УДК 681.3.06

**М. М. Биков, к. т. н.; доц., Д. Є. Балховський; І. В. Кузьмін, д. т. н., проф.**

# АНАЛІЗ СТАТИСТИЧНИХ ХАРАКТЕРИСТИК МОРФЕМ УКРАЇНСЬКОЇ МОВИ

*Розроблено алгоритм і програмне забезпечення для визначення статистичних і перехідних імовірностей морфем тексту. Проведено теоретичний аналіз проблеми побудови ефективної ієрархічної стратегії розпізнавання тексту, а також запропоновано процедуру побудови оптимального дерева класифікації образів тексту.*

**Ключові слова***: аналіз статистичних характеристик, морфеми, ефективна стратегія розпізнавання, приховані марковські мережі, оптимальна процедура класифікації.*

## Вступ

Практика розпізнавання рукописних символів показує, що використання тільки графічної інформації для їх опису не дозволяє отримати задовільні результати з погляду швидкості й надійності, тому виникає необхідність використання лінгвістичної інформації, яка міститься в текстовому документі [1]. Природно за таку інформацію використати контекстну, в якості якої може слугувати лексична і статистична інформація. Лексичну інформацію зручно використовувати тоді, коли елементами розпізнавання будуть морфеми – мінімальні змістовно розділені частини слова (наприклад, префікс, корінь, суфікс, закінчення). Тоді можна використати статистичну і лексичну інформацію про морфеми для побудови моделей слів тексту у вигляді прихованих марковських мереж (ПММ), що дозволило б застосувати відомі алгоритми розпізнавання на цих мережах. Крім того процедура сегментації на морфеми буде виконуватися значно рідше порівняно з посимвольним розпізнаванням.

## Постановка задачі

Авторами в роботі [2] було розроблено програмні засоби для побудови бази даних українських морфем для забезпечення можливості використання морфологічної інформації в задачі розпізнавання текстового документа. В результаті була побудована база даних, що містить понад 60 000 морфем української мови. Проте використання лише однієї бази даних без використання іншої статистичної інформації не дозволить оптимізувати процес розпізнавання текстів. Тому постає необхідність у вирішенні задачі визначення статистичних характеристик морфем у вигляді їх статичних і перехідних імовірностей. Використовуючи базу даних про статистичні характеристики, процес розпізнавання рукописних та інших текстів, написаних нетипізованими шрифтами, можна значно пришвидшити за рахунок модульної ієрархічної архітектури й апарату прихованих марковських мереж (ПММ). Такі мережі дозволяють після процедури сегментування тексту на морфеми та розпізнавання чергової морфеми в якості альтернативи наступної морфеми вибрати морфеми з найбільшими перехідними ймовірностями з бази даних. Тобто, зникає необхідність у порівняння графічного зображення морфеми (графеми) з усіма можливими еталонами. Прийняття рішень здійснюється при цьому вибором альтернативи з найбільшою сумарною ймовірністю. Такий підхід введення, обробки та розпізнавання текстів підвищує швидкодію та надійність усього процесу. Для реалізації вказаних ідей у цій роботі вирішуються задачі розробки ефективної стратегії розпізнавання текстового документа під час його введення в комп'ютер та процедури пошуку оптимального класифікатора текстових образів, а також розробки алгоритмів визначення й аналізу статистичних характеристик морфем української мови з метою їх використання на лексичному рівні розпізнавання.

## Теоретичний аналіз проблеми побудови ефективної ієрархічної стратегії розпізнавання тексту

Всякий текстовий документ можна розглядати не тільки як графічне зображення, а і як деякий носій мовної інформації, що використовується для її передачі в тій чи іншій комунікативній системі [3]. З такого погляду графіка тексту опосередкованим чином відображає різні інформаційні рівні, властиві комунікативному акту: прагматичний, семантичний, синтаксичний, лексичний, морфологічний, сигматичний і афективний [1]. Виникає питання: інформацію якого рівня і в якій послідовності потрібно використовувати в автоматизованому процесі введення і розпізнавання текстового документа, щоб отримати максимально можливу швидкість і мінімально можливі помилки й вартість. Для розв'язання цього питання автори в цій роботі пропонують нову технологію електронізації текстових документів, яка поряд з розпізнаванням графічних образів використовує часткове розуміння тексту. При цьому процес введення розглядається як процес взаємодії пристрою введення і мовного тезаурусу комп'ютерної системи розуміння тексту. Під час сканування зображення тексту пристрій введення виділяє чергову ознаку графеми, що належить до того чи іншого інформаційного рівня мови, яка використовується системою для зменшення ентропії про текстову одиницю й звуження кола кандидатів на прийняття рішення (розпізнавання). В роботі [3] автори запропонували формальну постановку задачі оптимізації процесу введення й обробки текстового документа, яка розглядає його у вигляді дерева класифікації текстових образів на різних інформаційних рівнях.

Оптимізація процесу розпізнавання образів текстового зображення здійснюється за інформаційним критерієм ефективності

$$Э_{\text{Р}} = \frac{\text{I}_{\text{Р}}}{C_{\text{Р}}}, \tag{1}$$

запропонованим в [3], де $I_p$ – кількість інформації, яку отримує система розпізнавання й розуміння тексту, визначається з урахуванням ентропійних властивостей текстових образів; $C_{\text{Р}}$ – вартість системи;

$$C_{\text{Р}} = C_{\text{Х}} + C_{\text{К}} \ , \tag{2}$$

де $C_{\text{Х}}$ – складність обчислення ознакового опису образів; $C_{\text{К}}$ – складність обчислень класифікації образів.

Оскільки складність $C_p$ системи розпізнавання є адитивною сумою складностей $C_i$ кожного з ієрархічних рівнів розпізнавання, а інформативність $I_p$ є неспадною функцією ймовірності правильного розпізнавання, то оптимальна стратегія є композицією алгоритмів розпізнавання, що максимізують відношення $\text{I}_i / C_i$ на кожному з рівнів. Послідовність композиції алгоритмів в оптимальній стратегії повинна відповідати послідовності розміщення рівнів дерева класифікації, які відповідають відповідним інформаційним рівням текстового документа.

Рішення проблеми належного вибору коефіцієнта розгалуження дозволяє в значній мірі звузити коло пошуку в оптимізаційній процедурі пошуку оптимального дерева рішень. В роботі [4] показано, що мінімізація сумісної помилки класифікації й часу класифікації дає межі коефіцієнта розгалуження $B_r$, що вибирається під час побудови оптимального дерева рішень:

$$2 \le B_r \le 5 \ . \tag{3}$$

Отже, встановлені властивості дерева рішень дозволяють звузити діапазон пошуку при вирішенні задачі визначення оптимального дерева класифікації текстових образів.

Рішення задачі побудови ефективної стратегії прийняття рішення у вигляді дерева класифікації можна здійснити з допомогою оптимізаційної процедури "керованого пошуку вперед з поверненням" [4]. У цій процедурі критерій (2) керує пошуком такої структури дерева рішень серед усіх можливих, в якій на кожному кроці пошуку вибирається та конфігурація вузлів, яка має найвище значення критерію. Для заданого вузла $\Omega_i^h$ дерева процедура пошуку виконується у вигляді такої послідовності кроків:

1. На основі вибраної ознаки $x^h \in X$ здійснюється одне з можливих розбиттів $\pi^h \in \Pi$ вузла $\Omega_i^h$ на підмножину вузлів-нащадків $\{\Omega_j\}, j = \overline{1,m}$. Ознака $x^h$ вибирається по "матриці розрізнюваності" таким чином, щоб коефіцієнт розгалуження лежав в межах, визначених в (3). Тут $h$ – рівень (висота) дерева класифікації, $X$ – апріорний алфавіт ознак.
2. Обчислюється значення критерію (1) для отриманої конфігурації вузла.
3. Повторюючи пункти 1 і 2, будують інші можливі розбиття і для них обчислюють значення критерія .
4. Визначають конфігурацію, для якої критерій має максимальне значення, і тим самим знаходять оптимальний набір ознак $\overline{B}_r$ для поточного вузла дерева й оптимальний крок алгоритму класифікації.

За "матрицю розрізнюваності" використовують таблицю попарної розрізнюваності графем тексту $w_i$ і $w_j$ за всіма ознаками їх опису з апріорного алфавіту ознак на основі вибраної в просторі ознак відстані $d_{ij}$.

### Аналіз статистичних характеристик морфем

Представимо графему слова на морфемному рівні дерева класифікації у вигляді послідовності ***O*** векторів обсервацій:

$$O = \vec{o}_1, \vec{o}_2, ..., \vec{o}_L \ , \tag{4}$$

де $\vec{o}_l$ – вектор зображення морфеми.

Задача розпізнавання слів тексту в такому випадку може розглядатися як обчислення максимуму правдоподібності

$$arg \ \max_i \{P(w_i / O)\} \ , \tag{5}$$

де $w_i$ є $i$-тим словом словника.

Згідно формули Байеса

$$P(w_i / O) = \frac{P(O / w_i) P(w_i)}{P(O)} \tag{6}$$

найбільш імовірна графема слова в зображенні тексту визначається імовірністю $P(O / w_i)$. Пряма оцінка сумісної умовної ймовірності $P(\vec{o}_1, \vec{o}_2, ..., \vec{o}_L / w_i)$ з корпусу тексту не практикується з причини астрономічної кількості можливих обсервованих послідовностей. У більшості випадків задачу оцінки щільності розподілу умовних імовірностей $P(O / w_i)$ замінюють простішою проблемою оцінки параметрів Марковської моделі генерації тексту $M$. Ця модель представляє автомат з скінченою кількістю станів, при встановленні стану $i$ генерується вектор зображення графеми $\vec{o}_l$ з імовірністю $b_i(\vec{o}_l)$. Крім того, перехід із стану $i$ в стан $j$ описується ймовірністю $a_{ij}$. Вибір найбільш імовірної графеми слова здійснюється шляхом знаходження найбільш правдоподібної послідовності станів:

$$P(O/M) = \max_{X} \left\{ a_{x(0)x(1)} \prod_{l=1}^{L} b_x(\vec{o}_l) a_{x(l)x(l+1)} \right\}, \quad (7)$$

де $b_x(\vec{o}_l)$ – імовірність обсервації вектора зображення морфеми $\vec{o}_l$, $a_{x(l)x(l+1)}$ – імовірність переходу від графеми $\vec{o}_l$ до $\vec{o}_{l+1}$, $X$ – множина станів, яку відтворює модель. З урахуванням просторової локалізації станів у марковській моделі тексту авторами запропонована модифікація цієї моделі, яка полягає в доповненні вимоги (7) вимогою:

$$\sum_{l=1}^{L} P_l(\vec{o}_{(l)}) = P(L), \quad (8)$$

де $P_l(\vec{o}_l)$ – математичне очікування довжини морфеми $\vec{o}_l$, $P(L)$ – математичне очікування довжини графеми слова $w_i$.

## Висновки

Отже, для реалізації на морфемному рівні алгоритму розпізнавання графеми слова тексту (7) необхідно визначити статистичні й перехідні імовірності морфем у тексті й статистичні характеристики довжин морфем і слів (їх зображень). У цій роботі розроблено алгоритм обчислення статичних і перехідних характеристик морфем на основі використання тестового корпусу тексту і розробленої авторами в [5] бази даних морфем української мови. Результати роботи алгоритму фіксуються у вигляді двох матриць, у першій з яких фіксуються статистичні імовірності морфем, у другій – перехідні імовірності. Таблиця 1 демонструє вигляд другої матриці.

Таблиця 1

**Матриця перехідних імовірностей морфем**

| Морфеми / Імовірності | Морфема 1 | Морфема 2 | ... | ... | ... | ... | Морфема N |
|---|---|---|---|---|---|---|---|
| Морфема 1 | 0 | P ($m_1/m_2$) | ... | ... | ... | ... | P ($m_1/m_N$) |
| Морфема 2 | P ($m_2/m_1$) | 0 | ... | ... | ... | ... | P ($m_2/m_N$) |
| ... | ... | ... | 0 | ... | ... | ... | ... |
| ... | ... | ... | ... | 0 | ... | ... | ... |
| ... | ... | ... | ... | ... | 0 | ... | ... |
| ... | ... | ... | ... | ... | ... | 0 | ... |
| Морфема N | P ($m_N/m_1$) | P ($m_N/m_2$) | ... | ... | ... | ... | 0 |

У цій таблиці прийнято позначення: $m_1$, $m_2$, .., $m_N$ – морфеми української мови; N – кількість морфем української мови; P ($m_i/m_j$) – імовірність переходу між і-ою та j-ою морфемами.

Алгоритм створення матриці перехідних імовірностей морфем української мови складається з наступних кроків:

1. Зчитування бази даних (БД) морфем української мови.
2. Зчитування масиву слів з тестового корпусу тексту, на якому буде побудована БД імовірностей.
3. Запуск циклу від першої до останньої морфеми (i = 1; i <= N) , де N – кількість морфем у БД.
4. Запуск внутрішнього циклу від першої до останньої морфеми (j = 1; j <= N) , де N – кількість морфем у БД.
5. Обнулюємо лічильник (k) знайдених і-ої та j-ої морфем, які йдуть одна за одною.
6. Запуск циклу від першого до останнього слова тексту (w = 1; w <= M), де M –

кількість слів у тексті.

7. Пошук у w-ому слові і-ої та j-ої морфеми.
8. Якщо морфеми знайдені та j-та морфема йде за і-ою морфемою, інкрементуємо лічильник k.
9. Повертаємося до пункту 5 для переходу на наступне слово.
10. По завершенні циклу 6 (пройдені усі слова та порахована сума (лічильник k) знайдених в усьому масиві слів і-ої та j-ої морфем, які йдуть одна за одною) визначаємо імовірність переходу між і-ою та j-ою морфемами: $P(m_i/m_j) = k / N$.
11. Записуємо визначену ймовірність $P(m_i/m_j)$ до бази даних.
12. По завершенні циклу 4 (пройдені усі j-морфеми) повертаємося до циклу 3.
13. По завершенні циклу 3 (пройдені усі і-морфеми) формуємо звіт та виходимо з програми.

Таблиця 2

**Приклад визначених імовірностей**

| Морфеми / Імовірності | роз | піз | нав | ан | н | я |
|---|---|---|---|---|---|---|
| роз | 0 | 0,0457 | 0,001 | 0,0255 | 0,0548 | 0,00023 |
| піз | 0,0652 | 0 | 0,0522 | 0,0453 | 0,0985 | 0 |
| нав | 0,0001 | 0 | 0 | 0,0781 | 0,0001 | 0,0012 |
| ан | 0,0001 | 0 | 0,00112 | 0 | 0,0268 | 0,0556 |
| н | 0 | 0 | 0 | 0,123 | 0 | 0,0434 |
| я | 0 | 0,002 | 0 | 0 | 0,0897 | 0 |

## СПИСОК ЛІТЕРАТУРИ

1. Пиотровский Р. Г. Текст машина, человек. — Ленинград: Наука", 1975. – 326 с.

2. Биков М. М. Використання інтелектуальних методів в розпізнаванні символів / М. М. Биков, Д. Є. Балховський, А. Раїмі // Інформаційні технології та комп'ютерна інженерія. – 2007. – № 2 (9). – С. 121 – 125.

3. Нова інформаційна технологія введення і оброблення текстових документів в автоматизованих інформаційно-пошукових системах // Автоматика-2008: доклади XV міжнародної конференції з автоматичного управління, 23 – 26 вересня 2008 р., – Одеса: ОНМА. – 992 с.

4. Биков М. М. Розробка ефективної стратегії прийняття рішень в комп'ютерних інтелектуальних системах / М. М. Биков // Вісник Хмельницького національного технічного університету. – 2005. – Ч.1. – Т. 2, № 2. – С. 22 – 30.

***Биков Микола Максимович*** **–** к. т. н., доцент, професор кафедри комп'ютерних систем управління, тел.: (0432)-598430, e-mail: nmbdean@ksu.vstu.vinnica.ua.

***Балховський Дмитро Євгенович*** **–** аспірант кафедри комп'ютерних систем управління, тел.: (0432)-598222, e-mail: vinbudya@yandex.ru.

***Кузьмін Іван Васильович*** **–** д. т. н., професор кафедри комп'ютерних систем управління, тел.: (0432)-598222, e-mail: nmbdean@ksu.vstu.vinnica.ua.

Вінницький національний технічний університет.