УДК 004.9

О.В. Бісікало, І.А. Кравчук

ВИКОРИСТАННЯ МОРФОЛОГІЧНОГО АНАЛІЗУ В ЗАДАЧАХ ДИСТАНЦІЙНОЇ ОСВІТИ

Анотація. В статті запропоновано підхід до використання у системах дистанційної освіти методу морфологічного аналізу, що базується на статистичному визначенні морфем зі словника мовних образів.
Ключові слова: аналіз тексту, морфологічний аналіз, отримання знань

Постановка проблеми

Дистанційна форма освіти набуває все більшої актуальності через необхідність забезпечення якісної підготовки, перепідготовки та підвищення кваліфікації максимальної кількості фахівців для всіх регіонів України з мінімальними витратами часових та фінансових ресурсів. Можливості дистанційної освіти, що базується на сучасних інформаційних і комунікаційних технологіях навчання й підвищення кваліфікації, відкривають нові перспективи для підвищення ефективності освітнього процесу.

В дистанційному навчанні знаходять застосування системи для автоматизованої обробки текстової інформації. Такі інформаційні системи використовують технології штучного інтелекту, що забезпечують оброблення множини текстів з метою формування бази знань предметної області [1]. Ефективність навчання в такому випадку зростає завдяки тому, що система в автоматизованому режимі попередньо відбирає для пізнавального процесу з великого обсягу інформації найбільш значущу, ключову.

Автоматична обробка текстів на природній мові розв'язує питання аналізу та синтезу на основних мовних рівнях – морфологічному, синтаксичному та семантичному. Морфологічний аналіз вхідних словоформ є початковим етапом аналізу природно-мовної конструкції (ПМК). Результати морфологічного аналізу є базовою основою для всіх наступних етапів автоматичного аналізу та синтезу тексту, проте в умовах навчальної системи всі етапи оброблення природномовного матеріалу повинні підкорятися дидактичним завданням, що

формулюються зазвичай на семантичному рівні. Отже, існує проблема інтеграції різнорівневих методів аналізу та синтезу ПМК з пріоритетом семантики предметної області з метою розв'язання задач пізнавальної діяльності.

Аналіз останніх досліджень та публікацій

В 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 якої лежить флективний аналіз слів, що базується на розбитті слів російської мови на флективні класи [2]. 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.

Всі етапи тісно взаємопов'язані між собою, оскільки невдале завершення другого і третього етапів свідчить про некоректне розбиття аналізованої словоформи на морфеми [3].

 ISSN 1562-9945

Другий напрямок морфологічного аналізу використовує інформацію, що міститься в кінцевих сполученнях літер (така інформація отримується в результаті попередньої статистичної обробки словника). Цей шлях також дає досить хороші для практичних цілей результати [4].

Третій 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].

Однією з таких моделей є модель Ю. П. Шабанова-Кушнаренко, що формалізує процеси російської мови за допомогою мови алгебри кінцевих предикатів. На основі такого універсального формалізму існує можливість математично описати будь-який аспект морфології російської мови [6].

Формулювання цілей дослідження

Виходячи з проведеного аналізу, метою дослідження є обґрунтування методів морфологічного аналізу, що були б найбільш близькими до природного способу отримання знань людиною, зокрема розуміння принципів побудови ПМК. Такі методи забезпечують перший етап аналізу текстової інформації, що складає основу бази знань для розв'язання задач дистанційного навчання.

Викладення основного матеріалу дослідження

До основних методів морфологічного аналізу відносять:

- методи з декларативною орієнтацією (на основі словника),
- методи з процедурною орієнтацією (на основі правил).

Для методів з декларативною орієнтацією характерна наявність повного словника всіх можливих словоформ для кожного слова з приписаною морфологічною інформацією. При цьому кожна словоформа забезпечується повною і однозначною морфологічною інформацією, куди входять як постійні, так і змінні морфологічні параметри [7]. Можна вважати, що в даному методі відсутній морфологічний аналіз як такий, оскільки завдання морфологічного аналізу в такому

випадку зводиться до пошуку потрібної словоформи в словнику та переписування з словника морфологічної інформації [5].

Перевагою декларативного методу є висока швидкість проведення морфологічного аналізу, а також точність результатів аналізу відносно всіх словоформ української мови, що внесені до словника.

Недоліком декларативного методу є великі затрати пам'яті, пов'язані з тим, що у кожного слова досить велика кількість словоформ. Інформація, що зберігається в словниках, є надлишковою. Ще однією проблемою є морфологічний аналіз слів, що не містяться в словнику словоформ.

При використанні методу з процедурною орієнтацією модуль морфологічного аналізу містить набір правил морфологічних перетворень. Кожне слово розділяється на основу і афікс (закінчення і, можливо, суфікс), словник містить тільки основи слів разом з посиланнями на відповідні рядки в таблиці можливих афіксів. Основний критерій при розбитті слова на основу і афікс – основа повинна залишатися незмінною у всіх можливих словоформах даного слова [8].

Перевагою процедурних методів є зменшення об'єму інформації, яку необхідно зберігати для виконання морфологічного розбору слова. Велика кількість слів української мови має одні й ті ж афікси, які не потрібно дублювати при зберіганні словника афіксів. Саме тому об'єм словника, що використовується при процедурних методах є значно меншим, ніж обсяг повного словника словоформ, що використовується в декларативних методах.

Але процедура морфологічного аналізу при цьому ускладнюється, оскільки зі словника основ необхідно по черзі вибирати всі основи, що збігаються з початковими літерами слова, що аналізується, і для кожної такої основи перебирати всі можливі для неї афікси. Аналіз вважається успішним у випадку точного збігу деякого варіанту «основа + афікс» з словом, що аналізується. При цьому, як правило, постійні морфологічні параметри визначаються основою слова, а змінні – афіксом [7].

Процедурний метод передбачає попередню систематизацію морфологічних знань про природну мову і розробку алгоритмів присвоєння морфологічної інформації окремій словоформі.

Складність природної мови призводить до того, що жоден з описаних методів не може охопити її повністю. Для української мови

 ISSN 1562-9945

відомо більше тисячі правил словотвору з безліччю виключень, що робить створення повного набору правил досить складним. Постійний розвиток мов і великий розмір словників робить неможливим «чисте» використання декларативного підходу. Тому в більшості сучасних систем використовують словник і набір правил для оброблення слів, що не містяться в словнику.

Іноді комбінований метод виділяють окремими методом морфологічного аналізу. За цим методом на першому етапі проводиться пошук по словнику словоформ, як при декларативному методі, і у випадку успішного пошуку аналіз на цьому завершується. В протилежному випадку використовується словник основ і процедурний метод аналізу [5].

Запропонований авторами метод морфологічного аналізу, який побудовано на використанні асоціативно-статистичного підходу до отримання знань [9], належить до процедурних методів. Особливістю методу є використання тезауруса мовних образів як онтології предметної області [10], що дає змогу для розв'язання задачі морфологічного аналізу закласти множину простих правил щодо побудови слів в українській мові. Реалізовано тезаурус за допомогою відношень

$$\mathrm{RE} = \left\{ \begin{array}{l} \mathbf{Im}\,\mathrm{age}, \mathrm{Assoc} - \mathrm{Twice}, \mathrm{Construct}, \mathrm{Event}, \\ \mathrm{Interrogative} - \mathbf{Pr}\,\mathrm{onoun}, \mathrm{Link}, \mathrm{Text}, \mathrm{Words}, \mathrm{Role} \end{array} \right\} \qquad (1)$$

де *Image* – мовні образи, *Assoc-Twice* – асоціативні пари образів, *Construct* – складові простих оповідних речень (синтагм), *Event* – події, *Interrogative-Pronoun* – питальні займенники, *Link* – типи асоціативних зв'язків, *Text* – тексти навчального контенту, *Words* – вербальні ознаки мовних образів, *Role* – роль мовного образу у події.

Надамо оцінку інформаційної збитковості отриманих в результаті застосування запропонованого методу відношень *Re*. З цією метою проведемо порівняння задіяних інформаційних ресурсів з відомими підходами.

Оскільки переважна більшість вхідної інформації для формування навчального контенту представляє собою множину текстових файлів, то потрібний для їх зберігання обсяг інформації є пропорційним загальній кількості слів n_v в усіх цих файлах. Запропонований підхід передбачає визначення n мовних образів, довжина яких пропорційна середній довжині слова, причому, в загальному випадку $n_v \geq n$, а при $n_v \to \infty$ за рахунок ефекту дублювання слів $n_v >> n$. За-

гальний обсяг інформації у Бт для відношень бази даних *Re* оцінимо з таких міркувань:

- загальна кількість записів у відношеннях *Interrogative-Pronoun*, *Link*, *Text*, *Role* дорівнює $n_{const} << n$, отже всі вони додають до БД обмежену деякою константою *E* [Бт] кількість інформації;
- кількість записів у відношенні *Image* дорівнює *n*, але в зв'язку з необхідністю зберігання поля з ідентифікаційним кодом та додаткових полів всього інформації потрібно $n \cdot (C' + log_{256} n)$, де C' – деяка константа [Бт];
- аналогічним чином оцінимо кількість інформації у відношеннях *Words*, *Event*, *Construct* як $n \cdot (C'' + log_{256} n)$, де C'' – константа [Бт], $C'' > C'$;
- найбільшу кількість інформації додає до БД відношення $Assoc - Twice$ з кількістю записів n^2, а саме $n^2 \cdot (A + 4 \cdot log_{256} n)$, де A – константа [Бт], $A < C'$.

Отже, сумарний обсяг інформації для відношень БД складає $V_\Sigma = A \cdot n^2 + B \cdot n^2 \cdot log_{256} n + C \cdot n + D \cdot n \cdot log_{256} n + E$, де A, B, C, D, E – константи [Бт]. Зрозуміло, що з моменту досягнення співвідношення $n_v > n^2$ кількість інформації V_Σ БД наблизиться до, а потім стане меншим від обсягу інформації, потрібної для зберігання множини вхідних текстових файлів.

Програмний модуль, що реалізує розроблений метод морфологічного аналізу, може бути ефективно використаний в навчальних системах. На рис.1 представлено отриману експериментальним шляхом залежність правильності визначення морфем від кількості слів *Words*, що формують тезаурус мовних образів.

Результати проведеного експерименту демонструють високу ступінь адекватності морфологічного аналізу навіть для невеликих обсягів текстового матеріалу та здатність розробленого алгоритму самоудосконалювати базу знань з морфології.

 ISSN 1562-9945

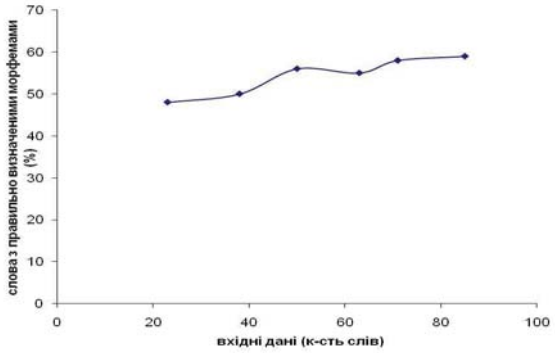

Рисунок 1 – Залежність правильності визначення морфем від кількості вхідних даних для проведення аналізу

Висновки та перспективи подальших досліджень

Використання запропонованого методу морфологічного аналізу в задачах дистанційної освіти може бути ефективним навіть при досить невеликих об'ємах вхідної інформації. З іншого боку, при досягненні певної межі обсягу інформації, потрібної для зберігання множини вхідних текстових файлів навчального контенту, кількість інформації бази даних *Re* наблизиться до, а потім стане меншим від цього первинного об'єму вхідної інформації.

До перспективних задач розвитку окресленого напрямку досліджень можна віднести автоматизацію визначення базових правил сполучення слів природної мови на основі асоціативно-статистичного підходу, що дасть змогу автоматизувати синтаксичний розбір речення.

ЛІТЕРАТУРА

1. Марчук Ю.Н. Термины в автоматической обработке текстов [Електронний ресурс] / Ю.Н. Марчук. – Режим доступу: http://www.sid.ir/fa/VEWSSID/J_pdf/68213831904.pdf.–Назва з екрану.
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. Золотов Е.В. Расширенные системы активного диалога / Е.В. Золотов, И.П. Кузнецов. – М.: Наука, 1982. – 115 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. Шабанов-Кушнаренко Ю.П. Теория интеллекта. Математические средства/Ю.П. Шабанов-Кушнаренко.–Харьков:Изд-во Харьковского унив-та,1984.–144 с.
7. Губин М.В. Влияние морфологического анализа на качество информационного поиска/М.В. Губин, А.Б. Морозов//Труды RCDL-2006. – 2006. – С. 224–228.
8. Дорохина Г.В. Модуль морфологического анализа слов русского языка/Г.В. Дорохина, А.П. Павлюкова//Искусственный интеллект.–2004.–№ 3. – С. 636–642.
9. 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.
10. Бісікало О.В. Методика побудови тезауруса навчальної системи на основі моделі образного мислення / О.В. Бісікало // Искусственный интеллект. – 2008. – № 4. – С. 730–735.