Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness

Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger

В статье используется усовершенствованный n-граммный подход для исправления ошибок типа non-word misspelling.

Постановка задачи: предложить варианты исправления non-word ошибок (ошибок в написании, при которых образуется несуществующее слово).

В классической n-граммной модели исправления ошибок близость между неправильно написанным словом и кандидатом на исправление рассчитывается как доля одинаковых n-грамм в общем числе n-грамм для этих двух слов. При этом порядок n-грамм не учитывается.

Для того, чтобы учитывать в оценке близости порядок n-грамм и включить в модель предположение о том, что ошибки редко встречаются в первой букве слова, авторы предлагают усовершенствовать классический подход. В предложенной модели в оценку близости ошибочного слова и кандидата на исправление включается число общих n-грамм в окне n-грамм заданной длины к общему числу n-грамм в обоих словах. Первая и последняя буква слов сравнивается отдельно. Таким образом, для расчета коэффициента близости учитываются только общие n-граммы, имеющие близкое расположение в слове.

Общий алгоритм: поиск слов с ошибками осуществляется с помощью словаря. Если слово отсутствует в словаре, для него составляется список возможных кандидатов на исправление, состоящий из слов длины на две буквы больше или меньше длины слова. Для каждого из кандидатов оценивается коэффициент его близости к изначальному слову на основе числа общих n-грамм в скользящем окне. Кандидат с наибольшим коэффициентом близости считается лучшим кандидатом на исправление.

Результаты работы алгоритма были проверены с помощью списка из 3975 английский слов с ошибками. Предложенный алгоритм на биграммах успешно исправил ошибки в 3334 словах (84%). Модель на триграммах показала более низкие результаты (73% правильных исправлений).

Эффективность алгоритма также была оценена с помощью сравнения результатов предложенной модели с другими алгоритмами на материале английского и португальского языка. Для английского языка в качестве тестовых данных были выбраны 120 английских слов с ошибками из Wikipedia. Сравнение проводилось со спелл-чекерами Google, Aspell и Microsoft Word. В основе Aspell лежит модель edit distance (определение близости между словами через количество операций вставки, удаления, замены символа или транспозиции двух соседних символов), дополненная сравнением по звучанию. Предложенный в статье алгоритм исправил 90% ошибок, Google – 88%, Aspell и Microsoft Word – 87,5%.

Для португальского языка результаты сравнивались с Aspell, TST (Ternary Search Trees) на 120 португальских словах с ошибками. TST алгоритм основан на словаре, представляющем возможность быстрого поиска по дереву для нахождения кандидатов на исправление. Предложенная n-граммная модель показала самое высокое качество – 80% правильных исправлений, TST – 65%, Aspell – 54%.

Предложенный в статье алгоритм демонстрирует хорошие результаты по исправлению ошибок для разных языков. Однако в этой модели не учитывается частотность кандидатов на исправление и не учитывается контекст ошибочного слова, как, например, в модели зашумленного канала. Возможно, качество алгоритма может повысится за счет учета языковой модели.