An Improved Error Model for Noisy Channel Spelling Correction

Задача

Авторы статьи ставят перед собой цель повысить качество спеллчекера в рамках модели “шумного канала”. Как объясняется во вводной части, решение задачи спеллчекинга в модели шумного канала состоит из моделирования источника (source), что в данном случае означает текст, который намеревался породить говорящий, и моделирования канала (channel), что означает представление о том, как люди ошибаются в написании. В статье авторы описывают попытку более продвинутого моделирования ошибок написания.

Метод

В статье описывается новый подход к понятию замены, а именно, вместо посимвольных замен в каждой строке авторы ввели замены подстрок. Такой подход позволяет учитывать вероятность замены строки в определённом контексте. Также предлагается учитывать позицию замены (значение из набора {начало слова, середина слова, конец слова}). Авторы утверждают, что это позволит значительно повысить качество модели ошибок, так как, например, оценка P(a|e) вне зависимости от позиции (вероятность появления строки a при намерении написать строку e) не позволяет предсказать вероятность замены так же хорошо, как P(ant|ent) с учётом позиции, так как эта подстрока чаще всего является суффиксом слова, в написании которого люди часто ошибаются.

Итак, для строки s, не принадлежащей словарю D, авторы подбирают максимальное значение вероятности (w | s)P(w | context), где w -- слово, котороен намеревался написать человек.

Результаты

Авторы представляют результаты для своей модели ошибки с контекстом от 0 до 5 символов, с учётом и без учёта позиции замены, а также с применением и без применения языковой модели. Эксперименты проводились с использованием 10 000-ного размеченного корпуса ошибок английского языка. 80 % корпуса использовалась для обучения, а 20 % для оценки.

Для модели без учёта позиции и модели языка, модель ошибки авторов даёт значительное улучшение по сравнению с моделью, описанной в (Church and Gale 1991), при контексте от 3 символов и больше. При применении информации о позиции результат начинает гораздо сильнее зависеть от размера окна, и увеличение окна, вплоть до 5 символов, даёт значительный прирост. В случае с использованием

языковой модели, модель ошибки, описанная в данной статье, даёт более гладкий, но всё же значимый прирост в качестве, по сравнению с (Church and Gale 1991).

Своё мнение о статье и подходе

На мой взгляд, предложенный авторами подход к моделированию ошибок написания действительно эффективен и представляет новый взгляд на понимание замены в спелл-чекинге. В качестве недостатка можно отметить увеличение вычислительных мощностей, которое требует такой подход.