KEYBOARD PATTERNS FOR SPELLING DETECTION AND CORREECTION

A Research Report

Submitted to

Assumption University

By:

MA. SHIELA C. SAPUL

December 2016

Table of Contents

Introduction

Spelling is considered as a one of the important part or element of any language, it is a tool that helps humans master the basics of a language and it is important on how we communicate. Knowledge of correct spelling facilitates good communication; misspelling can cause misunderstanding, confusion and serious problems. As humans we need to understand the written medium to effectively communicate.

The influence of text technology, email, IM services and social network make many people tend to forget the correct form of a word. There are also people who find spelling difficult because of poor motor skills. And many people pay a little attention to correct spelling.

Knowledge about what is a typical error helps in finding correct word. Errors are important source to understand better users' performance of a typing task.

Spelling errors can be due to phonetics such as homophones, spoonerism, and letter/word substitution. Groupings of the letters can be also sources of errors. Knowledge of spelling can be a cause of errors. Error involving diacritics and cedilla.

Typographical errors are mainly caused due to keyboard adjacencies. Not too many studies conducted for determining keyboard error patterns. Such factors can be keyboard adjacencies, shifting of key characters, sound and shape similarity of the keys. Keyboard mistyping can be cause due to spacebar issues, keyboard proximity.

Majority of the studies for spelling error detection uses the following methods: uses the four categories or types of errors such as insertion, deletion, substitution and transposition on how errors likely occurred. Matching words against the dictionary or a word list. In case the word is not available in the word list, character-based modelling is used. And comparing English common error mistakes with non-word and real word spelling errors. Majority of the spelling corpora are derived from dictionaries.

The researcher wants to investigate in identifying the sources of errors were error patterns may be extracted due to keyboard factors, identifying keyboard patterns and determining why people repeat certain types of typing mistakes will provide new information on how to effectively and accurately design and spell checker or spell correctors.

Review of Related Literatures

Spelling is very important because it improves or helps in reading, how we pronounce a word depends on how we read. Spelling detection and correction are two important tasks in spelling. Causes of spelling errors can be due to mispronunciation, non-native language speakers, not knowing spelling rule and typographical error or typos. Typos are mistakes in the typing process whether it is in a typewriter, computer or in a touch screen device. The researcher would like to focus on the spelling errors cause by typos using computer keyboard because of user behaviors and keyboard dynamics.

The following related studies provided the researcher ideas and concepts how different spell checkers perform and their performance. It provided knowledge in which areas of research there is a need to explore more or to improve.

Baba and Suzuki [1] collected their data using the logs from online users keystroke for two different languages, English and Japanese through the crowd sourcing infrastructure of Amazon Mechanical Turk (MTurk) and automatically deriving pre and post correction strings from them. And performed two comparative analyses: between corrected and uncorrected in English and between English and Japanese corrected errors. They extracted only words that included the use of backspace key. Deriving the pre correction involves comparing the prefix of the backspace usage with the substrings after error correction and considered that the prefix was spelled corrected into the substring which each longest with the smallest edit distance. For the post correction, they deleted the same number of characters preceding a sequence of backspace keys. Error types in English keystrokes and Japanese keystrokes are dominated by substitution. Deletion error type is the most common in the uncorrected English. In terms of errors due to visual similarities uncorrected English has the highest frequency rate. Corrected Japanese has the highest density rate in transposition errors in terms of the position of the errors within the words. Corrected English has the highest frequency ratio in deletion error types due to character repetition. In terms of the vowel and consonant error patterns, uncorrected and corrected English has almost the same frequency rate in the insertion error type. And corrected English and corrected Japanese have almost the same frequency rate in consonant to consonant substitution error.

The work of Rodrigues and Rytting [5] was to gather spelling error corpora using an online word-type game it was tested both on native language and English native players. It was costly because they acquire crowd outsourcing resources such as the Amazon Mechanical Turke to implement their game however it saves their time in collecting the game data.

The users were given an option to what type of keyboard they were using and it was administered in several countries; however only USA data was considered in their evaluation because likely of the size distribution of the users from this country. All non-Right handed and

non-QWERTY participants were also removed from the results. They utilized three spelling corpora to compare the errors made in their study and it shows that it collects similar errors from previous studies.

The keyboarding history module such as the type of keyboard or whether they were left handed or right handed and their geographical information was not fully utilized. Variable with the highest percentage rate was considered with only the reason that to reduce the amount of variables in the data so as their result will be comparable with other researches. It could have made comparable results between different keyboards and type of user (left-handed and right-handed) and the geographical locations of the players. These data could provide meaningful information to researches in the design of keyboards.

The function internal key logger data were not fully utilized to analyze probable cause of misspelled words or characters; it could have done analysis behavior of the users using the error patterns that can be identified in respect to keystrokes. They focus only in finding and storing the errors. And since it is a multiplayer game, there is a tendency that players may cheat. They only focused in identifying error types in insertion, deletion and substitution. They omitted transposition for they find it complex to determine the results. Deletion spelling error type has the highest frequency rate. On the three corpus used, the American typewritten has the highest rate because respondents only from the United States were considered.

The work of Tachibana and Mamoru [6] defined spelling errors as a typo from an incorrect keyboard operation and spelling confusion form an incorrect identification. Their goal is to extract and analyzed English spelling errors using word-typing game logs. A previous work was also done by Rodriguez and Rytting (2012) but it is quite costly because they utilized crowdsourcing resources and it is implemented in a multi-player environment and it is an uncorrected type-word game.

They implemented two word-type games for a single player but it does not involve outsourcing cost. The first word-type game is similar to Richards and Rytting, every time a user hits the enter key, his response is sent to a server, giving him no chance of correcting the error. The second word-type game used the same code with the first word-type game ,but with the exception that it will allow the user to modify or make corrections before hitting the enter key.

The results of their experiments were analyzed by determining the difference between the corrected and uncorrected spelling errors. They were able to extract correct pairs of intended word and the actual input.

They also emphasized the user's skills and personality, using the Pearson correlation coefficients between the times for one keystroke, the ratio of the corrected spelling errors and the ratio of uncorrected spelling errors for each user. They concluded that the faster the user types, the more likely it is to commit mistakes. However they failed to show that there is a correlation between the corrected and uncorrected spelling for each user, the users personality appears to

determine whether he wants to correct the errors. Comparison for the consonants and vowel substitute's errors shows the most frequent errors are found in consonant-consonant both for corrected and uncorrected likely due to incorrect keyboard operation. Adjacency errors among letters are due to the keyboard operations.

The work could have been improved if they have shown patterns of character or letters that is most frequently occurring and adjacent keys rate of errors were not given emphasis which could be used to improve spelling correction strategies. It would be better if the numbers of users for both games are the same.

A dataset for this study is available as a text file (typing_gamelogs.txt), it consist of 21468 English words for correctable word-type game, 19 users or players. And each player is only allowed to input 50 words. Also available is the data set for English sets based on Basic English (Ogden, 1930), consist of 8591 English words.

Another work of Kyongho, Wilson and Moon [4] focused on using typographical and orthographical spelling error correction. They use a dictionary lookup and two selection strategies using character distance and word frequency. They concluded that frequently used words are more prone to an error than words that are rarely used and the smallest character distance using the Pythagorean metric is chosen as the best corrector. They evaluated 20 different correctors for them to choose the best corrector. They found out that spelling correctors cannot correct multiple spelling errors and there are some words that were corrected wrongly. Spelling correction employing correction priority based on general error frequency showed better results than based on a specific error frequency.

Zeeshan, Ismaili, Shaikh, and Javaid [7] emphasized the need to determine the error trends and patterns in Sindhi language before developing a spell checker using a rule-based approach, they concluded that error trends are common to all languages but there are error patterns that are specific only to the Sindhi language. Error trends can be identified by classifying errors in the following forms; single error words, multiple error words, long words, short words errors, error that occurs in the first character of a word or as nth character of a word. Errors trends can be within one word and or as a result of the word boundary delimiter. They have identified that four types of errors in the Sindhi are due to typographic errors, cognitive or phonetic errors, visual errors and space related errors. They used the single error trend that usually occur due to the four types of errors; insertion, deletion, substitution of a single character, and the transposition among two characters in identifying the error trends in the Sindhi language. The most frequent for Sindhi language is substitution errors caused due to the shape similarity of the letters in Sindhi alphabet and also due to the similar pronunciation of various letters. The other type of error found in Sindhi language is the omission or deletion of space character at the word boundaries. From the studied presented for Sindhi it can be assumed that these results and error trends and patterns can be also apply to other languages that are similar to Persio-Arabic script.

Clawson, Rudnick, Lyons and Starner [2] focus on the study of the mini-QWERTY keyboard for detecting and correcting typing errors by analyzing features of the typing itself. They concluded that the probable cause of these errors is the relative size difference between the users' thumb and the small densely packed keys of the mini-QWERTY keyboard. Their goal is also to improve expert typing speeds and accuracy by automatically correcting the users' typing errors before they are displayed on screen. Using the pattern recognition they were able to reduce the number of off–by–one errors by 39.78% and the total errors by 26.41%.

Kane, Wobbrock, Harniss, and Johnson [3] they developed TrueKeys that will aid people with motor impairments who have difficulty typing using the desktop keyboards. The system performed more corrections than popular commercial and open source spell checkers by modeling word frequency, keyboard layout and typing error patterns that can automatically identify and correct typing mistakes. A word is first check from a dictionary of words to determine if it is known, if it is unknown true keys creates a list of correction candidates and rank them according to their word distance score, the word with the lowest score replaces the misspelled word. The physical keyboard layout is added to the word distance score because physical distance is used as a weighting factor for edit operations.

Summary

Baba and Suzuki [1] were able to identify five error patterns for detecting spelling errors for comparing corrected and uncorrected English spelling errors, and comparing corrected English and corrected Japanese spelling errors. These spelling error patterns are error types, visual similarities of characters in substitution errors, position of errors within the words, character repetition in deletion errors, consonants and vowels in insertion and substitution error types.

The study Rodrigues and Rytting [5] were able to identify three error patterns for detecting spelling errors. These spelling error patterns are character differences, left and right character context errors, and edit operations.

Tachibana and Mamoru [6] error patterns identified in their study are the use of backspace to correct a misspelled word, edit operations, adjacency of the keys, vowel and consonant relation.

Kyongho, Wilson and Moon [4] emphasized that error patterns can be derived from most frequently used words or terms, and the distance of the character from each other.

Zeeshan, Ismaili, Shaikh, and Javaid [7] USING rule based approach for identify spelling errors, they were able identify the error patterns from single words and multiple words, long words and short words, occurrence of the error as the first character in the word or nth character in the word and boundary errors due to omission or wrong execution of the space character.

Clawson, Rudnick, Lyons and Starner [2] they used pattern recognition technique in identifying spelling errors. They were able to identify error patterns due to off-by-one errors.

In the work of Kane, Wobbrock, Harniss, and Johnson [3] spelling error patterns can be derived from word frequency, due to the keyboard layout, typing error patterns, and run-on errors.

Majority of the related literatures mentioned used edit distance in the error detection and correction strategies. Three mentioned they used word-typed game strategy in collecting spelling errors.

The studies are quite interesting because of the emphasis on the typographical error where error patterns can be derived from the keyboard characteristics and dynamics and one interesting fact that spelling corpora can be derived from a typing game. Rule based can be used to detect spelling errors for correction. Comparative analysis can be done between corrected and uncorrected errors.

The studies failed to mention if punctuation marks and numbers are included in their error detection because there are also errors related with the use of punctuation marks and numbers, majority of their discussions are on the English alphabets only.

Some of the study failed to identify the cause of misspellings in the keyboard whether it is a vowel substitution errors, inaccurate double consonants, keyboard adjacency error rate, and space inaccuracy.

References

[1] Y.Baba and H.Suzuki, “How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 373–377, Jeju, Republic of Korea, 8-14 July 2012, Association for Computational Linguistics

[2] J.Clawson, K.Lyons, A.Rudnick, Jr., R.Iannucci, and T.Starner. Automatic whiteout++: correcting mini-QWERTY typing errors using keypress timing. In CHI ’08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pages 573–582. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-011-1

[3] S.Kane, J.Wobbrock, M.Harniss, and K.Johnson. “TrueKeys: Identifying and Correcting Typing Errors for People with Motor Impairments” IUI'08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain. Copyright 2008 ACM 978-1-59593-987-6/ 08/ 0001

[4] K.Min, W.H. Wilson and Y.Moon, “Typographical and Orthographical Spelling Error Correction”. 2000

[5] P. Rodrigues and C.a. Rytting, “Typing Race Games as a Method to Create Spelling Error Corpora”. 2012

[6] R.Tachibana and M.Komachi, “Analysis of English Spelling Errors in a Word-Typing Game”, Priority Areas, Tokyo Metropolitan University. 2016

[7] B.Zeeshan, I.A. Ismaili, A.Shaikh, AND W.Javaid, “Spelling Error Trends and Patterns in Sindhi” CIS Journal, VOL. 3, NO.10 Oct, 2012, ISSN 2079-8407

Recommendations

Baba and Suzuki [1] will use data and analysis results of their study to build both online and offline spelling correction models.

Rodriguez and Rytting [5] recommended that other researchers of other languages or non-native English speakers may use their game but using template and a word lists as stimuli.

Tachibana and Mamoru [6] suggested that researchers explore other applications such as the Duolingo to explore possible spelling errors.

Clawson, Rudnick, Lyons and Starner [2] a possible study using or train system with different mini-QWERTY keyboard. And perform experiments in a different context such as a typist is a non-expert and the blind typing condition. Conduct user evaluation comparing the results of typing speeds and accuracy without using the Automatic Whiteout correction system.

Kane, Wobbrock, Harniss, and Johnson [3] suggested a longitudinal study is needed to determine how users will adapt to TrueKeys and incorporate it into their everyday typing behavior. Consider also personalized model of a user's typing errors might enable TrueKeys to more accurately correct spelling errors. Improvement of the correction algorithm they used. The user interface might also be redesigned to be less intrusive and explore the application of TrueKeys to hand-held devices

Zeeshan, Ismaili, Shaikh, and Javaid [7] their results and error trends may be applied to other languages that are similar to Persio-Arabic script

Perform comparative analysis for English and Japanese corrected and uncorrected errors, the study only performed comparative analysis for corrected errors in English vs. Japanese.

It could have made comparison of the errors encountered between different keyboards and type of users (left-handed or right-handed). The key logger data was not fully utilized to determine probable patterns of the source of errors

The authors could have put emphasize in their data analysis and results by providing markers in the Sindhi language which causes the spelling error and markers indicating the corrected pattern or position of the character in the Sindhi word.

Conduct experiment comparing using the standard QWERTY keyboard and the mini-QWERTY keyboard to determine if results will vary.

Pythagorean theorem is used in determining the distance between two characters/letters in the keyboard as basis for the correction, but the study failed to provide discussion on how to compute distance if it is a deletion, substitution of multiple characters that are not in its correct position.