Google Research | UCL | MGH INSTITUTE OF HEALTH PROFESSIONS

LARGE LANGUAGE MODELS AS A PROXY FOR HUMAN EVALUATION IN ASSESSING THE COMPREHENSIBILITY OF DISORDERED SPEECH TRANSCRIPTION

ICASSP 2024 KOREA

Katrin Tomanek[1], Jimmy Tobin[1], Subhashini Venugopalan[1], Richard J.N. Cave[1,2], Katie Seaver[1,3], Jordan Green[1,3], Rus Heywood[1],
[1]Google Research, [2]Language and Cognition, UCL, [3]MGH Institute of Health Professions,
{katrintomanek, jtobin}@google.com

Introduction

WER treats all errors the same

- Word Accuracy and Word Error Rate (WER) are measures of syntactic accuracy and errors of an automatic speech recognition (ASR) model, but they don't measure comprehensibility.
- On atypical speech (e.g. disordered speech), WER is often >20 and sometimes >60 for certain etiologies and severities.
- Individuals with disordered speech may still benefit from an ASR model with relatively high WER, provided that meaning is preserved.
- We aim to create a system that will automatically assess the ability of an ASR model to convey the user's intended message.

Error Type	Predicted Transcript	Actual Transcript	Word Acc.
Deletion	Come right back _	Come right back please	0.75
	I have a *head*_	I have a headache	0.75
Contraction	*I'm* a bit overwhelmed	I am a bit overwhelmed.	0.60
Normalization	play *Beyoncé*	play Beyonce	0.50
	Okay *9:30 five*	Okay, nine thirty five.	0.50
Proper Noun	Here are TV shows by Hugh *Griffiths*	Here are TV shows by Hugh Griffith	0.86
	First do you know how the story ends	Faust, do you know how the story ends?	0.88
Repetition	What *are you* are you trying to say to me	What are you trying to say to me?	0.75

Method

Classifiers Predicting Meaning Preservation

- Logistic Regression model on BERTScore+WER
- Logistic Regression model on cosine similarity of sentence embeddings (SentT5, 11b)
- Prompt-tuned LLMs: Flan-T5-XXL (11b) and Flan-cont-PaLM (62b)

Example 1
Input Sequence
```
Ground truth: {no no there are fifteen hundred total}.
Transcription: {no no there are 50 energy total}.
Transcript preserves the meaning of the ground truth: {
```
Target Sequence
```
no}
```

Example 2
Input Sequence
```
Ground truth: {He's huggable and lovable and a good with people.}.
Transcription: {He's huggable and laughable and a good with people}.
Transcript preserves the meaning of the ground truth: {
```
Target Sequence
```
yes}
```

Model Deployment Decisions

- Personalized ASR models [2] need to be quality checked (usually manually by Speech & Language Pathologists) before deploying to users.
- Word Accuracy does not distinguish well between high and low quality models.
- LATTEScore (LLMs to Assess TranscripTion Error Score) gives better model quality assessment.

$$\text{LATTEScore} = \frac{\text{\# Predicted Meaning Preserved}}{\text{\# Total Examples}} \times 100$$

Dataset

Transcript Comprehensibility Dataset

- 4731 tuples of ground truth (from Euphonia corpus [1]) and (erroneous) ASR transcript along with human-rated meaning preservation label
- Significant inter-annotator agreement when assessing meaning preservation, Cohen's κ = 0.7

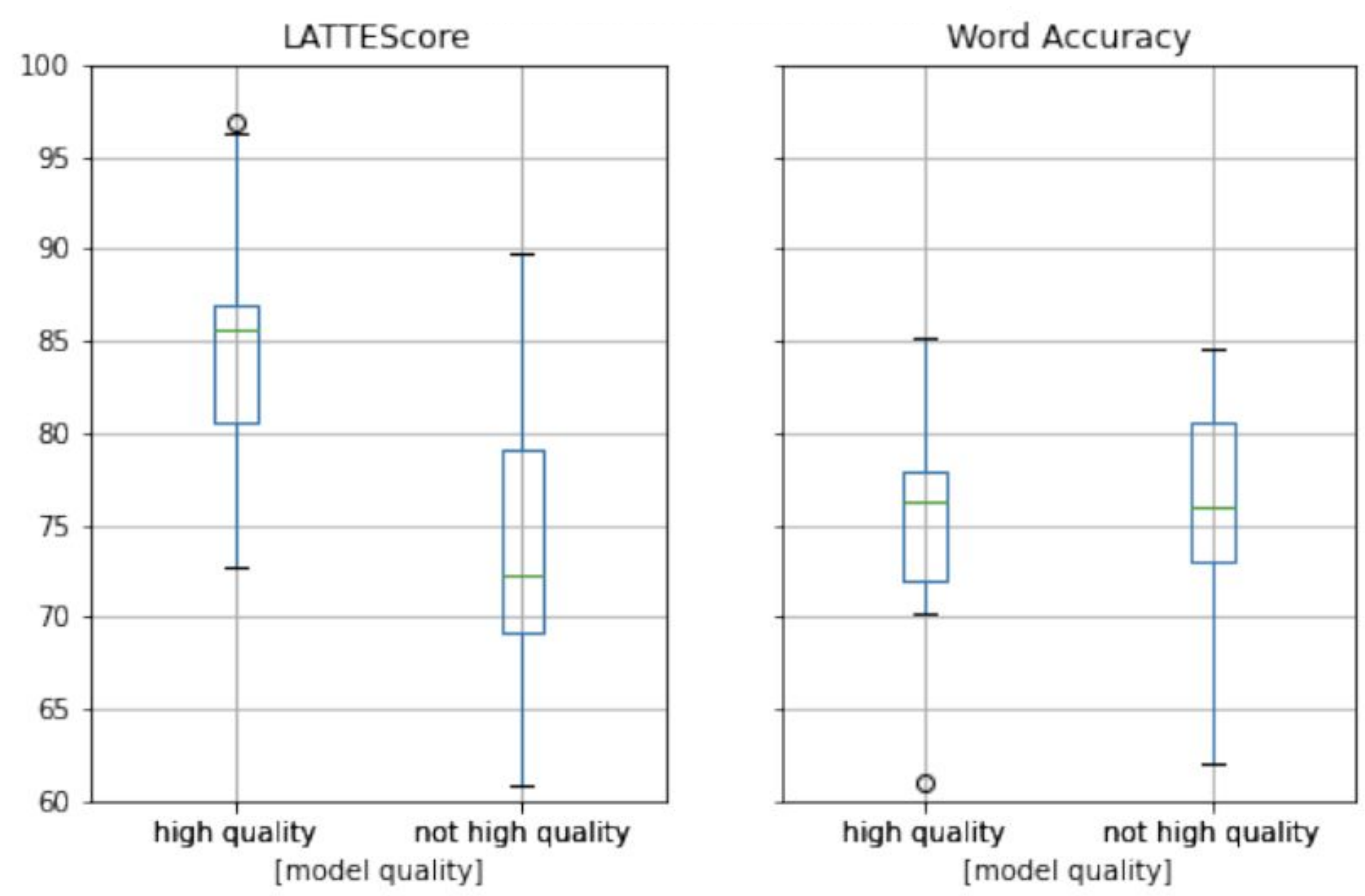
Error Severity	Meaning Preserved	Description	# Examples (%)	Example
0	yes	Meaning is completely preserved	900 (19%)	**G:** I would be fascinated to know your answers. **T:** I *will* be fascinated to know your answers.
1	yes	Some errors, but meaning is mostly preserved.	1145 (24%)	**G:** Yeah I have one basically every day. **T:** Yeah I have *I'm* basically every day.
2	no	Major errors, significant loss of intended meaning.	2686 (57%)	**G:** How large is that file? **T:** How large is a *funnel*?

Results

Classifiers Performance (ROC-AUC)

Approach	Full set (940)	Sliced by severity SEV (467)	MOD (302)	MILD (149)
BERTScore+WER	0.791	0.753	0.791	0.856
SentT5 Emb Sim	0.857	0.813	0.879	0.899
Flan-T5 XXL	0.878	0.836	0.923	0.890
Flan-cont-PaLM	**0.900**	**0.863**	**0.944**	**0.903**

LATTEScore to Distinguish Model Quality



Conclusion

- We propose a new approach to assess ASR model performance based on comprehensibility rather than syntax preservation.
- LLM-based classifiers perform very well in this task and outperform other classifiers.
- LATTEScore better predicts how useful a model will be to the end user.
- Beyond speech impairment, LLM-based classifiers can be useful for low-resource languages where human evaluation is challenging.
- Future work will explore using multi-lingual LLMs for zero-shot performance in other languages.

References
[1] MacDonald et al. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. Interspeech 2021
[2] Green et al. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. Interspeech 2021