Measuring Inviscid Text Entry Using Image Description Tasks

Figure 1: Sample image set

Mark D Dunlop
Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
mark.dunlop@strath.ac.uk

Emma Nicol
Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
emma.nicol@strath.ac.uk

Andreas Komninos
Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
andreas.komninos@strath.ac.uk

Prima Dona
KeyPoint Technologies
Botanical Gardens Road,
Kondapur, Hyderabad, India
pdona@keypoint-tech.com

Naveen Durga
ex KeyPoint Technologies
Botanical Gardens Road,
Kondapur, Hyderabad, India
naveen86@gmail.com

CHI'16 Workshop on Inviscid Text Entry and Beyond, 8 May 2016, San Jose, CA.

Abstract
We argue that measuring the Inviscid text entry rate requires new evaluation methods that support free-form text entry and that are not based on the traditional transcription/copy tasks. In this position paper we propose use of image description tasks and share some of our experiences of using this new language agnostic task type for free form text entry.

Author Keywords
Text Entry; Laboratory Tasks; Mobile Evaluation

ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/methodology.

Introduction
The text entry community has widely adopted a standard approach to studies in which users are asked to copy or transcribe a set of fixed phrases. The time they take and number of errors made are used as metrics to compare text entry within a study. To ensure study heterogeneity and allow comparison across studies, standard phrase sets are now widely used. The two most widespread are the MacKenzie and Soukoreff's original 500 short-phrases set [5] (e.g. *Have a good weekend*) and the Enron Mobile collection [7] of phrases that were written on mobiles (e.g. *Can you help me here?*). There are various other specific

collections such as an SMS corpus [1] and a child oriented corpus [2].

While the approach of fixed phrase copying gives strong internal consistency, reproducibility and heterogeneity advantages, the scenario of copying phrases is clearly not representative of most mobile text entry. Furthermore, the approach of prescribing the text to be typed does not support each user's natural typing style nor any learning/adapting that the keyboard has done to improve entry based on that individual user's language use. We argue that the short phrases and prescribed nature of standard text collections make them unsuitable for use in measuring the free flow inviscid text entry rate [4] and new alternatives need to be considered.

An alternative to copy tasks is to ask users to generate

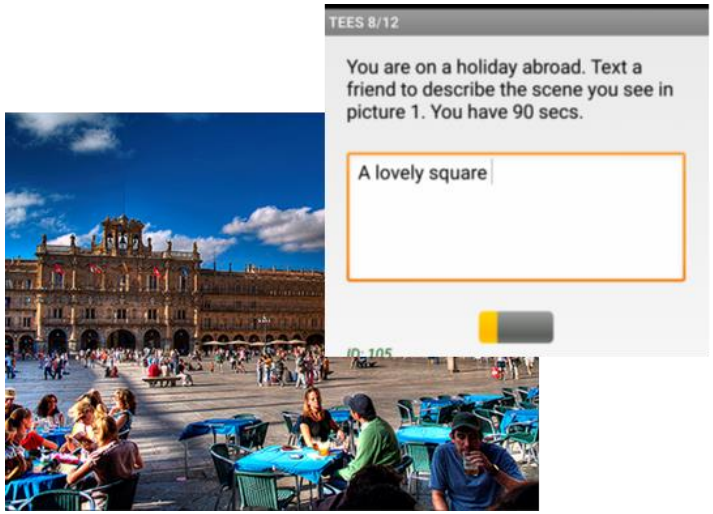

Figure 2: Sample description image and interface

text in composition tasks. Karat et al [3] compared copying sections of a novel with composing replies to scenarios and found composition speed was 58% of that for copying. More recently and inspired by mobile text entry evaluation, Vertanen and Kristensson [8] investigated complementing copy tasks with composition tasks by asking users to (a) reply to a message, (b) compose a message without scenario prompting and (c) compose with scenario prompting. They showed that composition tasks had an entry rate of 65-85% of the copy tasks depending on task type and that typed responses varied in length between 55% and 135% of copy tasks. They concluded that "providing participants with a simple instruction of creating a short message in the domain of interest was successful in getting participants to quickly invent and compose text. It does not appear necessary to provide participants with a specific situation or message in order to help them invent a message." Here we want to investigate an alternative form of prompting, asking users to describe an image or pair of images, in the hope of eliciting longer messages than traditional text entry study approaches in the user's own language.

Image Description Task

Our approach is to ask users to describe an image within a fixed time period and to ask them to type this into a visibly large, and conceptually unlimited, scrolling text field. We have tested several versions of our image description task and have tuned the method to two alternatives. Both variants are based around describing Creative-Commons images (e.g. Figure 1).

Single Image Task

The simpler approach is to give a fixed amount of time to users to describe a single image. Images are best

presented as A4 full colour printouts but on-screen is possible if not restrictive on a mobile device. Users are then asked to "*Please describe the image as if describing it to a friend: tell your friend about the scene and tell a story about the people in the scene. Think about the scenes and your story before you start typing. Use your imagination to elaborate on the image. You have 90 seconds once you start typing*." An alternative more focussed description can also be used (as per Figure 2). We found a time limit and large text entry area encouraged longer typing and that it was important to stress to users to think in advance and to use their imagination to encourage free-flow entry. Figure 2 shows a sample single image along with our evaluation tool (with 90 s timer below text entry area).

Multiple Image Task

An alternative to a single image is to ask users to describe two images from a set of three. This has the advantage of allowing some selection and reducing the risk of users not being able to think of a story about an individual scene. The rubric should be adjusted as follows: "*Please describe two of these images as if describing them to a friend: tell your friend about the scenes and tell a story about the people in each scene. Think about the scenes and your story before you start typing. Use your imagination to elaborate on the image. You have 3 minutes once you start typing and should split this between the two image descriptions*." Figure 1 shows a sample three image set.

Measuring Performance

Words per minute can be used as normally for copy tasks – using the time from first to last keystroke. However, it is also worth monitoring how much of the allocated time was used.

While the focus on much text entry experimentation is on speed of entry, accuracy is also important. For copy tasks, edit distance can be used a measure of accuracy of the final phrase [6]. For composition tasks correctness can be inspected manually (either by the researchers or crowd sourcing [8]), by simply counting out-of-dictionary word rates or by monitoring the input stream for text corrections [9].

Initial Results

In our initial study we recruited 14 Android users (13 aged <=25; 1 aged 46-55; 12 male; 2 female) to take part in a 10-day study in which they were asked to use a new keyboard for the study period and complete a set of tasks daily. The participants came into our laboratory on day 1 of the trial. In this session they ran a practice set of tasks using both image and text tasks with their new keyboard. They were then requested to use the new keyboard as their prime keyboard and complete a daily task set. Finally, they returned to the laboratory on day 10 for a last set of tasks and a debriefing group discussion. Participants were given a small gift token as thanks and the study was conducted under University of Strathclyde ethical approval. Users were asked to do tasks at their convenience but in a quiet location when they were unlikely to be disturbed. Each daily task sheet was composed of 12 tasks in three blocks of three copy tasks plus one image description task (total 9 text copy and 3 image description tasks per daily task set; image tasks take considerably longer so fewer were used in the study). In line with other studies, users were asked to enter the text quickly but accurately. Prompts were presented and text entered on our Android study client (Figure 2). Images were given to participants in advance in an A4 printed booklet with the on-screen prompt saying, for example,

"*Please describe two images from set 24 in three minutes*."

Table 1 shows the average word entry speed and submitted text lengths for image description and text transcription tasks. This shows a significant difference between the task types on speed (paired t-test, n=14, p<=0.01) with image description tasks around 78% of the text transcription task speed. For all image tasks the first-to-last keystroke time was over 179 s indicating users were still typing at timeout and used the full time available to them. In image tasks users entered an average 297 characters per task compared with an average length of text phrases at 25 characters (and reported 52 characters in straight composition tasks [8]).

	Length	WPM
Image	296.6 +/-46.9	19.6 +/-3.2
Copy	24.7 +/-0.4	25.3 +/-3.2

Table 1: Phrase length and entry speed comparison (mean +/- 95%c.i.)

The experimental system we used records the final phrase submitted along with indication of how many times backspace was used in that composition (to give an impression of how many corrections occurred). In our study the mean phrase Levenshtein string distance was 0.16, confirming a very low error rate (approx. 0.6% errors per character, dominated by missed words). We also spell-checked all submissions using a large English word list[1] augmented by adding identified out-of-dictionary words that were valid in Microsoft Word 2013 (UK English). In copy tasks, only 1 from the 5,671 words entered was out-of-dictionary (0.02%, an uncorrected compound *youresend*). In image task submissions 41 words were out-of-dictionary from the 19,994 words submitted (0.21%). While most errors were simple spelling errors, some were new words (e.g. *selfie*). While this process only checks that words typed were in the dictionary and does not reveal grammatical errors, or simply entering the wrong word, it does indicate a very low error rates as our keyboard did not support auto-correction. Crowdsourcing corrections could investigate this further (c.f. [8]).

[1] http://www.keithv.com/software/wlist/

Conclusions

Image description tasks allow fluid text entry that prevents the need for prescribing the words or language that is used. Our initial studies show that users type slower when describing but that they can easily fill 3 minutes with typing on a mobile to describe two images. Participants also liked the variation of task. As such we propose image description tasks as an addition to the current suite of transcription and straight composition tasks.

The image task set is available at: http://images.textentry.org.uk

Acknowledgements

Research was partly funded by the EPSRC under grant reference EP/K024647/1. We thank them and our participants. All images in the collection are creative commons – attributions at images.textentry.org.uk.

References

1. Tao Chen and Min-Yen Kan. 2012. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation* 47, 2: 299–335. http://doi.org/10.1007/s10579-012-9197-9
2. Akiyo Kano, Janet C Read, and Alan Dix. 2006. Children's Phrase Set for Text Input Method Evaluations. *Proceedings of the 4th Nordic Conference on Human-computer Interaction:*

Changing Roles, ACM, 449–452. http://doi.org/10.1145/1182475.1182534

3. Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 568–575. http://doi.org/10.1145/302979.303160
4. Per Ola Kristensson and Keith Vertanen. 2014. The Inviscid Text Entry Rate and its Application as a Grand Goal for Mobile Text Entry.
5. I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase Sets for Evaluating Text Entry Techniques. *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ACM, 754–755. http://doi.org/10.1145/765891.765971
6. R. William Soukoreff and I. Scott MacKenzie. 2001. Measuring Errors in Text Entry Tasks: An Application of the Levenshtein String Distance Statistic. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ACM, 319–320. http://doi.org/10.1145/634067.634256
7. Keith Vertanen and Per Ola Kristensson. 2011. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, ACM, 295–298. http://doi.org/10.1145/2037373.2037418
8. Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2: 8:1–8:33. http://doi.org/10.1145/2555691
9. Jacob O. Wobbrock and Brad A. Myers. 2006. Analyzing the Input Stream for Character- Level Errors in Unconstrained Text Entry Evaluations. *ACM Trans. Comput.-Hum. Interact.* 13, 4: 458–489. http://doi.org/10.1145/1188816.1188819