Ирина Соколова

Обзор статьи **“How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs”**

Авторы: Yukino Baba, Hisami Suzuki

Как появляются и исправляются орфографические ошибки? Исследование исправленных и неисправленных ошибок при помощи кейлоггинга

Вступление

В статье описывается сравнительное исследование ошибок, которые человек сразу же исправляет при наборе текста, и тех ошибок, которые остаются неисправленными в окончательном варианте текста. Данные о нажатиях клавиш записывались, из них восстанавливался первоначальный вариант слова с ошибкой и исправленный вариант. Анализ полученных данных показал значительные различия в типах ошибок, которые исправляют и не исправляют, а также разные виды ошибок для разных языков.

При наборе текста человек совершает как типографические ошибки (вызванные расположением букв на клавиатуре и движениями пальцев), так и когнитивные (вызванные фонетической или орфографической похожестью). Если человек уже в процессе набора текста замечает, что в слове есть ошибка, он может исправить её немедленно, но какие-то ошибки остаются неисправленными. Предыдущие исследователи обращали внимание преимущественно на неисправленные ошибки, вероятно, потому, что исправленные ошибки не записываются в виде текста. Однако есть по меньшей мере три причины изучать исправляемые в процессе набора ошибки. Во-первых, такие данные позволяют исследовать исправление ошибок самим автором текста, а не другим человеком или автоматическим методом. Во-вторых, такие исследования помогут создавать методы исправления ошибок уже в процессе набора текста, что сократит количество нажатий клавиш — это важно для таких языков, как китайский или японский, где ввод текста происходит с помощью транслитерации. В-третьих, данные об исправленных ошибках можно использоваться для улучшения методов исправления ошибок вообще.

Для исследования с помощью краудсорсинга были собраны данные об ошибках, исправляемых пользователями в процессе ввода текста. Данные были собраны для английского и японского языков; исправленные ошибки сопоставили с неисправленными; кроме того исследователи проанализировали типологию ошибок в английском и в японском. Также была выявлена ещё одна причина появления орфографических ошибок.

Сбор данных

Веб-сервис Amazon’s Mechanical Turk (MTurk) используется для краудсорсинга задач, которые проще выполнить человеку, чем компьютеру. Людей просили выполнить два типа заданий: описать картинку и придумать, что говорит изображённый на картинке человек или животное (задание “что говорят?”). Таким образом, испытуемые не знали, в чём смысл задания, а сами задания можно было использовать для разных языков. Пользователи не могли передвигать курсор и выделять текст и были вынуждены использовать клавишу Backspace для исправления ошибок. Для японского языка отключили функцию автоматического заканчивания слов.

В данных о нажатиях клавиш были выделены слова, в которых пользователи нажимали Backspace, и восстановлены строки до исправления и после. Чтобы получить строку после

исправления, из строки удалили количество символов равное последующему количеству Backspace'ов. Чтобы восстановить строку до исправления, сравнивали строку до Backspace'ов с подстроками после исправления ошибок и выбирали самую длинную подстроку с самым маленьким edit distance. Пары строк приводили к нижнему регистру и выбирали только те, где расстояние исправления было равно 2. Всего получилось 44104 пары строк для английского и 4808 пары для японского языка.

Для анализа также использовали базы часто встречающихся ошибок для английского языка с расстоянием 2 — всего 10608 пар строк.

Ошибки

Орфографические ошибки делят на 4 типа: удаление, вставка, замена и перестановка.

В статье называют следующие возможные причины возникновения ошибок. Физические: 1) движения пальцев, 2) расстрояние между клавишами. Визуальные: 3) похожесть символов, 4) позиция в слове, 5) повторение символа. Фонологические: 6) фонологическая схожесть символов/слов.

Для полученных результатов посчитали частоту определённых видов ошибок в виде отношения числа ошибок к общему числу подобных символов (например, частоту удаления согласных считали, разделив количество пропущенных согласных на общее количество согласных).

Исправленные и неисправленные ошибки в английском

Типы ошибок

Среди исправленных пользователями ошибок в английском чаще встречается замена, а среди неисправленных чаще встречается удаление: исследователи делают вывод, что замену символа легко заметить, а пропуск обычно не замечают. Эти выводы подтверждают данные других исследований для китайского языка.

Позиция ошибки в слове

Среди исправленных ошибок в английском чаще всего встречается удаление символа в начале слова, а вставка и замена встречаются и в начале слова, и в конце. Ошибки, оставшиеся неисправленными, чаще встречаются в середине слова. Таким образом, ошибки на краю слова исправляют чаще.

Эффект повторения символа

Удаление одного из двух повторяющихся символов значительно чаще встречается среди неисправленных ошибок по сравнению с удалением неповторяющегося символа. Среди исправленных ошибок такого эффекта нет. Таким образом, ошибки, которые больше бросаются в глаза, чаще исправляют в процессе набора текста.

Визуальная похожесть при замене символа

Замена одного символа на другой чаще встречается среди неисправленных ошибок, если символы визуально похожи. Среди исправленных ошибок такой тенденции нет.

Фонологическая похожесть при замене символа

Среди исправленных ошибок нет значительной разницы между количеством замен согласной на согласную и гласной на гласную, однако среди неисправленных ошибок намного чаще встречается замена гласной на гласную. Это согласуется с данными предыдущих исследований, которые показывали, что гласные несут больше лексической информации; возможно, влияет также то, что качество гласного в английском не всегда однозначно отражается на письме.

Ошибки в английском и в японском

Неисправленные ошибки в целом похожи в английском и в японском. Некоторые более характерные для японского языка типы ошибок объясняются фонологией и орфографией.

Перестановка в разных слогах

В английском чаще переставляют местами стоящие рядом буквы, а в японском — через одну. Вероятно, это связано с тем, что японское письмо слоговое, и люди печатают сразу слог.

Ошибки в гласных и согласных

В японском меньше вставок гласных по сравнению с согласными и меньше замен гласной на гласную. Возможно, это связано с тем, что в японском меньше гласных, они прозрачно отражаются на письме и реже становятся жертвами когнитивных ошибок.

Ошибки look-ahead и look-behind

И в английском, и в японском часто встречалась замена буквы на букву, которая уже появлялась в этом же слове или должна была появиться дальше, причём намного чаще встречался второй случай (look-ahead errors).

Выводы

Исследователи планируют использовать свой метод сбора данных и полученную информацию для разработки онлайн- и оффлайн-моделей исправления ошибок.

Моё мнение

Использованная авторами методика позволяет лучше понять, как появляются разные типы ошибок при набора текста, где и почему они встречаются чаще и какие есть различия для языков с разной фонологией и письменностью. Эти знания могут помочь разрабатывать системы проверки орфографии и немедленного исправления ошибок, что особенно важно для языков, которые пользуются транслитерацией на латиницу при наборе текста с клавиатуры.