УДК 681.3

О КРИТЕРИЯХ ОЦЕНКИ КАЧЕСТВА БАЗОВОГО СЛОВАРЯ ПРЕДМЕТНО-ОРИЕНТИРОВАННОЙ СИСТЕМЫ SPELLCHECKING'A И ВОЗМОЖНОМ АЛГОРИТМЕ ЕГО УЛУЧШЕНИЯ

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

Основой систем автоматического обнаружения ошибок пользователя в естественно-языковых словах (в общем случае - нерегулярных алфавитно-цифровых кодах) является базовый словарь БС допустимых («правильных») слов. Качество БС в первую очередь определяется двумя факторами: способностью обнаруживать как можно больше из наиболее вероятных (типовых) ошибок и способностью выдавать как можно меньше ложных сообщений об ошибке.

Первый фактор может быть оценен относительным количеством ρ слов БС, искаженных определенными типовыми ошибками (k-искажениями, $k = \overline{1, K}$) и совпавшими с реальными допустимыми словами.

Второй фактор оценим суммарной вероятностью π отсутствия востребованного слова в БС.

Словарь естественно-языковых слов и типовые ошибки пользователя имеют нерегулярную структуру, поэтому здесь упрощенные оценки, основанные на соотношениях мощностей множеств разрешенных (N) и всевозможных (q^n) комбинаций n символов в алфавите q практически неприемлемы [1].

Как показано в [1, 2], уменьшение объема словаря при прочих равных условиях ведет к уменьшению ρ за счет увеличения относительной избыточности представления слов и соответствующего уменьшения возможностей случайных совпадений ошибочных слов с допустимыми. С другой стороны, исключение из словаря слов с ненулевой вероятностью обращений увеличивает значение π.

Результаты моделирования для ряда русских и украинских словарей, приведенные в [1, 2], получены при использовании простого критерия «грубой 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.

j	p_j	A_j
1	0.2	576
2	0.2	316
3	0.15	676
4	0.15	516
5	0.1	311
6	0.1	428
7	0.05	328
8	0.05	119

Пусть $q = 10$, $N = 8$, $K = 1$ (k–искажения ограничены однократными транскрипциями).
Соответствующий гипотетический «словарь», упорядоченный по убыванию p_j, представлен в таблице.

В приведенном словаре не обнаруживаются однократные транскрипции 576↔676, 576↔516 и т.д. (см. рисунок).

Таблица иллюстрирует следующие положения:
1) Исключение нейтрального слова (в словаре это 119), k–искажения которого не вызывают совпадений

(т.е. необнаруживаемых ошибок) с реальными словами, не уменьшают значения ρ, но увеличивают значение π.

2) Для каждого слова A_l (например, 576), прямые k–искажения которого вызывают совпадения со словами A_s (в данном случае 676 и 516), существуют обратные k–искажения слов A_s (676, 516), совпадающие со словом A_l.

3) Исключение слова A_l (например, 576) уменьшает значение ρ за счет уменьшения количества совпадений обратных k–искажений слов A_s со словом A_l (здесь слов 676, 516).

Из приведенной структуры не обнаруживаемых k–искажений видно, что среднее абсолютное значение количества совпадений $\rho_{àáñ} = p_j \sum_{l,s} v_l^s = 1.5$.

Зададимся максимально допустимым значением $\pi_{\max} \le 0.2$ и будем искать слово, исключение которого даёт наименьшее отношение $\alpha_l = \Delta\pi \big/ -\Delta\rho_{àáñ}$. В гипотетическом словаре это слово $A_4(516)$, для которого $\Delta\pi = 0.15$, $-\Delta\rho_{àáñ} = 0.4$ и $\alpha = 0.375$.

Приведенные качественные рассуждения обобщает следующий критерий соответствия (~), который может быть положен в основу пошагового алгоритма решения задачи:

$$A_l \sim \min_l \alpha_l = \frac{p_l}{\Delta\rho_l},\ \Delta\rho_l = \sum_s p_s \sum_k P_k \big/ V_{ks}, \qquad (1)$$

где P_k - относительное количество каждого из рассматриваемых k-искажений слов БС; V_{ks} - полное количество всевозможных k-искажений слов A_s.

В приведенной постановке задачу можно рассматривать как некоторое обобщение задачи «о ранце» (Knapsack Problem [3]), а пошаговый алгоритм её решения на основе (1) – как разновидность «жадного» алгоритма (Greedy algorithm [4]), в котором в рюкзак помещаются предметы с максимальным отношением цены (в нашем случаи $\Delta\rho$) 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 БС.

Список литературы

1. Литвинов В.А., Майстренко С.Я. Хурцилава К.В. К оценке качества базового словаря в системе автоматического обнаружения ошибок пользователя и возможностей его улучшения // Материалы XIV международной научной конференции им. Т.А.Таран «Интеллектуальный анализ информации».-Киев.-2014.-14-16 мая.- С.139-143.

2. 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.

3. Knapsack problem [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/Knapsack_problem.

4. Задача_о_рюкзаке: жадный_алгоритм [Электронный ресурс]. – Режим доступа: http://traditio-ru.org/wiki/Задача_о_рюкзаке: жадный_алгоритм.