Evaluating Word Prediction: Framing Keystroke Savings

Keith Trnka and Kathleen F. McCoy
University of Delaware
Newark, DE 19716
trnka@cis.udel.edu

Abstract

Researchers typically evaluate word prediction using keystroke savings, however, this measure is not straightforward. We present several complications in computing keystroke savings which may affect interpretation and comparison of results. We address this problem by developing two gold standards as a frame for interpretation. These gold standards measure the maximum keystroke savings under two different approximations of an ideal language model. The gold standards additionally narrow the scope of deficiencies in a word prediction system.

1 Introduction

Word prediction is an application of language modeling to speeding up text entry, especially to entering utterances to be spoken by an Augmentative and Alternative Communication (AAC) device. AAC devices seek to address the dual problem of speech and motor impairment by attempting to optimize text input. Even still, communication rates with AAC devices are often below 10 words per minute (Newell et al., 1998), compared to the common 130-200 words per minute speech rate of speaking people. Word prediction addresses these issues by reducing the number of keystrokes required to produce a message, which has been shown to improve communication rate (Trnka et al., 2007). The reduction in keystrokes also translates into a lower degree of fatigue from typing all day (Carlberger et al., 1997).

Word prediction systems present multiple completions of the current word to the user. Systems generate a list of W predictions on the basis of the word being typed and a language model. The vocabulary is filtered to match the prefix of the current word and the language model ranks the words according to their likelihood. In the case that no letters of the current word have been entered, the language model is the sole factor in generating predictions. Systems often use a touchscreen or function/number keys to select any of the predicted words.

Because the goal of word prediction systems is to reduce the number of keystrokes, the primary evaluation for word prediction is keystroke savings (Garay-Vitoria and Abascal, 2006; Newell et al., 1998; Li and Hirst, 2005; Trnka and McCoy, 2007; Carlberger et al., 1997). Keystroke savings (KS) measures the percentage reduction in keys pressed compared to letter-by-letter text entry.

$$KS = \frac{keys_{normal} - keys_{with\ prediction}}{keys_{normal}} \times 100\%$$

A word prediction system that offers higher savings will benefit a user more in practice.

However, the equation for keystroke savings has two major deficiencies. Firstly, the equation alone is not enough to compute keystroke savings — actually computing keystroke savings requires a precise definition of a keystroke and also requires a method for determining how many keystrokes are used when predictions are available, discussed in Section 2. Beyond simply computing keystroke savings, the equation alone does not provide much in the way of interpretation — is 60% keystroke savings good? Can we do better? Section 3 will present two gold standards to allow better interpretation of keystroke savings.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 261–264,
Columbus, Ohio, USA, June 2008.

2 Computing Keystroke Savings

We must have a way to determine how many keystrokes a user would take under both letter-by-letter entry and word prediction to compute keystroke savings. The common trend in research is to simulate a "perfect" user that will never make typing mistakes and will select a word from the predictions as soon as it appears.

Implementation of perfect utilization of the predictions is not always straightforward. For example, consider the predictive interface in Microsoft Word™: a single prediction is offered as an inline completion. If the prediction is selected, the user may backspace and edit the word. However, this freedom makes finding the minimum sequence of keys more difficult — now the user may select a prediction with the incorrect suffix and correct the suffix as the optimal action. We feel that a more intuitive interface would allow a user to undo the prediction selection by pressing backspace, an interface which does not support backspace-editing. In addition to backspacing, future research in multi-word prediction will face a similar problem, analogous to the garden-path problem in parsing, where a greedy approach does not always give the optimal result.

The keystrokes used for training and testing word prediction systems can affect the results. We attempt to evaluate word prediction as realistically as possible. Firstly, many corpora have punctuation marks, but an AAC user in a conversational setting is unlikely to use punctuation due to the high cost of each key press. Therefore, we remove punctuation on the outside of words, such as commas and periods, but leave word-internal punctuation intact. Also, we treat capital letters as a single key press, reflecting the trend of many AAC users to avoid capitalization. Another problem occurs for a newline or "speak key", which the user would press after completing an utterance. In pilot studies, including the simulation of a speak key lowered keystroke savings by 0.8–1.0% for window sizes 1–10, because newlines are not able to be predicted in the system. However, we feel that the simulation of a speak key will produce an evaluation metric that is closer to the actual user's experience, therefore we include a speak key in our evaluations.

An evaluation of word prediction must address these issues, if only implicitly. The effect of these potentially implicit decisions on keystroke savings can make comparison of results difficult. However, if results are presented in reference to a gold standard under the same assumptions, we can draw more reliable conclusions from results.

3 Towards a Gold Standard

In trying to improve the state of word prediction, several researchers have noted that it seems extremely difficult to improve keystroke savings beyond a certain point. Copestake (1997) discussed the entropy of English to conclude that 50–60% keystroke savings may be the most we can expect in practice. Lesher et al. (2002) replaced the language model in a word prediction system with a human to try and estimate the limit of keystroke savings. They found that humans could achieve 59% keystroke savings with access to their advanced language model and that their advanced language model alone achieved 54% keystroke savings. They noted that one subject achieved nearly 70% keystroke savings on one particular text, and concluded that further improvements on current methods are possible. Garay-Vitoria and Abascal (2006) survey many prediction systems, showing a wide spectrum of savings, but no system offers more than 70% keystroke savings.

We investigated the problem of the limitations of keystroke savings first from a theoretical perspective, seeking a clearly defined upper boundary. Keystroke savings can never reach 100% — it would mean that the system divined the entire text they intended without a single key.

3.1 Theoretical keystroke savings limit

The minimum amount of input required corresponds to a perfect system — one that predicts every word as soon as possible. In a word *completion* system, the predictions are delayed until after the first character of the word is entered. In such a system, the minimum amount of input using a perfect language model is two keystrokes per word — one for the first letter and one to select the prediction. The system would also require one keystroke per sentence. In a word *prediction* system, the predictions are available immediately, so the minimal in-

put for a perfect system is one keystroke per word (to select the prediction) and one keystroke per sentence. We added the ability to measure the minimum number of keystrokes and maximum savings to our simulation software, which we call the *theoretical keystroke savings limit*.

We evaluated a baseline trigram model under two conditions with different keystroke requirements on the Switchboard corpus. The simulation software was modified to output the theoretical limit in addition to actual keystroke savings at various window sizes. To demonstrate the effect of the theoretical keystroke savings limit on actual savings, we evaluated the trigram model under conditions with two different limits — word prediction and word completion. The evaluation of the trigram model using word *completion* is shown in Figure 1. The actual keystroke savings is graphed by window size in reference to the theoretical limit. As noted by other researchers, keystroke savings increases with window size, but with diminishing returns (this is the effect of placing the most probable words first). One of

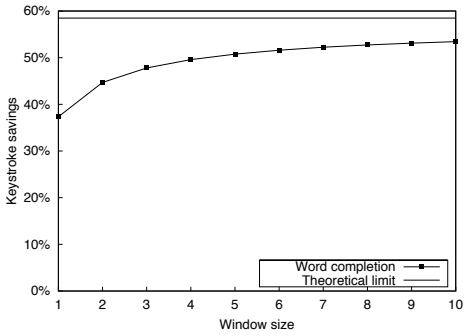

Figure 1: Keystroke savings and the limit vs. window size for word completion.

the problems with word completion is that the theoretical limit is so close to actual performance — around 58.5% keystroke savings compared to 50.8% keystroke savings with five predictions. At only five predictions, the system has already realized 87% of the possible keystroke savings. Under these circumstances, it would take a drastic change in the language model to impact keystroke savings.

We repeated this analysis for word *prediction*, shown in Figure 2 alongside word completion. Word prediction is much higher than completion, both theoretically (the limit) and in actual keystroke savings.

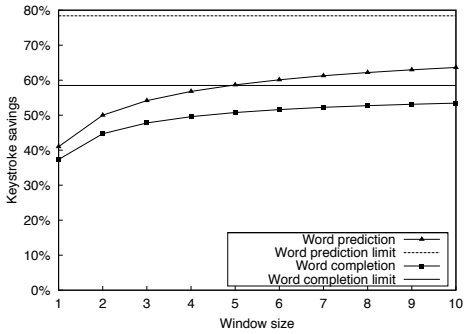

Figure 2: Keystroke savings and the limit vs. window size for word prediction compared to word completion.

Word prediction offers much more headroom in terms of improvements in keystroke savings. Therefore our ongoing research will focus on word prediction over word completion.

This analysis demonstrates a limit to keystroke savings, but this limit is slightly different than Copestake (1997) and Lesher et al. (2002) seek to describe — beyond the limitations of the user interface, there seems to be a limitation on the predictability of English. Ideally, we would like to have a gold standard that is a closer estimate of an ideal language model.

3.2 Vocabulary limit

We can derive a more practical limit by simulating word prediction using a perfect model of all words that occur in the training data. This gold standard will predict the correct word immediately so long as it occurs in the training corpus. Words that never occurred in training require letter-by-letter entry. We call this measure the *vocabulary limit* and apply it to evaluate whether the difference between training and testing vocabulary is significant. Previous research has focused on the percentage of out-of-vocabulary (OOV) terms to explain changes in keystroke savings (Trnka and McCoy, 2007; Wandmacher and Antoine, 2006). In contrast, the vocabulary limit gives more guidance for research by translating the problem of OOVs into keystroke savings.

Expanding the results from the theoretical limit, the vocabulary limit is 77.6% savings, compared to 78.4% savings for the theoretical limit and 58.7% actual keystroke savings with 5 predictions. The practical limit is very close to the theoretical limit

in the case of Switchboard. Therefore, the remaining gap between the practical limit and actual performance must be due to other differences between testing and training data, limitations of the model, and limitations of language modeling.

3.3 Application to corpus studies

We applied the gold standards to our corpus study, in which a trigram model was individually trained and tested on several different corpora (Trnka and McCoy, 2007). In contrast to the actual trigram model

Corpus	Trigram	Vocab. limit	Theor. limit
AAC Email	48.92%	61.94%	84.83%
Callhome	43.76%	54.62%	81.38%
Charlotte	48.30%	65.69%	83.74%
SBCSAE	42.30%	60.81%	79.86%
Micase	49.00%	69.18%	84.08%
Switchboard	60.35%	80.33%	82.57%
Slate	53.13%	81.61%	85.88%

Table 1: A trigram model compared to the limits.

performance, the theoretical limits all fall within a relatively narrow range, suggesting that the achievable keystroke savings may be similar even across different domains. The more technical and formal corpora (Micase, Slate, AAC) show higher limits, as the theoretical limit is based on the length of words and sentences in each corpus. The practical limit exhibits much greater variation. Unlike the Switchboard analysis, many other corpora have a substantial gap between the theoretical and practical limits. Although the practical measure seems to match the actual savings similarly to OOVs testing with cross-validation (Trnka and McCoy, 2007), this measure more concretely illustrates the effect of OOVs on actual keystroke savings — 60% keystroke savings when training and testing on AAC Email would be extraordinary.

4 Conclusions

Although keystroke savings is the predominant evaluation for word prediction, this evaluation is not straightforward, exacerbating the problem of interpreting and comparing results. We have presented a novel solution — interpreting results alongside gold standards which capture the difficulty of the evaluation. These gold standards are also applicable to drive future research — if actual performance is very close to the theoretical limit, then relaxing the minimum keystroke requirements should be the most beneficial (e.g., multi-word prediction). Similarly, if actual performance is very close to the vocabulary limit, then the vocabulary of the language model must be improved (e.g., cache modeling, adding general-purpose training data). In the case that keystroke savings is far from either limit, then research into improving the language model is likely to be the most beneficial.

Acknowledgments

This work was supported by US Department of Education grant H113G040051.

References

Alice Carlberger, John Carlberger, Tina Magnuson, M. Sharon Hunnicutt, Sira Palazuelos-Cagigas, and Santiago Aguilera Navarro. 1997. Profet, a new generation of word prediction: An evaluation study. In *ACL-97 workshop on Natural Language Processing for Communication Aids*.

Ann Copestake. 1997. Augmented and alternative NLP techniques for augmentative and alternative communication. In *ACL-97 workshop on Natural Language Processing for Communication Aids*, pages 37–42.

Nestor Garay-Vitoria and Julio Abascal. 2006. Text prediction systems: a survey. *Univ Access Inf Soc*, 4:183–203.

Gregory W. Lesher, Bryan J. Moulton, D Jeffery Higginbotham, and Brenna Alsofrom. 2002. Limits of human word prediction performance. In *CSUN*.

Jianhua Li and Graeme Hirst. 2005. Semantic knowledge in word completion. In *ASSETS*, pages 121–128.

Alan Newell, Stefan Langer, and Marianne Hickey. 1998. The rôle of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, 4(1):1–16.

Keith Trnka and Kathleen F. McCoy. 2007. Corpus Studies in Word Prediction. In *ASSETS*, pages 195–202.

Keith Trnka, Debra Yarrington, John McCaw, Kathleen F. McCoy, and Christopher Pennington. 2007. The Effects of Word Prediction on Communication Rate for AAC. In *NAACL-HLT; Companion Volume: Short Papers*, pages 173–176.

Tonio Wandmacher and Jean-Yves Antoine. 2006. Training Language Models without Appropriate Language Resources: Experiments with an AAC System for Disabled People. In *Eurospeech*.