Analytical Evaluation of the Impact of Phrase Set on Text Entry Rates

Kent Lyons
Nokia Research Center
200 S. Mathilda Ave.
Sunnyvale, CA, 94086 USA
kent.lyons@nokia.com

James Clawson
GVU
Georgia Institute of
Technology
Atlanta, GA, 30318 USA
jamer@cc.gatech.edu

CHI 2012, May 5–10, 2012, Austin, TX, USA.
Workshop on Designing and Evaluating Text Entry Methods

Abstract

Recently the text input community has seen a flurry of research attempting to refine the methods by which input technologies are rigorously evaluated. Specifically, researchers have investigated the validity of the phrases that study participants input when engaged in transcription typing tasks. The phrase set published by MacKenzie and Soukoreff has become the *de facto* standard for text entry evaluations since its publication. New phrase sets have recently emerged that attempt to address limitations of MacKenzie and Soukoreff's phrase set. In this paper, we present a preliminary investigation of whether the choice of phrase sets has an impact on potential typing performance by analyzing several different phrase sets using existing typing data from mini-qwerty keyboards.

Keywords

Text entry evaluation, corpus, phrase set

ACM Classification Keywords

H.5.2 [**Evaluation/methodology**]

General Terms

Measurement, Experimentation, Human Factors

Introduction

Mobile text entry evaluations traditionally are composed of a series of short transcription typing tasks wherein a participant is asked to input a phrase displayed on a screen "as quickly and accurately as possible." These tasks often employ one or more novel input devices or techniques. Performance measures are usually calculated such as a participant's words per minute or accuracy rates. In an effort to preserve internal validity, the text entry community has adopted a set of phrases introduced by MacKenzie and Soukoreff as the *de facto* standard stimuli for text entry evaluations [5]. The MacKenzie and Soukoreff phrase set contains 500 phrases that were designed to be of moderate length, easy to remember, and representative of written English.

There has been an increased interest in producing phrase sets for text entry evaluation that both maintain strict internal validity and attempt to address different issues of ecological validity unexplored by MacKenzie and Soukoreff. Kano *et al.* produced a phrase set that they demonstrated was suitable for use by children [2]. Vertanen and Kristensson introduced a new phrase set for text entry studies comprised of phrases taken from Blackberry mobile phone users in the Enron data set [7]. Paek and Hsu take a theoretical perspective for examining how representative a phrase set is of a corpus [6].

While on the surface, the exact choice of phrase set seems important, there is limited data exploring the impact of phrase set choice for text entry studies (Kristensson and Vertanen is a notable exception [3]). Here we conduct an analysis of different phrase sets on pre-existing text entry data obtained from mini-qwerty typing [1]. Using the keystroke level timing data from that study, we simulate the potential typing rates from different phrase sets.

Method

For this analysis, we use data from the evaluation of mini-qwerty typing rates [1]. This data set consists of typing data obtained on two different mini-qwerty keyboards. Each participant completed a total of twenty 20-minute typing sessions. There is complete typing data for 10 subjects on one keyboard ("Targus") and for 7 participants on another keyboard ("Dell") where every keystroke was logged with a time stamp. In this study, the MacKenzie and Soukoreff phrase set [5] was used but modified to replace capitalized letters with lowercase letters and to use American English spellings. For a complete description of the procedure and data we refer the reader to the original work [1]. From this data, we extracted the timing data for typing. Here we utilize only phrases that were successfully typed correctly on the first try. We leave the analysis of errorful typing as future work.

For each session, user and keyboard, we extract the time to enter each bigram typed. This results in a list of times (in seconds) for each bigram entered such as "he"=[0.17, 0.17, 0.20, 0.375, ...] which in turn informs us as to how long it took the user to get from "h" to "e" every time they came across that pair of letters in the given session. For the analysis we present here, we had between 470 and 1647 bigram times across all of the users depending on the session and keyboard. In some instances, bigrams that are present in the phrase set were not typed by a participant in a given session. Additionally, some bigrams from the new phrase sets do not exist in the original phrase set. In both of these situations we compute the average amount of time it takes the participant to get from any source key to a given target key (again on a per participant per session basis). This calculated value is used in place of the actual bigram time in cases where particular bigrams are missing from the data. Finally, if a given character was not entered in a session, we use the mean per-character typing rate of that session to approximate the time needed to input the given character.

Next, we use this data in a simulation. In the simulation we treat the original typing data as some unknown distribution and resample it based on our target phrase set. For each bigram in each phrase of the target phrase set, we randomly pick one of the times for that bigram in the original typing data and add it to a subtotal on a per participant basis. If the bigram timing is not present, we use the mean estimate described above. We then repeat this process for the rest of the characters in the phrase and the rest of the phrases in the set to generate typing times for each participant.

Results

We simulated typing 1000 phrases each for three different phrase sets using the data generated by the two keyboards in our original study. In a preliminary analysis we found that simulating more than 1000 samples did not meaningfully impact our findings. The results are shown in Table 1 with the mean typing rates in words per minute (WPM) across the participants and keyboards. "Original" is the actual typing rate of the participants extracted from the original mini-qwerty data. "MacKenzie" is running the simulation on the MacKenzie and Soukoreff phrase set [5] used in the study. This is a baseline test and should produce comparable results to the "original" condition given that it is the same phrase set. We also run the simulation for two new phrase sets: the Kano *et al.*'s phrase set for children [2] and the recommended "mem_bi" phrase set from Vertanen and Kristensson [7]. Furthermore, for this analysis we used timing data from both the first typing session where the participants were novices and the last (20th) typing session where the participants had 400 minutes of typing experience at the conclusion.

A Friedman test on the independent variable of phrase set yields statistically significant differences ($p < 0.01$) for our four conditions (Dell session 1, Dell session 20, Targus session 1, Targus session 20). Further analysis with Wilcoxon signed-ranked tests (Bonferroni corrected) looking for differences between any two given phrase sets does not find any statistically significant differences. While not conclusive, given the mean times it seems that the overall differences found are due to the original typing speeds being faster than the simulation results. If that is the case it could be because the participants' typing rate relies on more than just bigram timings but this hypothesis would require further investigation.

	Session 1	Session 20
Dell	Mean WPM (SD)	Mean WPM (SD)
Original	31.7 (4.8)	58.4 (8.5)
MacKenzie	30.3 (4.7)	51.6 (6.4)
Kano	30.3 (4.9)	52.6 (6.4)
Vertanen	30.3 (4.9)	54.3 (5.5)

	Session 1	Session 20
Targus	Mean WPM (SD)	Mean WPM (SD)
Original	38.3 (4.9)	60.3 (6.2)
MacKenzie	36.7 (4.4)	52.4 (11.5)
Kano	37.3 (4.9)	54.4 (11.9)
Vertanen	37.0 (5.4)	54.7 (11.7)

Table 1: Simulation from the two keyboards with mean typing rate in words per minute and (Standard Deviation).

Discussion and Future Work

While the pair-wise tests do not allow us to reject the null hypothesis that the phrase sets result in different text entry rates; from a pragmatic perspective, if there are differences between phrase sets they would seem to be relatively small in an absolute sense. All of the simulation results are within a few words per minute of each other. If this result holds with further investigation, it would imply that the choice of phrase set used in a study is not particularly critical as long as it is reasonably representative of the target language. Therefore, if there is a particular reason for selecting a phrase set (eg a target audience such as with Kano *et al.*, or for internal validity with prior work), the results found in the experiment are likely comparable with other text entry studies conducted with different phrase sets.

There are several opportunities for extending this work. It would be useful to understand the main effect difference found and determine if in fact the simulation results do differ from the original typing speeds (since we could not reject the null hypothesis in this analysis). Hopefully that investigation would allow us to update the simulation to be more accurate. In this work, we ignored the impact of errors as we only examined phrases originally typed correctly. A logical extension would be to update the simulation to account for phrases containing those errors. We could also update the simulation to use other strategies for deriving timing information for the missing bigrams from our source data. For the mini-qwerty data perhaps we could use Fitts' Law or a text entry model. In this work, we looked at two different mini-qwerty keyboards; it would be interesting to extend this work to other keyboards or text entry methods. For example, we could examine the impact of phrase set on a chording keyboard like the Twiddler [4]. It would also be useful to look at phrase sets for some specific target domains that differ more from traditional written English such as the Twitter phrase set described by Paek and Hsu [6]. Finally, it would be good to empirically test our findings and verify that the absolute differences in typing rates are small. A user study could be conducted whereby participants entered phrases from different phrase sets and the resulting typing rates are compared.

References

[1] E. Clarkson, J. Clawson, K. Lyons, and T. Starner. An empirical study of typing rates on mini-qwerty keyboards. In *CHI 2005 extended abstracts*, pages 1288–1291, 2005.

[2] A. Kano, J. C. Read, and A. Dix. Children's phrase set for text input method evaluations. In *Proceedings of NordiCHI 2006*, pages 449–452, 2006.

[3] P. O. Kristensson and K. Vertanen. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proceedings Intelligent User Interfaces 2012*, pages 29–32, New York, NY, USA, 2012. ACM.

[4] K. Lyons, T. Starner, D. Plaisted, J. Fusia, A. Lyons, A. Drew, and E. W. Looney. Twiddler typing: one-handed chording text entry for mobile phones. In *Proceedings of CHI 2004*, pages 671–678, 2004.

[5] I. S. MacKenzie and R. W. Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI 2003 extended abstracts*, pages 754–755, 2003.

[6] T. Paek and B.-J. Hsu. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In *Proceedings CHI 2011*, pages 2477–2480, 2011.

[7] K. Vertanen and P. O. Kristensson. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings MobileHCI 2011*, pages 295–298, 2011.