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

Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger

В данной работе описан подход к автоматическому исправлению ошибок в английском и португальском языках. Задача исследования состояла в создании независимого от языка подхода к исправлению ошибок в не-словах (non-words, словах, которых не существует в языке и которые появляются в результате опечаток), основанного на расширенной н-граммной модели. Модель присваивает веса возможным кандидатам на исправление, основываясь на лексических ресурсах и статистических мерах, и предлагает возможные замены.

Использованные технологии:

- MultiWordNet - мультиязычная лексическая база данных. Использовался авторами в качестве словаря;
- MultiSpell - алгоритм проверки орфографии, разработанный авторами работы.

Порядок выполнения действий в MultiSpell:
1) сравнить слова, поступившие на вход, со словами в словаре (слова из MultiWordNet) ;
2) если слово написано неправильно, для него извлекаются н-граммы. В качестве кандидатов на замену отбираются только те слова, которые на два символа короче или длиннее исходного слова;
3) вычисляется уровень близости для каждого кандидата. Лучшим признаётся наиболее похожий.

Идея алгоритма вычисления уровня близости слов: строки похожи, если в них много похожих н-граммов. Поэтому можно вычислить коэффициент сходства для каждых двух строк.: количество одинаковых н-грамм в словах a и b делить на количество уникальных н-грамм в обоих словах. Порядок н-грамм не учитывается. Первый и последний н-грамм заменяются на первую и последнюю букву и сравниваются независимо от н-граммов между ними. Сравниваются только близкие друг к другу н-граммы.

Применение:

MultiSpell был инкорпорирован в Sense Folder Framework, применяется для поиска документов по ключевым словам, а также в семантическом поиске. В последнем система предлагает схожие слова пользователю в том случае, если по его первоначальному запросу ничего не находится.

Оценка работы алгоритма:
1) на списке слов английского языка, в которых часто допускаются опечатки - правильно скорректировано 84% слов;
2) на списке из 120 неправильно написанных слов из Википедии проверялась эффективность MultiSpell в сравнении с Aspell, а также спеллчекерами MicrosoftWord и Google. Эффективность практически одинаковая, с небольшим преимуществом у MultiSpell.
3) для португальского: 120 слов, сравнение MultiSpell, Aspell и Ternary Search Trees (авторы не строят их самостоятельно, а сравнивают свою систему исправления ошибок с другой системой, написанной для португальского языка с использованием этого алгоритма). Использовался не MultiWordNet, а словарь для португальского. MultiSpell показал лучший результат с большим отрывом.

В данной работе авторы применили нестандартное решение для автоматического исправления орфографических ошибок, создав систему, показавшую хорошие результаты для английского и португальского языков. Авторы планируют протестировать алгоритм также на данных для других языков. Тем не менее, данный алгоритм никак не справляется с опечатками, при которых одни слова превращаются в другие, также присутствующие в словаре языка, но не подходящие по контексту. Кроме того, не описано, как обрабатываются сложные случаи, например, слова с пропущенными дефисами. По выдаче системы, предоставленной в статье, можно увидеть, что некоторые аффиксы в английском языке заменяются неправильно (algebraical заменяется на algebraically вместо algebraic), что указывает на всё-таки присутствующую необходимость осуществлять оценку контекста при выборе замены. Однако, в целом этот подход показался мне интересным, и я думаю, что для европейских языков подобные алгоритмы автоматического исправления ошибок действительно могут работать с неплохим результатом.