Improving Word-Recognizers Using an Interactive Lexicon with Active and Passive Words

Per Ola Kristensson
Cavendish Laboratory
University of Cambridge
JJ Thomson Avenue, CB3 0HE, Cambridge, UK
pok21@cam.ac.uk

Shumin Zhai
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120, USA
zhai@almaden.ibm.com

ABSTRACT

The words a user is likely to write comprise the user's active vocabulary. This vocabulary is considerably smaller than the passive vocabulary of words a user reads. We explore an interactive adaptive lexicon method that separates a large lexicon into active and passive sets, and gradually expands and adapts the active set to reflect the user's active vocabulary. The adaptation is achieved through lightweight interaction as a by product of actual use. The effectiveness of the technique is demonstrated through a computational experiment and a user study.

Author Keywords

lexicon, recognition, handwriting recognition, adaptive user interfaces

ACM Classification Keywords

H5.2. Information interfaces and presentation: User Interfaces – *input devices and strategies*; I5.1. Pattern Recognition: Implementation – *interactive systems*.

INTRODUCTION

A word-recognizer uses a lexicon to match the user's input against template words. Ideally this lexicon contains those words that a particular user needs to write, no more, no less. A too small lexicon does not cover all of the user's active vocabulary, causing frequent out-of-vocabulary errors. A too large lexicon, on the other hand, contains many more words than the user's active vocabulary, which enlarges the recognizer's search space, consequently increasing the likelihood of confusion errors [6].

It is therefore desirable to customize the word-recognizer's lexicon to the individual user's active vocabulary. The conventional method to achieve this goal is to create a personalized lexicon from the user's past writing during, or after, the installation of the recognition system (*e.g*. in [4]). It is often difficult to find enough written material from a user, since it is usually scattered over various devices, document formats, local and on-line email, chat sessions, etc. Users tend to be reluctant in spending the upfront attention and time overhead to initiate and participate in the personalization procedure. It is also challenging to filter out the 'noise' mixed in a user's writing (forwarded emails, signatures, misspellings, software messages, spam, etc.).

INTERACTIVE ADAPTIVE LEXICON

We explore a complementing approach to lexicon customization called the *interactive adaptive lexicon*. People use a smaller vocabulary when they write and speak (the *active vocabulary*) than the full vocabulary of words they understand (the *passive vocabulary*). Our adaptive lexicon reflects this by dividing the lexicon into *active* and *passive* sets. Initially only the most frequent, for example 10,000, words in English (or other languages) are set active, but over time the set of active words evolves to reflect the user's language use.

Examples

To illustrate how the proposed interactive adaptive lexicon works, we use a few examples. Assume our recognizer has a lexicon containing 50K words, with 10K of the top ranked words (by their frequency count in, for example, the American National Corpus) in the active, and the rest in the passive set. Furthermore assume four words with similar appearance (by a certain recognition criterion): *compete*, *complete*, *compile*, and *compiles* are all in the lexicon but *compete* and *complete* are in the active set (marked with white background in Figure 1d) and *compile* and *compiles* in the passive set (marked with gray background in Figure 1d). If the user writes *compete* in reasonably good form, the recognizer takes the written sample, classifies it against its (entire) lexicon and finds the closest matching candidates in the order of *compete*, *complete*, *compile*, and *compiles*. Being the top match and in the active set, *compete* is returned (displayed) to the user as the recognized word.

Now suppose the user writes the word "compile" in reasonably good form (Figure 1a). The word-recognizer takes the input and classifies it against its (entire) lexicon

IUI '08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.
Copyright 2008 ACM 978-1-59593-987-6/ 08/ 0001 $5.00

and finds the closest matching candidates in the order of *compile*, *compiles*, *complete*, and *compete*. Since the recognizer always returns *the top match in the active lexicon* to the user, *complete* is returned to the user (Figure 1b). Seeing *complete* rather than *compile* displayed, the user clicks on the displayed word *complete,* causing the recognition system to display an *N*-best list with words from the active set displayed in white, and words from the passive set displayed in gray background (Figure 1c). Seeing the intended word *compile* on the list, the user slides the pen down and selects it. The returned word is now changed to *compile*. At the same time, the recognition system learns that the word *compile*, although not frequent in common English, is part of the user's active vocabulary and moves the word *compile* into the active set (Figure 1e).

Hereafter anytime the user writes *compile* (in reasonably good form) the intended word will always be returned and the user does not need to go through the detour with the *N*-best list again.

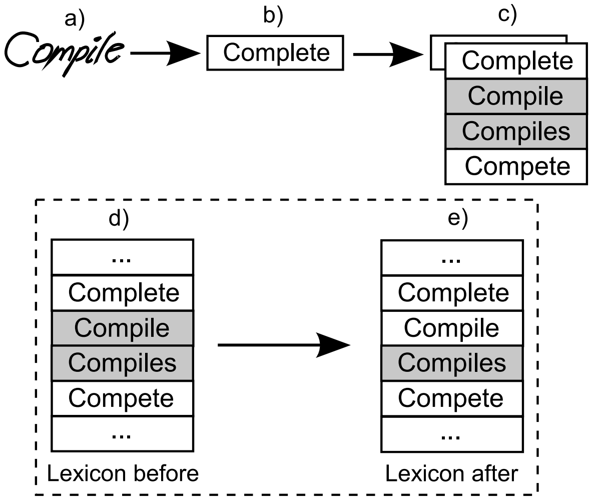

Figure 1. Graphical illustration of the interactive adaptive lexicon (see the Example subsection for an explanation).

To test its feasibility we have implemented the interactive adaptive lexicon concept in the ShapeWriter word-recognizer [4]. The system searches a lexicon with 60,000 words in less than 20 ms on a standard PIII 800 MHz PC.

Categories of Errors

Error and error correction are unavoidable in word-recognizers and are in fact frequent in all text input interfaces. For instance, the backspace key is one of the most frequently typed keys on the keyboard. However, the adaptive lexicon improves the user experience of a word-recognizer by making a tradeoff that handles these errors in a low cost manner to the user. This can be understood by analyzing the two primary categories of errors: confusion errors and out-of-vocabulary errors.

Confusion errors are caused by the recognizer confusing two words with each other. For instance if the user writes *the* (but not accurately enough) and the recognizer returns *then*, the recognizer confused the two words with each other. When a confusion error occurs, the user has to correct it by for example using an *N*-best list.

Out-of-vocabulary errors [2] (OOV errors) are caused by the recognizer's lexicon not containing the intended word. When an OOV error occurs, it carries a high cost for the user because invoking and searching the *N*-best list will still not result in finding the intended word. The user will assume the input was too inaccurate and may retry input. Of course, this will continue to fail, and eventually the user may realize that the intended word is out of the recognizer's vocabulary and resort to alternative means to add the word to the system's lexicon.

Using the adaptive lexicon we can reduce confusion errors by keeping the initial size of the active set in the lexicon relatively small. Adaptive lexicons can also reduce out-of-vocabulary errors because they can have a very large number of words in their passive set. Adaptive lexicons achieve these benefits at the small cost for the user to periodically activate passive words from the *N*-best list as a by-product of selecting the intended word.

In the interactive adaptive lexicon adaptation is achieved through light weight user interaction in use, rather than heavy weight upfront user overhead operation (e.g. letting the user select words in the lexicon before use) or completely automatic adaptation processes that may not tightly fit the user.

Out-of-vocabulary errors cannot be entirely eliminated because the total number of possible words is enormous. For example the Oxford Dictionary contains 291,500 lexeme entries [5]. One may also need to write non-dictionary words including names, jargons, and acronyms. However when writing a non-dictionary word, the writer is more likely to expect it to be out of the recognizer's vocabulary, thus prompting them to add it to the active set.

COMPUTATIONAL EXPERIMENT

Informal tests of the implemented adaptive lexicon technique in the ShapeWriter word-recognizer [4] have shown that it is indeed effective. However the efficacy of the method cannot be systematically measured empirically by a traditional user study alone. Instead, it also amounts to a set of analytical questions: With a relatively small number of active words (e.g. 7K), how often does the user have to invoke the *N*-best list to fetch a passive word? As the writer enters an ever growing text mass, how many passive words are activated? Would different users' set of active words be similar (in which case it would be better to include these common words among the initial active words)? Would the active set of words in the lexicon practically stop growing?

To answer these questions we performed a computational experiment based on real email texts. We scanned the five users in the Enron online email corpus [3] who had sent the most emails in the database. We parsed the sent email messages, ranked them by date and filtered out as much noise as possible, such as forwarded messages and email signatures. The number of emails sent by the individual users ranged from 4,407 to 8,926. The time intervals for the accounts spanned between 183-1036 days.

Procedure

We used a lexicon containing 57,392 words, obtained by scanning a large corpus of written email, novels, articles and forum postings. We defined a relative cutoff frequency threshold that divided the lexicon into an active set containing 7,037 words, and a passive set containing 50,355 words. The optimal size of the initial active set can be determined in the future without impacting the conclusions of the current inquiry. Note that the study is also independent of the word-recognizer used as long as it is lexicon-based.

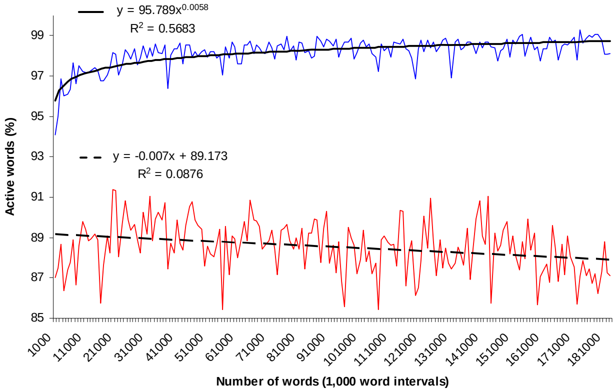

Figure 2. In chronological order and averaged over all five users, the percentage of every 1,000 words written that belonged to the adaptive active set is shown by the upper blue line. The lower red line shows a baseline of a fixed lexicon with 7,037 words. The adaptive lexicon and the baseline do not start at the same percentage because the first data point starts at 1000 words.

For each of the five users the email texts were extracted and tokenized into a chronological stream of words. For each word encountered the first time, its membership in the active or passive sets was recorded (counted), and if it was in the passive set it was promoted to the active set (upon first occurrence). Words not found in the lexicon were counted and registered in a separate list of *unknown tokens*. Thereafter these words were also transferred to the active lexicon set.

Results

Figure 2 is a plot of the percentage of written words in each 1,000 word interval, averaged over all five users, covered by the active set of an adaptive lexicon The figure demonstrates that the active set grows when the written text mass increases (upper plot). As a comparison, the lower plot shows a baseline where the adaptive lexicon is not in use, *i.e.* no activations takes place.

Table 1 and Table 2 show snapshots of the percentage distribution for the first and last 10,000 words each user wrote respectively. For all users, the active membership covered over 95% words written initially, and over 98% eventually. Furthermore, the active and passive sets combined represent 99.48% of each user's writing at the end. This means only 0.52% of the words written were true out-of-vocabulary words that the user would not have been able to write directly or access through the *N*-best list.

User	Active (%)	Passive (%)	Unknown (%)
User 1	97.42%	2.12%	0.46%
User 2	96.68%	2.23%	1.09%
User 3	95.24%	3.78%	0.98%
User 4	97.32%	1.68%	1.00%
User 5	95.1%	3.67%	1.23%

Table 1. The percentage of words that were active, passive or not included in the lexicon (unknown), for the first 10,000 words written by each user.

User	Active (%)	Passive (%)	Unknown (%)
User 1	99.0%	0.66%	0.34%
User 2	98.16%	1.19%	0.65%
User 3	98.81%	0.90%	0.29%
User 4	98.11%	1.42%	0.47%
User 5	98.01%	1.23%	0.76%

Table 2. The percentage of words that were active, passive or not included in the lexicon (unknown), for the last 10,000 words written by user.

Commonality between Users' Activated Passive Words

An interesting aspect is how many of the activated passive words in the lexicon were shared among the users. If many words were shared, this suggests that the relative frequency threshold we defined for active lexicon membership was set too high, excluding many words frequently used by users.

In total the five users wrote 16,484 words that were passive in the lexicon. The intersection of all the users' passive words contains only 282 words, or 1.7%. The largest pair-wise intersection between two users was 1549 words, or 9.4%. Hence the advantage of the adaptive lexicon is apparent. A slight increase of the active set in the lexicon would generate little benefit overall to all users.

USER STUDY

The computational experiment showed that the adaptive lexicon quickly converged to users' vocabulary. An open question is how much word recognition accuracy can increase when the recognizer searches a smaller lexicon. To address this question we carried out a user study. To "stress-test" recognition we asked participants to write as fast as possible. Since the input signal is increasingly noisier when users write faster and sloppier, this scenario is the hardest to improve upon by modifying the recognition algorithm alone. Therefore any accuracy increase here is very useful.

Procedure

We recruited five paid volunteers. Participants were each given a single phrase (4-6 words) taken randomly from the Enron email corpus [3]. They were asked to write the single phrase repeatedly for 10 minutes using the ShapeWriter word recognizer [4] with a 55K lexicon and an active lexicon set arbitrarily at 15K. The data collected was the users' pen traces.

Results

We investigated how accuracy changed as a function of lexicon size by taking the pen traces recorded from participants and re-running them through the recognizer. Figure 3 shows that accuracy steadily decreased from 71.5% (5K lexicon) to 55.6% (55K) lexicon. That is, the adaptive lexicon set at a 5K active lexicon threshold would increase accuracy by 15.9% compared to a baseline system with a 55K lexicon. With a 15K lexicon recognition accuracy was at 64.5%, 7% lower than for a 5K lexicon. We should caution the reader with two caveats to this analysis. First, the pen traces were collected from users under the instruction to push the recognizer as much as possible to produce a relatively high error rate. This may have resulted in pen traces deviating from normal use. Second, the "what-if" analysis of different lexicon thresholds cannot account for the recognizer's feedback impact on participants. In other words, the amount of errors seen may have influenced participants' behavior.

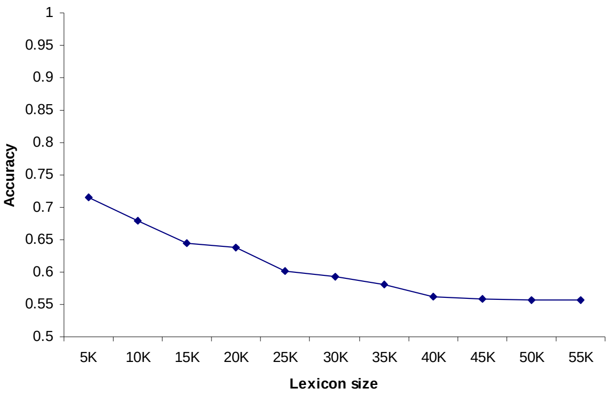

Figure 3. Accuracy as a function of lexicon size.

DISCUSSION

Our analysis of real-world email data indicates that the active set in the lexicon will most likely always increase slightly. The alternative method of personalizing the lexicon by examining users' past writing mentioned in the introduction would therefore result in more OOV errors than the adaptive lexicon. However, users' past writing can, if easily available, be searched for passive words that can be immediately activated.

Unlike traditional computational methods the adaptive lexicon relies on a user-interface and opens up many potential future contributions from the HCI field to pattern recognition. As an example, an interesting analysis would be to study the increased effort imposed on users to activate words, compared to the decreased effort and user frustration in handling recognition errors. Since most activations of passive words occur in the beginning, such an analysis is not necessarily in favor of the adaptive lexicon at the initial use stage. However, once an adaptive lexicon is configured for a user, it closely resembles that particular user's active vocabulary. At that point, the word-recognizer's lexicon contains the minimum number of distracter words and the user would rarely (if ever) consult the *N*-best list in order to activate a passive word.

CONCLUSIONS

In summary, our analytical experiment based on email texts shows that the active set of words covered initially 94-97% of words entered depending on the user, and grew to a size that covers 98-99% of the words entered. Further, our initial user study verified that recognition accuracy indeed decreased as a function of lexicon size.

The adaptive lexicon is a novel, straight-forward and easy to implement technique that can be used to reduce user frustration with pen/finger-based word-recognizers. The proposed method complements work on writer adaptation [1] and language modeling in handwriting recognition and can coexist with any such methods. Although we have demonstrated that the adaptive lexicon shows potential in aiding word-recognizers, much work is still ahead. For example, we are currently working on modeling estimated user effort and the likelihood of additional recognition errors as the size of the lexicon progressively increases.

ACKNOWLEDGMENTS

Parts of this work were conducted when Per Ola Kristensson was working at Linköping University and IBM Almaden Research Center. Per Ola Kristensson was supported by grants from Nokia and Ericsson Research Foundation.

REFERENCES

1. Connell, S.D. and Jain, A.K. Writer Adaptation for Online Handwriting Recognition. *IEEE Trans. PAMI*, 24(3), 2002, 329-346.
2. Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. The Vocabulary Problem in Human-System Communication. *Comm. ACM*, 30(11), 1987, 964-971.
3. Klimt, B. and Yang, Y. Introducing the Enron Corpus. *Proc. CEAS 2004*.
4. Kristensson, P.O. and Zhai, S. SHARK^2: A Large Vocabulary Shorthand-Writing System for Pen-Based Computers. *Proc. ACM UIST 2004*, 43-52.
5. Simpson, J. and Weiner, E. C. *Oxford English Dictionary*. Clarendon Press, 1989.
6. Xue, H. and Govindaraju, V. On the Dependence of Handwritten Word Recognizers on Lexicons. *IEEE Trans. PAMI*, 24(12), 2002, 1553-1564.