Recognition Errors and Recognizing Errors – Children Writing on the Tablet PC

Janet Read, Emanuela Mazzone, and Matthew Horton

Child Computer Interaction Group, University of Central Lancashire,
Preston, PR1 2HE, UK
{jcread, emazzone, mplhorton}@uclan.ac.uk
http://www.chici.org

Abstract. The paper describes a research study to determine the usability of handwriting recognition technology on a tablet PC for free writing by children. Results demonstrate that recognition error rates vary according to the metrics used, and the authors discuss how some of the errors are created concluding that the error rates say very little about what was happening at the interface and that with research of this type (novel interfaces and young users) researchers need to be immersed in the context in order to produce useful results.

1 Introduction

Over recent years there has been a significant increase in the published work relating to children and interaction design. However as the discipline of Child Computer Interaction (CCI) is still quite new [1] the methods used by researchers are generally derived from HCI and many of these have not been well tested with children. Using handwriting recognition for text entry on a tablet device is a relatively new form of interaction that has relied on evaluation methods from discrete text input and from speech recognition; the suitability of these methods for handwritten input have also not been well researched [2].

The tablet PC is a variation of the notebook PC incorporating a touch screen that can be written on by the user with a special stylus, in a similar way to writing on paper. This technology has recently been evaluated for use in learning environments and with children writing [3], [4]. Using a tablet PC, writing can be done in the user's regular script (handwriting) and software provided with the tablet PC is then able to change the writing into ASCII text (handwriting recognition) so that it can be manipulated in any text or word processing package.

It is common to evaluate the effectiveness of any text input method by measuring the accuracy of the process. The de-facto measure for the accuracy of any text input method is generated from two text strings; usually called the presented text (PT) and the transcribed text (TT). These two strings are compared, and each 'error' in the transcribed text is classified as either an insertion (I), a deletion (D) or a substitution (S). This measure can exist in two forms, as a word error rate (WER) (typically used in speech recognition) or as a character error rate (CER) (typically used in handwriting recognition as well as in discrete text input as is done at a keyboard) [5].

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 1096 – 1099, 2005.
© IFIP International Federation for Information Processing 2005

2 The Empirical Study

The small study that is described here was intended to determine the usability of the tablet PC for children writing. In particular, the intention was to look at the accuracy of the handwriting recognition that was supplied with the Windows Journal® application (as shipped with the tablet PC). Ten children aged 7 and 8 were recruited to the study that took part in school time. They came to the room individually and used the tablet PC to write their own stories using ideas that had already been developed in the classroom. Each child stayed for around fifteen minutes and the researchers, who were on hand to assist with any hardware problems, supervised the writing tasks. Children had the technology demonstrated to them before they began and had a chance to do a short piece of practise writing before they started writing their story.

2.1 Analysis and Results

There were three outputs from each instance of use. The first was a journal file that showed the writing of the child. This was used to generate an image of the child's writing; an example is seen in Fig. 1. The second output (PT, presented text) was created by the lead researcher and was a text file of what the child wrote (as seen in Fig. 3). A related text file (TT, transcribed text) was created from the journal file by the recognition software (shown in Fig. 4).

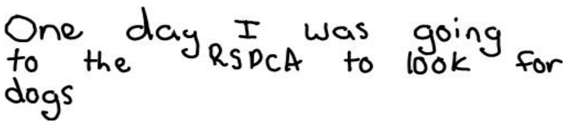

One day I was going
to the RSPCA to look for
dogs

Fig. 1. Writing as collected in the Windows Journal Application

Outputs PT and TT were aligned in two ways using minimum string distance (MSD) algorithms [6] and from these, two error rates, word error rate (WER) and character error rate (CER) were derived. Each was calculated in a similar way where: E*rror Rate* = *(S + I + D) / N* where N is the number of words or characters The error rates from the work of the children are shown in Table 1.

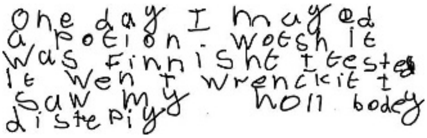
On average, around one in every six letters was inaccurately recognized. The average WER was 30%; this was considerably less favorable than the CER at 17%, and for around half of the children, the difference between the CER and WER was quite pronounced. Reasons for these discrepancies are briefly explored in the next section.

Table 1. Recognition rates from the text pairs (N = number of characters written)

Child	1	2	3	4	5	6	7	8	9	10	Avg
N	38	172	31	132	83	121	79	122	45	86	
WER (%)	0	19	37	28	60	24	42	21	36	35	30
CER (%)	0	10	41	13	25	12	11	7	42	12	17

3 Discussion

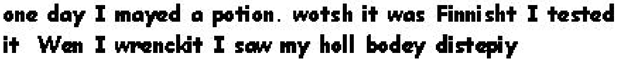
The discussion that follows uses (as an example) the writing from child number 5 (seen in Fig. 2.) to demonstrate some of the problems with the derivation of error rates.



One day I mayed
a potion. wotsh it
was finnisht I tested
it wen I wrenckit I
saw my holl bodey
distepiy

Fig. 2. The writing as it was done

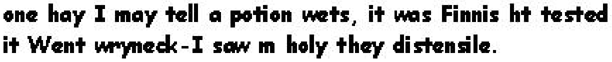
The first process that was carried out was the interpretation of this writing into text (to create a text string for the error rate calculations), and the resulting text (PT) can be seen in Fig. 3. There is a problem at this point as this is the '*researchers guess*' about the child's writing. For example, the character that follows 'wen' was assumed to be a capital I, but without using the contextual clues (provided by the sense of the sentence), it could be easily considered as a 't' and this would have had an impact on the recognition results.



one day I mayed a potion. wotsh it was Finnisht I tested
it Wen I wrenckit I saw my holl bodey distepiy

Fig. 3. The writing once it was interpreted (PT)

The recognizer that was used in this experiment uses a dictionary to assist in the recognition process. This has an effect on the recognition results (shown in Fig. 4) as, for instance, when the word 'distepiy' is recognized, the characters would individually make a word sufficiently close to the word 'distensile' to convince the recognition software that this is what was written.



one hay I may tell a potion wets, it was Finnis ht tested
it Went wryneck-I saw m holy they distensile.

Fig. 4. The writing once it was recognized (TT)

The teacher of the child (and the researcher) assumed that the child wrote 'distepiy' to mean 'disappear'; a phonetically designed matching algorithm (spell checker) would have had a much better chance of getting this word right. It is in a similar way that 'bodey' ('body') turns into 'they'.

4 Conclusion

From this very simple study it is evident that the reported error rate numbers fail to say it all. Firstly, the CER and the WER metrics were not consistent, it is easy to see how one or other of these figures could be reported and could present conflicting results. Secondly, the small investigation of a single child's writing demonstrates the impact of several factors, the included dictionary, the text creation task, the knowledge of the researcher and the diversity of the child population. It appears that for this study there was a real need for the researcher to be immersed in the context related to the single task as well the overall context of learning, school setting, and user experience, calligraphy, and child motivation [7].

Some of these findings translate into other studies; any study which relies on written (or to a lesser extent) spoken language with child users will be influenced by their developmental stage and the researchers knowledge and studies using other novel applications that 'borrow' metrics from related domains need to be investigated to determine the appropriateness of the metrics and to determine which metrics are most valid.

Further work that is planned in relation to this study includes an investigation of the impact of phrase choices when children use copied phrases for handwritten text input and a study with older children to determine whether the disparity between the CER and WER measures is reduced as children gain common knowledge in language.

References

1. Read, J.C., *The ABC of CCI.* Interfaces, 2005. **62**: p. 8 - 9.
2. Plamondon, R. and S.N. Srihari, *On-line and Off-Line Handwriting Recognition: A Comprehensive Survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**(1): p. 63 - 84.
3. McFall, R., E. Dahm, D. Hansens, C. Johnson, and J. Morse. *A Demonstration of a Collaborative Electronic Textbook Application on the Tablet PC.* in *World Conference on Educational Multimedia, Hypermedia and Telecommunications.* 2004: AACE.
4. Read, J.C. and M. Horton. *The Usability of Digital Tools in the Primary Classroom.* in *Ed-Media2004.* 2004. Lugano: AACE.
5. MacKenzie, I.S. and R.W. Soukoreff, *Text Entry for Mobile Computing: Models and Methods, Theory and Practice.* Human-Computer Interaction, 2002. **17**(2): p. 147 - 198.
6. MacKenzie, I.S. and R.W. Soukoreff. *A Character-Level Error Analysis for Evaluating Text Entry Methods.* in *NordiChi2002.* 2002. Aarhus, Denmark: ACM Press.
7. Nardi, B., *Context and Consciousness : Activity Theory and Human-Computer Interaction.* 1996, Cambridge, MA: MIT Press.