Open Challenge for Correcting Errors of Speech Recognition Systems

Marek Kubis[1] Zygmunt Vetulani[1] Mikołaj Wypych[2]
Tomasz Ziętkiewicz[1,2]

[1]Adam Mickiewicz University

[2]Samsung Poland R&D Institute

19 maja 2019

Outline

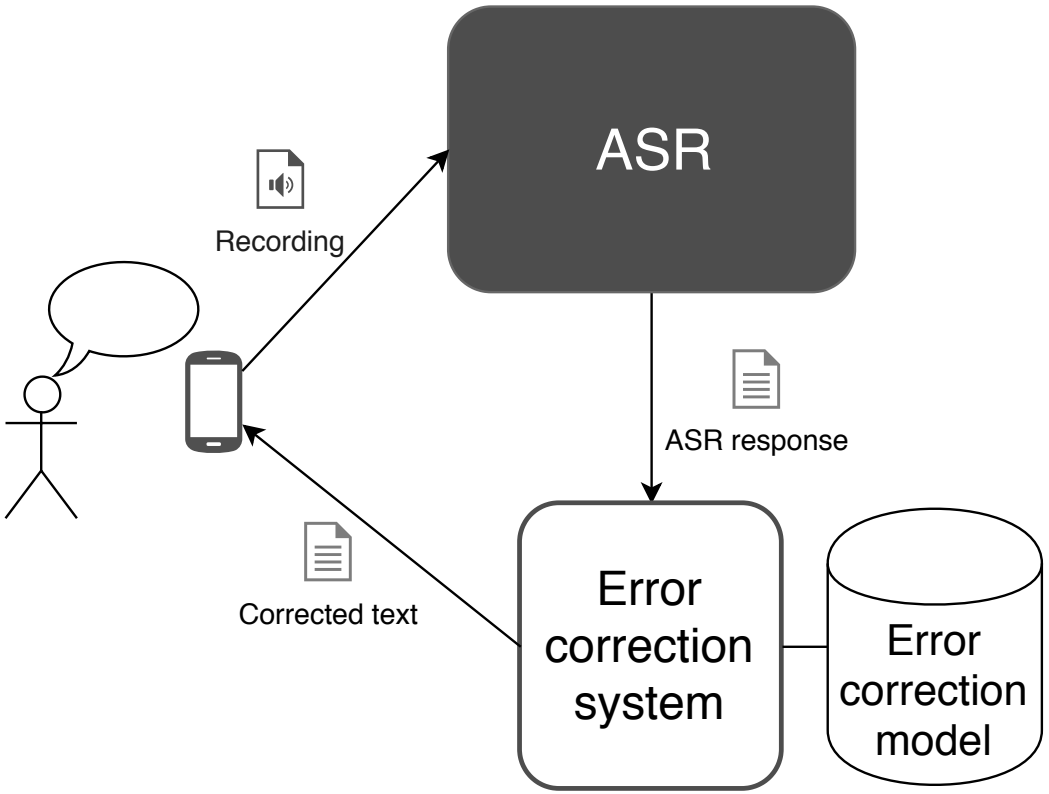
Goal

Investigate the methods of improving the performance of speech recognition systems on the basis of previously made errors.

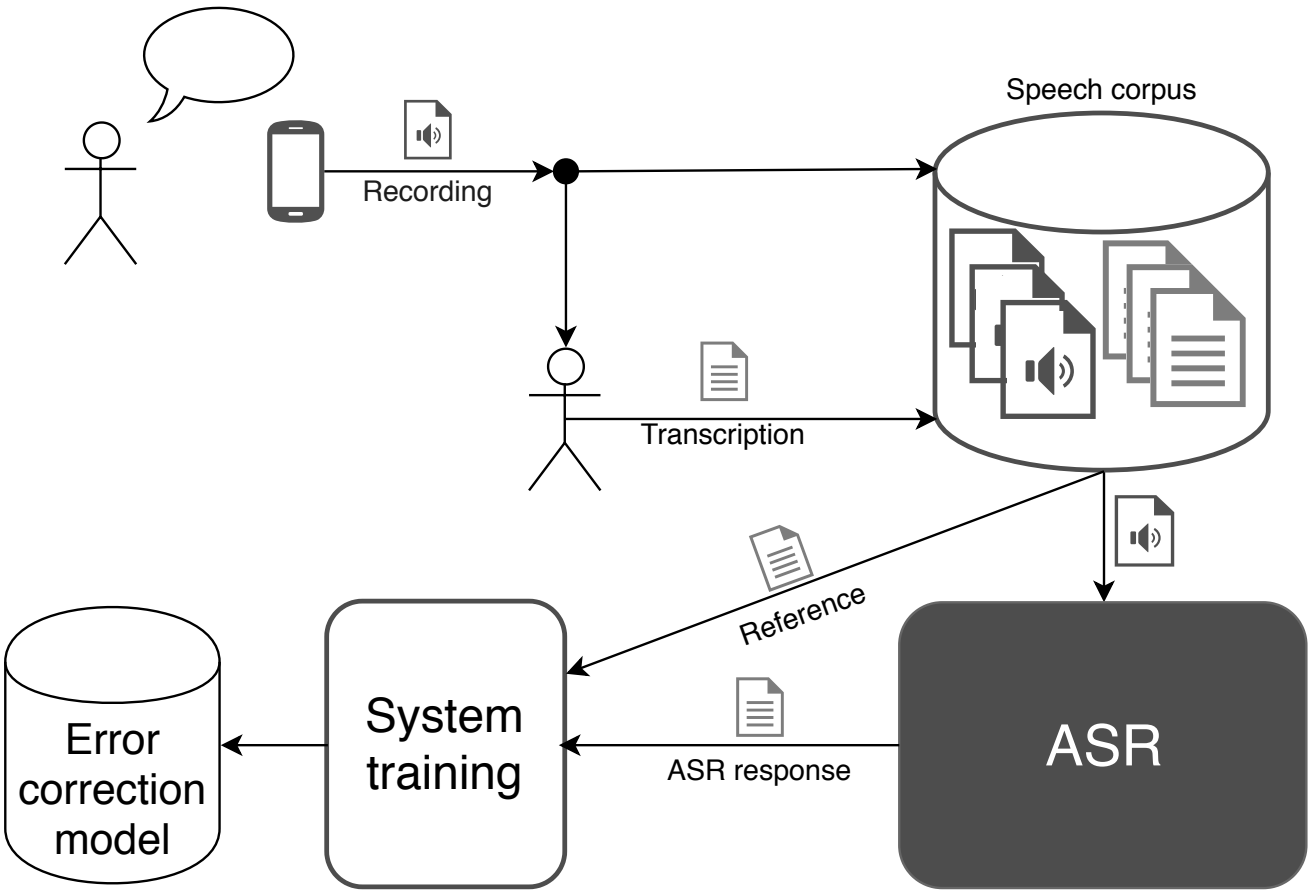
Goal

Develop a method that improves the result of speech recognition process on the basis of the (erroneous) output of an existing ASR system and the correct human-made transcription of voice recordings without access to audio data.

ASR error correction



Error correction model training



In order to make the challenge approachable by participants from outside the speech recognition community we provide the dataset that consists solely of:

1. *Hypotheses* – textual outputs of the automatic speech recognition system
2. *References* – transcriptions of sentences being read to the automatic speech recognition system

Dataset

- 9142 sentences from Polish Wikinews
- Recorded in a studio
- Transcribed by human annotators
- Recognized by an Automatic Speech Recognition system
- 8142:1000 train/test split

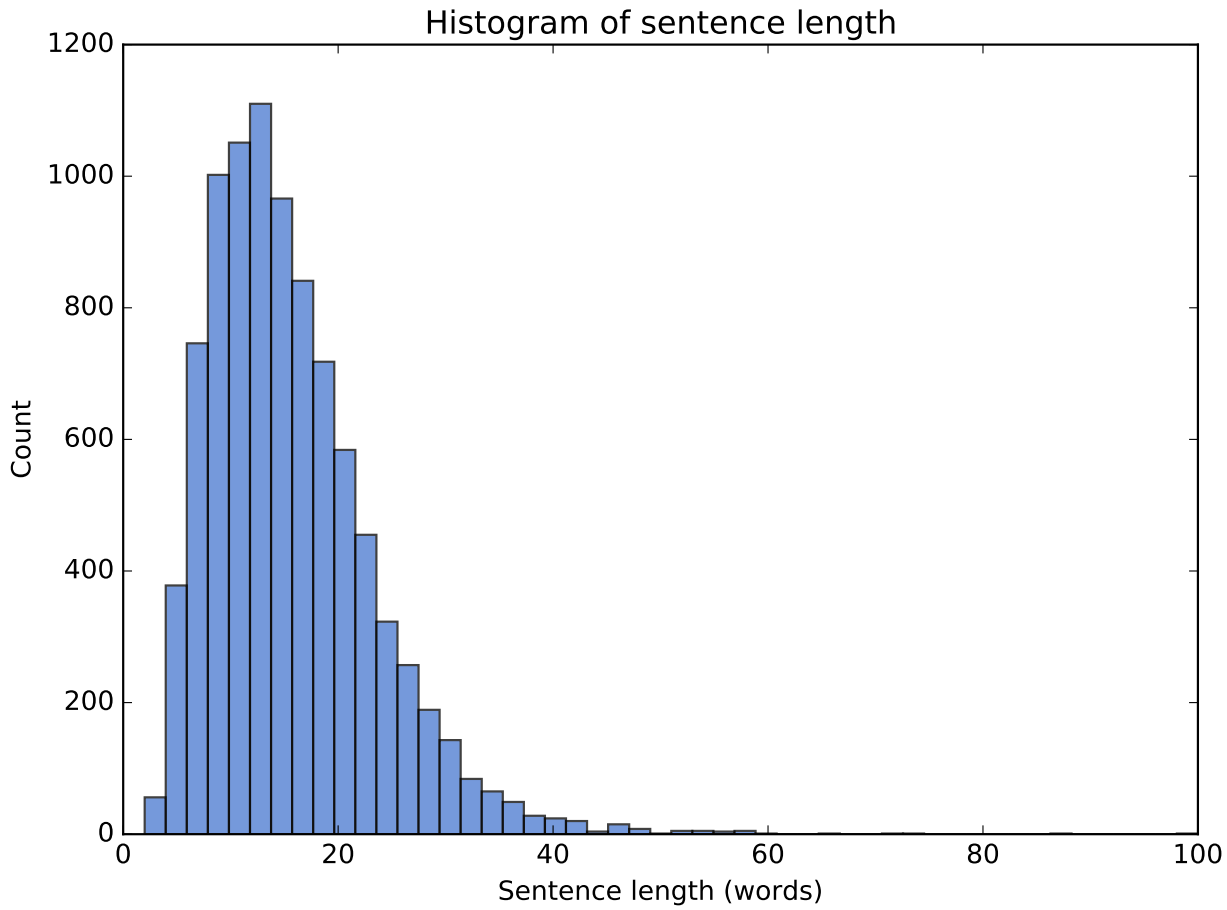
Dataset normalization

- all words are UPPERCASED
- punctuation marks (except for hyphens) are removed
- numbers and special characters are replaced by their spoken forms

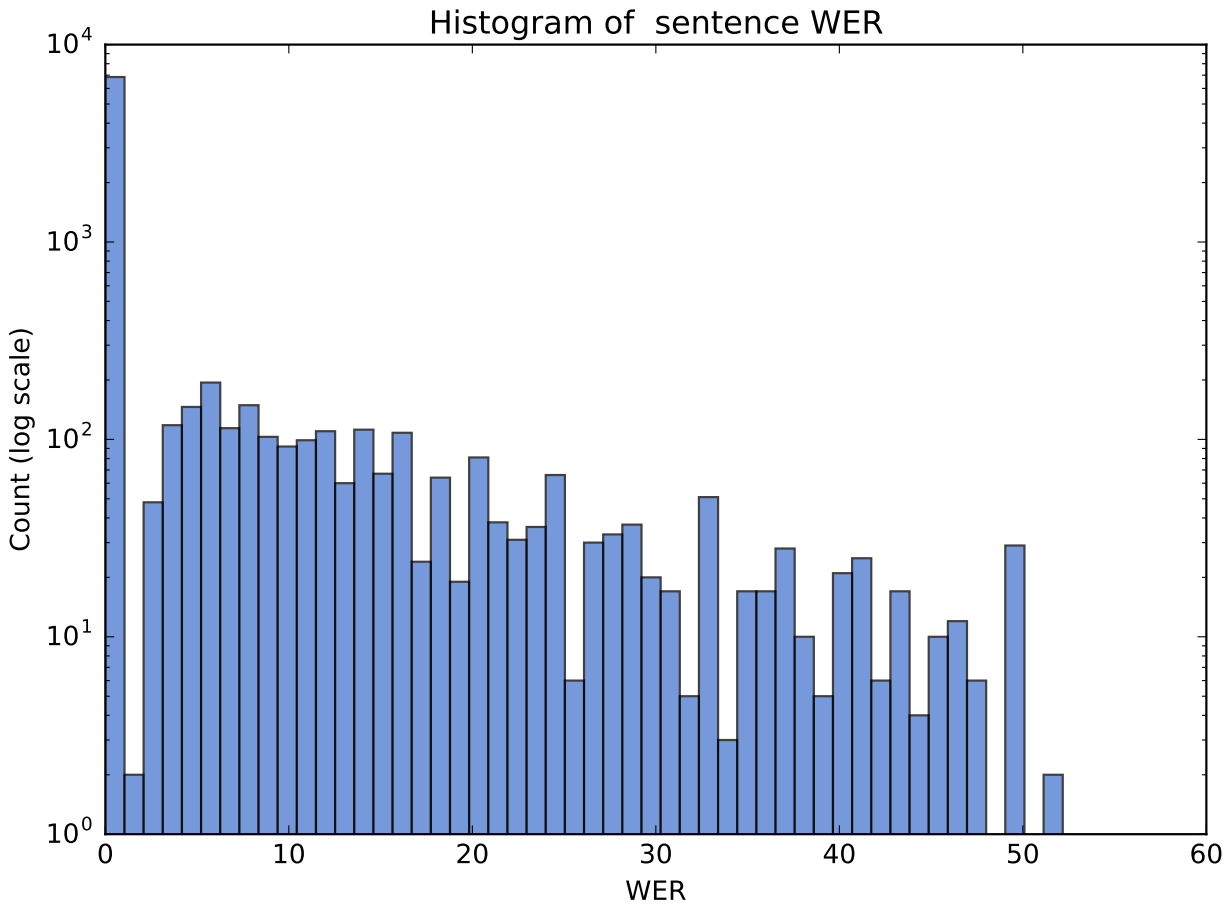
Datasets statistics

	Train set	Test set
Number of sentences	8142	1000
Average WER	3.94	4.01
Sentence Recognition Rate	0.74	0.75
Average utterance length (words)	15.40	15.10
Minimum utterance length (words)	2	3
Maximum utterance length (words)	100	48

Sentence length histogram



WER histogram



Example data

Example line from tsv file containing training dataset:

id	train-1
hypothesis	DWUDZIESTEGO CZWARTEGO KWIETNIA BIEŻĄCEGO ROKU ROZMAWIALI O WIKIPEDII INTERNECIE WSPÓŁPRACY KLASYFIKOWANIU WIEDZY KSIĄŻKACH I WŁASNOŚCI **LEKTURALNEJ**
reference	DWUDZIESTEGO CZWARTEGO KWIETNIA BIEŻĄCEGO ROKU ROZMAWIALI O WIKIPEDII INTERNECIE WSPÓŁPRACY KLASYFIKOWANIU WIEDZY KSIĄŻKACH I WŁASNOŚCI **INTELEKTUALNEJ**
source	https://pl.wikinews.org/w/index.php?curid=27343&actionaction=history
id	train-2
hypothesis	**EUROPA POWINNA JĄ** TEŻ ŻE SESJE PE W STRASBURGU SĄ DLA NICH UTRUDNIENIEM BO KOMISJA EUROPEJSKA I RADA UE Z KTÓRYMI PE CIĄGLE WSPÓŁPRACUJE MAJĄ SWOJE STAŁE SIEDZIBY W BRUKSELI
reference	**EUROPOSŁOWIE PRZYPOMINAJĄ** TEŻ ŻE SESJE PE W STRASBURGU SĄ DLA NICH UTRUDNIENIEM BO KOMISJA EUROPEJSKA I RADA UE Z KTÓRYMI PE CIĄGLE WSPÓŁPRACUJE MAJĄ SWOJE STAŁE SIEDZIBY W BRUKSELI
source	https://pl.wikinews.org/w/index.php?curid=21290&actionaction=history

Evaluation

- Evaluation performed on online competition platform Gonito
- `http://gonito.net/challenge/asr-corrections`
- Submissions made with git commits
- Results instantly visible to everyone after submission
- Multiple submissions possible

Evaluation metrics

- Word Error Rate
- Sentence Recognition Rate
- CharMatch

Word Error Rate

WER - Word Error Rate of hypothesis corrected by the proposed system, averaged over all tests sentences.

$$WER = \frac{S + D + I}{N = H + S + D}$$

where S = number of substitutions, D = number of deletions, I = number of insertions, H - number of hits, N - length of reference sentence.

Sentence Recognition Rate

SRR - Sentence Recognition Rate - sentence level accuracy of hypothesis corrected by the proposed system.
SRR is computed as ratio of the number of sentences with $WER = 0.0$ (correctly recognized sentences) to the number of all sentences in the corpus.

CharMatch

Introduced in [JGO17].
$F_0.5$-measure defined in as follows:

$$F_0.5 = (1 + 0.5^2) \times \frac{P \times R}{0.5^2 P + R}$$

Where: P is precision and R is recall:

$$P = \frac{\sum_i T_i}{\sum_i d_L(h_i, s_i)}, R = \frac{\sum_i T_i}{\sum_i d_L(h_i, r_i)}$$

Where: r_i - i-th reference utterance, h_i - i-th ASR hypothesis, s_i - i-th system output, $d_L(a, b)$ - Levenshtein distance between sequences a and b, T_i - number of correct changes performed by the system

CharMatch

T_i - number of correct changes performed by the system, calculated as:

$$T_i = \frac{d_L(h_i, r_i) + d_L(h_i, s_i) - d_L(s_i, r_i)}{2}$$

Where: r_i - i-th reference utterance, h_i - i-th ASR hypothesis, s_i - i-th system output, $d_L(a, b)$ - Levenshtein distance between sequences a and b, T_i - number of correct changes performed by the system

CharMatch

- penalizes the system for introducing new errors
- Prefers system that does not change anything over system that corrects one error and introduces another

Thank you

Thank you for your attention!

References I

Krzysztof Jassem, Filip Graliński, and Tomasz Obrębski, *Pros and cons of normalizing text with thrax*, Proceedings of the 8th Language & Technology Conference (Poznań, Poland) (Zygmunt Vetulani and Patrick Paroubek, eds.), Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2017, pp. 230–235.