**Introduction of the weight edition errors in the Levenshtein distance.**

1. Постановка задачи

В работе спелл-чекеров выделяют четыре типа ошибки при написании текстов:

- вставка: добавлен символ
- удаление: символ пропущен
- перестановка: изменение позиции между символами
- замена: замена одного символа на другой

Один из известных методов для спелл-чекеров -- расстояние Damerau-Levenshtein. В этом методе рассматривается только три операции редактирования (вставка, удаление, перестановка). Расстояние сравнивает два слова, вычисляя количество операций, которые необходимы для преобразования неправильного слова в правильное.

В этой статье авторы представляют новый подход, основанный на алгоритме Левенштейна. В обычном расстоянии Левенштейна, если у нас есть неправильно написанное слово и у него есть правильная пара на расстоянии один, то там может быть несколько кандидатов, у которых тоже шаг один. Метод авторов статьи решает проблему выбора кандидата при помощи ввода в формулу Левенштейна поправку на частоту той или иной ошибки..

2. Описание метода

Вся работа основана на арабском языке.

В своем подходе авторы статьи при расчете расстояния Левенштейна учитывают частоту ошибок трех типов. Для этого определили 3 матрицы:

- матрица частотности ошибки вставки
- матрица частотности ошибки удаления
- матрица частотности ошибки перестановки

$\mathcal{F}_{aj}(x_i)$ - частота ошибки добавления символа $x_i$ в слово

$\mathcal{F}_{sup}(y_j)$ - частота удаления символа $y_i$ из слова

$\mathcal{F}_{permut}(x_i / y_j)$ - частота ошибки перестановки символа $x_i$ с символом $y_i$.

Частота ошибки того или иного типа была вычислена при помощи четырех человек, которых просили писать текст. Получились следующие результаты:

| Editing operation | Number of errors | Total |
|---|---|---|
| Insertion | 202 | 1420 |
| Deletion | 295 | |
| Permutation | 923 | |

3. Результаты

Представленный в статье метод сравнили с работой обычного метода Левенштейна. Для сравнения двух методов использовали всего 190 ошибок.

Подытоживая все результаты, этот метод работает лучше для арабского языка (с которым работали авторы статьи), чем метод ливенштейна.

4. Свое мнение о статье и подходе

Сам по себе подход интересный. Возможно, мы даже попробуем использовать его в своем спеллчекере. Если найдем какую-нибудь статистику по опечаткам (или какой-нибудь корпус). Что качается статьи, в ней плохо раскрыто объяснение их метода. Там много математики и формул, но они не очень понятны человеку с плохим математическим бэкграундом. Так что эту часть авторы статьи могли бы расписать более подробно и снабжать математические формулы подробным описанием.