підходи до обробки документів: статичний, лінгвістичний.

В основі статистичного лежить гіпотеза, що слова, які найбільш часто зустрічаються в тексті, будуть відображати його зміст. Але статистичні методи представляють текст лише як просту множину слів, що є недостатнім для відображення його змісту.

Лінгвістичні методи дозволяють подолати цю перешкоду, оскільки вони представлять текст як набір слів, що розташовані в певній заданій послідовності [1].

Лінгвістичний аналіз включає в себе наступні етапи:

– графематичний;

– синтаксичний;

– морфологічний;

– семантичний.

Метою морфологічного аналізу є визначення морфологічних характеристик слова та його основної словоформи.

Таким чином, розробка методу морфологічного аналізу як складової лінгвістичної обробки тексту, який би використовував статистичний аналіз, поєднуючи переваги обох вищерозглянутих підходів до обробки природно-мовних текстів, є актуальним завданням.

Аналіз останніх досліджень та публікацій. В наш час проводиться велика кількість досліджень методів морфологічного аналізу.

На даний час виділяють кілька напрямків морфологічного аналізу:

– на основі поділу словоформи на основу та афікси з наступною перевіркою;

– на основі використанні інформації, що міститься в кінцевих сполученнях літер;

– на основі створення універсальних математичних моделей.

Перший напрямок морфологічного аналізу моделює класичну схему аналізу шляхом поділу словоформи на основу і афікси з наступною перевіркою на спільність закінчення з основою, що залишається. До цього напряму можна віднести модель морфологічного аналізу Г.Г. Бєлоногова.

В основі морфологічної моделі Г.Г. Бєлоногова [2] лежить флективний аналіз слів, що базується на розбитті слів російської мови на флективні класи. Хоча розробки Г.Г. Бєлоногова проводилися для російської мови, отримана модель може бути застосована і до української мови.

Морфологічний аналіз починається з пошуку лексеми у словнику готових словоформ. При успішному завершенні отримується код флективного класу, що відповідає даній лексемі і вказує на частину мови та синтаксичну функцію словоформи. У протилежному випадку словоформа піддається флективному аналізу, який включає в себе наступні етапи:

– ідентифікація морфем словоформи (послідовна перевірка можливостей входження в аналізовану словоформу кореня, суфікса, закінчення і префікса);

– визначення флективного класу словоформи (отримання коду флективного класу з таблиць сумісності кореня зі словотворчими афіксами);

– привласнення словоформі морфологічної інформації [3].

Модель морфологічного аналізу Ю.П. Шабанова-Кушнаренко [4] формалізує процеси російської мови за допомогою мови алгебри кінцевих предикатів. На основі такого універсального формалізму існує можливість математично описати будь-який аспект морфології російської, а значить і української, мови.

Ще один напрямок морфологічного аналізу намагається побудувати більш адекватні морфологічні

моделі. Цей напрямок викликаний прагненням подолати обмеженість існуючих алгоритмів морфологічного аналізу. Відомо, що вони орієнтувалися на тексти певної тематики і тому не повністю враховували всі особливості морфології. В межах даного напрямку створюються універсальні математичні моделі в формі відкритої системи рівнянь, що дозволяють шляхом обчислення здійснювати нормалізацію словоформ, отримання граматичної інформації і синтез словоформ [5].

Формулювання цілей дослідження. Метою даного дослідження є розробка концептуальних алгоритмів виокремлення коренів, префіксів, суфіксів та закінчень, що реалізовуються при виконанні морфологічного аналізу на основі асоціативно-статистичного підходу.

Основний матеріал дослідження

Авторами в [6] запропоновано підхід до морфологічного аналізу, що розроблений на основі аналогії з побудовою семантичної мережі під час аналізу текстової інформації.

Запропонований підхід до морфологічного аналізу передбачає аналіз вхідних даних у вигляді тезауруса мовних образів, що являє собою набір множин понять, що позначають спільне поняття.

Результатом такого морфологічного аналізу є побудова бази знань з морфології на основі виявлення в природномовних конструкціях певних закономірностей. Таким чином, на виході морфологічного аналізатора отримуються набори морфем проаналізованих вхідних слів. Отримані морфеми й будуть використовуватись для безпосереднього морфологічного аналізу будь-яких слів.

Для реалізації запропонованого методу, на основі аналізу структури слова флексійних мов розроблено наступні алгоритми:

1. Алгоритм для виокремлення закінчень.
2. Алгоритм для виокремлення коренів.
3. Алгоритм для виокремлення суфіксів.
4. Алгоритм для виокремлення префіксів.

Загальна схема отримання коренів, префіксів та суфіксів для одного слова з вхідних даних представлена на рис. 1. Розглянемо детально кожен з алгоритмів виокремлення морфем.

Алгоритм для виокремлення закінчень. Закінчення – частина слова, що служить для зв'язку слів у реченні, смислового навантаження воно не несе. Тому необхідним для виконання морфологічного аналізу є перелік можливих закінчень слів.

На першому етапі морфологічного аналізу кожне слово з вхідних даних перевіряється на наявність закінчення. Перевірка здійснюється шляхом порівняння кінцевих символів слова з переліком можливих закінчень. У випадку, коли слово містить закінчення, закінчення відсікається, і на подальших етапах аналізу використовується тільки основа слова.

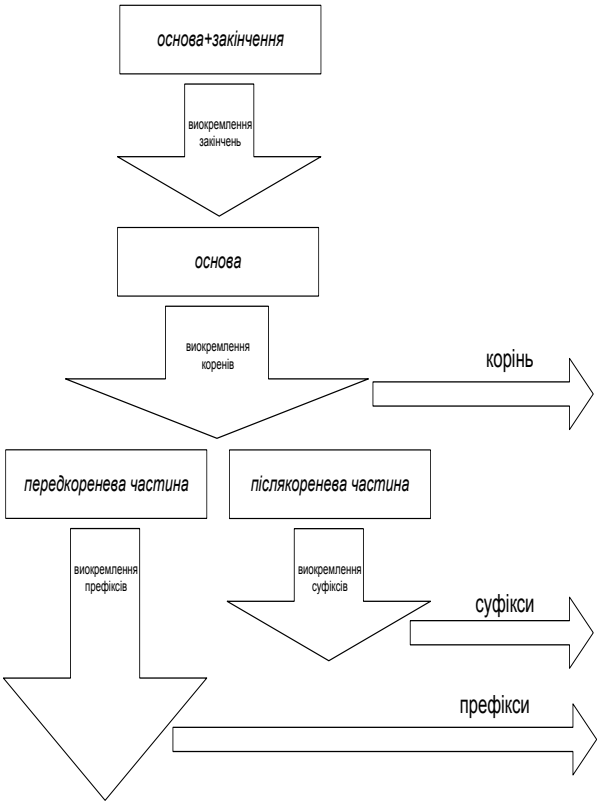

Рис. 1. Загальна схема виокремлення морфем на прикладі одного слова

Алгоритм для виокремлення коренів. Оскільки вхідні дані представлені у вигляді мовних образів,кожен з яких являється, по суті, набором спільнокореневих слів, то виокремлення коренів відбувається шляхом порівняння слів у межах одного мовного образу. Корінь визначається як найбільш довга спільна послідовність символів для всіх слів списку, що відповідають окремому мовному образу.

Алгоритм для виокремлення суфіксів та префіксів. В основу алгоритмів для виокремлення суфіксів та префіксів покладено прості правила:

1. Префікс – частина слова, що знаходиться перед коренем.

2. Суфікс – частина слова, що знаходиться після кореня.

Слово може мати кілька суфіксів та префіксів.

Для перевірки багатосуфіксальності та багатопрефіксальності використовуються статистична оцінка знайдених морфем, після якої остаточно визначаються афікси.

Результатом виокремлення коренів зі слів є набір частин слів, що знаходяться перед коренем (*preroot*) та після кореня (*postroot*). Суфікси визначаються на основі аналізу післякореневих частин сло-

ва.

Введемо поняття "коефіцієнт входження", що показує кількість входжень певної післякореневої або передкореневої частини.

Кожній післякореневій частині привласнюється початкове значення коефіцієнта входження, що дорівнює одиниці. Потім відшукується початкова післякоренева частина (*initial_postroot*), що складається з найменшої кількості слів. Після цього здійснюється порівняння решти післякореневих частин на входження до їх складу початкової. У випадку наявності в певній післякореневій частині (*current_postroot*) початкової (*initial_postroot*), коефіцієнт входження початкової частини слова збільшується на 1. При цьому початкова післякоренева частина (*initial_postroot*) вилучається з проаналізованої (*current_postroot*). Після цього обирається нова початкова післякоренева частина, процедура аналізу повторюється.

Суфікси визначаються за коефіцієнтами входження післякореневих частин. Розглянемо, як це відбувається, на прикладі. Візьмемо три післякореневі частини: *postroot1* = "*ськ*", *postroot2* = "*к*", *postroot3* = "*ськ*". В результаті аналізу отримано наступні потенціальні суфікси з відповідними коефіцієнтами входжень: "*к*" (3), "*сь*" (2), "*ськ*" (2). Перша післякоренева частина має коефіцієнт входження 2, і складається з двох потенційних суфіксів ("*к*" і "*сь*"), коефіцієнти входження яких 3 і 2 відповідно. Оскільки коефіцієнт входження потенційного суфікса "*сь*" співпадає з коефіцієнтом входження самої після кореневої частини "*ськ*", очевидно, що він не є суфіксом. Відповідно, "*к*" є суфіксом, оскільки його коефіцієнт входження більший, ніж у самої післякореневої частини.

Префікси визначаються на основі аналізу передкореневих частин слова за алгоритмом, аналогічним до виокремлення суфіксів.

Висновки

Запропоновані алгоритми для виокремлення морфем (корінь, префікс, суфікс, закінчення) у слові дозволяють формувати базу знань з морфології на основі аналізу тезауруса мовних образів та проводити морфологічний аналіз будь-яких слів на основі закладених простих правил побудови слова.

Використання статистичної оцінки знайдених потенційних префіксів та суфіксів та введення показника "коефіцієнт входження" дозволило коректно виокремлювати у слові кілька суфіксів та префіксів.

Список літератури

1. Калиниченко А.В. Сущность проблемы анализа текста в полнотекстовых поисковых системах. Подходы и пути решения / А.В. Калиниченко // Журнал научных публикаций аспирантов и докторантов. – 2010. – № 5.

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

3. Файн В.С. Распознавание образов и машинное понимание естественного языка / В.С. Файн. – М.: Наука, 1987. – 173 с.

4. Шабанов-Кушнаренко Ю.П. Теория интеллекта. Математические средства / Ю.П. Шабанов-Кушнаренко. – Х.: Изд-во Харьковского университета, 1984. – 144 с.

5. Найханова Л.В. Методы и алгоритмы трансляции естественно-языковых запросов к базе данных в SQL-запросы / Л.В. Найханова, И.С. Евдокимова. – Улан-Удэ: Изд-во ВСГТУ, 2004. – 148 с.

6.Квєтний Р.Н. Морфологічний аналіз слова на основі асоціативно-статистичного підходу / Р.Н. Квєтний, О.В. Бісікало, І.А. Кравчук // Вісник Черкаського державного технологічного університету. – 2010. – № 3. – С. 132–135.

Надійшла до редколегії 14.04.2011

Рецензент: д-р техн. наук, проф. В.М. Дубовой, Вінницький національний технічний університет, Вінниця.

КОНЦЕПТУАЛЬНЫЕ АЛГОРИТМЫ ВЫДЕЛЕНИЯ МОРФЕМ ДЛЯ РЕАЛИЗАЦИИ ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ ОБРАБОТКИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

О.В. Бисикало, И.А. Кравчук

В статье дается обзор существующих направлений морфологического анализа как составной части лингвистической обработки текстовой информации и предложены алгоритмы для выделения корней, суффиксов, приставок и окончаний, которые реализуются при выполнении морфологического анализа слов с использованием ассоциативно-статистического подхода к получению знаний. Входными данными для предлагаемых алгоритмов выделения морфем является тезаурус речевых образов. Введен показатель "коэффициент вхождения " для статистической оценки определения суффиксов и префиксов, что позволило корректно определять в слове несколько префиксов или суффиксов.

Ключевые слова: *обработка естественно текстов, морфологический анализ, тезаурус речевых образов.*

CONCEPTUAL ALGORITHMS OF ALLOCATION OF MORPHEMES FOR THE IMPLEMENTATION OF INFORMATION TECHNOLOGY PROCESSING NATURAL LANGUAGE TEXT

O.V. Bisikalo, I.A. Kravchuk

This paper provides an overview of current trends of morphological analysis as a part of the linguistic processing of textual information and algorithms for the retrieving of roots, suffixes, prefixes and ends, which are realized for the morphological analysis of words using the associative and statistical approach to knowledge, is proposed. The input data for proposed algorithms is thesaurus of language images. "Coefficient occurrence" index for statistical evaluation of the definition of prefixes and suffixes that will properly define the word a few prefixes or suffixe, is introduced.

Keywords: *processing of natural texts, morphological analysis, thesaurus of language images.*