УДК 007.001.362

М. М. Биков, к. т. н., доц.;
Д. Є. Балховський, асп.;
В. В. Ковтун, к. т. н.

ОЦІНКА ІНФОРМАТИВНОСТІ ГРАФЕМ ТЕКСТУ

Досліджено оцінки інформативності розпізнавання та розуміння текстового документу, виділено низку ефективних ознак та характеристик графем української мови, запропоновано технологію введення і оброблення текстових документів, що реалізує оптимальний розподіл процесу розпізнавання тексту між мікропроцесорним пристроєм введення і комп'ютером, вибрано критерій оцінки ефективності системи.

Вступ

Традиційні технології обробки текстових документів в інформаційно-пошукових системах передбачають виконання їх посимвольного розпізнавання під час уведення. З огляду на невисоку швидкодію сучасних засобів сканування і розпізнавння, насьогодні процедура введення тексту найбільшою мірою гальмує процесс електронізації текстів. Підвищення швидкості й точності цієї процедури можливі за рахунок застосування інтелектуальних технологій [1], які під час обробки інформації наслідують механізми діяльності мозку людини. Ці механізми передбачають значною мірою використання мовного тезаурусу мозку на основі значних ресурсів його пам'яті і ієрархічної паралельної обробки інформації. З урахуванням можливостей сучасної обчислювальної техніки, реалізація таких інтелектуальних технологій ускладнюється недостатністю знань про інформативність різних графічних елементів тексту (будемо називати їх графемами) для розуміння текстової інформації. В цій роботі наводяться результати досліджень з оцінки інформативності графем складу, морфеми, слова і літери у тій чи іншій позиції.

Інформативність графем текстового зображення

Дослідження інформативності графем тексту здійснювалось шляхом моделювання процессу уведення тексту у вигляді комунікативного акту між джерелом і приймачем інформації по каналу передачі з шумами. Передача інформації, зашифрованої текстовим документом, здійснюється в комунікативній системі, що включає в ідеалі такі компоненти (рис. 1).

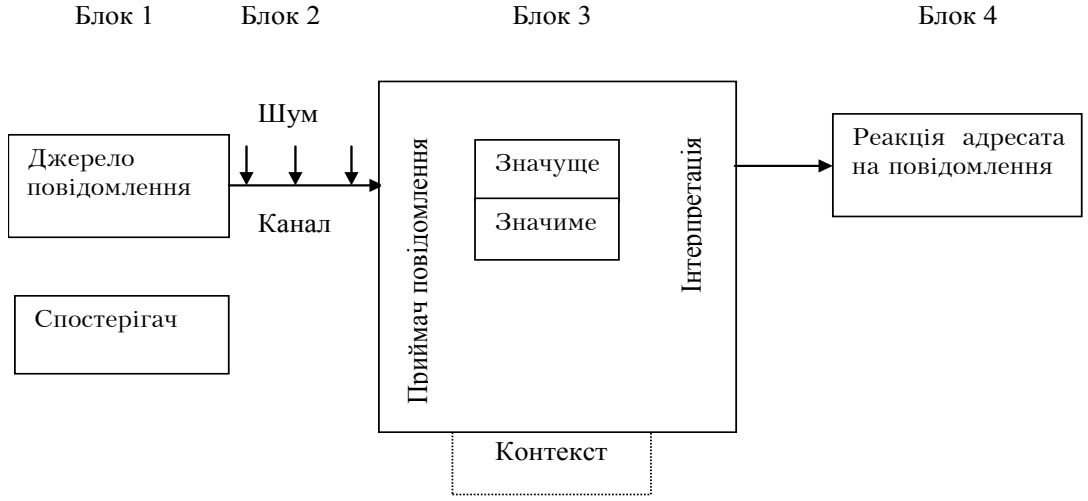

Рис. 1. Схема передачі інформації текстовим документом

Прообразом джерела текстового зображення для моделі є мікропроцесорний пристрій сканування і попередньої обробки, а приймачем — комп'ютерна система розпізнавання і розуміння тек-

сту з її мовним тезаурусом.

В процесі розв'язання проблем, пов'язаних з обробкою текстових документів (наприклад, їх розпізнаванням чи розумінням), в залежності від того, до яких аспектів знака відноситься інформаційна міра, розрізняють такі види інформації [2]:

а) прагматична інформація, що характеризує цінність повідомлення з точки зору тієї мети, яку переслідує в даний момент приймач повідомлення, і з точки зору вибору того рішення, яке є найкращим для досягнення цієї мети ;
б) семантична інформація, яка оцінює відношення, що виникають в процесі семіозису між значущим і десигнатом знака;
в) сигматична (лексична) інформація, що характеризує відношення значимого і денотата, яке описується словом;
г) синтаксична інформація, що є оцінкою приймачем тих обмежень, які накладаються на комбінаторику і частоту використання знаків;
д) морфологічна інформація, яка характеризує взаємозв'язок і відношення між окремими складовими слова;
е) графічна інформація, що міститься в зображенні символів, з яких складаються слова тексту.

У запропонованій моделі приймач сприймає на виході каналу ту чи іншу графему тексту і на основі попередньо прйнятої його частини і використання тезаурусу вгадує наступну графему тексту. Експеримент із вгадування дає можливість оцінити осмислену інформацію, що знаходиться в повідомленні. Ця осмислена інформація кількісно оцінюється не шеннонівською мірою інформації цього ж повідомлення, а різницею значень ентропії, що припадають на наступну за даним повідомленням частину тексту (одне значення отримане за умови, що зміст повідомлення невідомий, а інше – що цей зміст уже відомий вгадувачу, тобто у його тезаурусі).

Починаючи вгадувати текст, зразковий інформант оперує тезаурусом, що містить задану смислову інформацію. При цьому невизначеність початкової частини (у нашому випадку початкової букви) для вгадувача дорівнює H_1. Довідавшись першу букву тексту, вгадувач збільшує кількість інформації в тезаурусі, переводячи його в стан I_1, угадування другого, третього й т. д. частин тексту ще більше насичує тезаурус.

У міру нагромадження тезаурусом відомостей про текст невизначеність під час вгадування букв послідовно зменшується так, що

$$H_1 > H_2 > ... > H_{n-1} > H_{\text{п}} > ... > H_\infty. \quad (1)$$

Пам'ятаючи, що $H = I$, вираз (1) можна переписати у вигляді:

$$I_1 > I_2 > ... > I_{n-1} > I_n > ... > I_\infty. \quad (2)$$

Останній член нерівності (2) кількісно дорівнює тій інформації, яку отримує вгадувач при відгадуванні частини (графеми), як завгодно далеко віддаленої від початку тексту. Величину I_∞ ми будемо називати гранично синтаксичною інформацією зв'язного тексту.

Розглянемо ланцюги окремих значень I_n, що оцінюють величини тої інформації, яку отримує вгадувач з тексту. Ці значення виявляють тенденцію до зменшення залежно від росту значень n. Представимо кожний із цих ланцюгів не як послідовності дискретних значень ентропії, а як неперервні криві (точніше, як неперервні функції неперервного аргументу). Якщо відволіктися на час від неперіодичних коливань, відносячи їх за рахунок розкиду, то ці ланцюги значень можуть бути апроксимовані в першому наближенні теоретичною показниковою кривою (експонентою)

$$I_\xi = \left(I_0 - I_\infty\right) e^{-s\xi} + I_\infty, \quad (3)$$

де I – верхні або нижній теоретичні межі інформації в даній ділянці тексту; ξ – неперервний аргумент функції, що заміняє дискретні величини n; $I_0 = H_0$ – інформація абетки, а I_∞ – гранична інформація мови або її різновиду, величину якої представляє асимптотична крива I_ξ; e – основа натурального логарифму; s – розрахований спеціально для кожної кривої контекстний коефіцієнт.

Замінивши величину $I_\xi = I_n$ в формулі (2) правою частиною виразу (3) і зробивши деякі спрощувальні перетворення, отримаємо загальний вираз контекстної зв'язаності в даній частині тексту:

$$K_{\xi} = \left(I_0 - I_{\infty}\right)\left(1 - e^{s\xi}\right). \tag{4}$$

Визначаючи кількість правильно вгаданих графем відносно загальної кількості, обчислювались їх імовірності, що дає можливість підрахувати їх інформативність за формулою

$$\overline{I}_n^{(c)} = \sum P_{\lambda} I_n^{(c)}, \tag{5}$$

де P_{λ} — емпірична ймовірність тексту певної довжини λ; $I_n^{(c)}$ — інформація, що припадає на n-у графему в окремій схемі.

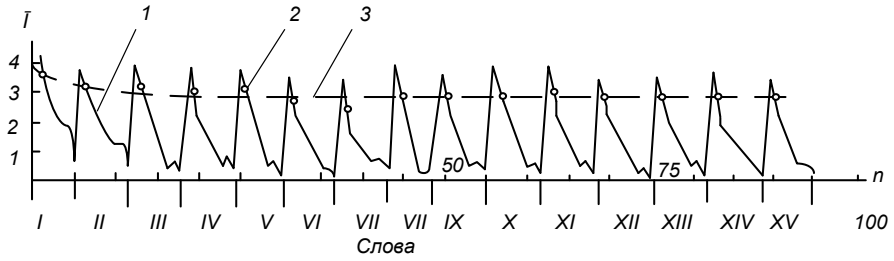

Рис. 2. Лексична схема усередненого українського тексту:
1 – полігон розподілу інформації з верхньої границі
з урахуванням достовірних продовжень

Отримана шляхом обробки результатів угадування по повній програмі схема українського тексту довжиною в 1500 слів показана на рис. 2.

У випадках, коли потрібно визначити розподіл інформації в слові на складовому й морфемному рівнях, результати вгадування групуються таким чином, щоб отримати розподіли частот спроб для вгадування першої й останньої літер складу або морфеми. Ця процедура, аналогічна перегрупуванню експериментального матеріалу під час побудови лексичної схеми тексту, дає можливість не тільки визначити загальну кількість інформації, що припадає на склад або морфему, але дозволяє також чисельно оцінити інформацію на складових і морфемних границях, з'ясовуючи при цьому особливості складового й морфемного членування слова.

Будуючи літерні розподіли, пропуск вважається останньою буквою слова. У складових схемах пропуск не враховується (як відомо, пропуск не входить до фонетичного складу). У морфемних схемах ураховуються лише ті пропуски, які є функцією нульових морфем. Оцінка інформативності графем в слові проводилася також на складовому й морфемному рівнях. На рис. 3 показані результати побудови узагальнювальних складових і морфемних схем українського позатекстового й текстового слова.

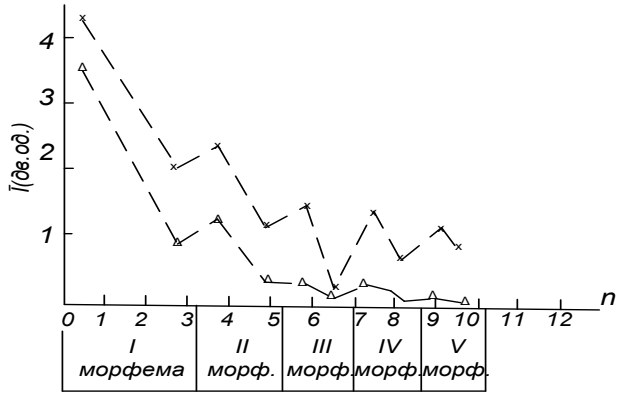

Рис. 3. Загальна морфемна схема розподілу
інформації російської мови: 1 — позатекстове слово;
2 — текстове слово

Морфемні схеми позатекстових слів у всіх мовах характеризуються послідовним чергуванням максимумів і мінімумів інформації. Максимуми інформації падають на першу, а мінімуми на останню букву морфеми. Іншими словами, по всій довжині позатекстового письмового слова виявляються чітко позначеними границі між морфемами. Ці границі збігаються з видорозділом між останньою літерою попередньої морфеми (мінімум інформації) і першою літерою наступної морфеми (максимум інформації). Якщо звернутися до текстових слів, то виявиться, що вони також виявляють морфемне членування, хоча воно й виражено тут значно слабкіше.

Адекватність отриманих результатів визначалась за результатами обробки вгадування тексту декількома експертами. Аналіз цих результатів дозволив з'ясувати такі інформаційні характеристики графем:

— для слів розподіл у тексті статистичної інформації має квантовий характер. Початок слів несуть максимуми інформації, у той час як середини й особливо подальші за ними пробіли виявляються або мало інформативними, або взагалі надлишковими;

— морфемні схеми позатекстових слів у всіх мовах характеризуються послідовним чергуванням максимумів і мінімумів інформації. Максимуми інформації падають на першу, а мінімуми на

останню букву морфеми. Іншими словами, по всій довжині позатекстового письмового слова виявляються чітко позначеними границі між морфемами; отже, можна стверджувати, що з погляду синтаксичної інформації слову в українській мові властива чітко виражена морфемна структура, що придушує складове членування слова.

Той факт, що знакова (морфемна) структура слова придушує синтагматичну систему фігур (тобто складів), проливає світло на взаємодію різних механізмів тезауруса нашого ідеального вгадувача. Як відомо, мова являє собою складний марковський процес лінійного проходження різних елементів мови — як фігур, так і знаків-символів. Іншими словами, ймовірносно-статистичні закономірності, що характеризують сполучуваність фігур, взаємодіють із імовірністю сполучуваності символів (морфем, слів, словосполучень і т.д.).

Висновки

Проведені в роботі дослідження показали, що найбільш інформативними для розпізнавання і розуміння текстового документу є графеми окремих літер, морфем, довжина в складах і літерах слів, графічні особливості напису окремих слів, статичні і перехідні ймовірності вказаних елементів. Тому для в провадження інтелектуальних технологій в обробку текстових документів нагальною проблемою є розроблення програмного забезпечення для створення бази даних, що містить інформацію з визначених характеристик.

СПИСОК ЛІТЕРАТУРИ

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

2. Пиотровский Р. Г. Текст машина, человек / Р. Г. Пиотровский. — Ленинград : Наука, 1975. — 326 с.

Рекомендована кафедрою комп'ютерних систем управління

Надійшла до редакції 21.10.08
Рекомендована до друку 20.11.08

Биков Микола Максимович — професор, ***Балховський Дмитро Євгенович*** — аспірант, ***Ковтун В'ячеслав Васильович*** — доцент.

Кафедра комп'ютерних систем управління, Вінницький національний технічний університет