Modeling the Effect of Speech Recognition Errors on Text Understandability for People who are Deaf or Hard of Hearing

R·I·T

Sushant Kafle, Matt Huenerfauth
Rochester Institute of Technology (RIT)
sushant@mail.rit.edu, matt.huenerfauth@rit.edu

Abstract

This research investigated the **impact of different inaccurate transcriptions** from an Automatic Speech Recognition (ASR) system on the **understandability of captions** for people who are Deaf or Hard-of-Hearing (DHH). Through a user study with **30 DHH users**, the effect of the presence of an error on a text's understandability for DHH users was studied, and **several linguistic features** were investigated to model this **relation accurately.**

Background

ASR system based captioning is an interesting prospect.

~40-50 X CHEAPER THAN *CART*	REALTIME CLOUD-POWERED ONLINE	SCALABLE AUTOMATIC AVAILABLE

But, they are still not fully accurate! The errors in the output tend to be confusing to the readers.

ASR OUTPUT
The ***climb*** meeting has been ***mood*** to Tuesday.

Minimum Bayes Risk (MBR) decoders allow ASR decoding process to **incorporate a loss function.** This loss function describes a task performance metric; one such popularly used metric is **Word Error Rate (WER)**.

WER might not always be an ideal metric to evaluate the output of an ASR system.

"The meeting has been moved to Tuesday."

ASR OUTPUT #1
The ***meet in*** has been ***move*** to Tuesday. ✓

ASR OUTPUT #2
The ***eating*** has been moved to Tuesday. ✗

ASR Output #1 has a greater number of errors than Output #2 but the understandability of ASR Output #1 may be higher than that of ASR Output #2.

Research Question

Can we **learn a custom loss function** that optimizes the comprehensibility of ASR output for DHH users?

- **Unlike WER, our loss function may provide a better measure of text understandability for this group of users.**

Preliminary User Study

Designed **20 short English text passages** (average length 177 words).

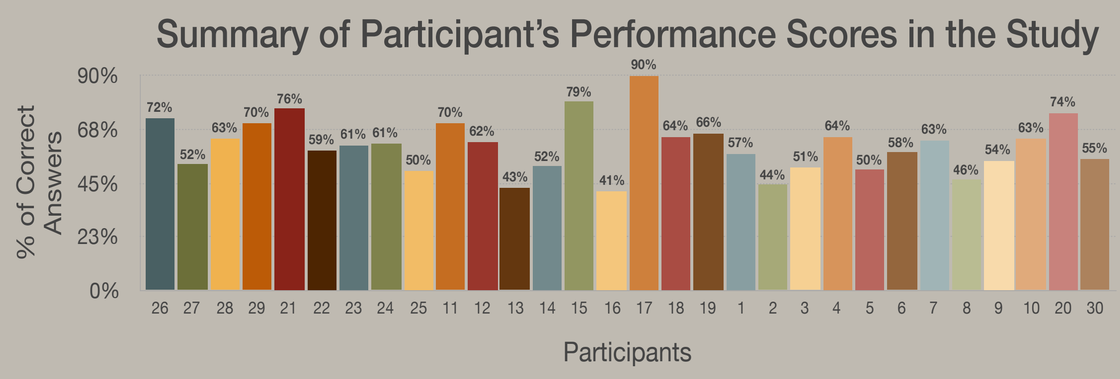
For each passage, **three questions** were designed. A question would only require inference from one sentence in the passage.

Comprehension passages for the study were generated by **inserting an ASR recognition error** in the sentence in the passage containing the answer to a question.

Participants

Recruited 30 **DHH participants** who were associate degree students at the National Technical Institute for the Deaf (**NTID**) at Rochester Institute of Technology (RIT).

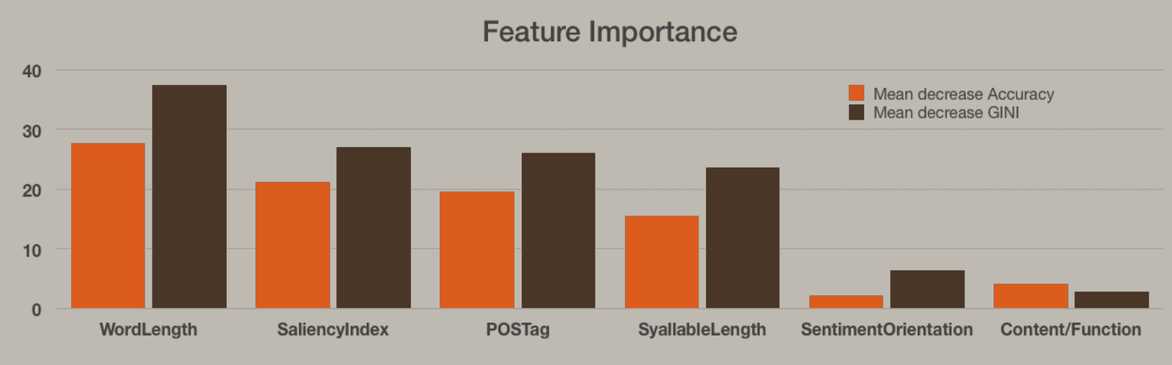
Each participant was given **10 different comprehension passages** to read. Each passage contained three multiple choice questions that needed to be answered in a time period of 70 minutes.



Modeling the Loss Function

We explored 6 features for modeling: **Word Length, Saliency Index, Part of Speech tag, Syllable Length, Sentiment Orientation and Content or Function word.**

Through features selection, three most contributing features were selected: **Word Length, Saliency Index, Part of Speech Tag.**



Our final Random Forest based model produced an accuracy of 62.04% (sigma = 4.41) on a 5-fold cross validation testing.

This material is based upon work supported by the National Technical Institute for the Deaf.