УДК 681.3

АЛГОРИТМ ПАРЕТООПТИМАЛЬНОГО ФОРМИРОВАНИЯ РЕФЕРЕНТНОГО СЛОВАРЯ СИСТЕМЫ ПРОВЕРКИ ОРФОГРАФИИ

В.А. Литвинов, С.Я. Майстренко, К.В. Хурцилава
Институт проблем математических машин и систем НАН Украины
e-mail: maistrsv@ukr.net

1. Введение

Центральным элементом систем проверки орфографии (СПО [1, 2]) является референтный орфографический словарь (РОС), содержащий «правильные» слова некой предметной области, с которыми сравниваются проверяемые слова. Качество РОС может быть определено двумя основными факторами [3]: относительным качеством ρ необнаруживаемых ошибок (назовем ρ показателем дисфункции словаря) и вероятностью φ ложных сигналов об ошибке. В [3] для ряда русских и украинских орфографических словарей приведены результаты моделирования алгоритма согласования критериев ρ, φ, полученные при простом исключении из РОС слов $\boldsymbol{A_j}$ с минимальными значениями вероятностей обращения $\boldsymbol{p_j}$. В [4] обосновывается "точечный" критерий исключения, оценивающий конкретный вклад потенциально исключаемых слов $\boldsymbol{A_l}$ в значения факторов качества:

$$\alpha = \min_{l} \frac{p_l}{\Delta p_l} \qquad (1)$$

В настоящем сообщении рассматривается пошаговый алгоритм паретооптимального формирования РОС (его коррекция путем исключения слов на основе критерия (1)) и результаты моделирования для некоторых словарей русского и украинского языков.

2. Алгоритм

Целью работы алгоритма является исключение из исходного (базового) РОС слов, которые в наибольшей степени уменьшают значение ρ и в наименьшей степени увеличивают значение φ. В такой постановке алгоритм можно рассматривать как разновидность "жадного" алгоритма (Greedy Algoritm [5]) решения задачи "о ранце" (Knapsack Problem [6]). Особенностью рассматриваемого алгоритма является изменение контролирующих свойств слов, оставшихся после выполнения каждого шага (исключения слова $\boldsymbol{A_l}$).

Основой учета контролирующих свойств РОС и расчета ρ является диагностическая таблица ДТ, в которой для каждого слова приводится:

1) характеристика востребованности;

2) количество и перечни слов, совпадающих с данным словом в случае его искажения каждой из рассматриваемых классов ошибок (транскрипции, пропуски, добавления, транспозиции);

3) служебные указатели, необходимые в процессе работы алгоритма.

Фрагмент ДТ для словаря русского языка приведен на рис. 1. Из приведенного фрагмента видно, что РОС не обнаруживает, в частности, переходы в слово "влечь" слов "слечь", "впечь", искаженных транскрипцией, слова "увлечь" искаженного добавлением

символа и слова "лечь", искаженного пропуском символа. В процессе работы алгоритма при прогнозировании последствий влияния на ρ, φ исключения из РОС очередных слов учитывается, что при исключении например, слова, "влечь" ошибки {"слечь", "впечь", "увлечь", "лечь"} ⇒ "влечь" будут обнаруживаться (поскольку слова "влечь" не будет в РОС), а ошибки "влечь" = {…} - нет (поскольку востребованность слова "влечь" никуда не девается, а "связанные" слова остаются в словаре).

curword	transk	transk_l	vstav	vstav_l	vipad	vipad_l	transp	transp_l
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'
выграненный	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	'плетёный','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	'дожидаться'	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.1. Фрагмент диагностической таблицы

Это свойство определяет принятое значение Δp_l в (1):

$$\Delta p_l = \sum_k p_k \sum_s \bar{p}_s \frac{v_{k_s}}{V_{k_s}}$$

где $\bar{p}_s \approx p_s n_s / \sum_j^N p_j n_j$ - вероятность искажения слова A_s длиной n_s символов;

$\mathbf{P}_k$ - относительное количество ошибок каждого из рассматриваемых классов;

$\frac{v_{k_s}}{V_{k_s}}$ - относительное количество необнаруживаемых искажений слов A_s.

В приведенном примере A_l - это слово "влечь", A_s - связанные с ним слова («слечь» и т. д.).

На начальном этапе работы алгоритма в некий промежуточный пул ПП из РОС перемещается m слов A_j с минимальными значениями p_j. Далее на каждом очередном шаге выполняется следующая последовательность действий.

1. Поиск в ПП слова, для которого $\alpha_l = \min$.
2. Исключение слова A_l и корректировка ДТ.
3. Расчет значений ρ, φ.
4. Пополнение ПП очередным словом A_j.

Физически РОС, ДТ и ПП являются единым целым, а принадлежность слова A_j к указанным областям определяются соответствующими признаками.

3. Результаты моделирования

На рис. 2 и 3 представлены результаты работы алгоритма для русского нормированного словаря Лопатина [3] (кривые ***p)***, и украиноязычной версии этого словаря (кривые ***y***). Принята экспоненциальная аппроксимация гипотетического ступенчатого распределения значений P_j с параметром λ, определяющем "крутизну" экспоненциальной кривой. Принятое на рис 2, 3 значение λ = 8/N характеризуется соотношением 20/80 (80% обращений к РОС охватывает всего 20% слов).

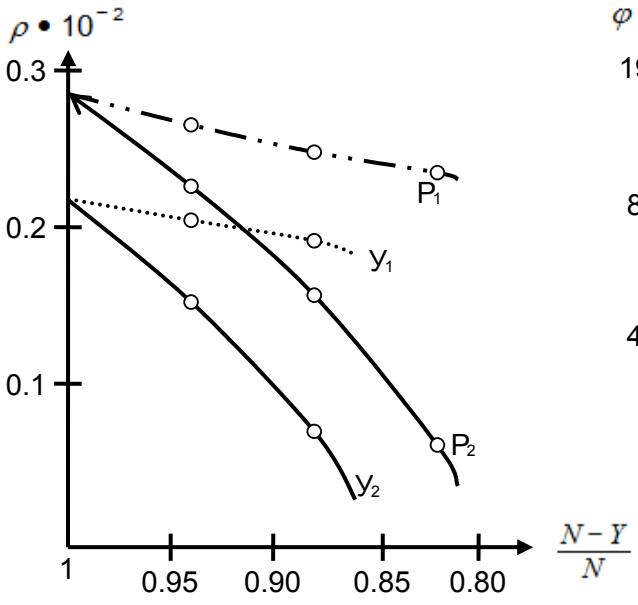

Рис. 2. Тренды значений показателя дисфункции ложной ошибки

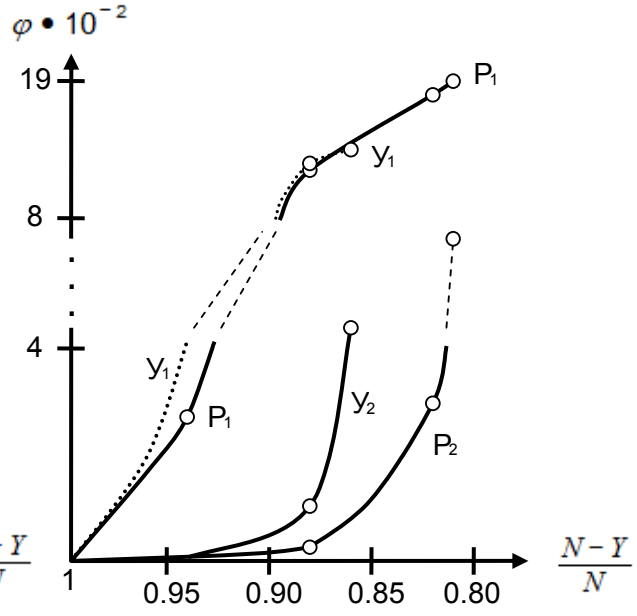

Рис. 3. Тренды значений вероятности

При «произвольном» уменьшении РОС объемом ***N*** слов на величину ***Y*** (кривые 1) слова исключались случайным образом, при «расчетном» (кривые 2) - в соответствии с результатами работы описанного выше алгоритма (точечным подбором).

Как видно из рис. 1, для $\lambda = \frac{8}{N}$ выборочное исключения 6% слов РОС приводит к снижению значения показателя дисфункции на 18% (русский словарь Лопатина) и 29% (русскоязычная версия словаря), а произвольное - всего на 5,8% и 6,1%. Соответствующие значения φ составляют 4*10^(-4) и 5,9*10^(-4) для выборочного исключения и 5,7*10^(-2) и 6,1*10^(-2) для случайного.

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 соответствующих решений с учетом особенностей конкретной СПО.

Литература

1. Системы проверки орфографии [Електронний ресурс]. - Режим доступа: http://compress.ru/article.aspx?id=9511.

2. Проверка орфографии [Електронний ресурс]. - Режим доступа: http://www.bestfree.ru/article/computer/spell-check.php.

3. Литвинов В.А., Майстренко С.Я., Хурцилава К.В. Оценка контролирующих свойств базового словаря допустимых слов в системе автоматического обнаружения ошибок пользователя// Математичні машини і системи. - 2014. - №2. - С. 65-70.

4. Литвинов В.А., Майстренко С.Я., Хурцилава К.В. О критериях оценки качества базового словаря предметно-ориентированной системы spellchecking'а и возможном алгоритме его улучшения //Матеріали наук.-практ. конференції з міжнародною участю «Системи підтримки прийняття рішень. Теорія і практика». – Київ, 2015. С. 139-140.

5. Задача о рюкзаке: жадный алгоритм [Електронний ресурс]. - Режим доступа: http://traditio-ru.org/wiki/Задача_о_рюкзаке:жадный_алгоритм/.

6. Knapsack problem [Електронний ресурс]. - Режим доступа: http:://en.wikipedia.org/wiki/Knapsack_problem.