Обзор статьи: **“Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness”** *(авторы: Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger)*

Постановка задачи: разработать улучшенный алгоритм для проверки правописания, основанный на сравнении неправильно написанного слова и возможных кандидатов на замену по n-граммам (кандидаты ранжированы на основе анализа лингвистических данных для конкретного языка). Согласно тексту статьи, данный метод является универсальным для любого языка.

При этом авторы концентрируются на исправлении несуществующих слов, появляющихся в тексте в результате единичных вставки, удаления, замены или перестановки (согласно работе Damerau, к таким словам относятся около 80% всех ошибок).

Описание метода *(перечисление названий использованных методов не прощаем, надо разобраться и кратко и ясно всё описать)*: за основу взят уже существующий метод сопоставления по n-граммам. Основная причина, по которой авторы обратились к данному методу, - универсальность его использования для любого языка. Изначально данный метод основывался на вычислении схожести двух слов по n-граммам, при этом порядок следования n-грамм не учитывался, в то время как в описываемом методе данный порядок учитывается.

Опираясь на работу Yannakoudakis и Fawthrop, авторы приняли во внимание, что, как правило, в слове с ошибкой первая буква верна, а также длина слова либо совпадает, либо различается на 1 с длиной правильно написанного слова. Поэтому при вычислении схожести слов вместо первого и последнего n-грамм берутся только первая и последняя буквы; они сопоставляются отдельно от n-грамм.

Далее авторы вводят понятие “окно n-грамм”, то есть количество n-грамм справа и слева в сопоставляемом слове относительно соответствующего n-грамма. Данный подход позволяет сравнивать только n-граммы, находящиеся в непосредственной близости. Таким образом, для вычисления схожести исправляемого слова и кандидата по n-граммам можно задавать длину n-грамма и размер “окна”. В основе вычисления общего коэффициента схожести лежит сумма отдельных сопоставлений n-грамм (совпадение - 1, несовпадение - 0). Авторы также учли

варианты для n-грамм на границах слова и случаи, когда сравниваемые слова отличаются по длине.

Для определения правильности слова используется словарь (авторы использовали словарь на основе MultiWordNet, для английского языка). Если слова в словаре нет, для него подбираются возможные кандидаты (длина кандидатов не отличается от исправляемого слова больше, чем на 2 буквы). Затем посредством описываемого выше метода сопоставления n-грамм выбирается оптимальный кандидат для замены. Этот алгоритм авторы назвали MultiSpell.

Результаты: MultiSpell тестировался на материале английского и португальского языка. Для английского использовался список наиболее частых опечаток. Эксперименты показали высокие результаты, причем биграммы оказались продуктивнее триграмм: 3334 исправленных слов для биграмм и 2900 - для триграмм (всего слов 3975). Также MultiSpell незначительно превзошел Aspell, Microsoft Word и Google.

Для португальского языка результаты также были высокими по сравнению с Aspell и Ternary Search Trees: 80% - Multispell, 65% - TST, 54% - Aspell.

Данный подход кажется мне очень перспективным, так как при высоких экспериментальных результатах он универсален для любого языка и при этом прост для реализации. Авторы планируют дальнейшее улучшение и оптимизацию данного алгоритма, а также эксперименты на других языках.