УДК 681.51:57

В.А. ЛИТВИНОВ, С.Я. МАЙСТРЕНКО, В.И. ХОДАК

ОБНАРУЖЕНИЕ И ИСПРАВЛЕНИЕ ОШИБОК ПОЛЬЗОВАТЕЛЯ ПО СЛОВАРЯМ ДОПУСТИМЫХ СЛОВ И СЛОВОСОЧЕТАНИЙ

Abstract: *The method of the checking, automatic or semiautomatic to identifications and corrections spelling and semantic user' s errors, which were made in step of primary shaping or entering the document, is considered. Models of the estimation of the probabilistic features, defining efficiency and area of the expedient using the method are offered.*
Key words: *errors' s automatic identification, spelling and semantic errors, dictionary of the possible words.*

Анотація: *Розглядається метод контролю, автоматичної або напівавтоматичної ідентифікації та виправлення орфографічних і смислових помилок користувача, що були зроблені на етапі первинного формування або вводу документа. Запропоновано моделі оцінки імовірнісних характеристик, які визначають ефективність та область доцільного використання методу.*
Ключові слова: *автоматична ідентифікація помилок, орфографічні та смислові помилки, словники допустимих слів.*

Аннотация: *Рассматривается метод контроля, автоматической либо полуавтоматической идентификации и исправления орфографических и смысловых ошибок пользователя, допущенных на этапе первичного формирования или ввода документа. Предложены модели оценки вероятностных характеристик, определяющих эффективность и область целесообразного применения метода.*
Ключевые слова: *автоматическая идентификация ошибок, орфографические и смысловые ошибки, словари допустимых слов.*

1. Введение

Автоматическое обнаружение, идентификация и исправление ошибок пользователя является важным фактором повышения уровня интеллектуализации интерфейса человек-компьютер. В [1] исследуются модели и характеристики общих методов и алгоритмов автоматической идентификации и исправления типовых ошибок пользователя на основе словаря допустимых слов. В теоретических моделях [1] и практических коммерческих программных продуктах подобного назначения [2–4] в качестве словарей подразумеваются и используются орфографические словари соответствующих языков, т.е. словари, определяющие правильное написание (представление) отдельных слов. Целью настоящей работы является обобщение и распространение методов и подходов [1] на ошибки, имеющие более сложную грамматическую и смысловую структуру, и построение моделей оценки вероятностных характеристик, определяющих возможности их обнаружения, автоматической идентификации и исправления.

2. Основные понятия и определения

Нами рассматривается система ввода, контроля достоверности, идентификации и исправления обнаруженных ошибок.

Для основных понятий и компонентов системы примем следующие определения и обозначения:

A_k^l – входное l-е слово (атрибут некоего информационного объекта) длиной n_k символов в алфавите q_k;

$A^l = \left(A_1^l \ldots A_k^l \ldots A_K^l \right)$ – входное словосочетание (кортеж из K атрибутов);

T_k^i – словарь допустимых значений k-го слова $(i = 1,...,N_k)$;

TT^j – словарь допустимых значений словосочетаний $(j = 1,...,N)$.

Примечание. Словари могут быть как реальными, так и виртуальными. Во втором случае виртуальный словарь может быть задан некоторым логико-арифметическим соотношением, определяющим правило построения допустимых слов и словосочетаний (например, избыточным кодом контроля по модулю и т.п.).

Орфографическую ошибку определим как переход $A_k^l \to \overline{A}_k^l$, в результате которого образуется значение, *отсутствующее* в словаре T_k^i. Такая ошибка обнаруживается в результате проверки допустимости отдельного взятого слова.

Смысловую ошибку определим как переход $A_k^l \to \overline{A}_k^l$, в результате которого образуется допустимое значение, *разрешенное* словарем T_k^i. Смысловая ошибка может быть обнаружена (или нет) только в результате проверки допустимости словосочетания в целом.

Примечание. Смысловая ошибка в приведенной трактовке может иметь двоякое происхождение:

1) как результат неумышленного искажения отдельных символов слова при формировании документа (сообщения) или его вводе;

2) как результат неправильного истолкования формируемого словосочетания и замены одного значения атрибута другим, тоже формально допустимым.

В качестве наглядного иллюстративного примера рассмотрим правильное словосочетание русского языка "кот бежит". Орфографическая ошибка "кот → крт" может быть обнаружена путем проверки допустимости значения "крт", отсутствующего в словаре. Смысловая ошибка "кот → кит" не обнаруживается на уровне орфографического контроля, но ошибка в словосочетании "кит бежит" – налицо. И, наконец, смысловая ошибка "бежит → лежит" не обнаруживается вообще (без более широкого контекстного анализа, но этот уровень контроля мы здесь не рассматриваем).

Статическая структура рассматриваемой системы приведена на рис. 1.

Сформулируем исходное правило построения TT^j:

Если $A_1^i ... A_k^i ... A_K^i \in TT^j$, то

$$\forall A_k^i \in T_k^i \ \left(i = 1,...,N_k; k = 1,...,K\right). \qquad (1)$$

Суть правила заключается в том, что допустимые словосочетания состоят исключительно из допустимых слов.

Примечание. Правило не носит абсолютного характера, а лишь ограничивает область рассматриваемых ситуаций случаями, наиболее типичными для представления данных. Для более сложных случаев, которые теоретически могут быть свойственны в частности, отношениям между элементами знаний, это правило может и не выполняться. Например, некие лица X, Y, Z вместе $\left(XYZ\right)$ могут быть совместимыми, а попарно $\left(XY, XZ, YZ\right)$ – нет.

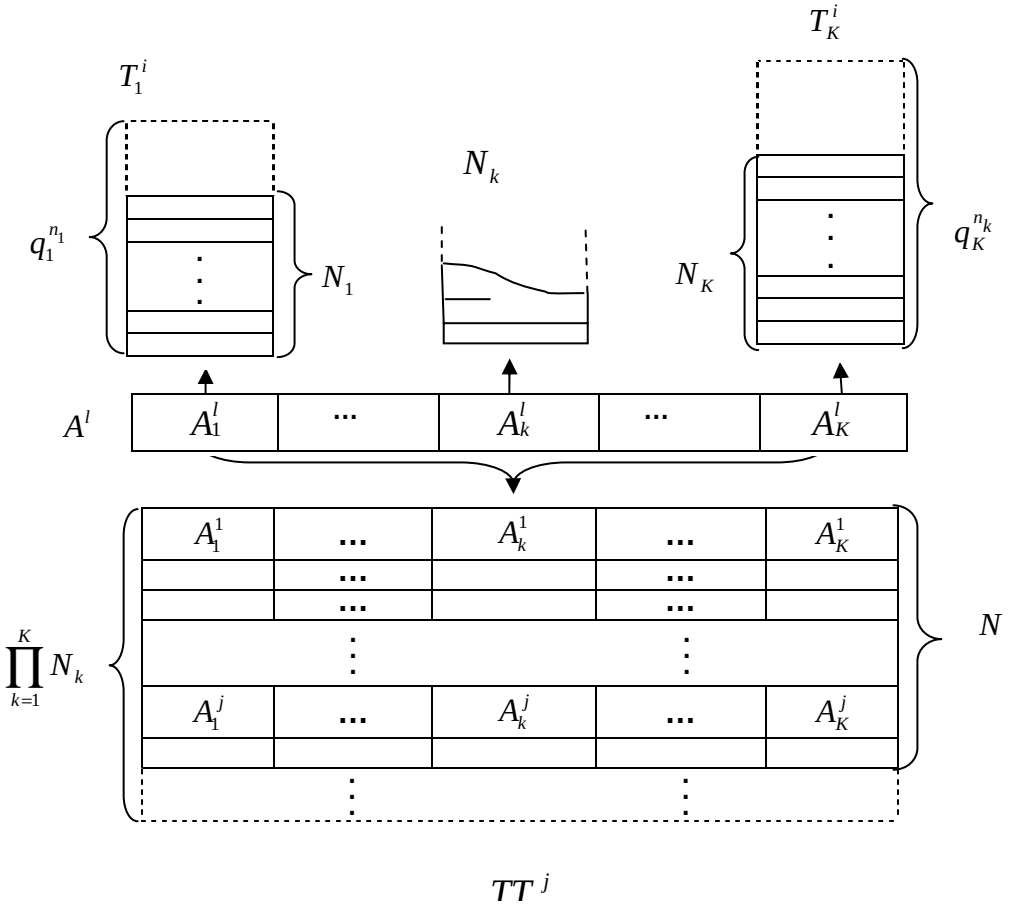

Рис. 1. Структура рассматриваемой системы

Из (1) вытекают следующие свойства проверяемых словосочетаний $A_1^l ... A_k^l ... A_K^l$ по отношению к возможным ошибкам.

1. Если $A_1^i ... A_K^i \notin TT^j$ и $\exists A_k^l \notin T_k^i$, то произошла орфографическая ошибка в слове A_k^l.

2. Если $A_1^i ... A_K^i \notin TT^j$ и $\forall A_k^l \in T_k^i$, то произошла смысловая ошибка из-за формально допустимого искажения неизвестного слова.

3. Если $A_1^i ... A_K^i \in TT^j$, то $\forall A_k^l \in T_k^i$, и ошибка в словосочетании отсутствует (или не обнаружена).

3. Общая схема контроля-коррекции словосочетания

Возможны два варианта этапности контроля-коррекции словосочетания.

В первом варианте вначале проверяются отдельные слова на наличие орфографических ошибок. Ошибки (при их наличии) обнаруживаются, идентифицируются и исправляются по алгоритмам моделей [1]. Затем выполняются контроль совместимости слов словосочетания, идентификации и коррекция смысловых ошибок.

Во втором варианте вначале производится контроль совместимости, а затем, в зависимости от результата, контроль отдельных слов и далее – идентификация и исправление орфографических, а затем смысловых ошибок.

Явная предпочтительность второго варианта определяется тем фактором, что ошибок следует ожидать далеко не в каждом словосочетании. Следовательно, во втором варианте большая часть контрольных проверок ограничится проверкой «группового» условия $A_1^l ... A_K^l \in TT^j$.

Общая схема алгоритма контроля-коррекции на основе второго варианта включает следующие этапы:

1. Проверка $A_1^l ... A_k^l \in TT^j$. Если результат положительный, то словосочетание считается безошибочным. Иначе – в словосочетании имеется ошибка; переход к п. 2.

2. Проверка условия $\forall A_k^l \in T_k^i$. Если результат отрицательный и $\exists A_l^k \notin T_k^i$, то произошла орфографическая ошибка в слове A_k^l. Она обрабатывается схемой АИК [1] с последующим переходом к п. 1. $(l := l+1)$. Иначе в словосочетании имеется смысловая ошибка. Переход к п. 3.

3. Идентификация смысловой ошибки и ее исправление (с участием или без участия пользователя). Переход к п.1.

Граф, отображающий структуру частных исходов-событий алгоритма, приведен на рис. 2.

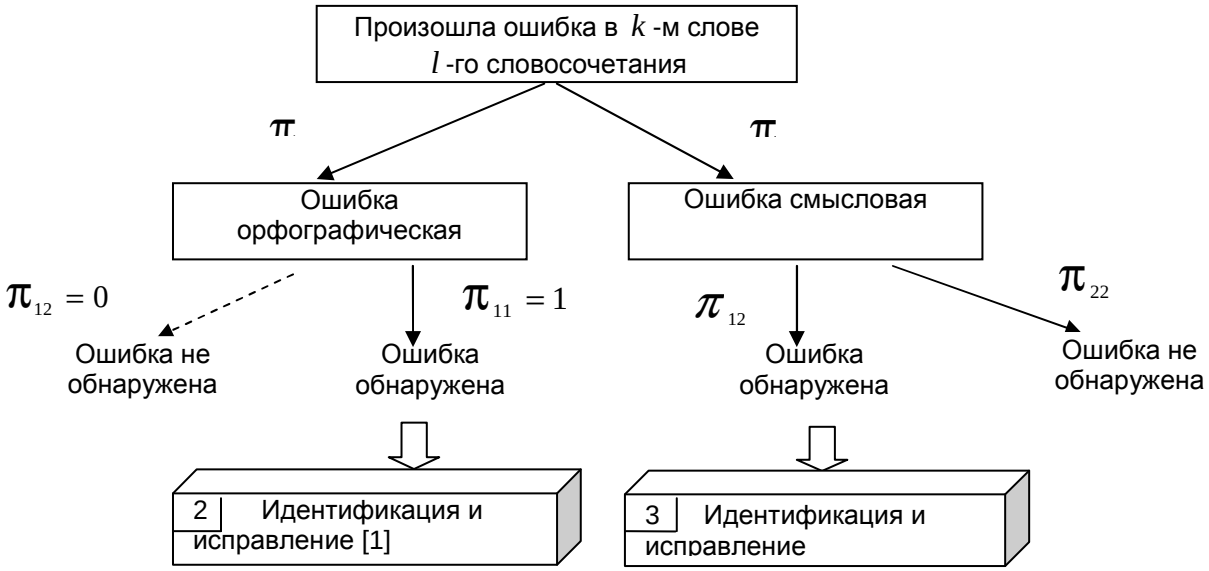

Рис. 2. Структура событий

При оценке вероятностей $\pi_1, \pi_2, \pi_{21}, \pi_{22}$ будем исходить из следующего допущения о их распределении:

а) множества допустимых значений слов A_k^i мощностью N_k среди множества их всевозможных значений мощностью $q_k^{n_k}$;

б) множества допустимых значений словосочетаний $A_k^j ... A_K^j$ мощностью N среди множества их всевозможных значений мощностью $\prod_{k=1}^{K} N_k$.

С учетом определения типов ошибок, принятого допущения о распределении вероятностей и свойств 1–2 получим,

$$\pi_1 = 1 - \frac{N_k}{q_k^{n_k}}; \quad \pi_2 = \frac{N_k}{q_k^{n_k}}; \quad \pi_{21} = 1 - \frac{N}{\prod_{k=1}^{K} N_k}; \quad \pi_{22} = \frac{N}{\prod_{k=1}^{K} N_k}. \tag{2}$$

Из приведенных соотношений видно, что контролирующая способность (относительное количество обнаруженных ошибок) для орфографического контроля равна:

$$D_{орф} = \pi_1 = 1 - \frac{N_k}{q_k^{n_k}},$$

а для совокупности орфографического и смыслового контролирующая способность будет равна:

$$D_{орф,см} = \pi_1 + \pi_2 \cdot \pi_{21} = \left(1 - \frac{N_k}{q_k^{n_k}}\right) + \frac{N_k}{q_k^{n_k}}\left(1 - \frac{N}{\prod_{k=1}^{K} N_k}\right) = 1 - \frac{N_k}{q_k^{n_k}} \cdot \frac{N}{\prod_{k=1}^{K} N_k}.$$

Примем для всех $k = 1,...K$ отношение $r = \frac{N_k}{q_k^{n_k}} = const$, $R = \frac{N}{\prod_{k=1}^{K} N_k}$.

Тогда $D_{орф,см} = 1 - rR$.

При $r < 1$ и $R < 1$ (а эти неравенства практически всегда достаточно глубоки) величина $D_{орф,см} \rangle\rangle D_{орф}$. Например, для $r = 10^{-2}$ и $R = 10^{-2}$ только орфографический контроль теоретически позволяет обнаружить 99% ошибок, а орфографический + смысловой – 99,99%.

4. Идентификация и исправление смысловых ошибок

Идентификация смысловых ошибок и оценка вероятностных характеристик этого процесса возможны на основе применения и исследования механизма [1] генерации обратных искажений ошибочного слова по словарям T_k^i.

Будем интерпретировать $K-$ кратное словосочетание как $K-$ символьное гиперслово в смешанном алфавите $N_1,...N_k,...N_K$, а смысловую ошибку в гиперслове – как однократную транскрипцию k-го гиперсимвола. Под гипервариацией будем понимать замену текущего значения слова A_k^l на очередное из словаря T_k^i. В контексте этих определений процесс автоматической (полуавтоматической) идентификации заключается в генерации гипервариаций в классе однократных транскрипций и проверке допустимости образованного гиперслова по словарю TT^j.

Полное количество генерируемых гипервариаций V_K определяется простым выражением:

$$V_K = \sum_{k=1}^{K} N_k - K. \tag{2}$$

В зависимости от используемого алгоритма разрешения возможной неоднозначности совпавшее гиперслово словаря TT^j может предлагаться пользователю для подтверждения корректировки либо исправляться автоматически. При условии сохранения допущения о случайном распределении значений словосочетаний в словаре TT^j и применении соотношений общей модели испытаний Бернулли вероятность $P(g,R,V_K)$ в точности g случайных совпадений определяется выражением

$$P(g,R,V_K) = C^g_{V_K} \cdot R^g \cdot (1-R)^{V_K-g}.$$

Вероятность одновременного искажения более одного слова будем считать пренебрежимо малой. В этом случае в терминах [1] "корректируемой" ошибкой является однократная транскрипция, и вероятность ее появления (при условии, что в словосочетании обнаружена смысловая ошибка) равна 1, так что все выражения [1] для вероятностей правильной P_{AK}, ложной $P_{ЛК}$ и ручной $P_{РК}$ коррекции соответственно упрощаются. Например, для наиболее простого (в смысле анализа) и перспективного для применения в рассматриваемом приложении алгоритма 3, требующего подтверждения предлагаемой корректировки пользователем,

$$P_{AK} = \pi(m),$$

$$P_{ЛК} \approx 0,$$

$$P_{РК} = 1-\pi(m),$$

где $\pi(m)$ определяет вероятность того, что среди m предложенных вариантов корректировки содержится правильный вариант. Как показано в [5],

$$\pi(m) = \sum_{g=0}^{m-1} C^g_{V_K} R^g (1-R)^{V_K-g-1} + \sum_{g=m}^{V_K-1} \frac{m}{g+1} C^g_{V_K-1} R^g (1-R)^{V_K-g-1}.$$

В таблице приведены иллюстративные результаты расчета значений $P_{AK} = \pi(m)$ для следующих данных:

$K = 3,\ N_1 = 10^2;\ N_2 = 5\cdot 10^2;\ N_3 = 10^3\ ;\ N = 5\cdot 10^4, 5\cdot 10^3, 5\cdot 10^2.$

В этом случае, как следует из (1) и (2), $V_k = 1597,\ R = 10^{-3}, 10^{-4}, 10^{-5}.$

Таблица. Иллюстративные результаты расчета значений P_{AK}

R	m				
	1	2	3	4	5
10^{-3}	0,4994	0,7964	0,9315	0,9857	0,9953
10^{-4}	0,9242	0,9960	0,9998	1,0000	1,0000
10^{-5}	0,9920	0,9999	1,0000	1,0000	1,0000

Как видно из таблицы, для данных значений K, N_k, N результаты полуавтоматической идентификации смысловой ошибки, допущенной при вводе, можно считать вполне приемлемыми –

правильное значение атрибута находится среди 3–5 альтернатив с вероятностью, весьма близкой к 1 (с точностью в пределах 5 знаков). Для ошибки, допущенной при формировании первичного документа, возможно только автоматическое исправление (алгоритмы 1, 2 [1]) или возврат документа на проверку и исправление к первоисточнику. Уверенное автоматическое исправление возможно только в случае таких сочетаний значений N_k, N, при которых $m \approx 1$.

Поскольку, как известно, среднее число "удачных" исходов для испытаний Бернулли равно RV_K (в наших обозначениях), то обобщенным ориентировочным критерием оценки применимости метода автоматической идентификации и исправления смысловых ошибок может служить неравенство $RV_k < \varepsilon$. Выражая левую часть через “первичные” параметры словарей N_k, N, получим:

$$\frac{N \cdot \sum_{k=1}^{K} N_k}{\prod_{k=1}^{K} N_k} < \varepsilon,$$

где ε – принятое допустимое относительное количество “ручных” (или ложных) исправлений.

Например, значение ε = 0,01 означает, что примерно на 100 случаев идентификации смысловой ошибки в 1 случае в дополнение к правильной вариации со словарем TT^j произойдет еще одно случайное ложное совпадение. В этом случае алгоритм 1 [1] выполнит правильную автоматическую коррекцию с вероятностью 0,5, а алгоритмы 2, 4 предложат идентифицировать и исправить ошибку "вручную".

5. Выводы

Совместный контроль орфографических и смысловых ошибок позволяет существенно повысить достоверность вводимой информации.

Для автоматической идентификации и исправления смысловых ошибок может быть использован метод генерации обратных искажений словосочетания и проверки их допустимости. Полученные в работе [1] соотношения вместе с выражениями (1)–(3) позволяют получить ориентировочные оценки результативности метода. В перспективе применение описанного подхода возможно и для более сложных ошибок, – в частности, сочетания кортежей таблицы (экземпляров входных форм). Это случай нуждается в отдельном исследовании.

СПИСОК ЛИТЕРАТУРЫ

1. Кузьменко Г.Є., Литвинов В.А., Майстренко С.Я., Ходак В.І. Алгоритми і моделі автоматичної ідентифікації та корекції типових помилок користувача на основі природної надмірності // Математичні машини і системи. – 2004. – № 2. – С. 134–148.
2. AfterScan. http://www/afterscan.com/ru.
3. http://www.abbyy.ru/products/handprint/WP_form_processing_65.pdf.
4. Редактор 1ДФ. http://octant.com.ua.
5. Литвинов В.А., Майстренко С.Я., Ступак Н.Б. Некоторые оценки вероятностных характеристик процесса автоматической идентификации ошибок пользователя на основе эталонного словаря // УСиМ. – 2001. – № 2. – С. 21–24.