Examining Knowledge Sources for Human Error Correction

Yongmei Shi

Department of Computer Science
and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD, USA
yshi1@umbc.edu

Lina Zhou

Information Systems Department

University of Maryland, Baltimore County
Baltimore, MD, USA
zhoul@umbc.edu

Abstract

A variety of knowledge sources have been employed by error correction mechanisms to improve the usability of speech recognition (SR) technology. However, little is known about the effect of knowledge sources on human error correction. Advancing our understanding of the role of knowledge sources in human error correction could improve the state of automatic error correction. We selected three knowledge sources, including alternative list, imperfect context, and perfect context, and compared their usefulness to human error correction via an empirical user study. The results showed that knowledge sources had significant impact on the performance of human error correction. In particular, perfect context was the best that could significantly reduce word error rate without increasing the processing time.

Index Terms: speech recognition, error correction, user study.

1. Introduction

With decades of efforts, significant progress has been made in speech recognition (SR) technology, leading to various speech-based applications. Due to its unsatisfactory performance, SR technology's promises in bringing convenience and efficiency for humans to interact with computers are seriously compromised by the laborious efforts and frustration experienced in detecting and correcting recognition errors [1]. To bridge the gap between what people expect from SR and what the technology can achieve, it is desirable to find effective ways to correct recognition errors. It is expected that improvements could be achieved by discovering and even automating humans' experience in speech recognition [2]. Advancing our understanding of human error correction could provide guidance for the development of automatic error correction methods and the design of system support for human error correction. Therefore, we focus on human error correction in this study.

Various interactive interface designs for manual error corrections have been proposed to reduce human efforts. Respeaking (e.g., [3]) and selecting from alternative hypotheses (e.g., [4, 5]) are two of the most popular methods. Multiple modality based mechanisms that attempt to prevent repeated errors by providing multiple input modals have also gained attention (e.g., [6, 7]). However, almost all interactive interfaces are designed for speakers themselves. Some commercial dictation software supports third-party correction by archiving and even aligning original audio files with text transcriptions. Nonetheless, in all the above situations, the person who proof-edits the transcript, be it the speaker or the third party, has access to the expected output by knowing what they said or by listening to the audio files.

Our goal in this study is unique in that we aim to evaluate and discover knowledge that could improve third-party error correction in situations where the expected output (in a non-text modality) is not available or is difficult to use. This is motivated by many real-world applications of SR technology such as business communications, interviews, and medical transcription. Brill et al. [8] conducted a user study to investigate what kind of linguistic knowledge humans used to improve SR output by allowing users to either select from the n-best list or to directly edit the output. The findings provided insights into the development of advanced linguistics based language models. Various knowledge sources from multiple aspects have been proposed to facilitate automatic error correction to date. However, there is little work on assessing the usefulness of such knowledge for human error correction. Therefore, we expect to shed light on the impact of knowledge sources on human error correction by conducting an empirical user study.

Based on the extant work on error correction and our observations, we chose the following three types of knowledge sources to support error correction in this study:

Alternative list of hypotheses Alternative hypotheses are the ranked list of alternative words for each output word. It has been widely used in both automatic (e.g.,

[9]) and manual (e.g., [4, 5]) error correction to provide more possible choices.

Imperfect context Imperfect context means that the contextual information is not perfect (e.g., containing recognition errors) and thus may be misleading. Sarma and Palmer [10] exploited imperfect SR output to derive co-occurrence statistics, which were used to correct errors for specified topic words and could possibly achieve high precision with reasonable recall.

Perfect context Perfect context is an improvement over the imperfect context by providing corrected contextual information, as appropriate. User corrections have been used to automatically adapt the lexicon and pronunciations (e.g., [11]) of the SR system to improve the word error rate of the following utterances. The corrections of the surrounding errors are expected to provide more accurate context and useful feedback for human error correction.

In the following sections, we first introduce the experimental system designed for this study, and then describe the experiment design in detail. Next, the experiment results are reported, followed by the conclusions of the study.

2. Experimental system

A prototype system was implemented to support this study.

An error must be detected before it could be corrected. To eliminate the possible confounding effect of error detection, we marked errors by highlighting the related words in red and asked participants to only correct the highlighted words. In addition, error corrections by the user or the system were highlighted in blue. Unlike substitution and insertion errors, deletion errors do not appear in the recognition output. To facilitate correcting deletion errors, we created a special symbol "[...]" to indicate the occurrence of deletion errors. Participants used the traditional input devices (i.e., mouse and keyboard) to make the corrections.

In addition to the baseline condition, which provides no additional knowledge, a prototype interface was developed for each of the other three conditions separately, which incorporates different types of knowledge. Details of the four types of experiment conditions and their interfaces are described as follows.

- baseline condition (C_0): contains only the sentences to be corrected. Participants could select an error and type its correction in the correction dialog box.
- alternative hypotheses condition (C_1): provides the alternative hypotheses of each error generated by the SR system. Word-level hypotheses were used. The interface of C_1 is similar to that of C_0 except that, in the correction dialog of C_1, participants could either choose one of the hypotheses provided or type in a new correction.
- imperfect context condition (C_2): provides both the preceding and following sentences of the sentence to be corrected according to their order of appearance in the original speech. Recognition errors in the surrounding sentences were preserved, which may provide inaccurate contextual information. The way to correct the errors in C_2 is the same as that in C_0.
- perfect context condition (C_3): as an extension of C_2, provides preceding and following sentences along with the corrections of all their errors. Figure 1 shows the representation of the contextual information, with the to-be-corrected sentence highlighted in bold and placed in the center. In addition, insertion errors in the surrounding sentences were crossed out; substitution errors were both crossed out and followed with their corrections; and deletion errors were corrected by inserting the missing words in the corresponding position.

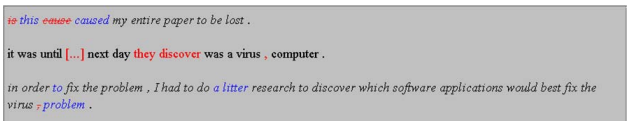

Figure 1: Contextual representation in perfect context condition (C_3)

3. Research methodology

Within-subject design was used in this experiment. Each participant experienced all the 4 conditions. To support direct pair-wise comparison between conditions, a series of randomized condition sequences have been developed, each of which contained two occurrences of each condition. Given 4 conditions, the length of each condition sequence was set to 8. The assignment of condition sequences to participants was randomized to avoid any carry-over effect of condition.

3.1. Data selection

Thirty-two sentences were randomly selected from a dictation corpus that was generated by IBM ViaVoice under high-quality conditions from the spontaneous dictation of twenty-seven speakers [12, 1]. All of the speakers were native but not professional English speakers.

The dictation corpus includes dictations on nine topic scenarios. The data were selected in the unit of sentence. To keep the topic diversity in the selected sentences, we chose two to four sentences from each scenario. To account for

the variation in language usage by different speakers, we chose at most two sentences from each speaker. Moreover, no two sentences were from the same scenario by the same speaker. All the selected sentences were prepared in four different formats in correspondence to the four conditions. Four sentences are randomly assigned to each occurrence of a condition by balancing sentence length and word error rate. The order of the sentences in each condition occurrence was randomly assigned.

3.2. Participants

Twenty-four students were recruited from a mid-sized university on the east coast of the United States for this study. They were all native English speakers and from thirteen different majors. Fourteen participants were female, and ten participants were male.

3.3. Procedure

The study was conducted in a controlled lab environment. Participants first completed training on the experimental system until they felt comfortable with the system.

During the experiment, the participants corrected recognition errors in sentences by following the assigned condition sequence. After completing the sentences in one condition, participants were asked to fill out a questionnaire on their perception of the current condition in relation to the previous condition.

The actual time spent in correcting errors for each sentence was recorded by the system automatically.

3.4. Measures

Two objective metrics were used to measure the performance of error correction:

- RWERR (Relative word error rate reduction). This metric was used to measure how well a participant performed on error correction. RWERR was defined as:
  $RWERR = \frac{WER_{original} - WER_{correction}}{WER_{original}}$
- Time (Time spent for correcting errors). This metric was used to measure how fast a participant corrected the errors. Time was measured in seconds.

Both RWERR and time were measured at the sentence level. RWERR and time for a condition were measured by averaging the results of all the sentences in that condition.

4. Results and analyses

The effects of the selected knowledge sources were evaluated with one-way repeated-measure analyses using RWERR and time as the dependent variables and condition setting as the independent variable. The descriptive statistics of the RWERR and time are reported in Table 1.

Table 1: Descriptive statistics of the RWERR and Time

Condition	RWERR		Time (seconds)	
	Mean	Std.	Mean	Std.
C_0	0.43	0.18	80.62	26.82
C_1	0.45	0.20	105.10	46.42
C_2	0.56	0.17	95.68	38.56
C_3	0.61	0.16	93.58	35.22

4.1. Analysis on RWERR

The ANOVA result showed that knowledge sources played a significant role in error correction, $F(3, 69) = 7.659, p < 0.005$. To examine the relative merit of each condition, multi-pair contrast comparison was conducted, and the results are shown in Table 2.

Table 2: Multi-pair comparison results

Pair		RWERR		Time	
C_i	C_j	Mean $(C_i - C_j)$	Sig.	Mean $(C_i - C_j)$	Sig.
0	1	-0.02	0.785	-24.48	0.017
0	2	-0.13	0.018	-15.06	0.023
0	3	-0.18	0.000	-12.96	0.091
1	2	-0.12	0.012	9.41	0.243
1	3	-0.16	0.002	11.52	0.090
2	3	-0.05	0.285	2.10	0.797

Compared with the baseline condition (C_0), participants were able to correct more errors in C_2 ($p < 0.05$) and C_3 ($p < 0.01$) when contextual information was provided. Perfect context (C_3) ($mean = 0.61$) provided more useful information than imperfect context (C_2) ($mean = 0.56$), but the improvement was not significant ($p = n.s.$). Compared with C_0, alternative list (C_1) did not lead to improvements on RWERR ($p = n.s.$).

4.2. Analysis on time

The ANOVA result showed that the condition of different knowledge sources had significant impact on the time for error correction, $F(3, 69) = 3.456, p < 0.05$. The results of multi-pair contrast comparison of different conditions are shown in Table 2.

Given alternative list (C_1), participants spent significantly more time than the baseline (C_0) ($p < 0.05$). The same

time pattern also occurred to the imperfect context (C_2) ($p < 0.05$). Although participants spent more time under the perfect context (C_3) than baseline condition, the difference was not significant ($p = n.s.$).

5. Conclusions

To break the barrier for the wide adoption of SR technology, we examined various knowledge sources for the third-party human error correction via an empirical user study. Three types of knowledge sources were selected for comparison. The findings of this study can be summarized as follows:

- Perfect contextual information provided the most useful knowledge among all the knowledge sources being tested. Humans can achieve significant word error rate reduction without significantly increasing the time. This suggests that the feedback to user correction is important to error correction.

- Imperfect contextual information is helpful to improving the accuracy of human error correction. However, that is achieved at the cost of time.

- Alternative hypotheses do not seem to be an ideal knowledge source for human error correction. The alternative words consume significantly more time, without improving word error rate. One possible explanation is that alternative words may sometimes be misleading. Another explanation is that the lengthy list of alternative words may have caused cognitive overload on human users.

The findings of this research have implications to the design and development of systems in support of human error correction. An adaptive error correction system that could provide more accurate context by dynamically learning from user corrections will be promising.

This study raises several issues that are worth to explore in the next step. For example, with the questionnaire data collected from this study, we would analyze participants' subjective perception of different knowledge sources. We would also investigate the effect of sentence difficulty on the performance of human error correction.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant # 0328391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

7. References

[1] A. Sears, J. Feng, K. Oseitutu, and C.-M. Karat, "Hands-free, speech-based navigation during dictation: Difficulties, consequences, and solutions," *Hum. Comput. Interact.*, vol. 18, no. 3, pp. 229–257, 2003.

[2] S. Dusan and L.R. Rabiner, "On integrating insights from human speech perception into automatic speech recognition," in *Proc. Interspeech*, 2005, pp. 1233–1236.

[3] A.E. McNair and A. Waibel, "Improving recognizer acceptance through robust, natural speech repair," in *Proc. ICSLP*, 1994, pp. 1299–1302.

[4] H. Lieberman, A. Faaborg, W. Daher, and J. Espinosa, "How to wreck a nice beach you sing calm incense," in *Proc. IUI*, 2005, pp. 278–280.

[5] J. Ogata and M. Goto, "Speech repair: Quick error correction just by using selection operation for speech input interfaces," in *Proc. Interspeech*, 2005, pp. 133–136.

[6] S. Oviatt and R. VanGent, "Error resolution during multimodal human-computer interaction," in *Proc. ICSLP*, 1996, vol. 1, pp. 204–207.

[7] B. Suhm, B. Myers, and A. Waibel, "Multimodal error correction for speech user interfaces," *ACM Trans. Comput.-Hum. Interact.*, vol. 8, no. 1, pp. 60–98, 2001.

[8] E. Brill, R. Florian, J. Henderson, and L. Mangu, "Beyond n-grams: Can linguistic sophistication improve language modeling," in *Proc. COLING-ACL*, 1998, vol. 1, pp. 186–190.

[9] L. Mangu and M. Padmanabhan, "Error corrective mechanisms for speech recognition," in *Proc. ICASSP*, 2001, vol. 1, pp. 29–32.

[10] A. Sarma and D. Palmer, "Context-based speech recognition error detection and correction," in *Proc. HLT-NAACL*, 2004, pp. 85–88.

[11] D. Yu, M.-Y. Hwang, P. Mau, A. Acero, and L. Deng, "Unsupervised learning from user's error correction in speech dictation," in *Proc. ICSLP*, 2004, pp. 1969–1972.

[12] J. Feng and A. Sears, "Using confidence scores to improve hands-free speech based navigation in continuous dictation systems," *ACM Trans. Comput.-Hum. Interact.*, vol. 11, no. 4, pp. 329–356, 2004.