Титков Даниил

Automatic spelling correction for russian social media text

Sorokin A.A. & Shavrina T.O.

Для разбора была выбрана эта статья, потому что показатели качества проверки у программы авторов в соревновании Диалога-2016 были значительно выше, чем у других участников.

Использованный авторами подход основан на работе (Shaback, 2007) и состоит в применении для расчета вероятностей возможных исправлений линейной регрессии, которой в качестве признаков подаются вероятности, полученные с помощью моделей разного уровня – модели языка, модели ошибок и т.д. Система работает на уровне предложений. Рассмотрим работу системы подробнее. Она состоит из нескольких этапов.

1. Генерация кандидатов (возможных исправлений) для предложения.
    a. Поиск возможных исправлений для всех слов в предложении. В этот список входят всех слова с редакторским расстоянием равным 1 от каждого слова и само это слово, причем вне зависимости от того, есть ли оно в словаре или нет. Словарь в системе хранится в виде префиксного дерева, что позволяет быстро осуществлять поиск по нему. Кроме того, в список возможных исправлений для каждого слова добавляются фонологически близкие слова из словаря. Кроме того, добавляются некоторые фиксированные исправления наподобие «ваще -> вообще». Генерация исправлений производится как для каждого слова в отдельности, так и

для каждой пары стоящих рядом слов (для исправления ошибочных вставок пробелов).
    b. Так как при таком методе количество возможных предложений получается очень большим, кандидаты сразу же ранжируются на основании моделей ошибок и языка и отсекаются, если имеют плохие показатели.
2. Переупорядочивание кандидатов. После получения начального списка предложений-кандидатов с помощью линейной регрессии им приписываются новые ранги. Регрессионная модель была обучена на обучающих данных, предоставленных «Диалогом». В качестве признаков использовались такие: количество слов в предложении, значения, полученные из моделей ошибок и языка, количество исправленных слов, количество неизвестных слов, количество исправлений в словах разного типа. Те признаки, которые рассчитывались для каждого слова отдельно, суммировались, чтобы получить значение признака для всего предложения.

В результате применения данного метода авторы получили наиболее эффективную систему из показанных на соревновании на «Диалоге-2016» (F1 = 0.75).

Непонятным осталось, почему авторы позиционируют свою систему как нацеленную на исправление ошибок именно в социальных медиа. К специфике социальных медиа из описанного в статье можно отнести разве что словарь типичных исправлений («ваще» -> «вообще»).