

Assessing The Effectiveness of Pen-Based Input Queries

Stephen Levin
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2664

s.levin@sheffield.ac.uk

Paul Clough
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2664

p.d.clough@sheffield.ac.uk

Mark Sanderson
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2648

m.sanderson@sheffield.ac.uk

ABSTRACT

In this poster, we describe an experiment exploring the effectiveness of a pen based text input device for use in query construction. Standard TREC queries were written, recognised, and subsequently retrieved upon. Comparisons between retrieval effectiveness based on the recognised writing and a typed text baseline were made. On average, effectiveness was 75% of the baseline. Other statistics on the quality and nature of recognition are also reported.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *search process*; H.5.2 [Information Interfaces and Presentation]: User Interfaces – *input devices and strategies*; I.7.5 [Document and Text Processing]: Document Capture – *graphics recognition and interpretation*.

General Terms

Measurement, Performance, Experimentation.

Keywords

Tablet PC, handwriting recognition.

1. INTRODUCTION

Stylus-based handwriting recognition technology - although researched and available for many years - became more known with the recent release of Microsoft's Tablet PC system [4]. Given the potential ubiquity of such a well promoted product, it was decided to explore its utility to IR, starting with a study of pen-based queries.

Central to the use of pen-based input as a query input mechanism to an IR system is the notion of converting the query into a searchable form. Such ideas have been explored in the past with different media and text: Kupiec et al. [1] examined spoken queries and many have explored cross language (CL) retrieval [2, 3], where queries are translated to be searchable. This study aimed to determine the Tablet PC's potential usefulness as a query input mechanism.

2. METHODOLOGY

In total, 20 individuals participated in a study that took place over a week during February 2003, at the University of Sheffield. The

participants were of various ages and nationalities and had different educational backgrounds, although all had been in further education. It was anticipated that including participants from many different nationalities would result in a wider variety of handwriting samples.

Participants were asked to complete an initial questionnaire aimed at collecting information, including whether they had used writing recognition tools in the past, and how they rated their own handwriting. After a brief training task, participants were then asked to write the titles of TREC-6 ad hoc topics 301-350, into a "notepad" application. Upon completing the task, participants were interviewed to determine: how easy they found it to get to grips with, how they felt it compared to using a pen and paper and whether they altered their style of writing at all when using the system.

To determine the effect of handwriting recognition on retrieval effectiveness, the Tablet PC's recogniser was applied to all the participants' hand-written queries and the textual output (the recogniser's best guess for each word) was used as input to a BM25-based IR system. For every query, precision at rank 20 was calculated. A baseline of retrieval using typed queries was also computed

3. RESULTS

Baseline precision at 20 was 0.33 - on average, 6 relevant documents retrieved in the top 20. Across the 20 participants, the average precision at 20 was 0.247 (mean) and 0.259 (median), meaning that on average the pen-based IR system operated at 75% (mean) and 78.5% (median) of the baseline.

Of the 20 participants, 17 were significantly worse than the baseline, using a t-test ($p < 0.05$). Results from 2 of the 20 participants were lower than 50% of baseline, the lowest precision at 20 score being 0.07 (21.2% of the baseline). The highest of the significant results was 0.28 (84.8% of the baseline performance). Upon inspection, it was found the precision at 20 scores appeared to be fairly similar from participant to participant for each topic, suggesting that topics were consistently recognised either correctly or incorrectly for most participants.

Figure 1 shows the differences in precision at 20 between the baseline and the average score across all participants, for each topic. Topics 326 ("ferry sinkings"), 324 ("argentine british relations") and 331 ("world bank criticism") differ from the baseline by more than 30%, whilst the majority of topics differ by less than 10%. Figure 1 also shows cases where the difference is negative, i.e. retrieval was better than the baseline. For those queries, using all the words in the query retrieved fewer relevant

documents in the top 20 than if only some of the words were used. In most cases, terms were incorrectly recognised as words not found in the index and were therefore ignored, improving retrieval effectiveness.

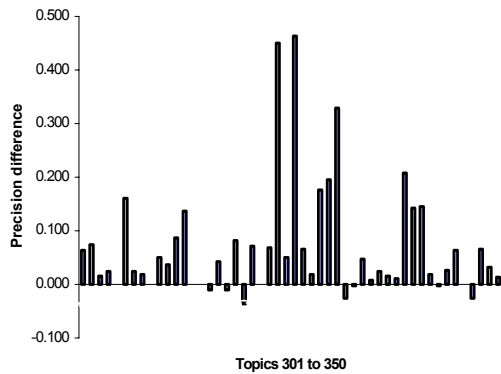


Figure 1. Difference in precision at 20 for each topic in the case of the baseline and average score across all participants

As well as retrieval effectiveness, accuracy of individual word recognition was assessed. The 50 topics contained a total of 136 words. We computed the number of words in the participants’ recognised queries that matched the original query terms. On average, across all participants, 77.2% words were recognised correctly, the worst case being participant 1 at 43.4%, and the best being participants 3 and 9 at 89%. Perhaps unsurprisingly, retrieval effectiveness was better as more words were recognised (confirmed using a non-parametric bivariate correlation, Spearman’s rho, giving $\rho(0.724)$ at a 99% confidence level).

‘Easiest’	# correct	‘Hardest’	# correct
telescope	20	sinkings	0
disease	20	iraq	1
endangered	20	iran	3
and	20	polygyny	3
unexplained	20	fiber	5
health	20	hubble	5
accidents	20	british	8
adoptive	20	alzheimer’s	8
dangerous	20	mail	9
and	19	faxes	10
journalistic	19	ban	10

Table 1. Top 10 easy and hard words, with the number of participants for which the word was correctly recognised

Baseline word	Example written word after recognition
sinkings (326)	[sinking],[sinking’s],[linking],[sin things]
iraq (330)	[trans],[turn],[van],[cran]
iran (330)	[wag],[trend],[vang],[fang]
polygyny (316)	[polygamy]
fiber (320)	[fibber],[fibre],[then],[offie]
hubble (303)	[bubble],[hurtle],[hobble]
british (324)	[briskish],[flitch],[brutish]
alzheimer’s (339)	[aleheimers],[alzheimers],[olzheimers5]
e-mail (344)	[email][e-moort],[e-more]
faxes (317)	[taxes],[farer],[faxed]
ban (340)	[can],[bar],[has],[been]

Table 2. Hardest topic words and alternates from recogniser

Table 1 lists the top 10 terms, each of which were found in at least 19 of the participant’s queries (i.e. the ‘easy’ to recognise words).

Table 1 also shows the 10 query terms that were recognised across the least number of participants. Some of the words occurred more than once, as certain terms occurred in numerous phrases and each occurrence was regarded as unique. Table 2 lists various examples of how the 10 ‘hardest’ words were incorrectly recognised.

From a usability perspective, comments made by participants during the interview sessions suggested that the system was highly usable. All participants agreed that it was straightforward and easy to get to grips with. However, users observed several differences between using the system and a normal pen and paper, the most common being the low resistance between pen and screen. Some participants remarked that this presented difficulty at first, with one participant commenting that it caused them to write larger than they normally would, but most said that they quickly became accustomed to the difference; two stated that they actually preferred the lack of resistance as it meant having to apply less pressure. Several participants also commented that the height of the Tablet PC caused them to raise their arm and subsequently write at an unusual angle. However, despite these comments, the majority of participants claimed that their style of writing was not altered, although several remarked that the resulting output appeared untidier than usual.

4. CONCLUSIONS

An experiment testing the effectiveness of a pen-based input system as a means of entering queries was conducted. Using the writing of 20 subjects and 50 topics from the TREC test collection, it was found that using just the writing recogniser’s “best guess” words produced retrieval effectiveness that was 75% of that expected from queries that were typed in. Despite variation in writing styles, some words were found to be consistently recognised, and others consistently not.

5. FUTURE WORK

With certain changes to the retrieval system, it is anticipated that effectiveness can be improved further, in a manner similar to some spoken document and CL retrieval systems [2, 3], where word hypotheses from the recogniser are grouped as synonyms. Multiple hypotheses could be added to the search query using the InQuery synonym operator, hopefully increasing search recall without harming precision.

6. REFERENCES

- [1] Kupiec, J., Kimber, D. and Balasubramanian, V. (1994): Speech-based retrieval using semantic co-occurrence filtering, in *Proceedings of ARPA Human Language Technology Workshop*: 373-377.
- [2] Ballesteros, L., Croft, W.B. (1998): Resolving ambiguity for cross-language retrieval, in *Proceedings of ACM SIGIR*: 64-71.
- [3] Pirkola, A. (1998): The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval, in *Proc. of ACM SIGIR*: 55-63.
- [4] Windows XP Tablet Edition Home Page. <http://www.microsoft.com/windowsxp/tablet/default.asp> [Site visited 2nd May 2003].