12. Критерії оцінки захищеності інформації в комп'ютерних системах від несанкціонованого доступу: НД ТЗІ 2.5-004-99. — [Чинний від 1999.04.28]. — К. : ДСТСЗІ СБУ, 1999. — № 22. — (Нормативний документ системи технічного захисту інформації).

13. Литвак Б. Г. Экспертная информация. Методы получения и анализа. – М.: Радио и связь, 1982. – С. 23–28.

14. Горніцька Д. А. Визначення коефіцієнтів важливості для експертного оцінювання у галузі інформаційної безпеки / О. Г. Корченко, Д. А. Горніцька, В. В. Волянська // Захист інформації. – Київ, 2012. – №1. – С.108-121.

Надійшла: 12.07.2012 р.

Рецензент: д.т.н., професор Конахович Г.Ф.

УДК 004.056.53 **Бабенко Т.В., Сушко С.О.**

ПРО ЕНТРОПІЮ УКРАЇНСЬКОЇ МОВИ

У статті виконано аналіз підходів до формалізації природних мов та викладено результати експериментальних досліджень теоретико - інформаційних характеристик різних стилів української мови, визначена її надлишковість та приведено результати порівняння ентропії української мови з іншими природними мовами.

Ключові слова: ентропія, мова, дослідження, параметри.

У даний час існують різні підходи до формалізації природних мов сформульовані у вигляді математичних конструкцій, але не існує універсальної моделі мови, що дозволяла б з достатньою точністю апроксимувати реальну мову. Також залишається відкритим питання створення моделі, спроможної оцінити природність мови. Як відомо, ця задача є актуальною при вирішенні проблем оптимізації роботи пошукових систем у мережі Інтернет, задач криптографічного аналізу та інших.

Однією із задач, що потребує вирішення при синтезі моделі відповідної мови є визначення її теоретико-інформаційних характеристик, зокрема ентропії. Як відомо, знання ентропії відповідної мови є важливим при дослідженні асимптотики кількості осмислених відкритих текстів фіксованої довжини, при обчисленні відстані єдиності шифрів, для виявлення атипових зразків даних, що можуть нагадувати шкідливий код, при проведенні частотного аналізу, зокрема, ентропією оцінюють складність пароля в комп'ютерній індустрії.

Доцільно відзначити, що спроби обчислити ентропію із задовільною точністю виконувались для багатьох мов. Так, ентропію англійської мови досліджував класик теорії інформації К.Шеннон [1]. Радянський академік Піотровський Р.Г. із співробітниками одержали багато цікавих інформаційно-статистичних параметрів російської та інших мов колишнього СРСР [2–4]. Деякі дослідження статистичних властивостей української мови проводились в Інституті мовознавства ім. О.О. Потебні НАН України.

У роботі [5] зроблена спроба оцінити для української мови значення умовних ентропій розподілу ймовірностей біграм, триграм і чотириграм. Нажаль, у своїх розрахунках автори обмежились аналізом виключно українських та зарубіжних (у перекладі) літературних творів при цьому загальний обсяг текстового матеріалу незначно перевищив 12 млн. символів. Відповідно в зазначених роботах були отримані виключно інформаційні характеристики української абетки.

У даному досліджені авторами поставлено за мету оцінити значення ентропії української мови базуючись на вивченні текстів, що представляють всі стилі сучасної української мови.

Розглянута нами імовірнісна n-грамна модель мови спирається на так званий принцип незалежності від «давньої» історії: якщо у загальному вигляді ймовірнісна модель дозволяє обчислити ймовірність того, що s-грама $a_{i_1}a_{i_2}...a_{i_s}$ може існувати в українській мові, то в n-грамній моделі робиться припущення, що

$$P(a_{i_j}\left|a_{i_1}a_{i_2}...a_{i_{j-1}}\right.)\approx P(a_{i_j}\left|a_{i_{j-n+2}}...a_{i_{j-1}}\right.)\cdot$$

А відтак, ймовірність $P(a_{i_1}a_{i_2}...a_{i_s})$, записана у вигляді добутку умовних ймовірностей вигляду

$$P(a_{i_1}a_{i_2}...a_{i_s})=P(a_{i_1})\cdot P(a_{i_2}\left|a_{i_1}\right.)\cdot...\cdot P(a_s\left|a_{i1}a_{i_2}...a_{i_{s-1}}\right.),$$

може бути оцінена через добуток ймовірностей:

$$P(a_{i_j}\left|a_{i_{j-n+1}}...a_{i_{j-1}}\right.)$$

Отже, у такому випадку модель зводиться до скінченної множини ймовірностей, кожну з яких можна оцінити, обчисливши частоту повторюваності відповідних n-грам.

Для дослідження ентропії була сформована база текстів, що належали до художнього, наукового, ділового, розмовно-побутового і публіцистичного стилів, зокрема тексти з інтернет-сайтів низки електронних газет. Об'єм текстової бази перевищив 100 млн. слововживань (~800 Мб) і містив тексти, що відображають сучасний стан української мови, зокрема розмовну українську мову.Дослідження зібраного мовного матеріалу виконували за методикою описаною в [6]. Для фільтрації знаків пунктуації і цифрових даних та підрахунку n-грам використовувався відкритий програмний модуль «Gramms» [5].

У ході експериментів були обчислені частоти повторюваності у текстах: 1) 34 уніграм (33 літери абетки та пропуск між словами як самостійний знак); 2) 1156 можливих біграм; 3) 39304 триграм; 4) 1336336 чотириграм; 5) 45435424 п'ятиграм; 6) 1544804416 шестиграм; 7) 52523350144 семиграм; 8) 1785793904896 восьмиграм; 9) 60716992766464 дев'ятиграм.

Оскільки в текстах на будь-якій мові найбільш вживаним є пропуск між словами, то на першому етапі частоти букв і n-грам підраховувались з врахуванням пропуску, а на другому етапі – без його врахування. За статистичним визначенням ймовірності обчислені частоти вважалися наближеними значеннями ймовірностей появи *n*-грам у відкритих текстах українською.

Ентропія української мови оцінювалась за допомогою послідовних наближень. За Шенноном в якості першого наближеного значення ентропії мови вибрано абсолютну ентропію мови $Í_0=\log_2 m=\log_2 34$, тобто максимальну ентропію окремих незалежних букв (m – потужність алфавіту). Друге наближене значення ентропії української мови – це умовна ентропія $H_1=-\sum_{i=1}^{m}p(a_i)\log_2 p(a_i)$ посимвольної імовірнісної моделі відкритих текстів, в якій імовірність $p(a_i)$ букв збігається з частотою повторюваності букв в українській мові.

Наявність у природних мовах додаткових закономірностей зумовлює подальше зменшення ступеня невизначеності (ентропії) однієї букви мови. Тому за третє наближене значення ентропії української мови було вибрано умовну ентропію H_2 імовірнісного розподілу біграм, поділену на два (бо нас цікавить ентропія на один знак алфавіту). Іншими словами, $H_2/2$ визначає середню інформацію, що міститься в букві українського тексту, якщо відома попередня буква. Аналогічно четверте наближене значення ентропії мови – це умовна ентропія H_3 імовірнісного розподілу триграм, поділена на три і т.д.

Дослідження К. Шеннона показали, що найбільш точні наближення до ентропії букви для осмислених текстів дають відношення $\frac{H_r}{r}$ при $r = 5,6,...$ Із зростанням r ентропія зменшується і при $r \to \infty$ прагне до границі H_∞, яка й приймається за ентропію мови H_m, тобто

$$\lim_{r\to\infty} \frac{H_r}{r} = H_m.$$

Обчислені наближені значення ентропії української мови зведені у представлено в таблиці 1. Оскільки в текстах будь-якою мовою найбільш вживаним є пропуск між словами, то на першому етапі частоти букв і n-грам підраховувались з врахуванням пропуску, а на другому етапі – без.

Таблиця 1

Значення умовної ентропії української мови

Наближене значення Ентропії	З урахуванням пропуску між словами (бітів/символ)	Без урахування пропуску між словами (бітів/символ)
H_0	5,09	5,04
H_1	4,51	4,52
$H_2 / 2$	3,47	3,60
$H_3 / 3$	3,14	3,26
$H_4 / 4$	2,78	2,93
$H_5 / 5$	1,58	1,62
$H_6 / 6$	1,41	1,55
$H_7 / 7$	1,26	1,47
$H_8 / 8$	1,25	1,45
$H_9 / 9$	1,25-1,40	1,45-1,62

При вилученні пропуску, по-перше, зменшується кількість букв алфавіту, а по-друге, зростають частоти повторюваності букв і n-грам.

Як результат $H_0^{(з\ проп.)} > H_0^{(без\ проп.)}$.

Якщо пропуск розглядати додатковою «буквою» алфавіту, та ще й такою, що порівняно з рештою має дуже велику ймовірність, приходимо до зменшення наближених значень ентропії: $H_n^{(з\ проп)} < H_n^{(без\ проп)}$.

У випадку наближення ентропії за допомогою розподілу дев'ятиграм розрахунки проводились окремо за текстами різної тематики і різних стилів сучасної української мови.

Аналіз отриманих результатів дозволяє зробити висновок, що ентропія H української мови складає 1,25-1,40 бітів/символ без урахуванням пропуску між словами і 1,45-1,62 бітів/символ з його урахуванням. За цих умов надлишковість української мови $D = 1 - \frac{H}{\log_2 m} \cdot 100\%$ складає 72,5 – 75,4 %.

При цьому порівняння отриманих результатів ентропії української мови з аналогічними результатами, що отримані для російської та європейських мов дозволяє стверджувати:

1). Надлишковість української мови знаходиться на тому ж рівні (близько 70 %), що й надлишковість інших мов. Це обумовлено біосоціальною природою мови. Такий рівень надлишковості слугує захистом мовної комунікації від фізичних і психолінгвістичних вад;

2). Коливання надлишковості пов'язані із зміною тематики, професійної і стилістичної орієнтації текстів. Мовні і художні тексти показують низький рівень надлишковості (72 % і нижче), а публіцистична та науково-технічна мова має надлишковість на рівні 75 % (інколи до 85 %).

ЛІТЕРАТУРА

1. Шеннон К. Работы по теории информации и кибернетике. – М.: Наука, 1973. – С.68 – 128, 236 – 273, 441 – 483.

2. Белоногов Г.Г., Фролов Г.Д. Эмпирические данные о распределении букв в русской письменной речи // В сборнике «Проблемы передачи кибернетики». – 1963. – Вип. 9. – С. 287 – 305.

3. Пиотровский Р.Г. Информационные измерения языка. – Л. Наука, 1968. – С. 17 – 81.

4. Статистика речи. Сборник. Отв. редактор Пиотровский Р.Г. – Л.: Наука, 1968. – С.50 – 60, 228 – 230.

5. Кригін М.Ю., Широков В.А. Дослідження інформаційно-статистичних властивостей українського тексту. Математичні машини і системи, 2000, №1, с.120 -127.

6. Сушко С.О., Фомичова Л.Я., Барсуков Є.С. «Частоти повторюваності букв і біграм у відкритих текстах українською мовою» . – Захист інформації. –2010. – №3, С. 94-102.

Надійшла: 27.07.2012 р.

Рецензент: д.т.н., професор Петров О.С.

УДК 004.056.53:004.492.3 (045)

Гнатюк С.О., Волянська В.В., Карпенко С.В.

СУЧАСНІ СИСТЕМИ ВІРТУАЛЬНИХ ПРИМАНОК НА ОСНОВІ ТЕХНОЛОГІЇ HONEYPOT

У цій статті проведено аналіз існуючих систем віртуальних приманок на базі технології honeypot. Аналіз показав еволюцію honeypot-систем від Low-Interaction Honeypots до найсучасніших Gen III Honeynets і вказав на недоліки існуючих рішень. Крім того, проведено класифікацію honeypot-систем за ознаковим принципом. У подальшому ці результати можна використати для розробки honeypot-систем з метою підвищення ефективності роботи систем управління інцидентами інформаційної безпеки.

Ключові слова: віртуальна приманка, технологія honeypot, виявлення вторгнень, honeynet, ознаковий принцип класифікації.

Вступ. Достатньо довго в протистоянні «напад-захист» практикувалася своєрідна покрокова стратегія – зловмисники користувалися однією «діркою» захисту і з часом її закривали, тоді вони шукали іншу – її згодом також закривали і т.п. Такий аналог гри в шахи, де партія може тривати як завгодно довго, вимагає від захисту колосальних витрат часу і ресурсів, тим паче в нападника майже завжди є можливість адекватно відреагувати на захисні заходи. Зважаючи на це, сторона захисту повинна «грати на попередження», цим самим мінімізуючи ризик вторгнення. Саме реалізація такої ідеї лежить в основі використання віртуальних приманок – так званих, *honeypot-систем* (від англ. – «горщик з медом»). Мета їх функціонування – бути атакованими або сканованими зловмисниками для вивчення стратегії останніх, визначення кола їх засобів, за допомогою яких можуть бути нанесені удари по реальних об'єктах безпеки. Метод реалізації віртуальної приманки не принциповий – це може бути як спеціально розгорнута цілісна мережа так і один єдиний емульваний мережевий сервіс, основним і першочерговим завданням якого є зацікавлення (привернення уваги) порушника [1].

Концепція віртуальних приманок бере свій початок з робіт К. Столла і Б. Чесвіка [2, 3]. Ця концепція була реалізована в ряді ранніх продуктів (Deception Toolkit, CyberCop Sting, BackOfficer Friendly [4, 5]). Подальше удосконалення і розширення сфери застосування даної технології пов'язане з оформленням у 1999 р. проекту Honeynet Project [6]. Завдяки роботам таких фахівців як Л. Спітцнер [4], Н. Провос [5], Ф. Коен [7], Е. Балас [8], М. Рош [9] концепція віртуальних приманок оформилася в конкретну технологію з власною сферою застосування архітектурою і інструментарієм. Протягом