Э.В. Нагорнович (Минск, МГЛУ)

ЛОГИ ПОИСКОВЫХ СИСТЕМ КАК ОСНОВА ФОРМАЛИЗАЦИИ ПРОЦЕДУРЫ АВТОМАТИЧЕСКОЙ КОРРЕКЦИИ ОРФОГРАФИЧЕСКИХ ОШИБОК В ПОИСКОВЫХ ЗАПРОСАХ ПОЛЬЗОВАТЕЛЕЙ

В идеале каждая компьютерная система обработки текста должна быть оснащена средствами автоматизированного или даже автоматического исправления допущенных человеком ошибок. К настоящему времени методы обнаружения ошибок разработаны только применительно к орфографическим ошибкам, поскольку большинство ошибок в тексте являются именно таковыми. Известны три основных метода автоматического обнаружения орфографических ошибок в письменном тексте: статистический, полиграммный и словарный [1]. В языках, обладающих низкой флективностью, где не учитываются синтаксические связи проверяемых слов, наибо-

лее эффективным считается словарный метод обнаружения ошибок. Более того, для данного метода неважно на каком языке и с использованием какого алфавита будет записан текст. Главное, чтобы был эталонный словарь правильно записанных слов/словосочетаний и соответствующий алфавит. При словарном методе все входящие в текст словоформы, после упорядочения или без него, в своем исходном текстовом виде или после морфологического анализа, сравниваются с содержимым заранее составленного машинного словаря. Если словарь такую словоформу допускает, она считается правильной, иначе предъявляется контролеру. Он может оставить слово, как есть; оставить его и вставить в словарь, так что далее подобное слово будет опознаваться системой как верное; заменить слово в данной текстовой позиции; потребовать подобных замен по всему дальнейшему тексту; отредактировать слово вместе с его окружением. Результаты неоднократных исследований показали, что словарный метод экономит труд человека и ведет к минимуму ошибочных действий — пропуска текстовых ошибок, с одной стороны, и отнесения правильных слов к сомнительным единицам, с другой. Часто дополнительно к словарному методу применяют полиграммный метод. При этом все встречающиеся в тексте двух- или трехбуквенные сочетания (биграммы и триграммы) проверяются по таблице их допустимости в данном естественном языке. Если словоформа не содержит недопустимых полиграмм, она считается правильной, в противном случае — сомнительной и предъявляется человеку для визуального контроля и, если нужно, исправления.

В данной работе в качестве основы для создания эталонного словаря был использован фрагмент лога запросов франкоязычных пользователей системы поиска технической информации, сведений и данных в области химии, физики, биологии медицины, строительства и транспорта. Анализ лога запросов показал, что он содержит всю необходимую для исправления ошибок информацию: частоту употребления слов (позволяет заменить слово на более частотное), сочетаемость слов (проверяется сочетаемость слова с соседними словами), показатели замены слов (можно заметить, как изменяют слово сами пользователи). Количество экземпляров в эталонном словаре (S) составило 1100 единиц, со средней длиной экземпляра (L) равной 12 буквам. В алфавит были включены буквы французского алфавита (A–Z, Éé, Ââ, Êê, Îî, Ôô, Ûû, Àà, Èè, Ùù, Ëë, Ïï, Üü, Ÿÿ, Çç, œ) и цифры (0–9). Размер данного алфавита (A) составил 51 символ.

Исследование лога франкоязычных запросов показало, что наиболее частотным видом ошибок являются однобуквенные ошибки, которые можно классифицировать следующим образом:

1) вставка буквы, например, *inclinaisaon* вм. *Inclinaison;*
2) замена буквы, например, *logociel* вм. *Logiciel;*
3) перестановка двух соседних букв, например, *porjet* вм. *projet;*
4) пропуск буквы, например, *raccordemnt* вм. *raccordement.*

Процедура автоматической коррекции ошибок в однословных и многословных франкоязычных запросах состоит из следующей последовательности шагов.

Шаг 1: Проверка входного значения (цепочки алфавитно-цифровых символов, разделенных пробелами) по эталонному словарю.

Шаг 2: Проверка входного значения по эталонному словарю с помощью использования его различных вариаций.

Шаг 3: Проверка входного значения по дополнительному словарю с ошибками.

Шаг 4: Проверка отдельных лексем составного входного значения и восстановление сокращений.

Отметим, что сначала входное значение проверяется по эталонному словарю (шаг 1). В случае наличия входного значения в словаре оно считается правильным и не требует дальнейшей проверки на ошибки. Основным шагом процедуры автоматической коррекции ошибок является шаг 2, в основе которого лежит метод исправления орфографических ошибок с помощью перебора, предложенный в работе [2].

Рассмотрим вариант данного метода для опечатки типа «замена буквы». В процессе поиска опечатки входное значение будет модифицироваться. Получаемые модификации назовем гипотезами входного значения (далее просто гипотезами). Эталонный словарь, по которому будет происходить поиск, должен быть упорядочен по алфавиту.

Сначала входное значение проверяется по упорядоченному эталонному словарю с целью нахождения места, где входное значение больше предыдущего и меньше последующего экземпляра словаря. Обозначим эти позиции как *Prev* и *Next*. Возможны варианты, когда входное значение не входит в словарь. Тогда в качестве *Prev* и *Next* выбирается, соответственно, первая или последняя единица словаря.

Далее производится побуквенное сравнение слова на позициях *Prev* и *Next* с входным значением с целью определения позиции, в которых слова *Prev* и *Nex*t не совпадают с входным значением. Пусть это будут позиции *P* и *N*. Затем выясняется, какая из этих позиций наболее удалена, т.е. находится дальше от начала слова. Условно обозначим большую позицию символом *D*. Так как обнаружилось несовпадение входного значения со словом из словаря, то, следовательно, цепочка букв входного значения с первой по букву, стоящую на позиции *D*, отсутствует в эталонном словаре в качестве начала какой-либо единицы словаря. Значит, данная цепочка символов содержит ошибку. Следовательно, никакие исправления на участке входного значения за позицией *D* не приведут к нужной гипотезе. Поэтому перебор ограничивается варьированием букв на позициях, не превышающих *D*. В качестве первой варьируемой позиции *V* выбирается максимальная позиция *D*.

В процессе выполнения цикла перебора на варьируемую позицию *V* последовательно подставляются буквы (цифры или знаки) алфавита. При каждой подстановке символа на позицию *V* полученная гипотеза входного

значения ищется в эталонном словаре. Если поиск завершился удачно, т.е. было найдено такое значение, данная гипотеза сохраняется в специальном выходном стеке, а поиск продолжается, поскольку возможны и другие правильные варианты. После прохождения всего алфавита для выбранной позиции *V* эта позиция уменьшается на единицу и осуществляется новое выполнение цикла перебора. При *V = 0* процесс перебора завершается. При этом в выходном стеке будут находиться все правильные варианты входного значения.

Выше был описан процесс автоматической коррекции только одного типа ошибки — замены буквы. Рассмотрим, как представленная процедура может быть обобщена с целью исправления любого типа однобуквенных ошибок. После определения варьируемой позиции *V* вместо замены буквы можно производить вставку, что приведет к процедуре коррекции ошибок типа «пропуск буквы». Для выявления лишней буквы требуется поочередно удалять буквы, начиная с позиции *V* и двигаясь к началу входного значения. Для нахождения ошибок, связанных с перестановкой букв, можно переставлять соседние буквы либо буквы через одну, начиная с позиций *V+1* и *V+2* и двигаясь к началу входного значения. Необходимо отметить что, для перебора не обязательно использовать буквенный алфавит. Это может быть любой набор символов, пригодный для данного эталонного словаря.

Важно оценить, какое количество операций строковых сравнений будет выполняться при использовании данной процедуры в случае обнаружения пяти выше перечисленных типов ошибок. Условно обозначим количество символов в используемом алфавите символом *A*, а количество экземпляров правильных единиц в эталонном словаре — символом *S*. Количество операций строковых сравнений при поиске одной гипотезы в словаре (*I*) может быть различным в зависимости от выбранной процедуры поиска. Для обнаружения ошибки типа «замена буквы» понадобится *A*I*N* операций строковых сравнений, где *N* — расстояние (количество символов) от начала входного значения до максимальной позиции расхождения *D*. В итоге для всех типов рассматриваемых ошибок будет определено следующее количество операций строковых сравнений:

1) замена буквы — A*I*N;
2) пропуск буквы — A*I*N;
3) лишняя буква — I*N;
4) перестановка соседних букв — I*N;
5) перестановка букв через одну — I*N;

общее количество *Sum = (2*A+3)*I*N*.

Существует возможность существенно сократить общее количество операций сравнения за счет изменения не только входного значения, но и единиц эталонного словаря.

ЛИТЕРАТУРА

1. Андреев, А. Автоматизация обнаружения и исправления опечаток в названиях географических объектов для системы семантического контроля документов электронной библиотеки / Электронные библиотеки: перспективные методы и технологии, электронные коллекции. [Электронный ресурс]. — Режим доступа : http://www.rcdl2007. pereslavl.ru/papers/paper_25_v1.pdf. — Дата доступа: 20.04.2014.

2. Гельбух, А. Ф. Исправление орфографических ошибок с помощью перебора, управляемого морфологическим словарем / Научно-техническая информация. Сер. 2. Информационные процессы и системы. — 1993. — № 5. — С. 23–30.