УДК 681.3

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

КРИТЕРИИ И МОДЕЛИ ОЦЕНКИ КОРРЕКТИРУЮЩИХ СВОЙСТВ РЕФЕРЕНТНОГО ОРФОГРАФИЧЕСКОГО СЛОВАРЯ ПРИ АВТОМАТИЧЕСКОМ ИСПРАВЛЕНИИ ТИПОВЫХ ОШИБОК ПОЛЬЗОВАТЕЛЯ

[*]Институт проблем математических машин и систем НАН Украины, г. Киев, Украина
[**]Национальный университет пищевых технологий, г. Киев, Украина

Анотація. *Розглядається логіко-імовірнісна модель корекції виявлених помилок у системі перевірки орфографії, формується критерій оцінки коригувальних властивостей конкретного орфографічного словника. Пропонується імітаційна і аналітична модель для оцінки значень показників правильної і помилкової корекції по відношенню до типових помилок користувача, наводяться результати моделювання для обраних словників російської та української мов. Обговорюється взаємозв'язок між значеннями показників коригувальних і контролюючих властивостей.*
Ключові слова: *помилки користувача, спелл-чекінг, орфографічний словник.*

Аннотация. *Рассматривается логико-вероятностная модель коррекции обнаруженных ошибок в системе проверки орфографии, формируется критерий оценки корректирующих свойств конкретного орфографического словаря. Предлагается имитационная и аналитическая модель для оценки значений показателей правильной и ложной коррекции по отношению к типовым ошибкам пользователя, приводятся результаты моделирования для выбранных словарей русского и украинского языков. Обсуждается взаимосвязь между значениями показателей корректирующих и контролирующих свойств.*
Ключевые слова: *ошибки пользователя, спелл-чекинг, орфографический словарь.*

Abstract. *The logical and probability model of detected error correction in spell checking system is considered; the criterion of assessment of the correction properties of the reference spelling dictionary is being formed. Simulation and analytical models for assessment of values of the right and false correction in relation to the typical typing errors is proposed, results of modeling for the selected dictionaries of the Russian and Ukrainian languages are given. The correlation between values of indicators of the correction and verification properties is discussed.*
Keywords: *typing errors, spell checking, reference spelling dictionary.*

1. Введение

В настоящее время функция проверки орфографии является обязательной компонентой функционала текстовых редакторов, поисковых систем, почтовых клиентов и т.п. В [1] предложен возможный подход к улучшению контролирующих свойств референтного орфографического словаря (РОС), выраженных через относительное количество необнаруживаемых типовых ошибок. Вместе с обнаружением орфографических ошибок многие общие и специализированные текстовые редакторы и другие программы обработки текстов наряду с проверкой орфографии слов предлагают функцию автоматического и полуавтоматического исправления ошибок. Эффективность реализации такой функции должна определяться, с одной стороны, скоростью обработки ошибочного слова и поиска в РОС наиболее подходящего «правильного» слова для исправления, а с другой, качеством коррекции, связанным с возможной ошибочностью выбранного слова.

Известные доступные прикладные исследования в области автоматического исправления ошибок направлены в первую очередь на алгоритмические аспекты проблемы скорости и качества коррекции.

Краткий обзор основных алгоритмов автоматического исправления и нечеткого поиска (fuzzy string search) на основе оценки известных расстояний Левенштейна и Дамерау-Левенштейна [1, 2] приведен в [3, 4]. Типовым решением при выборе алгоритмов для кон-

кретной реализации является использование фонетических алгоритмов [5, 6]. Исследованиям различных модификаций фонетических алгоритмов и альтернативных алгоритмических подходов и систем [7, 8] посвящен ряд публикаций в постсоветских и зарубежных источниках. Вопросам же оценки потенциальных корректирующих свойств самих РОС, в контексте оценки ожидаемого качества коррекции, практически не уделяется внимания.

С целью частичного заполнения отмеченного пробела предлагаемая статья развивает подход [1, 9] в направлении моделирования и оценки характеристик, определяющих потенциальное качество конкретного словаря по отношению к автоматическому исправлению заданных типовых ошибок.

2. Общие положения. Логико-вероятностная модель коррекции

Примем следующие обозначения:

A_j – слово РОС $(j = 1..N)$;

$\overline{A}_j$ – слово РОС, искаженное ошибкой;

$d(A_j^i, \overline{A}_j)$ – функция расстояния, определяющая в некоторой метрике орфографическую близость слова $\overline{A}_j$ и слов РОС $(i = 1..N)$;

$F_1(A_j^i, \overline{A}_j)$ – функция предварительного выбора, определяющая множество слов РОС, для которых $d(A_j^i, \overline{A}) < d_{max}$;

$\hat{A}_j^l$ – слова РОС, для которых $d(\hat{A}_j^l, \overline{A}_j) = \min\limits_i d(A_j^i, \overline{A}_j)$; $l = 1..z$; $z = 0,1,\ldots$, для $z = 0$ таких слов не найдено;

$F_2(\hat{A}_j^l \to A_j)$ функция предпочтения, определяющая выбор из z слов конкретного слова $\hat{A}_j^l$ для корректировки (замены) ошибочного слова $\overline{A}_j$.

В результате корректировки ошибочного слова $\overline{A}_j$ возможно следующее:

– ошибка $A_j \to \overline{A}_j$ не обнаружена (финальное событие $S_{jно}$, вероятность исхода $Q_{jно}$);

– ошибка $A_j \to \overline{A}_j$ обнаружена, (событие S_0) найдено одно или более слов-кандидатов $\hat{A}_j^{il}(m \ge 1)$, функция $F_2(\hat{A}_j^l, A_j)$ определила правильное решение, и корректировка выполнена правильно (финальное событие $S_{jпк}$, вероятность исхода $Q_{jпк}$);

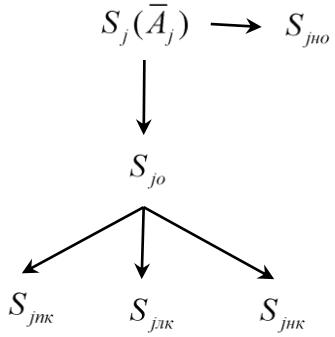
– ошибка $A_j \to \overline{A}_j$ обнаружена, $z \ge 1$, функция $F_2(\hat{A}_j^{il} \to A_j)$ определила ошибочное решение, и корректировка выполнена ложно (финальное событие $S_{jлк}$, вероятность исхода $Q_{jлк}$);

– ошибка $A_j \to \overline{A}_j$ обнаружена, не найдено ни одного $(z = 0)$ слова-кандидата, для которого $d(A_j^i, \overline{A}_j) < d_{\max}$, корректировка не производится (финальное событие $S_{jнк}$, вероятность исхода $Q_{jнк}$).

Цель построения и анализа конкретной логико-вероятностной модели заключается в определении для конкретного РОС значений вероятностей соответствующих исходов, определяющих корректирующие свойства РОС для отдельных слов и словаря в целом.

При реализации процесса коррекции возможны различные решения, определяющие выбор функций расстояний и предпочтений. Для оценки свойств РОС конкретизируем обобщенную модель (рис. 1) для следующих условий.

1. Определяя функцию предварительного выбора, разделим всевозможные ошибки $\bar{A}_j$ на две группы: ансамбль «корректируемых» специфических ошибок ***K***, для которых $d < d_{\max}$, и остальные ошибки («некорректируемые», или произвольные). Традиционно (и в соответствии с составляющими показателя расстояния Дамерау-Левенштейна) отнесем к ансамблю ***K*** типовые орфографические ошибки пользователя – однократные транскрипции E_1, вставки E_2 и удаления E_3 символа, смежные транспозиции E_4.



2. Примем

$$d_0(A_j^i, \bar{A}_j) = \begin{cases} 0 \text{ для } A_j^i \equiv \tilde{\tilde{A}}_j^i, \text{ где } \tilde{\tilde{A}}_j^i \text{ есть вторичное} \\ \quad \text{искажение (вариация 1, 9) слова } \bar{A}_j, \\ d_{\max} \text{ в противном случае} \end{cases} \quad (1)$$

Рис. 1. Обобщенная модель событий

Равенство $d_0(A_j^i, \bar{A}_j) = 0$ означает, что вариация $\tilde{\tilde{A}}_j^i$ совпадает со словом A_j^i. В рамках принятых условий расстояние Дамерау-Левенштейна минимально для слов A_j^i, с которыми совпадает вариация $\tilde{\tilde{A}}_j^i$ в классах $E_1 - E_4$ ансамбля корректируемых ошибок.

3. Для функции предпочтения определим наихудшее решение ─ равновероятный выбор из z совпадений (например, выбор первого же совпадения). Поскольку генерируются все вариации ошибочного слова, по крайней мере, одно совпадение здесь обеспечено, то есть $z \geq 1$.

Логико-вероятностная модель, конкретизированная для принятых условий, приведена на рис. 2.

Приняты следующие дополнительные обозначения для частных событий:

S_{jo} и $S_{jно}$ – ошибка обнаружена / не обнаружена;

S_{j1} и $S_{j2} = \bar{S}_{j1}$ – ошибка принадлежит / не принадлежит ансамблю ***K*** соответственно;

S_{j1}^k – ошибка принадлежит классу E_k;

S_{j11} и $S_{j12} = \bar{S}_{j11}$ – ошибка однозначна $(z = 1)$ / неоднозначна $(z > 1)$;

$S_{jпк1}^k$ – ошибка класса E_k корректируется однозначно правильно;

S_{j121}^k – фактической многозначной ошибке класса E_k соответствует первое $(l = 1)$ из совпадений;

$S_{j122}^k = \bar{S}_{j11}^k$ – фактической многозначной ошибке класса k соответствуют совпадения с $l = 2 \ldots Z$.

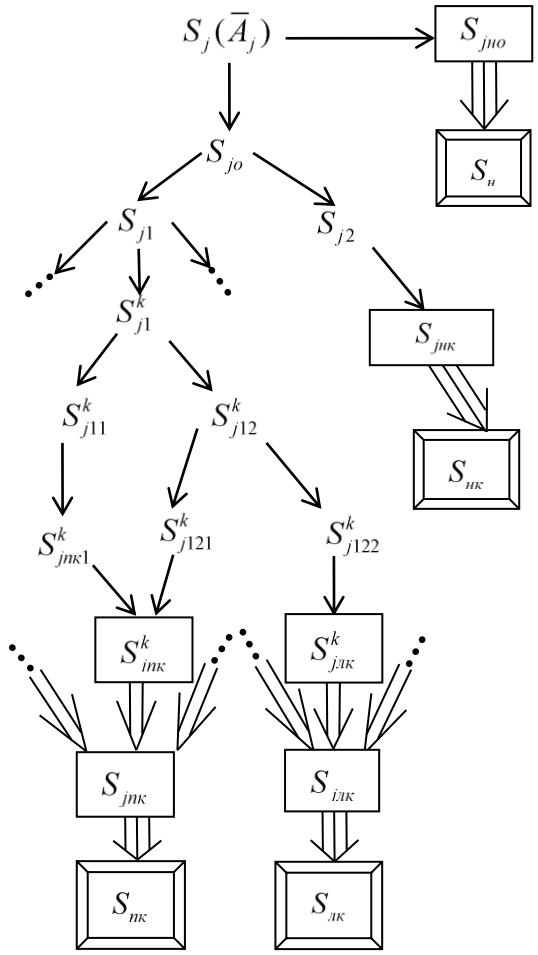


Рис. 2. Логико-вероятностная модель определения корректирующих свойств РОС моделирования

Таким образом,

$S^k_{jпк1} = S_{jo} \wedge S^k_{j1} \wedge S^k_{j11}$,

$S^k_{jпк} = (S_{jo} \wedge S^k_{j1} \wedge S^k_{j11}) \vee (S^k_{jo} \wedge S^k_{j1} \wedge S^k_{j12} \wedge S^k_{j121})) =$

$(S_{jo} \wedge S^k_{j1}) \wedge (S^k_{j11} \vee (S^k_{j12} \wedge S^k_{j121}))$,

$S^k_{jлк1} = S^k_{jo} \wedge S^k_{j1} \wedge S^k_{j12} \wedge S^k_{j122}$,

$S_{пк,\,лк} = \bigvee_j \bigvee_k S^k_{jпк,\,jлк}$.

3. Натурно-имитационная модель коррекции

Натурно-имитационное моделирование процесса искажения и коррекции слов РОС основано на генерации для каждого слова A_j возможных корректируемых ошибок ансамбля ***K***, на проверке обнаруживаемости ошибки, генерации для каждой ошибки возможных вариантов коррекции (обратных искажений) и поиске совпадений в словаре. При этом вероятности промежуточных и финальных событий определяются де-факто для конкретного словаря через соответствующие количества совпадений. Схема моделирования приведена на рис. 3.

Дополнительные обозначения на рис. 3 имеют следующий содержательный и количественный смысл (через $Q_{индекс}$ обозначены вероятности событий $S^k_{индекс}$ схемы (рис. 2):

ν_{jks} – суммарное количество совпадений вариаций $\tilde{\tilde{A}}_{jks}$ ошибочного слова $\bar{A}_{jk}$, искаженного обнаруживаемой ошибкой *s* класса E_k $(\nu_{jks} \geq 1)$;

$\pi_{jks} = \dfrac{1}{\nu_{jks}}$ – вероятность правильной коррекции обнаруживаемой ошибки *s* класса E_k в слове A_j;

$Q^k_{jопк} = \pi_{jk} = \frac{1}{V_{jk}} \sum_{s=1}^{v_{jk}} \pi_j k_s$ – вероятность правильной коррекции произвольной обнаруживаемой ошибки класса E_k в слове A_j. Здесь V_{jk} – суммарное количество всевозможных обнаруживаемых ошибок класса E_k в слове A_j;

$Q_{jопк} = \pi_j = \sum_k \pi_{jk} P_k$ – вероятность правильной коррекции обнаруживаемых ошибок в слове A_j $(\sum_k P_k)$;

$Q_{jопк} = \frac{1}{N} \sum_k \pi_j$ – вероятность правильной коррекции обнаруживаемых ошибок ансамбля **K** в целом по словарю.

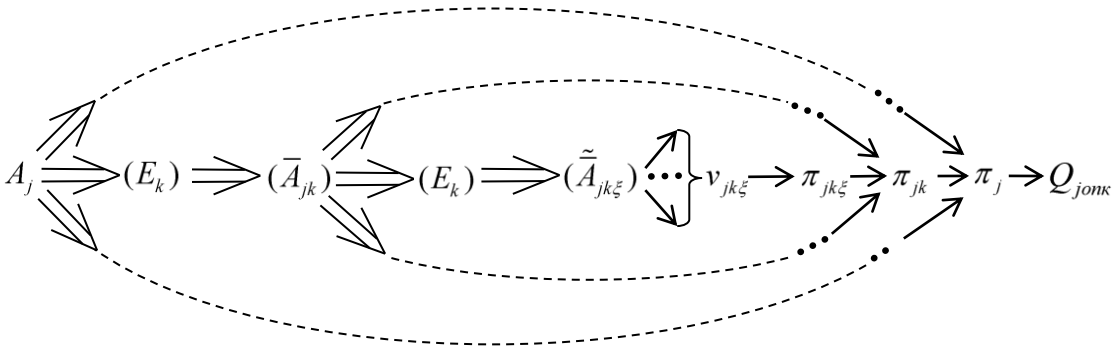


Рис. 3. Схема моделирования

Из очевидных соображений для обнаруживаемых ошибок и принятых допущений $Q_{олк} = P - Q_{опк}$ и для всех ошибок $Q_{пк} < Q_{опк}(1-Q_{но}), Q_{лк} = Q_{олк}(1-Q_{но}), Q_{нк} = 1-P.$

Примеры возможных частных исходов при определении значений π_{jks} для конкретного слова $A_j \coloneqq$ *арак*:

$\bar{A}_{jks} \coloneqq$ *мрак*, ошибка не обнаруживается;

$\bar{A}_{jks} \coloneqq$ *аеак*, $\tilde{\bar{A}}_{jk\xi} \coloneqq \{$*арак*$\}$, $z = 1$, ошибка корректируется однозначно;

$\bar{A}_{jks} \coloneqq$ *прак*, $\tilde{\bar{A}}_{jk\xi} \coloneqq \{$*арак, брак, мрак, трак, рак, парк*$\}$, $z = 6$, при случайном выборе ошибка корректируется правильно с вероятностью 1/6 и ложно с вероятностью 5/6;

$\bar{A}_{jks} \coloneqq$ *аарк*, $\tilde{\bar{A}}_{jk\xi} \coloneqq \{$*барк, карк, марк, парк, тарк, арак*$\}$, $z = 6$, при случайном выборе ошибка корректируется правильно с вероятностью 1/6 и ложно с вероятностью 5/6;

Поскольку при моделировании генерируются все возможные корректируемые ошибки и все варианты их исправления, результаты моделирования (в частности, значения вероятностей $Q_{пк}$ и $Q_{лк}$) полностью характеризуют корректирующие свойства данного конкретного РОС.

Моделирование проведено для набора словарей и значений P_k, принятых в [1, 9] в контексте анализа их контролирующих свойств (дисфункции обнаружения ошибок). В связи с относительно высокой вычислительной трудоемкостью процесса генерации ошибок и вариантов их исправления обработке подвергались случайным образом сформированные выборки объемом 20000 слов (с оценкой соответствующих доверительных вероятностей). Результаты моделирования приведены в табл. 1.

Таблица 1. Результаты моделирования для натурно-имитационной коррекции

Словарь	$Q_{опк}$	$Q_{олк}$	$Q_{пк}$	$Q_{лк}$	$Q_{но}$	$Q_{нк}$
«Словарь русской литературы» $N = 161730$	0,7549	0,1443	0,7410	0,1416	0,0184	~01
«Словарь Лопатина» $N = 150213$	0,8282	0,0709	0,8233	0,0706	0,0060	"
«Словарь Зализняка» $N = 92555$	0,8281	0,0710	0,8236	0,0706	0,0054	"
«Словарь Лопатина» усеченный $N = 84575$	0,8518	0,0474	0,8483	0,0472	0,0038	"
Украинская версия усеченного «Словаря Лопатина» $N = 84575$	0,8610	0,0382	0,8585	0,0381	0,0028	"

Доверительные интервалы для получения средних общих значений $Q_{опк}$, $Q_{олк}$, вычисленные на основе допущения о близком к нормальному закону распределения частных значений $Q_{jопк}$, $Q_{jолк}$ с вероятностью 0,99, составляют $\pm 0,5\%$ для словаря «Русской литературы», $\pm 0,3\%$ для словарей «Лопатина» и «Зализняка» и $\pm 0,2\%$ для усеченных словарей.

Из данных табл. 1 видно, что корректирующие свойства, так же, как и контролирующие, заметно различаются для разных словарей. Так, для словаря «Русской литературы» из 1000 произвольных ошибок не обнаруживается 18,4 ошибок, правильно корректируется 741 ошибка и ложно – 141 ошибка. Соответствующие значения для усеченного словаря Лопатина составляют 2,5, 850 и 47.

Разброс значений $Q_{пк}$, $Q_{лк}$ для разных словарей объясняется двумя факторами. С одной стороны, словарь меньшего объема при прочих равных условиях должен обладать более высокими значениями $Q_{пк}$ и меньшими $Q_{лк}$ за счет большего значения относительной избыточности представления слов и соответствующего уменьшения возможностей совпадения генерируемых вариантов исправления ошибок с реальными словами словаря. Так, для словаря Лопатина объемом 92555 слов значение $Q_{пк} = 0,8233$, а для усеченного (случайным образом) этого же словаря объемом 84575 $Q_{пк} = 0,8483$. С другой стороны, играют роль и чисто лингвистические факторы (язык, тезаурус). Так, для украинской версии усеченного словаря Лопатина, имеющего тот же объем и тот же набор слов, что и русскоязычная версия, $Q_{пк} = 0,8594$. В целом, как видно из данных табл. 1, существует явно высокая степень корреляции между значениями $Q_{но}$ и $Q_{лк}$. Этот фактор в сочетании с отмеченным влиянием относительной избыточности словаря дает основания для следующих предварительных выводов:

– словарь, оптимизированный (по Парето) в отношении контролирующих свойств [1], обладает и лучшими корректирующими свойствами;

– показатель относительной избыточности словаря может быть использован в качестве основы для оценки его корректирующих свойств.

4. Натурно-аналитическая модель корректирующих свойств

Остановимся подробнее на смысле упомянутого в предыдущем разделе понятия «относительная избыточность словаря» и его количественной связи с контролирующими и корректирующими свойствами. Рассмотрим идеализированный гипотетический словарь объемом N слов одинаковой длины n символов в алфавите q.

На рис. 4 показана линейная модель такого словаря, в которой q^n «активных» ячеек обозначают всевозможные значения комбинаций n символов, а выделенные ячейки A_j обозначают комбинации, соответствующие реально существующим словам $(j=1\ldots N)$.

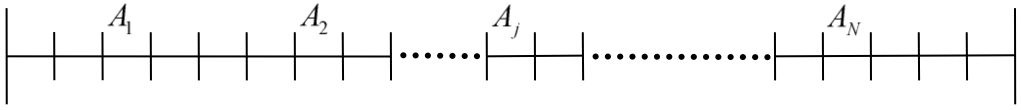


Рис. 4. Линейная интерпретация словаря

Акт проверки «правильности» слова A_j, искаженного произвольной ошибкой $A_j \to \bar{A}_j$, можно рассматривать здесь как акт опустошения ячейки A_j и «бросания» комбинации $\bar{A}_j$ на регистр ячеек. В предположении случайного характера распределения активных ячеек в интервале $0 \div q^n - 1$ вероятность попасть комбинацией $\bar{A}_j$ в занятую ячейку равна $r = N \big/ q^n$, а относительную избыточность словаря C можно оценить как $C = 1 - r = 1 - \dfrac{N}{q^n}$.

Чем больше N при прочих равных условиях, тем больше r, и тем хуже и контролирующие свойства (выше вероятность случайного совпадения ошибочного слова с реальным существующим), и корректирующие свойства (больше количество равноправных вариантов корректировки ошибочного слова, в частности, вариантов полного совпадения).

Для рассматриваемой функции (1) и гипотетического идеализированного словаря возможна вероятностная оценка количества случайных совпадений произвольного ошибочного слова (генерируемой вариации) со словарем на основе модели независимых испытаний Бернулли и соответствующей формулы биномиального распределения:

$$P(g,r,V) = C_V^g r^g (1-r)^{V-g}, \qquad (2)$$

где $P(g,r,V)$ – вероятность получения в точности g случайных совпадений в результате V испытаний, в каждом из которых вероятность благоприятного исхода равна r;

C_g^V – число сочетаний из V по g.

Однако для реального словаря такая оценка значений $Q_{пк}$, $Q_{лк}$ является слишком грубой, так как испытания не являются однородными: генерируемые вариации так же, как и слова словаря, имеют разную длину и различную "лексикографическую уязвимость" в смысле возможностей взаимных совпадений.

Для повышения степени адекватности модели (2) регистр (рис. 4) следует рассматривать в двух измерениях (номер ячейки и длина ячейки), а значения V и r – индивидуально для каждого слова словаря и вариации ошибочного слова.

Предположим заданными вероятность β_{1j} – совпадения со словарем ошибочного слова и вероятность β_{2j} совпадения вариации ошибочного слова $\tilde{\tilde{A}}_j$. Тогда в соответствии с логическими выражениями для событий рис. 2 и модели испытаний (2) мы можем записать следующие выражения для вероятностей частных событий:

$$Q_{jно} = \beta_{1j} \cdot P,$$

$$Q_{jпк} = (1-Q_{jно}) \cdot P[(1-\beta_{2j})^{V_j-1} + \sum_{g=1}^{V_j-1} \frac{1}{g+1} \cdot P(g, \beta_{2j}, V_j - 1)\,],$$

$$Q_{jпк} = (1-Q_{jно}) \cdot P \cdot \sum_{g=1}^{V_j-1} \frac{g}{g+1} \cdot P(g, \beta_{2j}, V_j - 1)\ ,$$

$$Q_{jнк} = (1-Q_{jно})(1-P).$$

При выводе выражений учтено, что из z возможных совпадений проверяемых слов одно определенно правильное, соответствующее искаженному слову $\bar{A}_j$, и g случайных совпадений – ложные. Правильная коррекция имеет место в случае, если $g = 0$ (вероятность события равна $P(0, \beta_{2j}, V_j - 1) = (1-\beta_{2j})^{V_j-1}$) или если из $g+1$ вариантов будет сделан правильный выбор (вероятность $\frac{1}{g+1}$).

Для определения величин β_{1j} и β_{2j} рассмотрим следующую принятую интерпретацию зависимости значений вероятности $\beta_{xj}(x)$ совпадения со словарем x раз искаженного типовой ошибкой слова A_j (рис. 5).

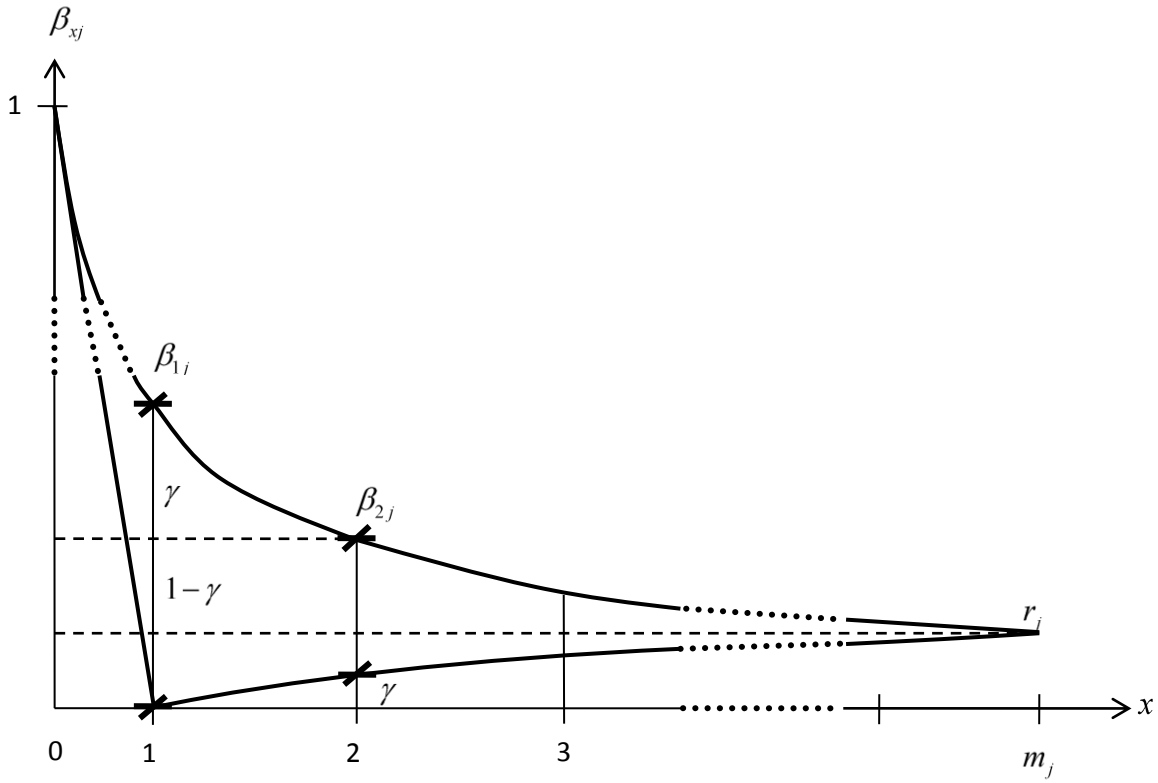


Рис. 5. Графическая интерпретация зависимости значений вероятности совпадения со словарем

Если $x = 0$, $\beta_{0j} = 1$, так как неискаженное слово совершенно определенно совпадает со словарем.

Если $x=1$, величина β_{1j} равна относительному количеству совпадений слов $\bar{A}_j$, искаженных типовыми ошибками. Эта величина определяется «прямым» путем с помощью имитационной модели [1, 9].

Если $x=m_j>>1$, величина β_{mj} асимптотически стремится к значению

$$r_j=\frac{\hat{N}(n_j)}{q^{nj}},$$

где $\hat{N}\ n_j$ – количество слов словаря длиной $n_j\pm1$.

На основании предыдущих рассуждений для $x=2$ положим

$$\beta_{2j}:=\beta_{1j}-(\beta_{1j}-\frac{\hat{N}(n_j)}{q^{nj}})\cdot\gamma,$$

где коэффициент γ определяет крутизну падения кривой $\beta_{xj}(x)$.

Для расчетов по модели принято

$$\beta_{1j}=\frac{\sum_k v_{kj}}{V(n_j)},$$

где v_{kj} – количество совпадений со словарем слова $\bar{A}_j$, искаженного типовой ошибкой E_k;

$V(n_j)=V_j$ – суммарное количество всевозможных типовых ошибок слова длиной n_j символов;

$$\hat{N}(n_j)=N(n_j)\frac{V_1(n_j)+V_4(n_j)}{V(n_j)}+N(n_j-1)\frac{V_2(n_j)}{V(n_j)}+N(n_j+1)\frac{V_3(n_j)}{V(n_j)},$$

где $N(n_j)$, $N(n_j-1)$, $N(n_j+1)$ – фактическое количество слов длиной n_j, n_j-1, n_j+1, $\gamma=0,87$.

В выражении для $\hat{N}(n_j)$ учтено изменение длины слова $\bar{A}_j$, искаженного пропусками и вставками символов, а значение коэффициента γ подбиралось в процессе моделирования (по траектории $0,9\to0,88\to0,86\to0,87$).

Результаты моделирования приведены в табл. 2.

Таблица 2. Результаты моделирования для натурно-аналитической коррекции

Словарь	$Q_{опк}$	$Q_{пк}$	$Q_{лк}$	$Q_{но}$
«Словарь русской литературы» $N=161730$	0,7490	0,7355	0,1470	0,0184
«Словарь Лопатина» $N=150213$	0,8373	0,8323	0,0628	0,0600
«Словарь Зализняка» $N=92555$	0,8383	0,8338	0,0614	0,0054
«Словарь Лопатина» усеченный $N=84575$	0,8608	0,8576	0,0394	0,0038
Украинская версия усеченного «Словаря Лопатина» $N=84575$	0,8698	0,8674	0,0300	0,0028

Как видно из данных табл. 2, результаты расчетов по аналитической модели близки к результатам табл. 1. Так, отклонение значений основного показателя корректирующих свойств ($Q_{пк}$) составляет $0,75\%$ для «Словаря русской литературы» и не превышает $1,2\%$ для остальных словарей. При этом обработка словаря требует на порядки меньше времени.

Так, для используемого маломощного одноядерного компьютера и последовательной схемы моделирования время обработки одного слова по имитационной модели 1 составляло 6 с, а по аналитической модели – 0,04 с. Кроме того, отклонение могло бы быть еще меньше (до 1%) при более тщательном подборе значений γ. Из сравнительных данных испытаний и их интерполяционных оценок следует, что оптимальное значение γ, соответствующее минимальному суммарному отклонению, находится в пределах $\gamma = 0,865 - 0,867$. Существенно, что отклонение результатов мало зависит от лингвистической структуры и содержания словарей, их объемов и рассматриваемых языков. Подобная «устойчивость» дает основания для подтверждения правомерности принятого подходя к построению аналитической модели. В свою очередь, это означает, что аналитическая модель, может быть положена в основу оценки корректирующих свойств по отношению к более сложным ошибкам, находящимся на расстоянии Дамерау-Левенштейна большем, чем типовые ошибки.

5. Выводы

1. Представленные модели могут быть положены в основу инструмента сравнительной оценки потенциальных корректирующих свойств конкретного орфографического словаря по отношению к типовым ошибкам пользователя. При этом аналитическая модель может служить для предварительных решений, а имитационная – для уточненных оценок, полнота которых определяется учетом всех возможных типовых ошибок и вклада каждого слова в итоговое значение $Q_{пк}$. При известных вероятностях искажения слова A_j этот вклад может быть соответствующим образом взвешен.
2. Существует высокая степень корреляции между значениями показателей контролирующих и корректирующих свойств ($Q_{но}$ и $Q_{пк}$). С одной стороны, это дает основания полагать, что словари, улучшенные в отношении контролирующих свойств [1], обладают и лучшими корректирующими свойствами. С другой, говорить о некоем общем показателе «орфографической уязвимости» словаря по отношению как к отдельным типовым ошибкам, так и их кратным комбинациям. Если принять за основу принятую интерпретацию зависимости значений вероятности совпадения со словарем от кратности типовой ошибки (рис. 4), интерпретацию, правомерность которой предварительно подтверждают отмеченные в разд. 3 результаты моделирования, то в качестве такого общего показателя могло бы быть принято значение вероятности совпадения произвольного слова, искаженного однократной типовой ошибкой. Количественная оценка возможной связи этого показателя с корректирующими свойствами словаря при иных функциях F_1, F_2 нуждается в отдельном исследовании.

СПИСОК ИСТОЧНИКОВ

1. Литвинов В.А. Дисфункция референтного словаря системы проверки орфографии и подход к ее снижению / В.А. Литвинов, С.Я. Майстренко, К.В. Хурцилава // Математичні машини і системи. – 2017. – № 2. – С. 39 – 48.
2. Расстояние Дамерау-Левенштейна [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Расстояние Дамерау – Левенштейна.
3. Нечёткий поиск в тексте и словаре [Электронный ресурс]. – Режим доступа: https://habrahabr.ru/post/114997/.
4. Расстояние Левенштейна в MySQL и алгоритмы нечёткого поиска средствами PHP [Электронный ресурс]. – Режим доступа: https://habrahabr.ru/post/342434/.
5. Фонетические алгоритмы [Электронный ресурс]. – Режим доступу: https://habrahabr.ru/post/114947/.
6. Phonetic Algorithms [Електронний ресурс]. – Режим доступу: https://deparkes.co.uk/2017/12/01/phonetic-algorithms/.

7. Hodge V.J. A comparison of standard spell checking algorithms and a novel binary neural approach / V. J. Hodge, J. Austin // IEEE Transactions on Knowledge and Data Engineering. – 2003. – P. 1073 – 1081.
8. de Amorim R.C. Effective Spell Checking Methods Using Clustering Algorithms [Електронний ресурс] / R.C. de Amorim, M. Zampieri. – Режим доступа: http://www.aclweb.org/anthology/R13-1023.
9. Литвинов В.А. Оценка контролирующих 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.

Стаття надійшла до редакції 12.04.2018