Are Some Speech Recognition Errors Easier to Detect than Others?

Yongmei Shi
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250
yshi1@umbc.edu

Lina Zhou
Department of Information Systems
University of Maryland, Baltimore County
Baltimore, MD 21250
zhoul@umbc.edu

Abstract

This study investigates whether some speech recognition (SR) errors are easier to detect and what patterns can be identified from those errors. Specifically, SR errors were examined from both non-linguistic and linguistic perspectives. The analyses of non-linguistic properties revealed that high error ratios and consecutive errors lowered the ease of error detection. The analyses of linguistic properties showed that ease of error detection was associated with changing parts-of-speech of reference words in SR errors. Additionally, syntactic relations themselves and the change of syntactic relations had impact on the ease of error detection.

1 Introduction

Speech recognition (SR) errors remain as one of the main impediment factors to the wide adoption of speech technology, especially for continuous large-vocabulary SR applications. As a result, lowering word error rate is the focus of SR research which can benefit from analyzing SR errors. SR errors have been examined from various perspectives: linguistic regularity of errors (McKoskey and Boley, 2000), the relationships between linguistic factors and SR performance (Greenberg and Chang, 2000), and the associations of prosodic features with SR errors (Hirschberg et al., 2004). However, little is understood about patterns of errors with regard to ease of detection.

Analyzing SR errors can be helpful to error detection. Skantze and Edlund (2004) conducted a user study to evaluate the effects of various features on error detection. Our study is different in that it investigates the relationships between the characteristics of SR errors and their ease of detection through an empirical user study. Given two SR systems with the same word error rates, the output of one system could be more useful if its errors are easier to detect than the other. Accordingly, SR and its error detection research could focus on addressing difficult errors by developing automatic solutions or by providing decision support to manual error detection.

2 Experiment

A laboratory experiment was carried out to evaluate humans' performance in SR error detection.

2.1 Experimental Data

Speech transcripts were extracted from a dictation corpus on daily correspondence in office environment generated using IBM ViaVoice under high-quality condition (Zhou et al., 2006).

Eight paragraphs were randomly selected from the transcripts of two task scenarios based on two criteria: recognition accuracy and paragraph length (measured by # of words). Specifically, the overall recognition accuracy (84%) and the length of a medium-sized paragraph (90 words) of the corpus were used as references.

The selected paragraphs consist of 36 sentences. Sentence lengths range from 9 to 38 words, with an average of 20. For error detection, SR output instead of references is a better base for computing

Proceedings of NAACL HLT 2007, Companion Volume, pages 157–160,
Rochester, NY, April 2007.

error rates because SR output but not reference transcripts are accessible during error detection. This may result in fewer number of deletion errors because when one SR error maps to several reference words, it is counted as one substitution error. Based on this method, there are totally 140 errors in the selected data: 104 substitution, 31 insertion, and 5 deletion errors. The error ratio, defined as the ratio of the number of errors to the number of words in output sentence, ranges from 4.76% to 61.54%.

2.2 Task and Procedure

Participants were required to read error annotation schema and sample transcripts prior to the experiment, and could attend the experiment only after they passed the test on their knowledge of the schema and SR output.

Each participant was asked to detect errors in all eight paragraphs. All sentences in the same paragraphs were presented all at once. The paragraphs were presented with different methods, including three with no additional information, three with alternative hypotheses, and two with both dictation scenario and alternative hypotheses. The sequence of paragraphs and their presentation methods were randomized for each participant.

Ten participants from a mid-sized university in the U.S. completed the study. They were all native speakers and none of them was professional editor.

3 Analysis and Discussion

In this section, we analyze the relationship between characteristics of SR errors and ease of error detection. We characterize errors with non-linguistic and linguistic properties and further break down the latter into parts-of-speech and syntactic relations.

3.1 Ease of Error Detection

The ease of detecting an error was defined as the number of participants who successfully detected the error. When computing the ease of error detection, we merged all the data by ignoring the presentation methods. The decision was made because a repeated measure ANOVA of recall failed to yield a significant effect of presentation methods ($p = n.s.$). The recall was selected because it measures the percentage of actual errors being detected and the focal interest of this study was actual errors. The average recalls of error detection of three presentation methods were very close, ranging from 72% to 75%.

The ease values fell between 0 and 10, with 0 being the least ease when all participants missed the error and 10 being the most ease when everyone found the error. To improve the power of statistical analyses, errors were separated into 3 groups using equal-height binning based on their ease values, namely 1 for low, 2 for medium, and 3 for high (see Table 1). The overall average ease value was 2.15.

Level of Ease	Ease Values	# of Errors
Low (1)	0-5	39
Medium (2)	6-8	41
High (3)	9-10	60

Table 1: Grouping of ease values

3.2 Non-linguistic Error Properties

Three non-linguistic error properties, including error ratio, word error type, and error sequence (in isolation or next to other errors) were selected to examine their relationships with ease of error detection.

Two-tailed correlation analyses of error ratio and ease of detection showed that the Pearson correlation coefficient was -0.477 ($p < 0.01$), which suggests that it is easier to detect errors in sentences with lower error ratios.

One way ANOVA failed to yield a significant effect of error type on ease of detection ($p = n.s.$). Nonetheless, mean comparisons showed that insertion errors were less easy to detect ($mean = 2.03$) than deletion errors ($mean = 2.20$) and substitution errors ($mean = 2.18$). Users may have difficulty in judging extra words.

Among the 140 errors, about half of them (i.e., 71) were next to some other errors. One way ANOVA revealed a significant effect of error sequence on ease of detection, $p < 0.05$. Specifically, isolated errors ($mean = 2.33$) are easier to detect than consecutive errors ($mean = 1.97$).

3.3 Part-Of-Speech(POS)

SR output and reference transcripts were analyzed using Brill's part-of-speech tagger (Brill, 1995). To alleviate data sparsity problem, we adopted second-

level tags such as NN and VB. The POSes of SR errors as well as POS change patterns between reference words and SR errors were analyzed.

Table 2 reports the average eases of detection for difference POSes on all the errors, substitution errors only, and insertion errors only. Deletion errors were not included because they did not appear in SR output. Only those POSes with frequency of at least 10 in all the errors were selected.

POS	All	Substitution	Insertion
NN	2.03	2.00	2.25
VB	2.30	2.41	1.67
CC	2.21	2.38	1.83
IN	2.22	2.27	2.00
DT	1.80	2.25	1.50

Table 2: Ease of detection for different POSes

It was easier to detect verbs that were misplaced than verbs that were inserted mistakenly ($p < 0.1$ in one-tailed results). This is because an additional verb may change syntactic and semantic structures of entire sentence. Similar patterns held for both CC and DT ($p < 0.1$ in one-tailed results). The less ease in detecting DT and CC when they were inserted than replaced is due in part to the fact that they play significant syntactic roles in constructing a grammatical sentence. Further, ease of detecting DT was lower than the average ease of all errors ($p < 0.1$ in one-tailed results).

Only substitution errors were applicable in POS change analysis. POS change was set to 'Y' when the POSes of an SR error and its corresponding reference word were different, and 'N' when otherwise. This resulted in 69 Ys and 35 Ns. One way ANOVA results yielded a significant effect of POS change on ease of detection ($p < 0.05$). Specifically, it was easier to detect errors that had different POSes ($mean = 2.32$) from their references than those that shared the same POSes ($mean = 1.91$). This is partly due to the requirements of semantic and even discourse information in detecting errors from the same POSes.

3.4 Syntactic Relations

Both SR output and reference transcripts were parsed using minipar (Lin, 1998), a principle-based parser that can generate a dependency parse tree for each sentence. The dependency relations between SR errors and other words in the same sentence were extracted as the syntactic relations of SR errors. The same kinds of relations were also extracted for corresponding reference words.

Three types of properties of syntactic relations were analyzed, including the number of syntactic relations, syntactic relation change, and errors' patterns of syntactic relations.

Table 3 reports descriptive statistics of ease of detection for SR errors with varying numbers of syntactic relations. The average number of syntactic relations for all errors was 1.64. Analysis results showed that it was easier to detect errors with no syntactic relations than those with one relation ($p < 0.05$). The analysis of correlation between the number of syntactic relations and the ease of detection yielded a very small Pearson correlation coefficient ($p = n.s.$). They suggest that errors that do not fit into a sentence are easy to detect. However, increasing the number of syntactic relations does not lower the ease of detection.

# of Syntactic Relations	Mean	Std Deviation	Frequency
0	2.40	0.695	35
1	1.98	0.883	51
2	2.21	0.918	19
3	2.00	0.791	17
> 3	2.22	0.808	18

Table 3: Ease of detection for numbers of relations

Same as POS change, only substitution errors were considered in syntactic relation change analysis, and the values of the syntactic changes were set similarly. By dividing the syntactic relations into head and modifier according to whether the words served as heads in the relations, we also derived syntactic changes for head and modifier relations, respectively.

Two-way ANOVA analyses of head and modifier syntactic relation changes yielded a significant interaction effect ($p < 0.05$). A post-hoc analysis revealed that, when the modifier syntactic relations were the same, it was easier to detect errors that did not cause the change of head syntactic relations than

those causing such changes ($p < 0.05$).

Table 4 reports descriptive statistics of ease of detection in terms of syntactic relations of SR errors that occurred at least 5 times. Two relations were presented in the "Syntactic Relations" column. The first one is the relation in which errors played the head role, and the second one is the relation that errors served as a modifier. "None" indicates no such relations exist.

Syntactic Relations	Mean	Std Deviation	Frequency
none none	2.40	0.695	35
none subj	2.70	0.675	10
none det	1.78	0.833	9
none punc	2.00	0.926	8
none nn	2.00	1.000	7
none pcomp-n	2.33	1.033	6
mod pcomp-n	1.20	0.447	5
none obj	1.80	0.837	5

Table 4: Ease of detection for syntactic relations

It is shown in Table 4 that it is easier to detect if an error is the subject of a verb (subj). A typical example is the "summary" in sentence "summary will have to make my travel arrangement ... ". All the participants successfully detected "summary" as an error. In contrast, "mod pcomp-n" was difficult to detect. Manual scrutinizing of the data showed that such errors were nouns that both have some other words/phrases as modifier (mod) and are nominal complements of a preposition (pcomp-n). For example, for "transaction" in sentence "I'm particularly interested in signal transaction in ... ", 80% participants failed to detect the error. It requires domain knowledge to determine the error.

4 Conclusion and Future Work

This study revealed that both high error ratio and consecutive errors increased the difficulty of error detection, which highlights the importance of SR performance. In addition, it was easier to detect SR errors when they had different POSes from corresponding reference words. Further, SR errors lacking syntactic relations were easy to detect, and changes in syntactic relations of reference words in SR errors had impact on the ease of error detection. The extracted patterns could advance SR and automatic error detection research by accounting for the ease of error detection. They could also guide the development of support systems for manual SR error correction.

This study brings up many interesting issues for future study. We plan to replicate the study with automatic error detection experiment. Additional experiments would be conducted on a larger data set to extract more robust patterns.

Acknowledgement

This work was supported by the National Science Foundation (NSF) 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 NSF.

References

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565.

Steven Greenberg and Shuangyu Chang. 2000. Linguistic dissection of switchboard-corpus automatic speech recognition systems. In *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*.

Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43:155–175.

Dekang Lin. 1998. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

David McKoskey and Daniel Boley. 2000. Error analysis of automatic speech recognition using principal direction divisive partitioning. In *Proceedings of ECML*, pages 263–270.

Gabriel Skantze and Jens Edlund. 2004. Early error detection on word level. In *Proceedings of Robustness*.

Lina Zhou, Yongmei Shi, Dongsong Zhang, and Andrew Sears. 2006. Discovering cues to error detection in speech recogntion output: A user-centered approach. *Journal of Management of Information Systems*, 22(4):237–270.