

УДК 681.3

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

ДИСФУНКЦИЯ РЕФЕРЕНТНОГО СЛОВАРЯ СИСТЕМЫ ПРОВЕРКИ ОРФОГРАФИИ И ПОДХОД К ЕЕ СНИЖЕНИЮ

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

Анотація. Розглядається модель процесу чекінга в системі перевірки орфографії і співвідношення, що визначають показник дисфункції референтного словника. Формулюється задача узгодження критеріїв формування референтного словника, пропонується алгоритм її вирішення, наводяться результати моделювання алгоритму на конкретних словниках української і російської мов.

Ключові слова: помилки користувача, спелл-чекінг, референтний словник, задача «про ранець».

Аннотация. Рассматривается модель процесса чекинга в системе проверки орфографии и соотношения, определяющие показатель дисфункции референтного словаря. Формулируется задача согласования критериев формирования референтного словаря, предлагается алгоритм ее решения, приводятся результаты моделирования алгоритма на конкретных словарях украинского и русского языков.

Ключевые слова: ошибки пользователя, спелл-чекинг, референтный словарь, задача «о ранце».

Abstract. The checking process model in the system of orthography spellchecking and correlations which determine the index of dysfunction of the referential vocabulary is considered. The task of agreeing the criteria for forming the referential vocabulary is stated; its solution algorithm has been proposed; the results of algorithm modeling have been cited for specific vocabularies of the Ukrainian and Russian languages.

Keywords: user errors, spellchecking, referential vocabulary, knapsack task.

1. Введение

В настоящее время системы проверки орфографии (СПО) являются как востребованным самостоятельным продуктом (ОРФО, Language Tool и др.), так и обязательным компонентом текстовых редакторов, поисковых систем, почтовых клиентов, электронных словарей и т.п. [1, 2]. Центральным элементом таких систем – референтный орфографический словарь (РОС), содержащий «правильные» слова некой предметной области, с которыми сравниваются проверяемые слова.

Имеющаяся в доступных источниках проблемно-ориентированная информация отражает, в основном, знания об инструментах СПО – функционал, технология, используемые словари, особенности использования и т.п. Лингвистические и лексикографические аспекты формирования РОС для СПО пока остаются без видимого внимания исследователей, и пользователи располагают, в целом, лишь предоставляемыми возможностями обновления РОС (добавления, исключения и т.п.).

В предлагаемой статье, развивающей тематику [3], представлено изложение возможного подхода к определению качества РОС на основе оценки показателя его дисфункции, связанной с принципиальной невозможностью полного обнаружения ошибок в проверяемых словах.

2. Модель процесса чекинга

Определим качество РОС двумя факторами:

– способностью обнаруживать как можно больше наиболее вероятных (типовых) ошибок;

– способностью выдавать как можно меньше ложных сообщений об ошибках.

Первый фактор оценим показателем дисфункции РОС, который определим через значение p относительного количества слов, искаженных определенными ошибками и совпавших с другими, реально допустимыми словами; соответствующие ошибки системой не обнаруживаются.

Второй фактор оценим вероятностью отсутствия востребованного слова в РОС.

Для уточнения понятия дисфункции и вопросов, связанных с ее количественной оценкой, рассмотрим следующую модель искажения слов и обнаружения ошибок (чекинга).

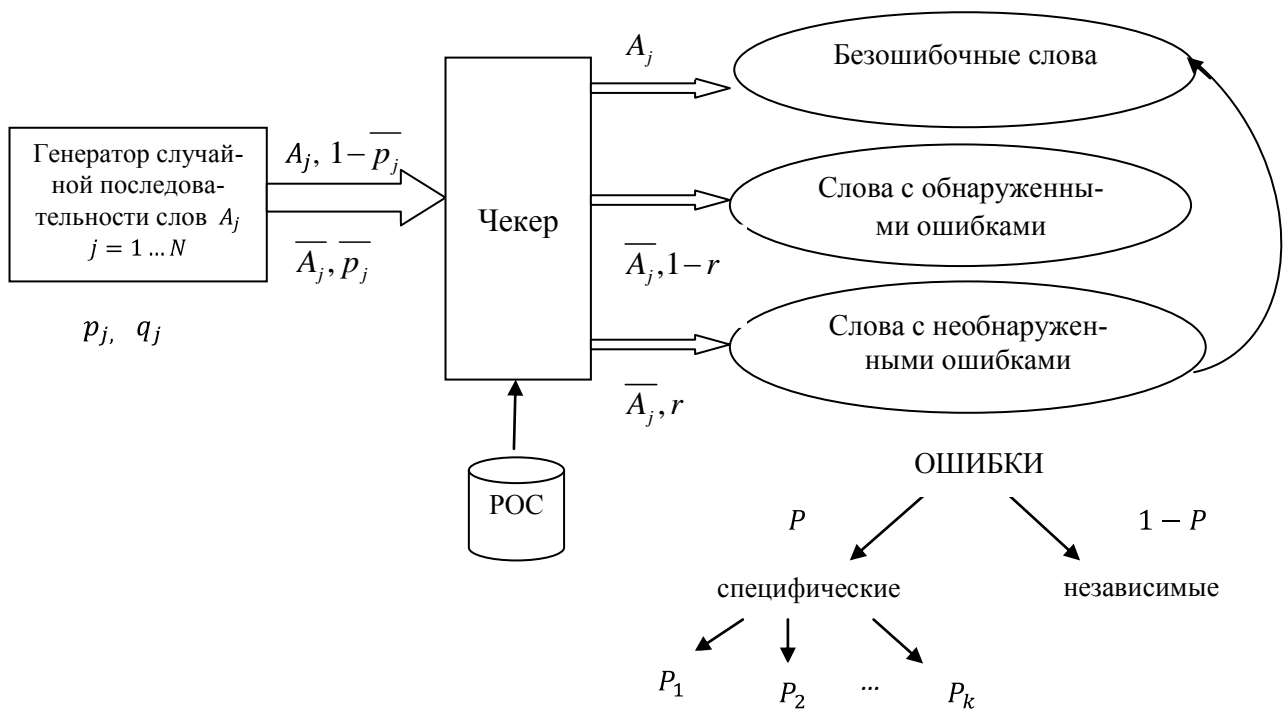


Рис. 1. Схема вероятностной модели искажения слов

Для пояснения принятых обозначений рассмотрим следующую интерпретацию функций представленных на рис. 1 объектов.

Представим генератор слов A_j в виде некой урны, в которой находятся некие фишки со словами A_j ; значение p_j определяет относительное количество фишек с конкретным словом A_j . При заместительном извлечении фишек из урны (генерации последовательности проверяемых слов) фишки могут повредиться, причем конкретная фишка A_j повреждается ($A_j \rightarrow \bar{A}_j$) с вероятностью q_j .

В выходном потоке слов генератора относительное количество \bar{p}_j поврежденных фишек \bar{A}_j определяется произведением $\bar{p}_j = p_j \cdot q_j$ (с точностью до условий нормировки). Чекер сравнивает фишки A_j, \bar{A}_j с эталонными и отделяет поврежденные, причем некоторая часть повреждений (r) не распознается из-за того, что поврежденная фишка \bar{A}_j совпадает с эталонной A_j . Показатель дисфункции r определяет относительное количество не-

обнаруженных повреждений (ошибок) по отношению к полному количеству поврежденных фишек (ошибочных слов).

Определим величину r через свойства слов A_j, \bar{A}_j словаря РОС. Обозначим через r_j относительное количество необнаруживаемых ошибок в слове \bar{A}_j по отношению ко всем возможным ошибкам в этом слове. Тогда

$$r = \sum_{j=1}^N r_j \cdot \bar{p}_j. \quad (1)$$

Ошибки в словах \bar{A}_j разделим на 2 класса: определенные (специфические) ошибки с относительной долей P и произвольные ошибки с относительной долей $(1 - P)$. Контролирующие свойства РОС по отношению к произвольным всевозможным ошибкам принципиально не поддаются учету и анализу, поскольку, если не накладывать ограничений на характер рассматриваемых ошибок, можно утверждать, что для каждого слова A_j существуют ошибки, переводящие его в любое другое слово. Поэтому ограничимся рассмотрением показателя дисфункции ρ по отношению к первому классу ошибок, для общности разделив его на K подклассов с относительной долей, равной P_k , с нормирующим условием $\left(\sum_k P_k = P \right)$.

Тогда для (1) примем

$$r_j \rightarrow \rho_j = \sum_k \frac{\rho_{kj} \cdot P_k}{P}$$

и

$$r \rightarrow \rho = \sum_j \sum_k \bar{p}_j \cdot \frac{\rho_{kj} \cdot P_k}{P}, \quad (2)$$

где ρ_{kj} – относительное количество необнаруживаемых ошибок подкласса k в слове A_j по отношению к всевозможным ошибкам данного подкласса.

Из очевидных соображений

$$\rho_{kj} = \frac{v_{kj}}{V_{kj}}, \quad (3)$$

где v_{kj} – количество совпадений слова \bar{A}_j , искаженного ошибками подкласса k , с другими неискаженными словами A ;

V_{kj} – полное количество всевозможных искажений слова A_j в подклассе ошибок k (количество вариаций \tilde{A}_{kj}) [3].

Таким образом, выражения (2), (3) определяют принятый смысл и значение показателя дисфункции произвольного РС и произвольной функции q_j .

Для процесса последовательного ввода (передачи) символов проверяемых слов можно принять известное допущение о прямой зависимости вероятности искажения слова от его длины:

$$q_j \approx \pi_c \cdot n_j \cdot \mu,$$

где π_c – статистическая вероятность искажения произвольного символа в процессе последовательного ввода;

n_j – количество символов в слове A_j ;

μ – нормирующий множитель.

Тогда с учетом нормирующего условия $\sum_j q_j = 1$.

$$q_j = \frac{n_j}{\sum_{j=1}^N n_j}. \quad (4)$$

В частном случае для $n_j = \text{const} = n$ получим $q_j = \frac{1}{N}$.

Если еще и $p_j = \frac{1}{N}$, то нормированное значение $\overline{p_j}$ тоже равно $\frac{1}{N}$.

3. Задача согласования критериев формирования РОС и подход к ее решению

В настоящее время для одних и тех же предметных областей (в частности, понятий и слов украинского, русского и других естественных языков) существуют различные готовые орфографические словари, отличающиеся широтой охвата тезауруса (объемами) и контролирующими свойствами. Так, например, для русского языка в [3] были рассмотрены свойства «Словаря Зализняка» [4] ($N=92555$ для оригинального словаря и $N=84575$ – для украинской версии), «Словаря Лопатина» [5] ($N=150213$ и $N=135401$ – украинская версия), «Словаря русской литературы» [4] ($N=161730$ и $N=1292440$ – украинская версия). Проведенное моделирование показало, что из 1000 ошибочных слов, искаженных типовыми ошибками пользователя в процессах тайпинга, первый словарь не обнаруживает 5,4 слов, второй – 6,0, третий – 18,4. Показано, что уменьшение объема словаря при прочих равных условиях противоположным образом влияет на факторы качества – значения ρ и φ . С одной стороны, за счет увеличения относительной избыточности представления слов и соответствующего уменьшения случайных совпадений ошибочных слов с допустимыми уменьшается показатель дисфункции ρ . С другой стороны, исключение из РОС слов с ненулевой востребованностью (вероятностью обращения) увеличивает значение φ .

Представим тезаурус T некоей предметной области, состоящий из двух частей: действительной (видимой) и мнимой (скрытой). Действительная часть представляет собой конкретный реальный РОС, а мнимая (МРС) – часть слов T , не представленных в РОС, но потенциально востребованных.

Задачу согласования критериев формирования РОС сформулируем, как задачу исключения и перемещения в МРС слов исходного базового словаря БРС, которые больше других уменьшают ρ и меньше других увеличивают φ . Иными словами, речь идет о построении парето-оптимальной траектории значений ρ , φ по мере перемещения выбранных слов БРС \rightarrow МРС с целью возможного выбора приемлемого сочетания ρ , φ .

Результаты моделирования для ряда русских и украинских словарей, приведенные в [3], получены при использовании простого критерия «тупой силы» – исключения слова A_j с минимальным значением вероятности обращения p_j . В силу простоты критерия эти результаты иллюстрируют, скорее, возможное существование задачи совершенствования БС за счет исключения малозначимых слов, чем ее решение.

С целью формирования «точечного» критерия, оценивающего конкретный вклад потенциально исключаемых слов в значения факторов качества, рассмотрим следующий пример.

Таблица 1. Исключение слов

j	p_j	ρ_j	A_j
1	0,2	0,2	57
2	0,2	0,2	316
3	0,15	0,15	676
4	0,15	0,15	516
5	0,1	0,1	311
6	0,1	0,1	428
7	0,05	0,05	328
8	0,05	0,05	119

Пусть $n_j = const = n$, $N = 8$, $K = 1$ (k -искажения ограничены однократными транскрипциями).

Соответствующий гипотетический «словарь», упорядоченный по убыванию p_j , представлен в таблице.

Как видно из таблицы, в приведенном словаре не обнаруживаются однократные транскрипции $576 \leftrightarrow 676$, $576 \leftrightarrow 516$ и т.д.

Таблица иллюстрирует следующие положения.

1. Исключение нейтрального слова (в словаре это 119), k -искажения которого не вызывают совпадений (то есть необнаруживаемых ошибок с реальными словами, не уменьшают значения ρ , но увеличивают значение φ . Следовательно, исключение нейтральных слов не входит в парето-оптимальные решения.

2. Для каждого слова A_j^s (например, 576) необнаруживаемые прямые k -искажения которого ($A_j \Rightarrow A_s$) вызывают совпадения со словами A_s (в данном случае 676 и 516), существуют необнаруживаемые обратные k -искажения слов A_s^l (676, 516), совпадающие со словом A_j .

Это положение справедливо и для остальных классов рассматриваемых типовых ошибок тайнинга. Так, легко показать, что любой прямой транспозиции соседних символов в слове A_j^s соответствует обратная транспозиция в слове A_s^l ; вставка символа в слове A_j^s соответствует пропуску слова в слове A_s^l .

3. Исключение слова A_j^s (например, 516) уменьшает значение ρ за счет уменьшения количества совпадений обратных k -искажений слов A_s^l со словом A_j (здесь слов 676, 316).

Для пояснения этого положения рассмотрим потенциальные последствия исключения слова, например, 311. В результате ошибка $316 \rightarrow 311$ будет обнаруживаться, но поскольку слово 316 остается в РС, ошибка $311 \rightarrow 316$ остается необнаруживаемой.

Из приведенной в табл. 1 структуры совпадений k -искажений видно, что среднее абсолютное значение количества необнаруживаемых ошибок равно $\rho_{abc} = \sum_j \overline{p_j} \sum_s v_j^s = 1,5$.

Будем искать слово, исключение которого с учетом положения 3 дает наименьшее отношение

$$\alpha_l = \frac{\Delta P}{\Delta \rho_{abc}}$$

В гипотетическом словаре табл. 1 это слово A_4 (516), для которого $\Delta P = 0,15$, $\Delta \rho_{abc} = 0,4$ и $\alpha_l = 0,375$. В результате исключения слова A_4 из РС $\rho_{abc} = 1,1$ и $\varphi = 0,15$.

Приведенные качественные рассуждения обобщает следующий критерий соответствия (\sim), который может быть положен в основу пошагового алгоритма решения задачи выбора для исключения слова A_l :

$$A_l \sim \min_l \alpha_l = \frac{P_l}{\Delta \rho_l}, \Delta \rho_l = \sum_k p_k \sum_s \overline{p_s} \cdot \frac{V_{ks}}{V_{ks}}, \quad (5)$$

где V_{ks} – полное количество всевозможных k -искажений слов A_s .

В приведенной постановке задачу можно рассматривать как некоторое обобщение задачи «о ранце» (Knapsack Problem [5]), а пошаговый алгоритм её решения на основе (1) – как разновидность «жадного» алгоритма GA (Greedy algorithm [6]), в котором в рюкзак помещаются предметы с максимальным отношением цены (в нашем случае $\Delta \rho$) к весу (p_l). Рассматриваемая задача отличается от классической Knapsack Problem тем, что там цена и вес предметов остаются постоянными в процессе укладки рюкзака, а в нашем случае цена предметов, оставшихся после частичной загрузки рюкзака, может меняться в зависимости от того, что было загружено перед этим. Последнее связано с тем, что исключение слова A_l изменяет распределение последствий возможных k -искажений в оставшейся части РОС.

4. Алгоритм решения задачи

В основу алгоритма решения поставленной задачи положен алгоритм GA, адаптированной к отмеченным особенностям, – зависимости параметров (цены и веса) от шага решения.

Общая схема алгоритма приведена на рис. 2. Приведенные на рис. 2 дополнительные обозначения имеют следующие смысл и значения.

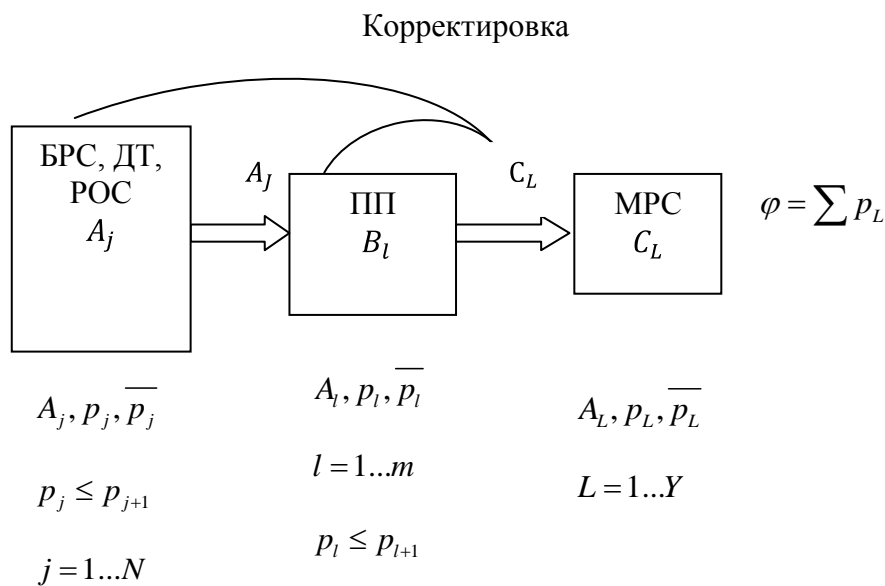


Рис. 2. Обобщенная схема алгоритма

ПП – промежуточный пул объемом m слов, в который помещаются слова A_j – кандидаты на исключение на данном шаге;

ДТ – диагностическая таблица дисфункций слов A_j ;

B_l – слова ПП;

C_L – исключенные слова.

На начальном шаге в ПП из БРС перемещается m слов с минимальными значениями p_j . Далее на каждом очередном шаге выполняется следующая последовательность действий.

1. Поиск в ПП слова $B_l \equiv C_L$, для которого выполняется (4), и перенесение его в МРС.
2. Исключение слова $A_l \equiv B_l \equiv C_L$ из БРС и корректировка параметров ДТ.
3. Расчет значений ρ , φ .
4. Пополнение пула очередным словом, переход к п. 1.

Физически БРС, ПП и МРС являются единым целым, объединенным вместе с диагностической таблицей ДТ. В частности, БРС представляет собой упорядоченный список слов A_j . ДТ содержит информацию о возможных совпадениях слов A_j , искаженных рассматриваемыми типовыми ошибками, с другими словами, а ПП и МРС реализуются соответствующими метками – признаками.

На рис. 3 приведен пример фрагмента интегрированного БРС для словаря Лопатина.

curword	transk	transk_I	vstav	vstav_I	vipad	vipad_I	transp	transp_I
полевать	5	'пожевать', ... , 'полетать'	1	'поплевать'	1	'плевать'	1	'оплевать'
барн	2	'барк', 'барс'	3	'баран', ... , 'барон'	0		0	
шкатулка	0		0		0		0	
обследовательский	0		0		0		0	
ярд	1	'ярь'	1	'лярд'	0		1	'ряд'
выгранный	0		0		0		0	
перепляс	0		1	'вперепляс'	0		0	
румянка	1	'румынка'	0		1	'румяна'	0	
жиреть	0		1	'ожиреть'	0		0	
прочитаться	3	'пропитаться', ... , 'прочищаться'	0		0		0	
садоводство	0		0		0		0	
ксилитный	0		0		0		0	
снизанный	2	'унизанный', 'слизанный'	0		0		0	
плетённый	1	'пленённый'	4	'вплетённый', ... , 'уплетённый'	2	'плетённый', 'плетёный'	0	
неведомый	1	'невесомый'	0		0		0	
денационализация	0		0		0		0	
либерально	0		0		0		0	
неизданный	0		0		0		0	
голенной	0		0		0		0	
неслёживающийся	0		0		0		0	
влечь	2	'слечь', 'впечь'	1	'увлечь'	1	'лечь'	0	
братоубийца	0		0		0		0	
латентный	1	'патентный'	0		0		0	
утайщица	0		0		0		0	
беннеттитовые	0		0		0		0	
ожидаться	1	'ожигаться'	1	'дождаться'	0		0	
спеваться	0		0		0		0	
дефиниция	0		0		0		0	
рукопожатие	0		0		0		0	
пневмомеханический	0		0		0		0	
зарыть	2	'забыть', 'замыть'	1	'закрыть'	0		0	

Рис. 3. Фрагмент интегрированного БРС для словаря Лопатина

Результаты моделирования алгоритма приведены в табл. 2–5. Данные таблицы имеют следующий смысл.

Таблица 2. Словарь Лопатина ($\lambda = \frac{8}{N}$)

$Y \cdot 10^3$	$(N - Y) \cdot 10^3$	$\frac{N - Y}{N}$	Случайное уменьшение			Расчетное уменьшение		
			$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-2}$	$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-4}$
0	84,57	1	0,277	1,0	0	0,277	1,0	0,0
5,08	79,50	0,94	0,261	0,942	5,660	0,225	0,812	4,00
10,16	74,42	0,88	0,246	0,888	11,909	0,156	0,563	28,38
15,24	69,35	0,82	0,232	0,838	17,901	0,062	0,224	299,36
16,16	68,41	0,81	0,229	0,827	19,064	0,035	0,126	632,099

Таблица 3. Словарь Лопатина ($\lambda = \frac{24}{N}$)

$Y \cdot 10^3$	$(N - Y) \cdot 10^3$	$\frac{N - Y}{N}$	Случайное уменьшение			Расчетное уменьшение		
			$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-2}$	$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-4}$
0	84,57	1	0,269	1,0	0,0	0,269	1,0	0,0
5,08	79,50	0,94	0,252	0,937	6,229	0,221	0,822	0,000014
10,16	74,42	0,88	0,236	0,877	12,525	0,149	0,554	0,003772
15,24	69,35	0,82	0,221	0,822	17,957	0,053	0,197	9,72321
16,18	68,40	0,81	0,219	0,814	19,009	0,011	0,041	229,8810

Таблица 4. Украинская версия словаря Лопатина ($\lambda = \frac{8}{N}$)

$Y \cdot 10^3$	$(N - Y) \cdot 10^3$	$\frac{N - Y}{N}$	Случайное уменьшение			Расчетное уменьшение		
			$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-2}$	$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-4}$
0	84,57	1	0,214	1,0	0,0	0,214	1,0	0,0
5,08	79,50	0,94	0,201	0,939	6,077	0,152	0,710	5,881
10,16	74,42	0,88	0,188	0,879	12,150	0,070	0,327	108,187
11,65	72,93	0,86	0,185	0,864	14,001	0,027	0,126	449,207

Таблица 5. Украинская версия словаря Лопатина ($\lambda = \frac{24}{N}$)

$Y \cdot 10^3$	$(N - Y) \cdot 10^3$	$\frac{N - Y}{N}$	Случайное уменьшение			Расчетное уменьшение		
			$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-2}$	$\rho^{(Y)} \cdot 10^{-2}$	$\rho^{(Y)} \div \rho^{(N)}$	$\varphi \cdot 10^{-4}$
0	84,57	1	0,214	1,0	0,0	0,214	1,0	0,0
5,08	79,50	0,94	0,203	0,949	5,487	0,150	0,701	0,000065
10,16	74,42	0,88	0,192	0,897	11,314	0,069	0,322	0,663742
15,24	69,35	0,82	0,187	0,874	13,122	0,009	0,042	1749,524

Параметр λ определяет «крутизну» экспоненциальной кривой, аппроксимирующей ступенчатое распределение плотности вероятности обращений к словам БРС. Для $\lambda = \frac{8}{N}$

распределение характеризуется отношением 20/80 (80% обращений к БС охватывают всего 20% слов), а для $\lambda = \frac{24}{N}$ – отношением 10/90 (с аналогичным смыслом) [3].

При «случайном» уменьшении БРС слова «переносились» в МРС случайным образом, при «расчетном» – в соответствии с результатами работы описанного выше алгоритма (точечным подбором).

Моделирование проводилось для словаря Лопатина (табл. 2, 3) и его украиноязычной версии (табл. 4, 5). Абсолютную результативность работы алгоритма иллюстрируют столбцы 7 (значения показателя дисфункции) и 9 (суммарная вероятность ложных сигналов об ошибке), а относительную – столбцы 5 и 8. В частности, для $\lambda = \frac{8}{N}$ выборочное исключение 6% слов БРС приводит к снижению значения показателя дисфункции на 18% (русский словарь Лопатина) и 29% (украиноязычная версия словаря), а произвольное – всего на 5,8% и 6,1%. Соответствующие значения ρ составляют $5,7 \cdot 10^{-4}$ и $5,9 \cdot 10^{-4}$ для выборочного исключения и $5,7 \cdot 10^{-2}$ и $6,1 \cdot 10^{-2}$ для случайного.

5. Выводы

Построенные модели и приведенный алгоритм дают возможность для конкретного словаря, избранного в качестве базового при формировании РОС системы проверки орфографии, получить данные о значении ожидаемого показателя дисфункции и возможностях его уменьшения за счет приемлемого повышения вероятности ложного сигнала об ошибочности слова. Такие данные могут быть полезны для принятия обоснованных решений с учетом особенностей конкретной СПО. К соответствующим решениям относятся:

- выбор готового базового словаря, описывающего заданную предметную область;
- решение о возможности и целесообразности его улучшения на основе критериев ρ , φ с последующим формированием рабочего референтного словаря РОС;
- частные решения при адаптивном сопровождении РОС (пополнения, исключения) на основе критерия (5).

Точность принимаемых решений зависит от точности знаний о востребованности слов РОС. Грубая оценка дисфункции словаря может быть произведена для $p_j = \frac{1}{N}$ и $q_j = \frac{n_j}{\sum n_j}$. С другой стороны, возможно и повышение точности рассмотренных моделей за счет дополнительного учета вероятностной структуры конкретных ошибок в рамках одного класса, например, транскрипций «соседних» (по клавиатуре) и «отдаленных» символов.

СПИСОК ЛИТЕРАТУРЫ

1. Системы проверки орфографии [Электронный ресурс]. – Режим доступа: <http://compress.ru/article.aspx?id=9511>.
2. Проверка орфографии [Электронный ресурс]. – Режим доступа: <http://www.bestfree.ru/artide/computer/spell-check.php>.
3. Литвинов В.А. Оценка контролируемых свойств базового словаря допустимых слов в системе автоматического обнаружения ошибок пользователя / В.А. Литвинов, С.Я. Майстренко, К.В. Хурцилава // Математичні машини і системи. – 2014. – № 2. – С. 65 – 70.
4. Словари русского языка [Электронный ресурс]. – Режим доступа: <http://speakrus.ru/dict>.
5. Словарь Лопатина [Электронный ресурс]. – Режим доступа: http://royallib.ru/book/lopatin_vladimir/russkiy_orfograficheskiy_slovar.html.
6. Knapsack problem [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/Knapsack_problem.

7. Задача о рюкзаке: жадный алгоритм [Электронный ресурс]. – Режим доступа: http://traditio.ru.org/wiki/Задача_о_рюкзаке:жадный_алгоритм.

Стаття надійшла до редакції 12.04.2017