УДК 681.3

В.А. ЛИТВИНОВ[*], С.Я. МАЙСТРЕНКО[*], К.В. ХУРЦИЛАВА[*]

ОЦЕНКА КОНТРОЛИРУЮЩИХ СВОЙСТВ БАЗОВОГО СЛОВАРЯ ДОПУСТИМЫХ СЛОВ В СИСТЕМЕ АВТОМАТИЧЕСКОГО ОБНАРУЖЕНИЯ ОШИБОК ПОЛЬЗОВАТЕЛЯ

[*]Институт проблем математических машин и систем НАН Украины, Киев, Украина

Анотація. *Розроблено імітаційну модель спотворень слів і виявлення помилок користувача. Наводяться результати моделювання для словників російської та української мов. Отримано оцінки реальних контролюючих властивостей словника, що дозволяють вирішити задачу оцінки його якості.*
Ключові слова: *помилки користувача, спелл-чекінг, достовірність даних, імітаційне моделювання.*

Аннотация. *Разработана имитационная модель искажений слов и обнаружения ошибок пользователя. Приводятся результаты моделирования для словарей русского и украинского языков. Получены оценки реальных контролирующих свойств словаря, позволяющие решить задачу оценки его качества.*
Ключевые слова: *ошибки пользователя, спелл-чекинг, достоверность данных, имитационное моделирование.*

Abstract. *A simulation model of distortion of words and detection of user errors has been developed. The simulation results for dictionaries of the Russian and Ukrainian languages have been demonstrated. Assessment of the real controlling properties of the dictionary, allowing to evaluate its quality, has been made.*
Keyword: *user errors, spell-checking, the accuracy of the data, simulation modeling.*

1. Введение

Основой системы «проверки правописания» при вводе естественно-языковых данных (в общем случае – нерегулярных алфавитно-цифровых кодов) является базовый словарь допустимых слов (БС). Контролирующие свойства БС определяются вероятностью необнаружения ошибки в результате случайного совпадения искаженного слова с некоторым посторонним допустимым словом. Грубая оценка значения $\pi^{(0)}$ такой вероятности может быть основана на предложении о случайном характере искажений входного слова и сопоставлении мощности $Q_з$ запрещенных (отсутствующих в БС) комбинаций символов и Q_p допустимых [1, 2]:

$$\pi^{(0)} \approx \frac{Q_p}{Q_p + Q_з} \approx \frac{N}{q^n}, \qquad (1)$$

где N – количество слов БС, q – алфавит символов БС, n – среднее количество символов в слове.

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 вероят-

ISSN 1028-9763. Математичні машини і системи, 2014, № 2

ные простые искажения могут дать значительно большее количество ложных совпадений с реально существующими словами и, соответственно, намного худшую результативность контроля.

В статье рассматриваются вопросы оценки реальных контролирующих свойств БС (на примере словарей русского и украинского языков) и возможные пути их улучшения.

2. Имитационная модель искажений и обнаружения ошибочных слов

Введем исходные понятия, термины и обозначения. Под общей диагностической (контролируемой) способностью словаря c будем понимать относительное количество ошибок (всевозможных искажений слов), не совпадающих ни с каким другим допустимым словом, то есть обнаруживаемых ошибок.

Частную диагностическую способность c_k определим как относительное значение c для ошибок определенного класса k. Соответственно под общей и частной дисфункцией словаря будем понимать показатели ρ и ρ_k, определяющие относительные количества недиагностируемых ошибок. Введенные показатели связаны следующими очевидными соотношениями: $c = 1 - \rho$, $c_k = 1 - \rho_k$. Структура имитационной модели, предназначенной для моделирования процесса искажений, обнаружения ошибочных слов для заданного словаря, и определения значений ρ_k, приведены на рис. 1.

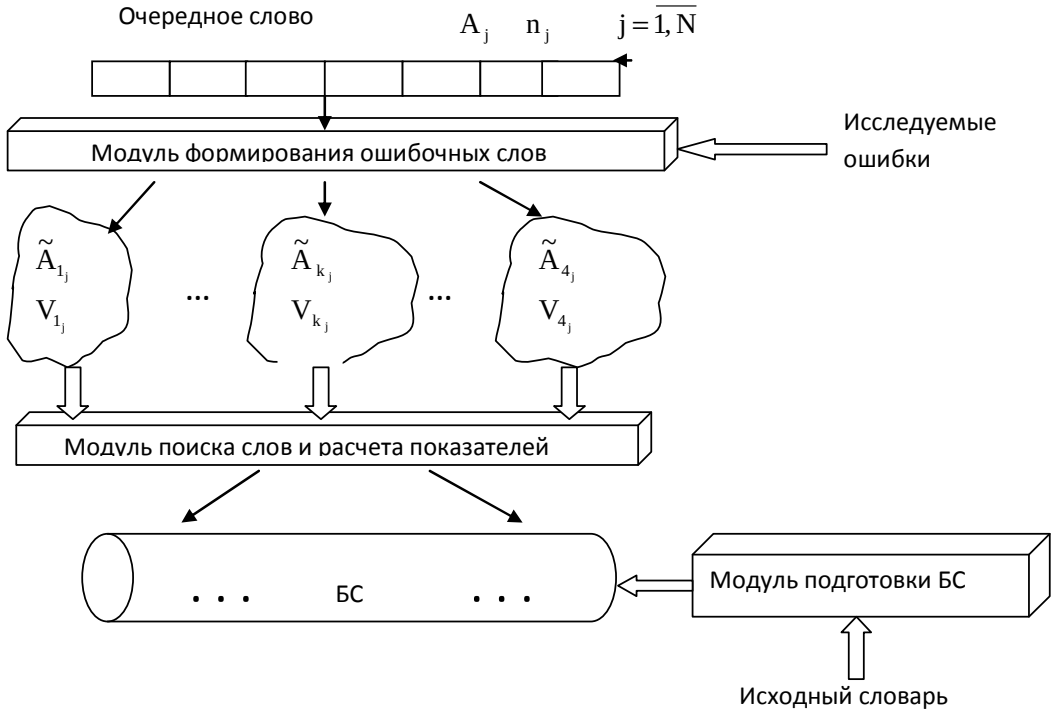

Рис. 1. Структура имитационной модели

На рисунке приняты следующие обозначения:

$\tilde{A}_{k_j}$ – множество ошибок класса k в слове j, V_{k_j} – мощность множества $\tilde{A}_{k_j}$.

Полновыборочные $(j = 1...N)$ эксперименты проведены со словами трех словарей русского языка («Словарь Зализняка» [3] (СЗр – русский, СЗу – украинская версия), «Словарь Лопатина» [4] (СЛр – русский, СЛу – украинская версия), «Словарь русской литературы [3] (СРЛр – русский, СРЛу – украинская версия) и адаптированными украино-

язычными версиями указанных словарей, сформированными путем русско-украинской конвертации. Исследованы следующие основные классы типовых ошибок тайпинга: однократные транскрипции $(k=1)$, добавление символа $(k=2)$, выпадение символа $(k=3)$, транспозиция соседних символов $(k=4)$. Для ориентировочных оценок вероятностей P_k ошибок этих классов взяты значения, приведенные в [5].

Результаты моделирования приведены в табл. 1, 2.

Таблица 1. Словари русского языка

k	P_k	Словарь Зализняка $N=92555$ $\overline{n}=9,61$		Словарь Лопатина $N=150213$ $\overline{n}=10,06$		Словарь русской литературы $N=161730$ $\overline{n}=8,44$	
		$\overline{V}^k$	$\rho_k 10^2$	$\overline{V}^k$	$\rho_k 10^2$	$\overline{V}^k$	$\rho_k 10^2$
1	0,56	307,6	0,39	321,8	0,41	269,9	1,2
2	0,16	350,2	0,06	364,8	0,07	311,4	0,27
3	0,12	9,6	2,14	10,06	2,16	8,4	8,8
4	0,06	8,6	0,95	9,06	1,55	8,9	1,2
$\overline{\rho}_k$	0,9	–	0,54	–	0,6	–	1,84

Таблица 2. Словари украинского языка

k	P_k	Словарь Зализняка $N=84575$ $\overline{n}=9,49$		Словарь Лопатина $N=135401$ $\overline{n}=9,93$		Словарь русской литературы $N=1292440$ $\overline{n}=8,31$	
		$\overline{V}^k$	$\rho_k 10^2$	$\overline{V}^k$	$\rho_k 10^2$	$\overline{V}^k$	$\rho_k 10^2$
1	0,56	313,1	0,28	327,7	0,28	274,2	1,0
2	0,16	356,6	0,04	371,6	0,04	316,5	0,15
3	0,12	9,5	1,39	9,93	1,40	8,3	5,2
4	0,06	8,5	0,91	8,93	1,22	7,3	1,1
$\overline{\rho}_k$	0,9	–	0,38	–	0,41	–	0,77

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

1. Дисфункция контроля ошибок рассмотренных классов значительно (на порядки) превышает идеализированные значения $\pi^{(0)}$. Это является следствием того, что кластеры (цепочки взаимных искажений слов) типа <код> ⇒ <пол, мол, гол, фол, вол, тол, дол > дают гораздо большее количество совпадений со словарем, чем, например, случайный маловероятный гипотетический переход <кол> ⇒ <крах>. В результате контролирующая способность словарей как русского, так и украинского языка, значительно ниже, чем можно было бы предположить, исходя из (1).

2. Различные словари имеют заметно отличающиеся контролирующие свойства. Так, из 1000 случайных ошибочных слов из словарей, искаженных ошибками 1, 2, 3, 4 (в указанной пропорции), в среднем не обнаруживаются 5,4 ошибки для Словаря Зализняка и 18,4 ошибки для Словаря русской литературы. Исходя из полученных данных, можно предположить, что разброс значений $\overline{\rho}_k$ для исследованных словарей определяется как чисто лингвистическими факторами (язык, структура), так и разницей в объемах. При этом необходимо отметить, что уменьшение объема словаря при прочих равных условиях прогнозированно должно вести к уменьшению $\overline{\rho}_k$ за счет явного увеличения относительной

избыточности представления слов и соответствующего уменьшения возможностей случайных совпадений ошибочных слов с допустимыми. (Это свойство отмечено и иллюстрировано примерами в [2]). С другой стороны, исключение из словаря слов с ненулевой востребованностью увеличивает вероятность ложных сообщений об ошибках.

Для оценки зависимости контролирующих свойств словаря от указанных факторов следует разделить их совместное влияние на значения ρ_k.

3. Непрерывно-дискретная модель БС

Непрерывно-дискретная модель БС, построенная в развитие анализа отмеченных выше особенностей, основана на допущении об экспоненциальном характере функции $\rho(x)$, аппроксимирующей плотность гипотетического распределения востребованности слов БС (вероятностей обращения к словам):

$$\rho(x) = c\lambda \exp(-\lambda x),$$

где c – нормирующий множитель.
Определяя c из уравнения

$$\int_0^N c\lambda \exp(-\lambda x)\, dx = 1,$$

получаем

$$\rho(x) = \frac{\lambda \exp(-\lambda x)}{1 - \exp(-\lambda N)}.$$

Значения λ определяют крутизну падающей функции $\rho(x)$ и определяются здесь из уравнения

$$\int_0^{\alpha N} \rho(x)\, dx = \beta, \qquad (2)$$

где α и β – параметры, соответствующие принципу Парето с принятыми количественными соотношениями $(\alpha < 1,\ \beta < 1,\ \alpha + \beta = 1)$.

Смысл значений параметров α и β заключается в следующем: $(100\,\alpha)$ процентов слов БС востребованы в $(100\,\beta)$ процентов случаев обращений к БС. Для соотношений 20/80 и 10/90 значения α=0,2; 0,1 и β=0,8; 0,9.

Решение (2) для указанных значений α, β дает значения $\lambda \approx {}^{8}\!/_{\mathrm{N}}$ и ${}^{24}\!/_{\mathrm{N}}$ соответственно.

Назовем «усеченным» исходный БС, из которого исключаются Δ_i слов (порция усечения) с наименьшими вероятностями обращений p_j:

$$p_j = \int_{j-1}^{j} p(x)\, dx, \quad j = 1...N.$$

$$\Delta_i = i\delta N, \quad i = 1,2,\ldots,$$

где $\delta << 1$.

Цель моделирования заключается в экспериментальной оценке зависимостей $f\left(\lambda,\Delta,P^{(\Delta)},\overline{\rho}^{(\Delta)}\right)$, где $P^{(\Delta)}$ есть суммарная вероятность обращения к словам порции усечения, а $\overline{\rho}^{(\Delta)}$ – взвешенная по типовым ошибкам пользователя суммарная вероятность пропуска ошибочного слова усеченного БС:

$$P^{(\Delta)}=\int_{N-\Delta_i}^{N} p(x)\,dx,\ \ \overline{\rho}^{(\Delta)}=\sum_k \rho_k P_k .$$

Рис. 2 иллюстрирует геометрический смысл значений величины $P^{(\Delta)}$ (для большей наглядности масштабы экспонент искажены).

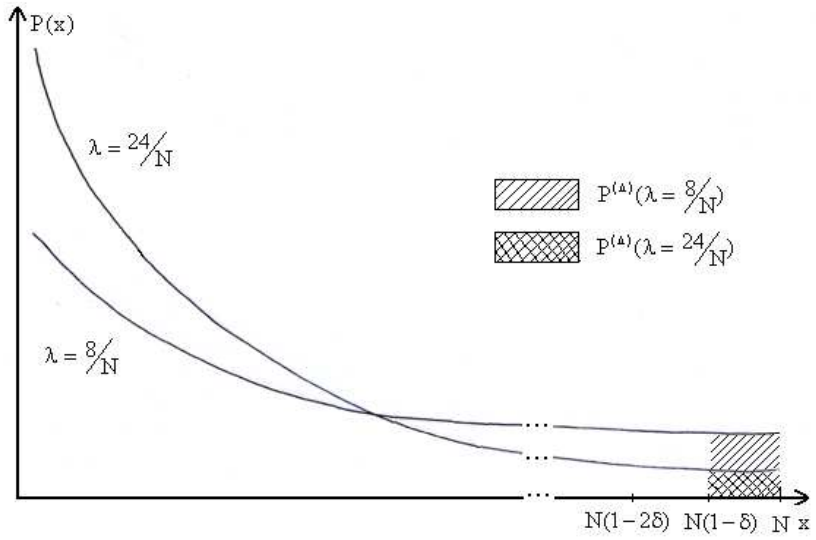

Рис. 2. Геометрический смысл величины $P^{(\Delta)}$

Таким образом, смысл принятого названия модели заключается в том, что распределение вероятностей обращений описывается непрерывной функцией, описывающей дискретные значения p_j для отдельных слов; дискретный характер носят и процессы имитационного моделирования искажений слов и обнаружения ошибок.

Для нивелирования влияния объема словаря на значения $\overline{\rho}^{(\Delta)}$ проведено нормирование словарей до минимального объема N=84 570, «покрывающего» все 6 рассматриваемых словарей; нормирование осуществлялось путем исключения слов случайным образом.

В табл. 3 приведены конкретные данные, полученные в процессе моделирования ($\delta=0.06$), а на рис. 3 – обобщенные результаты, более наглядно отражающие общие тенденции количественных зависимостей между существенными параметрами.

Таблица 3. Результаты моделирования

$\Delta\cdot10^{-3}$	$(N-\Delta)\cdot10^{-}$	$\overline{\rho}^{(\Delta)}\cdot10^{2}$						$P^{(\Delta)}\cdot10^{4}$ $\left(\lambda=8/N\right)$	$P^{(\Delta)}\cdot10^{9}$ $\left(\lambda=24/N\right)$
		СЗр	СЛр	СРЛр	СЗу	СЛу	СРЛу		
0	84,57	0,50	0,38	1,12	0,39	0,28	0,84	0	0
5,08	79,50	0,48	0,36	1,05	0,37	0,27	0,79	2,1	0,12
10,16	74,42	0,45	0,35	0,99	0,35	0,26	0,75	5,4	0,63
15,24	69,35	0,42	0,33	0,92	0,33	0,25	0,70	10,8	2,80
20,32	64,27	0,40	0,31	0,86	0,31	0,23	0,65	19,5	11,96
…	….	…	…	…	…	…	…	…	…
N	0	0	0	0	0	0	0	10^4	10^9

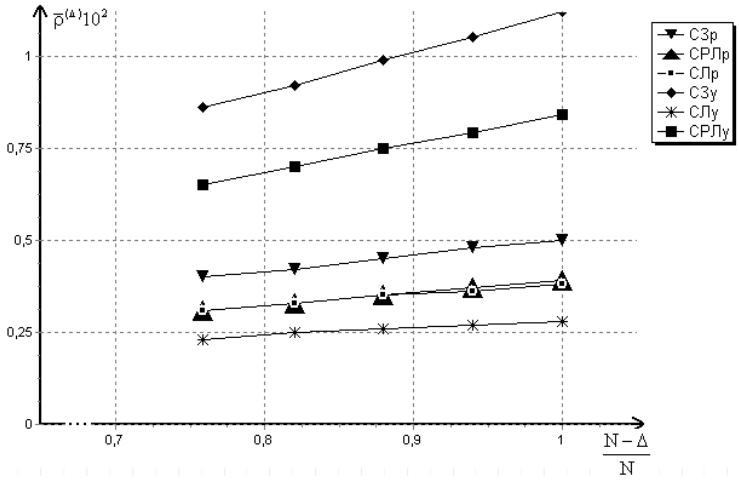

Рис. 3. Количественные зависимости между существенными параметрами

4. Заключение

1. Из приведенных данных видно, что уменьшение объема словаря противоположным образом влияет на факторы его качества (контролирующая способность и вероятность ложных сообщений об ошибке) и в какой именно степени влияет. Приемлемость компромиссного парето-оптимального решения зависит как от абсолютных значений $\bar{\rho}^{(\Delta)}$, $P^{(\Delta)}$, так и от их относительной значимости в конкретном случае. Поэтому моделирование, подобное проведенному выше, для реальных дискретных значений p_j конкретного словаря может дать полезную информацию для принятия обоснованных решений относительно задачи выбора исходного словаря за основу БС и возможной коррекции его объема.

2. Более точную информацию для решения упомянутой задачи могло бы дать «точечное» моделирование с оценкой конкретного вклада потенциально исключаемых слов в значения факторов качества. Например, слово, не являющееся «мишенью» ни для какого ошибочного слова, явно не влияет на итоговую контролирующую способность словаря. Описанная имитационная модель может служить инструментальной основой для такого точечного моделирования.

3. Что касается видимых из приведенных данных «преимуществ» украиноязычных словарей в смысле значений ρ (не имеющих, впрочем, практического значения для рассматриваемой задачи, так как язык БС не является предметом выбора), то одно из возможных объяснений заключается в принятом способе их формирования путем русско-украинской конвертации. Авторам пока не удалось найти подходящих украинских словарей в свободном доступе в формате, приемлемом для проведения соответствующих исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Литвинов В.А. Экспериментальная оценка эффективности автоматического обнаружения типовых ошибок пользователя по словарям русского и украинского языков / В.А. Литвинов, С.Я. Майстренко, О.П. Юденко // Міжнар. наук.-техн. конф. «Системний аналіз та інформаційні технології» SAIT 2012, (Киев, 24 апреля 2012 р.). – Киев, 2012. – С. 374.

2. Литвинов В.А. Контролирующая способность методов автоматического обнаружения типовых ошибок пользователя по словарям русского и украинского языков / В.А. Литвинов, С.Я. Майстренко // Матеріали наук.-практ. конф. з міжнар. участю «Системи підтримки прийняття рішень. Теорія і практика», (Київ, 3 червня 2013 р.). – Київ, 2013. – С. 46 – 48.

3. Словари русского языка [Электронный ресурс]. – Режим доступа: http://speakrus.ru/dict.

4. Словарь Лопатина [Электронный ресурс]. – Режим доступа: http://royallib.ru/book/ lopatin_vladimir/russkiy_ orfograficheskiy_slovar.html.

5. Литвинов В.А. Контроль достоверности и восстановления информации в человеко-машинных системах / В.А. Литвинов, В.В. Крамаренко. – Київ: Техніка, 1986. – 200 с.

Стаття надійшла до редакції 10.12.2013