Word, Syllable and Phoneme Based Metrics Do Not Correlate with Human Performance in ASR-Mediated Tasks

Anne H. Schneider[1], Johannes Hellrich[2], and Saturnino Luz[3]

1 School of Computing Science, University of Aberdeen, Scotland, a.schneider@abdn.ac.uk
2 JULIE Lab, Friedrich Schiller University Jena, Germany, johannes.hellrich@uni-jena.de
3 School of Computer Science and Statistics, Trinity College Dublin, Ireland, luzs@scss.tcd.ie

Abstract

Human conversation is full of errors, ill-formed sentences, false starts and hesitations. This complicates matters for NLP components, whereas humans are well able to filter out the relevant content. The way WER is calculated does not take into account the ability of a human listener to compensate for small mistakes. We collected sixteen ASR mediated dialogues using a map task scenario. The material was assessed extrinsically through measures like time to task completion and intrinsically through WER. We discovered a lack of correlation between extrinsic and intrinsic measures. However, more forgiving metrics based on phonemes and syllables do not seem to correlate better with human performance.

Experimental Setting

- 16 participants (adult German native speakers) performed the map task in groups of two
- 4 maps with eight to twelve landmarks, labelled in German
- Each participant executed a standard initial model training with the ASR system
- Participants were separated in two different rooms with a computer and the respective maps
- ASR as input for instructor's directions into a chat field (no further editing possible), follower could talk back freely
- Experiment was audio recorded and transcribed
- After the follower reached the goal, roles were swapped and the experiment was repeated with another map

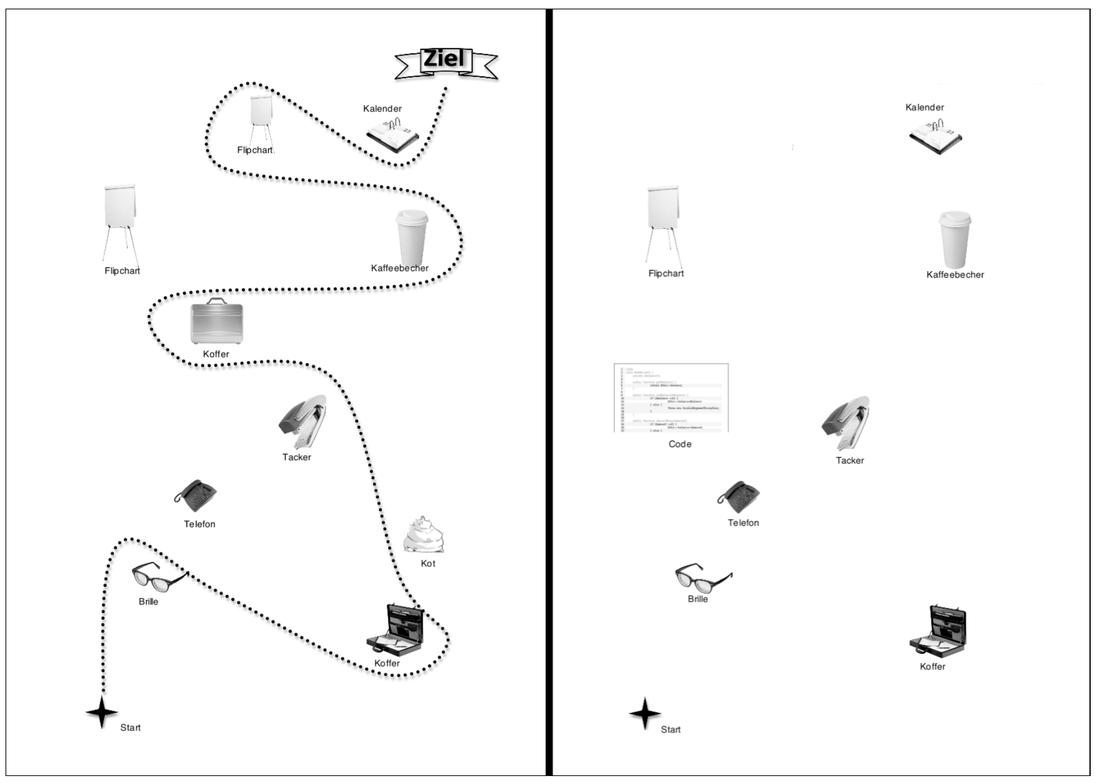

Figure 1: Example of instructor and follower map

reference word	ASR output
Compound mistakes	
geschlossenen Koffer	Geschlossenenkoffer
Richtung Saloon	Richtungssalon
Richtung Katze	Richtungskatze
Richtung Kaktus	Richtungskaktus
Decomposition mistakes	
Indianerzelt	Indianer Zelt
Linksbogen	links Bogen
Vorderbeinen	vorder Beinen

Table 1: Selection of frequent ASR errors.

Data Analysis

- WER was calculated and several extrinsic factors were measured, e.g. time to task completion and task success (correctness of path)
- Analysing the transcripts showed that many errors were of low relevance for human understanding, e.g. compounding mistakes

Alternatives to WER

- Character error rate, one version considering word boundaries (CER) and one not (CER2)
- Sound error rate, based on phoneme-classes determined by a German Soundex variant, considering word boundaries (SER) or not (SER2)
- Syllable error rate (SylER), similar to WER, based on automatically segmented syllables

	SylER		CER		CER2		SER		SER2	
Time	−0.09	(0.72)	−0.18	(0.49)	−0.20	(0.45)	0.02	(0.94)	−0.19	(0.43)
Task	−0.26	(0.33)	−0.30	(0.24)	−0.28	(0.28)	−0.17	(0.53)	−0.27	(0.26)
WER	0.95	(< 0.01)	0.85	(< 0.01)	0.84	(< 0.01)	0.99	(< 0.01)	0.86	(< 0.01)

Table 2: Overview of the Pearson coefficient for the correlation of the intrinsic evaluation methods (SylER, CER, CER2, SER, and SER2) with time to task completion (time) and task success (task). P-values in brackets.

From our results we see that WER is not a good indicator of time to task completion or route drawing accuracy.
However, we could not show a significantly higher correlation with the extrinsic results for the three alternative measures, i.e. syllable, character and sound error rate

This research is supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie)