WER we are
and WER we think we are

Piotr Szymański, Piotr Żelasko, Mikołaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banasiak, Łukasz Augustyniak, Jan Mizgajski, Yishay Carmiel

Word Error Rate (WER)

WER is a standard metric used in ASR evaluation.

$$WER = \frac{Sub + Ins + Del}{N_{ref}}$$

Example

Reference: *so she cannot eat seafood*

Hypothesis: *so she get the easy food*

$$WER = \frac{3 + 1 + 0}{5} = 80\%$$

Word Error Rate (WER)

WER is a standard metric used in ASR evaluation.

$$WER = \frac{Sub + Ins + Del}{N_{ref}}$$

Example

Reference: *so she cannot eat seafood*

Hypothesis: *so she get the easy food*

$$WER = \frac{3 + 1 + 0}{5} = 80\%$$

Measuring semantics?
Problem for another time!

What level of WER would you expect in ASR text?

What level of WER would you expect in ASR text?

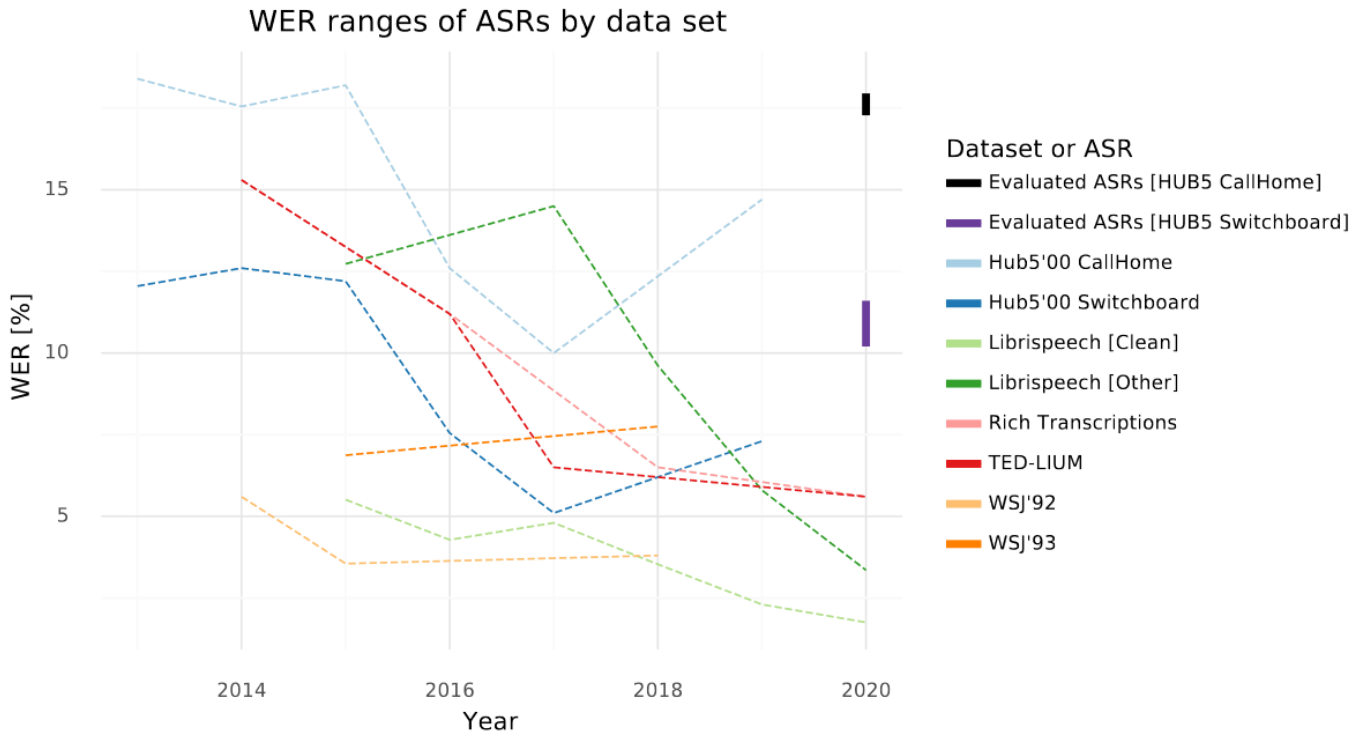
We asked researchers at ACL 2019:

- Most said: *something between 2-5 %*
- Some said: *it's solved right? close to 0%*
- Less than ten researchers said:

  it depends on the domain and problem and can be quite high

WER can be low in Academic SOTA

- Librispeech at 1.7-2%
- WSJ at 3-7%
- TED Talks at 5%
- HUB5 Switchboard 5%
- HUB5 Callhome 9%



Source: Papers with code, wer_we_are

Internal Conversational Call Corpus (CCC)

- 50 call center conversations
- recorded at 8~kHZ using standard modern telephony quality,
- 8.5 hours of audio, including 2.2 hours of speech,
- 1595 agent and 1361 customer utterances,
- average utterance consists of 10 words,

Experimental setup

- 3 different state-of-the-art commercial ASR solutions representative in terms of what is available on the market,
- when a given ASR vendor offered such an option, we used a telephonic speech model,
- we report the WER of evaluated ASRs on our CCC dataset and on the HUB5'00 Switchboard and CallHome evaluation subset to allow comparison to the state of the art on publicly available data.

WER is very high in commercial systems

Academic SOTA HUB'05

- Switchboard 5%
- Callhome 9%

Commercial ASRs

- Switchboard 10-12%
- Callhome 16-19%

Commercial ASR WER is high in spontaneous phone conversations

ASR	CCC	SWBD	CallHome
ASR 1	17.9	11.62	17.69
ASR 2	19.2	11.45	18.6
ASR 3	16.5	10.2	15.85

Table 1: WER [%] comparison on benchmarks

Commercial ASR WER is high across domains in spontaneous phone conversations

	ASR 1	ASR 2	ASR 3
Booking	21.19	22.16	20.95
Finance	16.82	18.46	15.83
Insurance 1	18.01	20.20	17.84
Insurance 2	15.25	17.11	13.73
Telecomm.	19.75	23.31	17.62
Agent	16.97	17.83	16.49
Customer	17.87	20.99	16.48

Table 2: Internal benchmark WER [%]

WER does the difference come from?

Productization:

- Strike a balance between **accuracy**, **latency** and **scalability**.
- **Streaming (on-line)** models required (e.g., no BLSTM).
- Language model has to be both **general** to account for variety of users' utterances and **specific** to include domain-related language.

Research:

- **Minimize WER at all cost**, rarely report real-time-factor.
- Usually, **no streaming constraint** is considered.
- Language model **can overfit** the corpus-specific domain

A call to action

- preparing new audio and transcript datasets with rich annotations including: POS tags, dependency structure, entity spans, sentiment annotations, question and answer pairs, dialogue and discourse annotation and augmenting existing corpora with NLP annotations

A call to action

- developing methods and tools for improving ASR acoustic and language model training and adapting NLP models and pipelines to conversational applications

A call to action

- developing tools that allow collecting conversational speech and recording real, spontaneous conversations that could be crowdsourced and released openly like Librispeech

A call to action

- organizing crowd-sourcing collection efforts similar to Mozilla CommonVoice where users could donate their phone calls and/or transcriptions

A call to action

- designing open solutions that can serve as common benchmarks for joint ASR+NLP tasks to monitor the progress of the field
- constructing new ASR quality measures, based on more richly annotated data, to better evaluate various aspects of transcription quality.

Thank you for your attention!

niedakh@gmail.com
pzelasko@jhu.edu