Самойленко Ксении

How Difficult is it to Develop a Perfect Spell-checker?

A Cross-linguistic Analysis through Complex Network Approach

Авторы статьи решили предложить свое решение для улучшения работы спелл-чекеров. Известно, что одна из главных особенностей в разработке методов автоматической проверки правописания — существование ошибок типа NWE (non-word errors — когда из-за ошибки или опечатки появляется несуществующее слово (иронично, что пока я писала этот текст, Word предложил заменить ‘NWE’ на ‘NEW’)) и RWE (real word errors — из-за ошибки в тексте появляется реально существующее слово, но отличающееся от задуманного; из-за этого смысл текста может измениться либо вообще исчезнуть).

Спелл-чекеры, в которых не учитывается контекст, не могут распознать RWE, так как формально слово написано правильно, за исключением того факта, что оно не подходит по смыслу. Точно так же они не могут исправлять ошибки типа RWE с высокой степенью эффективности, т.к. в некоторых случаях выбор одного из кандидатов на «правильное» слово зависит от контекста. Соответственно, сложность в создании идеального спелл-чекера с точки зрения авторов зависит от вероятности того, ошибка будет принадлежать к типу RWE и что на слово-результат NWE будет похоже более одного «правильного» слова.

В этой связи авторы в своей небольшой работе решили попробовать формализовать ряд показателей, связанных с проблемами спелл-чекинга и представить ряд мер, которые могли бы усовершенствовать качество оценки работы спелл-чекеров.

Свой подход авторы назвали SpellNet. Они применили его для трех языков: английского, бенгальского и хинди. SpellNet призван помочь в формализации орфографических характеристик языка. Смысл SpellNet заключается в следующем: строится сеть — граф, которая по сути является моделью лексикона определенного языка. Вершины этого графа — слова, ребра — связи между ними. Каждой связи присваивается определенный вес: он показывает, насколько слова близки друг другу орфографически.

На мой взгляд, такое моделирование интересно тем, что позволяет удобным и понятным образом рассчитать определенный «круг» потенциальных кандидатов на замену для слов с ошибкой (в основном для ошибок типа NWE, так как RWE все-таки сильно зависят от контекста).

Авторы ввели еще несколько мер для своей модели, в честности, *степень вершины* — количество связанных с ней ребер. Средняя степень для определенного языка может быть принята как вероятность RWE в этом языке.

Авторы обращают внимание на то, что при наличии полного списка допустимых слов в языке, обнаружить ошибку типа NWE довольно просто, но исправить ее не всегда получается успешно. Как правило, спелл-чекеры предлагают список возможных претендентов на правильное слово среди тех, которые наиболее похожи на ошибочно написанное. Но чем больше ошибок в слове и чем оно длиннее, тем больше список таких «кандидатов». Авторы предлагают посмотреть на эту проблему следующим образом: оценить вероятность того, что два случайно выбранных соседа одного слова связанны между собой. Это соотношение отображает, насколько плотно связаны слова в лексиконе данного языка; если оно высоко, значит, ошибку будет сложно исправить из-за большого количества орфографических соседей.

Помимо изложенных выше наблюдений, авторы изучили SpellNet для английского, хинди и бенгальского языков. Списки слов были получены из публичных источников, для построения графа взяли 10000 наиболее частотных. Оказалось, вероятность ошибки типа RWE для языка хинди значительно выше, чем для бенгальского (в котором она соответственно выше, чем для английского). Отдельно оценивалось соотношение уровней для двух связанных слов: оно также помогает определить вероятность ошибки RWE (т.е. чем больше в языке похожих по написанию слов, тем она выше). Аналогичная градация языков соблюдается и для вероятности ошибки NWE, она связана с весом кластеров (точнее сказать, плотности связей между словами в кластерах) в сети. Эти результаты и разные вероятности ошибок в данных языках могут являться как следствием особенностей их орфографических систем и в частности размера алфавита, кроме того, определенное влияние могло оказать то, что в хинди орфография очень сильно связана со звучанием слова, тогда как в английском эта связь самая слабая. Авторы предполагают, что это наблюдение заслуживает дальнейшего изучения.

На мой взгляд, модель, предложенная авторами, является довольно красивой и логичной и предоставляет широкие возможности для формализации и оценки взаимосвязей между словами в языке и их похожести друг на друга. Любопытным является наблюдение о зависимости сложности исправления ошибок и силы принципа «как слышится, так и пишется» в конкретном языке. С другой стороны, эта модель мало влияет на проблему контекста при подборе «кандидата» на замену. В идеале такую модель было бы интересно наложить на некую семантическую онтологию, отражающую смысловые связи между словами.