РЕАЛИЗАЦИЯ ВЫЯВЛЕНИЯ И ИСПРАВЛЕНИЯ ОРФОГРАФИЧЕСКИХ ОШИБОК В ТЕКСТЕ

Агеев И.В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: jason4eg@gmail.com

Реальностью сегодняшнего дня стали электронные издания, число которых постоянно увеличивается. Библиотеки, не имеющие в фонде тех или иных электронных изданий и предоставляющие к ним доступ через Интернет, уже с полным на это основанием включают их библиографические описания в свои каталоги и предоставляют их пользователям.

При создании различных по назначению баз данных производится ввод текстовой информации, осуществляемый двумя способами – набором вручную или сканированием. В обоих случаях возможны орфографические ошибки.

Современные текстовые редакторы (например, MS Word), как правило, проводят автоматизированную проверку орфографических ошибок. Это требует вмешательства пользователя. Автоматическая коррекция орфографических ошибок может быть более эффективным средством минимизации опечаток и их исправлений при создании текстовых файлов.

В данной работе была поставлена задача, рассмотреть существующие методы автоматической коррекции орфографических ошибок.

Орфографически ошибочным словом называется буквенная цепочка, полученная некоторым преобразованием из словоформы некоторой лексемы, принадлежащей естественному языку. Под исправлением ошибки в таком слове называется установление исходной словоформы. Исходная словоформа определяется неоднозначно. В данной постановке задача исправления ошибки называется также задачей полного словарного исправления. Результатом попытки исправления ошибки в пределах некоторого класса преобразований может быть также установление невозможности ее исправления, то есть несуществования в словаре словоформы, из которой данная образуется цепочка путем какого-либо преобразования заданного класса [1].

Опечатками называются ошибки, связанные с поверхностным, буквенным представлением слова, то есть с искажениями непосредственно буквенной цепочки, представляющей словоформу [1].

С целью объяснения и исправления ошибок выделяется некоторый класс элементарных искажений. Например, замена одной буквы на другую, перестановка гласной буквы через согласную, сдвиг руки на одну позицию при набивке части слова на клавиатуре. Сложными называются ошибки, являющиеся комбинацией нескольких элементарных. Например, замена двух букв в одном слове. Элементарное искажение называется локальным, если оно по определению затрагивает небольшой отрезок буквенной цепочки, например,

не больше трех букв. Цепочка называется словом с одиночной ошибкой, если она содержит только одно элементарное искажение [1].

Одним из важнейших классов элементарных искажений является класс однобуквенных ошибок, включающий в себя:

1. Замена одного символа на другой (83%)
2. Удаление символа (8%)
3. Добавление лишнего символа (4.5%)
4. Перестановка двух символов (4.5%)

В процессе анализа методов выявления и устранения орфографических ошибок были выделены следующие методы:

- метод n-грамм
- метод спел-чекера

Метод n-грамм основан на предположении, что похожие слова обладают достаточным количеством общих подстрок длины n (n-грамм). При создании индекса для каждого слова составляется список содержащихся в нем n-грамм, который сохраняется в инвертированном виде. При такой организации данных для каждого указателя в инвертированном списке n-грамм нужно в произвольном порядке считывать ключевые слова из словаря, но скорость произвольного доступа во много раз меньше чем последовательного. Еще одна проблема связана с поиском по коротким терминам, когда изменение одной буквы приводит к «непопаданию» слова в выборку. Метод основан на том факте, что каждом языке существует строго ограниченный набор допустимых сочетаний символов.

Метод спел-чекера часто применяется в системах проверки орфографии (т.е. в spell-checker'ax), там, где размер словаря невелик, либо же где скорость работы не является основным критерием. Он основан на сведении задачи о нечетком поиске к задаче о точном поиске. Из исходного запроса строится множество «ошибочных» слов, для каждого из которых затем производится точный поиск в словаре. Алгоритм может быть легко модифицирован для генерации «ошибочных» вариантов по произвольным правилам, и, к тому же, не требует никакой предварительной обработки словаря, и, соответственно, дополнительной памяти. Можно генерировать не всё множество «ошибочных» слов, а только те из них, которые наиболее вероятно могут встретиться в реальной ситуации, например, слова с учетом распространенных орфографических ошибок или ошибок набора.

Список литературы

1. *Гельбух А.Ф.* Эффективно реализуемая модель морфологии флективного естественного языка. /Гельбух А.Ф. – Москва, 1994. – 77 с.