ДЕРЖАВНА СЛУЖБА ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ УКРАЇНИ

УКРАЇНА **(19) UA** **(11) 75250** **(13) U**

(51) МПК
G06F 17/27 (2006.01)

(12) ОПИС ДО ПАТЕНТУ НА КОРИСНУ МОДЕЛЬ

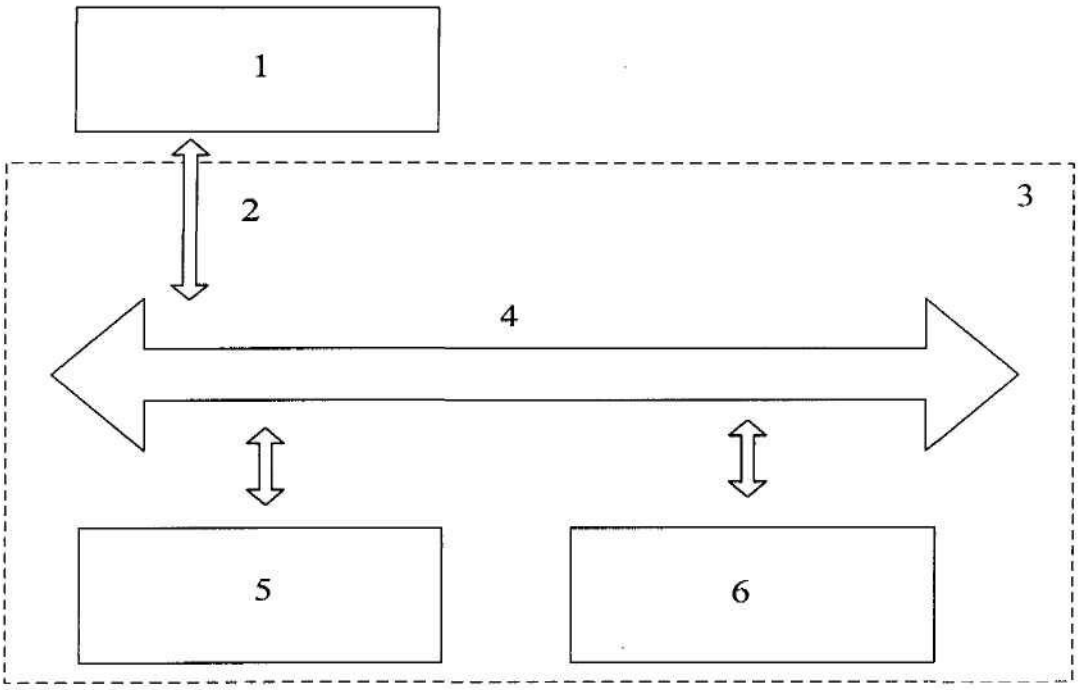
(21) Номер заявки: **u 2012 05834**
(22) Дата подання заявки: **14.05.2012**
(24) Дата, з якої є чинними права на корисну модель: **26.11.2012**
(46) Публікація відомостей про видачу патенту: **26.11.2012, Бюл.№ 22**

(72) Винахідник(и):
Бісікало Олег Володимирович (UA),
Кравчук Ірина Анатоліївна (UA)
(73) Власник(и):
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ,
Хмельницьке шосе, 95, м. Вінниця, 21021 (UA)

(54) СПОСІБ МОРФОЛОГІЧНОГО АНАЛІЗУ НА ОСНОВІ АСОЦІАТИВНО-СТАТИСТИЧНОГО ПІДХОДУ ДО ОТРИМАННЯ ЗНАНЬ

(57) Реферат:

Спосіб морфологічного аналізу на основі асоціативно-статистичного підходу до отримання знань, в якому визначають можливі корені кожної основи, та використовують корінь слова разом з інформацією про префікси та суфікси. Вхідними даними для процесу є словник мовних образів з записами у вигляді наборів спільнокореневих слів, який розміщений на зовнішньому електронному носії інформації. Словник надходить в блок морфологічного аналізу по шині вхідних даних. Обмін даними з процесором здійснюють через канал зчитування та передачі даних.



UA 75250 U

Корисна модель належить до інформаційних технологій і може використовуватись для виконання морфологічного аналізу в системах лінгвістичної обробки природномовних текстів.

Відомий морфологічний аналізатор, що виконує морфологічний аналіз вхідного слова (патент США № 5323316, м. кл. G06F 015/38, опубл. 21.06.1994).

Спосіб полягає у тому, що формальні лінгвістичні засоби аналізують вхідне слово, починаючи з першого символу і далі по одній літері. Засоби, що знаходять основу, працюють, починаючи від початку вхідного слова і продовжуючи через вхідне слово по одній літері, щоб знайти основу. Засоби, що знаходять основу, починають своє завдання розпізнавання на n-й літері вхідного слова, пропускаючи перші n-1 літери вхідного слова, де n є цілочисельною змінною, яка приймає різні значення в залежності від розміру вхідного слова. Засоби, що знаходять суфікс, продовжують працювати з частиною вхідного слова, залишаючи після основи одну літеру, щоб знайти суфікси. Засоби, що знаходять основу та суфікс, використовують дворівневу морфологічну модель для виконання своїх функцій. Дворівнева морфологічна модель є похідною від набору правил правопису, які перетворюють рядок в лексичний рядок, і це здійснюється одним автоматом.

Вказаний спосіб має той недолік, що він не виокремлює префікси вхідних слів, а лише основу та суфікси.

Найбільш близьким є спосіб автоматизованого морфологічного аналізу структури слова (патент США № 5251129, м. кл. G06F 17/27, опубл. 10.05.1993).

Спосіб полягає у створенні інформації про форму кожного слова тексту шляхом розбиття слова на потенційні пари основа-суфікс і основа-префікс. Тоді, використовуючи правила правопису, визначають можливі корені кожної основи. Для кожного можливого кореня варіант слова отримують зі словника. Для кожного слова отримують словотворчу інформацію для кожного афікса. Потім словотворчу інформацію представляють у словниковій статті. Для кожної пари використовують оптимізаційні операції для визначення ймовірності правильності кожної пари основа-суфікс і основа-префікс. Спосіб забезпечує чітке визначення у текстовому представленні відношень між словами, які є похідними від спільного кореня, шляхом використання кореневого слова разом з інформацією, що міститься в повній формі, в якій слово з'являється, тобто, використання префікса, суфікса і кореня.

Вказаний спосіб має той недолік, що морфологічний аналіз ефективно виконується лише для тих слів, інформацію про які попередньо розміщено у словнику, що вимагає залучення кваліфікованих експертів-лінгвістів.

В основу корисної моделі поставлено задачу підвищення ефективності виконання морфологічного аналізу слів тексту.

Поставлена задача вирішується тим, що в способі морфологічного аналізу на основі асоціативно-статистичного підходу до отримання знань, в якому вхідними даними є попередньо побудований словник мовних образів з записами у вигляді наборів спільнокореневих слів, що розміщений на зовнішньому електронному носії інформації, яку обробляє блок морфологічного аналізу, формують базу знань з морфології на основі правил, що отримані шляхом попереднього аналізу структури слова флексійних мов, попередньо відділяють основу слова, після чого визначають можливі корені кожного слова з вхідного словника мовних образів та префікси і суфікси в блоці морфологічного аналізу, обмін даними здійснюють через канал зчитування та передачі даних, суфікси та префікси визначають за допомогою коефіцієнтів входження потенційного суфікса/префікса в послідовності символів, отриманих за результатами аналізу, що зберігаються в оперативному запам'ятовувальному пристрої та постійному запам'ятовувальному пристрої.

Спосіб морфологічного аналізу на основі асоціативно-статистичного підходу до отримання знань має програмно-апаратну реалізацію у вигляді блока морфологічного аналізу, схема якого представлена на кресленні.

Схема для реалізації способу морфологічного аналізу на основі асоціативно-статистичного підходу складається з зовнішнього носія інформації 1, шини вхідних даних 2, персонального комп'ютера 3 та його складових - каналу зчитування та передачі даних 4, центрального процесору 5, запам'ятовувальних пристроїв 6 - оперативний запам'ятовувальний пристрій (ОЗП) та постійний запам'ятовувальний пристрій (ПЗП).

Спосіб здійснюється наступним чином.

На вхід каналу зчитування та передачі даних 4 надходить словник мовних образів, що містить записи у вигляді наборів спільнокореневих слів, та список можливих закінчень, що розміщені на зовнішньому електронному носії інформації 1. Частина словника мовних образів має вигляд:

1	мова	мовний	безмовний	мовленнєвий	мовлення	
2	вода	підводний	безводний	надводна	підводна	водний
3	друк	роздрукований	друкарня	друкарський	друкарка	друкар

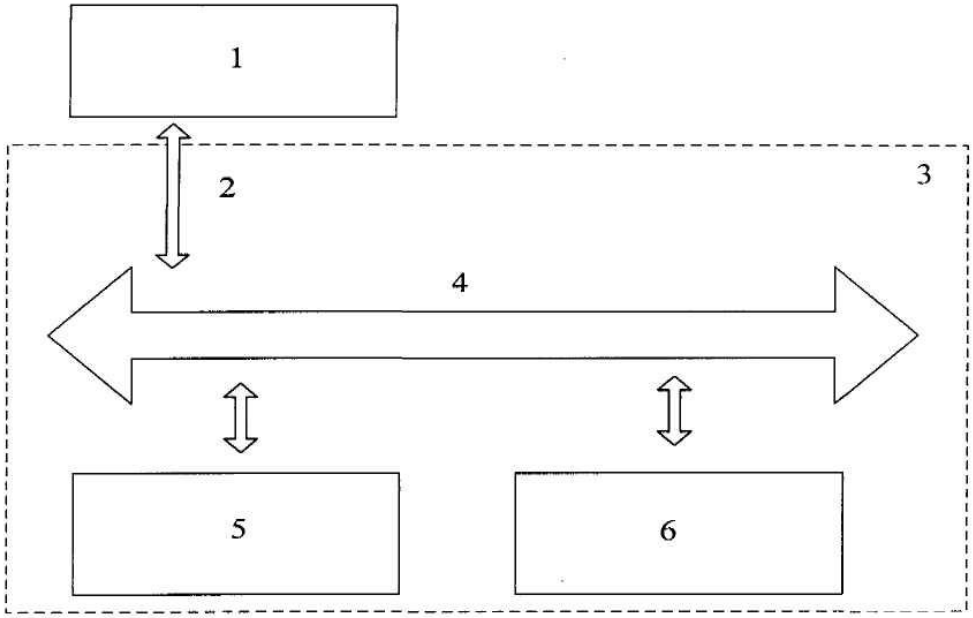
Список закінчень має вигляд: "ий, а, я, …"

У блоці морфологічного аналізу на рівні центрального процесору 5 кожний елемент словника порівнюють з переліком можливих закінчень, що надходять по шині вхідних даних 2. У випадку співпадіння останніх символів слова зі словника з можливим закінченням, закінчення відсікають, після чого отримане слово без закінчення є основою і зберігається в блоці пам'яті 6 для подальшого аналізу. В блоці морфологічного аналізу на рівні центрального процесору 5 порівнюють слова без закінчень (основи), що вже зберігаються в блоці пам'яті 6 з метою визначення найбільш довгої спільної послідовності символів для всіх слів списку, що відповідають окремому мовному образу. Визначена послідовність символів є коренем. На даному етапі слово як послідовність символів розбивають на 3 частини - символи, що розташовані до кореня, символи, що розташовані після кореня, які передають в блок пам'яті 6 для зберігання до подальшої обробки, та, власне, корінь. Потім в блоці морфологічного аналізу на рівні центрального процесору 5 визначають суфікси слів. Для цього символи, що розташовані після кореня, передають з блока пам'яті 6. Кожному післякореневому набору символів привласнюють початкове значення коефіцієнта входження, що, по замовчуванню, дорівнює одиниці. Коефіцієнт входження показує кількість входжень певного набору післякореневих символів. Після чого визначають початкову послідовність символів як найкоротшу послідовність. Після цього у блоці аналізу на рівні центрального процесору 5 порівнюють решту післякореневих наборів символів з початковою послідовністю. У випадку співпадіння символів початкової послідовності з символами поточної послідовності, коефіцієнт входження для початкової послідовності збільшують на 1. При цьому символи початкової послідовності видаляють з проаналізованої. Після цього вибирають нову наступну початкову післякореневу послідовність символів з блока пам'яті 6 і повторюють процедуру аналізу. Після виконання аналізу всіх післякореневих послідовностей, що містяться у блоці пам'яті 6, за коефіцієнтами входження визначають суфікси. Суфіксом визначають послідовності символів, коефіцієнт входження для якого більший, ніж у самої післякореневої послідовності, до складу якої вони входять.

Префікси визначають в блоці аналізу на рівні центрального процесору 5, використовуючи передкореневі послідовності символів, що зберігаються в блоці пам'яті 6, аналогічно до визначення суфіксів.

ФОРМУЛА КОРИСНОЇ МОДЕЛІ

Спосіб морфологічного аналізу на основі асоціативно-статистичного підходу до отримання знань, в якому, використовуючи закладені правила, визначають можливі корені кожної основи та використовують корінь слова разом з інформацією, що міститься в повній формі, в якій слово з'являється, тобто, використовують префікси та суфікси, який **відрізняється** тим, що вхідними даними для морфологічного аналізу є попередньо побудований словник мовних образів з записами у вигляді наборів спільнокореневих слів, що розміщений на зовнішньому електронному носії інформації та надходить в блок морфологічного аналізу по шині вхідних даних, базу знань з морфології формують на основі правил, що отримані шляхом попереднього аналізу структури слова флексійних мов, попередньо відділяючи основу слова, після чого визначають можливі корені кожного слова з вхідного словника мовних образів та префікси і суфікси в блоці морфологічного аналізу на рівні центрального процесору, обмін даними з яким здійснюють через канал зчитування та передачі даних; суфікси та префікси визначають за допомогою коефіцієнтів входження потенційного суфікса/префікса в послідовності символів, отримані за результатами аналізу, які зберігають в оперативному запам'ятовувальному та постійному запам'ятовувальному пристроях.



Комп'ютерна верстка Л.Литвиненко

Державна служба інтелектуальної власності України, вул. Урицького, 45, м. Київ, МСП, 03680, Україна

ДП "Український інститут промислової власності", вул. Глазунова, 1, м. Київ – 42, 01601