How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs. *Yukino Baba, Hisami Suzuki*

Постановка задачи

Задача этого исследования – изучение «исправленных ошибок» (corrected errors), то есть ошибок, которые пишущий исправляет прямо во время написания. Этот тип ошибок отсутствует в финальной версии текста, что заметно осложняет их изучение. В ходе исследования были выделены разные типы исправленных ошибок, была измерена их частота.

Изучение этого типа ошибок позволит, к примеру, улучшить качество спелл-чекеров, усовершенствовать исправление опечаток.

Описание метода.

Были собраны материалы для английского и японского языков. Это делалось с помощью Amazon's Mechanical Turk (MTurk), платформы, на которой можно проводить краудсорсинговые исследования. Был разработан специальный интерфейс, в котором была отключена возможность выделять слова и части слова с помощью мыши. Участникам показывались фотографии и предлагалось два задания – описать то, что они видят, или предположить, что животное или человек на фото могут говорить в этот момент.

Записывалось то, что вводится через клавиатуру с учетом бэкспейса (backspace), таким образом, что в слове с исправлением вводимый текст принимал вид «g-i-b-backspace-backspace-backspace-b-i-g».

Данные извлекались парами слов **до-исправления** и **после-исправления**. Слово после-исправления получали, удалив столько символов перед нажатием бэкспейса, сколько раз он был нажат. Слово до-исправления получалось после сравнения части слова перед бэкспейсом со словом после исправления и смотрели, какая именно часть слова была исправлена, а после дополняли. Все пары были переведены в нижний регистр.

Также были использованы собраны и составлены парами неисправленные английские ошибки, из Википедии и SpellGood.

Выделено четыре типа ошибок: *удаление, вставка, замена, перестановка*.

В предыдущей работе авторов были выделены потенциальные физические причины ошибок:

1. Контроль движений пальцев или рук;
2. Расстояние между клавишами;
3. Визуальное сходство букв;
4. Место в слове;
5. Повтор одной и той же буквы;
6. Фонологическое сходство букв или слов.

Исследователями было проведено сравнение трех типов ошибок – неисправленных английских, исправленных английских, исправленных японских по типу, потенциальной причине ошибки, месте в слове (длина слова оценивалась в процентах).

Результаты

Исправленные и неисправленные ошибки:

- Исправленные ошибки чаще встречаются на краю слова, неисправленные в середине.
- В неисправленных ошибках чаще встречаются ошибки типа удаление, а в исправленных замена.
- Ошибки типа удаления в местах, где буквы повторяются, встречаются заметно чаще в исправленных ошибках.
- Звуковое или визуальное сходство букв влияет на появление ошибок, и они чаще остаются неисправленными.

Ошибки в английском и ошибки в японском:

- В японском встречаются ошибки перестановки слогов, что связано с его слоговым алфавитом.
- В японском реже встречаются ошибки замены гласных.

Ошибки, совершенные по аналогии с левой частью слова (look-ahead) и с правой (look-behind): чаще встречаются look-ahead ошибки, то есть происходит замена на букву, которая должна появиться дальше.

Своё мнение о статье и подходе.

В статье представлен интересный способ изучения исправленных ошибок – в готовом тексте их выделить невозможно, так как у автора есть возможность в любой момент исправить свою ошибку. Изучение исправляемых ошибок, позволяет увеличить количество данных для классификации ошибок, возникающих из-за физических причин, но не исправленных, в том числе опечаток.

С опорой на эти данные спелл-чекеры могут быть улучшены, автоматически исправляя ошибки прежде, чем это сделает человек, а также исправляя ошибки тех же типов, которые остались незамеченными.

Сравнение языков с разными типами алфавитов (японским и английским) кажется мне полезным, потому что это позволяет выделить типы ошибок, характерных для определенного языка.

Я считаю, что это очень интересная и перспективная тема и с удовольствием приняла бы участие в подобном исследовании для русского языка или посмотрела бы на результаты такого исследования.