

УДК 681.3

# КОНТРОЛИРУЮЩАЯ СПОСОБНОСТЬ МЕТОДОВ АВТОМАТИЧЕСКОГО ОБНАРУЖЕНИЯ ТИПОВЫХ ОШИБОК ПОЛЬЗОВАТЕЛЯ ПО СЛОВАРЯМ РУССКОГО И УКРАИНСКОГО ЯЗЫКОВ

В.А. Литвинов, С.Я. Майстренко

*Институт проблем математических машин и систем НАН Украины*

e-mail: maistrsv@ukr.net

Одним из основных методов автоматического контроля и идентификации ошибок пользователя при вводе символьных данных в процессе обработки информации является проверка введенного слова по базовому словарю допустимых слов (БС). Грубая оценка результативности такого контроля  $\pi^{(0)}$  (вероятности необнаружения ошибки) может быть основана на предположении о случайном характере искажений входного слова и сопоставлении мощности  $Q_p$  множества допустимых (реально существующих) слов с мощностью  $Q_z$  запрещенных (отсутствующих в БС) комбинаций символов [1]:

$$\pi^{(0)} = \frac{Q_p}{Q_p + Q_z} = \frac{N}{q^n}, \quad (1)$$

где  $N$  - количество слов БС,  $q$  - алфавит символов БС,  $n$  - количество символов в слове. Близким по назначению является показатель MSD Error Rate [2].

Для БС с целенаправленно введенной избыточностью и относительно равномерным (случайным) распределением  $N$  реальных слов среди  $q^n$  всевозможных значений комбинаций  $n$  символов в алфавите  $q$  - в частности, для кодовых справочников - оценка (1) может быть достаточно близка к истине. Для естественно-языковых слов (слов в текстовом редакторе, ключевого слова в поисковой системе и т.п.) и специфических искажений, вызванных типовыми ошибками пользователя, допущения о случайном характере распределений значений слов и их возможных искажений не выполняются. Здесь наиболее вероятные простые искажения дают, как оказывается, значительно большее количество ложных совпадений с реально существующими символами и, соответственно, намного худшую результативность.

Для анализа и оценки реальной контролирующей способности словаря по отношению к заданным видам искажений (в рассматриваемом случае - типовым ошибкам пользователя) разработана соответствующая имитационная модель искажений слов и обнаружения ошибок.

На рисунке приняты следующие обозначения:

$\tilde{A}_{ij}$  - множество ошибок класса  $j$  в слове  $i$ ,  $V_{ij}$  - мощность множества  $\tilde{A}_{ij}$ .

Цели эксперимента состояли в определении:

- реальной вероятности (относительного количества)  $\pi_j$  необнаруживаемых типовых ошибок пользователя  $E_j$  для конкретных словарей русского и украинского языков;

- зависимости интегральной вероятности  $\pi$  необнаружения произвольной ошибки в зависимости от объема  $N$  словаря, варьируемого путем «усечения» исходного словаря,

$$\pi = \pi_j \cdot p_j \quad (2)$$

Полновыборочные эксперименты проведены с ошибочными словами двух полных словарей русского языка (Словарь Зализняка [4] исходным объемом 92555 слова,

«Словарь русской литературы» [4] исходным объемом 161 730 слов) и двумя соответствующими словарями украинского языка, сформированными путем русско-украинской конвертации (исходными объемами 84575 и 129244 слов).

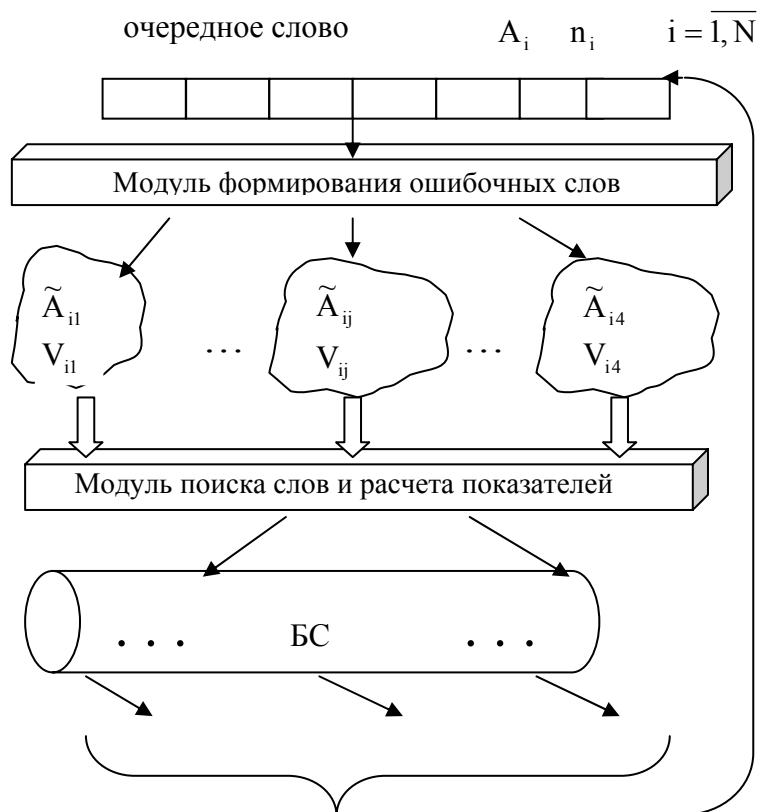


Рис. Имитационная модель

Исследованы следующие классы типовых ошибок:  $j=1$  – однократные транскрипции;  $j=2$  – вставки символов;  $j=3$  – выпадения символов;  $j=4$  – транспозиции соседних символов.

В таблицах 1, 2 приведены результаты экспериментов.

Таблица 1

| $j$ | $P_j[3]$ | Словарь Зализняка |             |         | Словарь Зализняка – укр. |             |         | Словарь Русской литературы |             |         | Словарь Русской литературы – укр. |             |         |
|-----|----------|-------------------|-------------|---------|--------------------------|-------------|---------|----------------------------|-------------|---------|-----------------------------------|-------------|---------|
|     |          | $\bar{n}_j$       | $\bar{V}_j$ | $\pi_j$ | $\bar{n}_j$              | $\bar{V}_j$ | $\pi_j$ | $\bar{n}_j$                | $\bar{V}_j$ | $\pi_j$ | $\bar{n}_j$                       | $\bar{V}_j$ | $\pi_j$ |
| 1   | 0.56     | 9.61              | 307.5704    | 0.0039  | 9.49                     | 313.1139    | 0.0028  | 8.44                       | 269.9527    | 0.0123  | 8.31                              | 274.2541    | 0.0100  |
| 2   | 0.16     | 9.61              | 350.1819    | 0.0006  | 9.49                     | 356.6022    | 0.0004  | 8.44                       | 311.3886    | 0.0027  | 8.31                              | 316.5648    | 0.0015  |
| 3   | 0.12     | 9.61              | 9.6116      | 0.0214  | 9.49                     | 9.4883      | 0.0139  | 8.44                       | 8.4360      | 0.0882  | 8.31                              | 8.3107      | 0.0522  |
| 4   | 0.06     | 9.61              | 8.61157     | 0.0095  | 9.49                     | 8.4883      | 0.0091  | 8.44                       | 7.4360      | 0.0175  | 8.31                              | 7.3107      | 0.0111  |

Данные, приведенные в таблицах, иллюстрируют следующие особенности контролирующих свойств исследованных словарей.

Таблиця 2

| Словарь Зализняка |                            | Словарь Зализняка – кр. |                            | Словарь Русской литературы |                            | Словарь Русской литературы – укр. |                            |
|-------------------|----------------------------|-------------------------|----------------------------|----------------------------|----------------------------|-----------------------------------|----------------------------|
| N                 | $\pi = \sum_1^4 \pi_j p_j$ | N                       | $\pi = \sum_1^4 \pi_j p_j$ | N                          | $\pi = \sum_1^4 \pi_j p_j$ | N                                 | $\pi = \sum_1^4 \pi_j p_j$ |
| 92555             | 0.0055                     | 84575                   | 0.0039                     | 161730                     | 0.0191                     | 129244                            | 0.0128                     |
| 82555             | 0.0049                     | 74575                   | 0.0035                     | 151730                     | 0.0179                     | 119244                            | 0.0118                     |
| 72555             | 0.0044                     | 64575                   | 0.0031                     | 141730                     | 0.0167                     | 109244                            | 0.0110                     |
| 62555             | 0.0038                     | 54575                   | 0.0027                     | 131730                     | 0.0157                     | 99244                             | 0.0099                     |
| 52555             | 0.0034                     | 44575                   | 0.0023                     | 121730                     | 0.0145                     | 89244                             | 0.0090                     |
| 42555             | 0.0027                     | 34575                   | 0.0019                     | 111730                     | 0.0133                     | 79244                             | 0.0081                     |

1. Вероятность пропуска типовых ошибок  $\pi_j$  рассмотренных классов значительно (на порядки) превышает идеализированное значение  $\pi^{(0)}$ . Это является следствием того, что кластеры (цепочки взаимных искажений слов) типа <кол>  $\Rightarrow$  <пол, мол, гол, фол, вол, тол, дол> дают гораздо большее количество совпадений со словарем, чем, например, случайный маловероятный гипотетический переход <кол>  $\Rightarrow$  <пртх>. В результате, значения  $\pi$  для реальных словарей значительно выше  $\pi^{(0)}$ , а СА – показатель значительно ниже, чем можно было бы предположить, исходя из (1). Аналогичное положение и с исследованными словарями украинского языка.

2. Существует предвидимая корреляция между объемом БС и значением  $\pi$ .

3. Наиболее уязвимым типом ошибок (в смысле возможностей их обнаружения) являются выпадения символов.

В целом, имитационная модель и получаемые с ее помощью данные могут служить основой для априорных оценок ожидаемой результативности системы автоматического обнаружения, идентификации и коррекции [5] ошибок (spellchecking'a), а также возможной постановки и решения задачи оптимизации рабочего естественно-языкового БС с учетом частот использования слов.

#### Список литературы

1. Модель анализа и оценки эффективности методов логического контроля информации / Кузьменко Г.Е., Литвинов В.А., Литвинова А.Н., Майстренко С.Я. // Математичні машини і системи. – 2002. - №1. – С. 49-55.
2. Typing: [http://en.wikipedia.org/wiki/Typing#Key\\_strokes\\_per\\_character\\_.28KSPC.29](http://en.wikipedia.org/wiki/Typing#Key_strokes_per_character_.28KSPC.29).
3. Литвинов В.А., Крамаренко В.В. Контроль достоверности и восстановление информации в человеко-машинных системах. – Киев: Техніка, 1986. – 200 с.
4. Словари русского языка, [http://speakrus.ru/dict/Толковый словарь по информатике](http://speakrus.ru/dict/Толковый_словарь_по_информатике). <http://www.bitoman.ru/download/slow/112384/9d4c2.html>.
5. Алгоритми і моделі автоматичної ідентифікації та корекції типових помилок користувача на основі природної надмірності / Кузьменко Г.Е., Литвинов В.А., Майстренко С.Я., Ходак В.І. // Математичні машини і системи – 2004.- №2. - С.134-148.
6. Имитационная модель автоматического обнаружения типовых ошибок пользователя в естественно-языковом интерфейсе / Литвинова А.Н., Майстренко С.Я., Юденко О.П. // XIX Міжнародна конференція з автоматичного управління Автоматика/Automatics, Національний університет харчових технологій, Київ, 26-28 вересня, 2012.