Open Challenge for Correcting Errors of Speech Recognition Systems

Marek Kubis[*], **Zygmunt Vetulani**[*], **Mikołaj Wypych**[†], **Tomasz Ziętkiewicz**[†,*],

[*]Adam Mickiewicz University
Umultowska 87, 61-614 Poznań, Poland
{mkubis,vetulani}@amu.edu.pl

[†]Samsung Poland R&D Institute
Pl. Europejski 1, Warszawa, Poland
{m.wypych,t.zietkiewic}@samsung.com

Abstract

The paper announces the new long-term challenge for improving the performance of automatic speech recognition systems. The goal of the challenge is to investigate methods of correcting the recognition results on the basis of previously made errors by the speech processing system. The dataset prepared for the task is described and evaluation criteria are presented.

1. Introduction

The rise in popularity of voice-based virtual assistants such as Apple's Siri, Amazon's Alexa, Google Assistant and Samsung Bixby imposes high expectations on the precision of automatic speech recognition (ASR) systems. Scheduling a meeting at the incorrect time, sending a message to a wrong person or misinterpreting the command for the home automation system can cause severe losses to the user of a virtual assistant. The problem is even more apparent in case of deep-understanding systems supposed to work in very difficult audibility condition, and where ASR errors can appear fatal for the end users. This is the case of systems for crisis situation management (e.g. Vetulani et al., 2010) where low quality and emotional voice input can generate a real challenge for speech recognition systems. Furthermore, successful integration of ASR solutions with the very demanding AI systems will depend on the degree of being able to take into consideration the non-verbal elements of utterances (prosody). Hence, despite significant improvements to the speech recognition technology in recent years it is now even more important to search for new methods of decreasing the risk of being misunderstood by the system.

One of the methods that can be used to improve the performance of a speech recognition system is to force the system to learn from its own errors. This approach transforms the speech recognition system into a self-evolving, auto-adapting agent. The objective of this challenge is to investigate to what extent this technique can be used to improve the recognition rate of the speech processing system.

In order to make the challenge approachable by participants from outside the speech recognition community and to encourage contestants to use a broad range of machine learning and natural language engineering methods that are not specific to the processing of spoken language, 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.

Thus, the goal of the contestants is to develop a method that improves the result of speech recognition process on the basis of the (erroneous) output of the ASR system and the correct human-made transcription without access to the speech recordings.
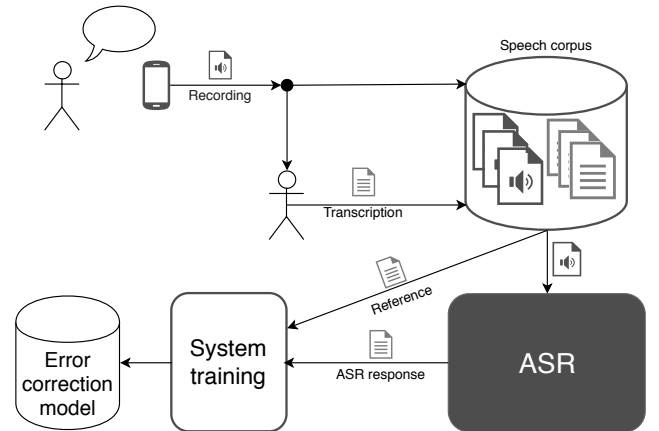

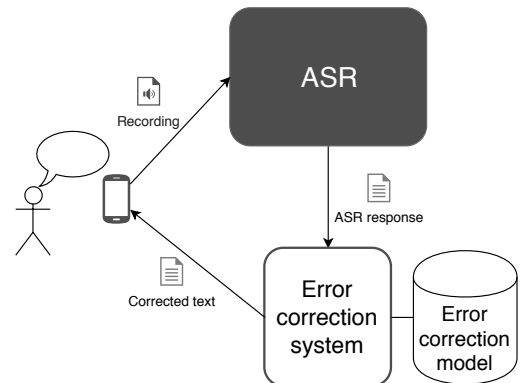Figure 1: Error correction model training

Figure 2: ASR error correction

2. Related work

2.1. Shared tasks

To our best knowledge this is the first open ASR error correction task. However, there were plenty of tasks targeting similar problems. They can be divided into two categories: speech translation tasks and grammatical error correction tasks. Representatives of the former category are 2 of 3 tasks conducted at 7th International Workshop on Spoken Language Translation (Paul et al., 2010). Tasks 1 and 2 provided sentences in source language in two forms: ASR recognition results (with errors) and correct recognition results (transcription without errors). Task of the participants was to translate the source text in both forms to the target language. Participants were provided with 3 training corpora composed of 86225 (Task 1), 19972 and 10061 (Task 2) sentence pairs. *CoNLL-2013 Shared Task on Grammatical Error Correction* (Ng et al., 2013) and *BEA 2019 Shared Task: Grammatical Error Correction* (Christopher Bryant and Briscoe, 2019) are examples of the second group of tasks. In these tasks participants are given parallel corpora of texts written by native or non-native English students, containing grammatical, punctuation or spelling errors and their manually corrected versions. The goal of the proposed system is to correct previously unseen texts. Training corpora in these tasks consist of 38785 and 57151 pairs of sentences respectively.

2.2. ASR error correction systems

(Errattahi et al., 2018) provide review of some ASR error detection and correction systems together with description of ASR evaluation metrics. (Cucu et al., 2013) propose error correction using SMT (Statistical Machine Translation) model. The SMT model is trained on relatively small parallel corpus of 2000 ASR transcripts and their manually corrected versions. At evaluation time the model is used to "translate" ASR hypothesis into it's corrected form. The system achieves 10.5% relative WER[1] improvement by reducing the baseline ASR system's WER from 11.4 to 10.2 . (Guo et al., 2019) describe ASR error correction model based on LSTM sequence-to-sequence neural network trained on large (40M utterances) speech corpus generated from plain-text data with text to speech (TTS) system. In addition to the spelling correction model authors experiment with improving results of end-to-end ASR system by incorporating the external language model and with combination of the two approaches. The proposed system achieves good results (19% relative WER improvement and 29% relative WER improvement with additional LM re-scoring, with baseline ASR WER of 6.03) but requires large speech corpus or high-quality TTS system to generate such corpus from plain text.

3. Dataset description

In order to develop the dataset for the task we decided to select 9142 sentences from Polish Wikinews (Wikimedia Foundation, 2019) and ask two native speakers of Polish (male and female) to read them to the speech recognition system. Dataset samples consist of transcription of

---
[1]Word Error Rate, see Section 4.

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

Table 1: Datasets statistics.

the sentence being read juxtaposed with the textual output captured from the system. Both references and hypotheses are normalized according to the following rules:

- words are uppercased,

- all punctuation marks except hyphens are removed,

- numbers and special characters are replaced by their spoken forms.

The dataset is divided into two sets. The training set consists of 8142 utterances randomly sampled from the dataset. The test set contains the rest of the samples.

The training and test set items consist of:

1. `id`: sample identifier

2. `hyp`: ASR hypothesis - recognition result for the sample voice recording

3. `ref`: reference - human transcription of the sample recording,

4. `source`: copyright, source and author attribution information.

Exemplary dataset items are shown in the appendix. The entire training set and test set (with the exception of reference utterances) are available for download via Gonito online competition platform (Graliński et al., 2016).
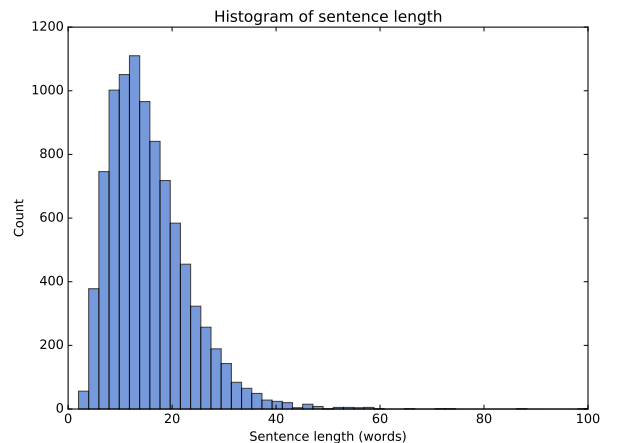

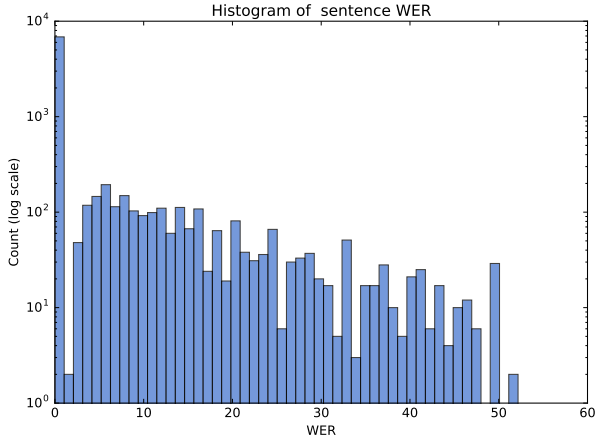Figure 3: Histogram of sentence length

Figure 4: Histogram of Word Error Rate

4. Evaluation

For the purpose of evaluation the contestants are asked to submit their results using the Gonito (Graliński et al., 2016) challenge available at `https://gonito.net/challenge/asr-corrections`.

The submission consists of single *out.tsv* file containing result of running the proposed system on *in.tsv* file, containing ASR system output. Both files contain one sentence per line. The output file should be aligned with the input.

The submissions are evaluated using geval tool (Graliński, 2019), part of the Gonito platform available also as a standalone tool. Submissions are evaluated using three metrics:

- WER - Word Error Rate of hypothesis corrected by the proposed system, averaged over all tests sentences. WER is defined as follows:

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

  where: S = number of substitutions, D = number of deletions, I = number of insertions, H - number of hits, N - length of reference sentence. See (Morris et al., 2004) for in-depth explanation.

- SRR - Sentence Recognition Rate - sentence level accuracy of hypothesis corrected by the proposed system. SRR is defined as ratio of the number of sentences with $WER = 1.0$ (correctly recognized sentences) to the number of all sentences in the corpus.

- CharMatch - $F_{0.5}$ - introduced in (Jassem et al., 2017). $F_{0.5}$-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, calculated as:

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

5. References

Christopher Bryant, Øistein Andersen, Mariano Felice and Ted Briscoe, 2019. Building educational applications 2019 shared task: Grammatical error correction. https://www.cl.cam.ac.uk/research/nl/bea2019st/.

Cucu, Horia, Andi Buzo, Laurent Besacier, and Corneliu Burileanu, 2013. Statistical Error Correction Methods for Domain-Specific ASR Systems. In Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, and Bianca Truthe (eds.), *Statistical Language and Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Errattahi, Rahhal, Asmaa El Hannani, and Hassan Ouahmane, 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37.

Graliński, Filip, Rafał Jaworski, Łukasz Borchmann, and Piotr Wierzchoń, 2016. Gonito.net - Open Platform for Research Competition, Cooperation and Reproducibility. In *Branco, António and Nicoletta Calzolari and Khalid Choukri (eds.), Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*.

Graliński, Filip, 2019. Geval. https://gitlab.com/filipg/geval.

Guo, Jinxi, Tara N Sainath, and Ron J Weiss, 2019. A spelling correction model for end-to-end speech recognition. *arXiv preprint arXiv:1902.07178*.

Jassem, Krzysztof, Filip Graliński, and Tomasz Obrębski, 2017. Pros and Cons of Normalizing Text with Thrax. In Zygmunt Vetulani and Patrick Paroubek (eds.), *Proceedings of the 8th Language & Technology Conference*. Poznań, Poland: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Morris, Andrew C., Viktoria Maier, and Phil D. Green, 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

Ng, Hwee Tou, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault, 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*. Sofia, Bulgaria: Association for Computational Linguistics.

Paul, Michael, Marcello Federico, and Sebastian Stücker, 2010. Overview of the IWSLT 2010 Evaluation Campaign. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon (eds.), *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*.

Vetulani, Zygmunt, Jacek Marciniak, Tomasz Obrębski, Grażyna Vetulani, Adam Dąbrowski, Marek Kubis, Jędrzej Osiński, Justyna Walkowska, Piotr Kubacki, and Krzysztof Witalewski, 2010. *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego*. Poznań, Poland: Wydawnictwo Naukowe UAM. English title: Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application.

Wikimedia Foundation, 2019. Wikinews. https://pl.wikinews.org. Access date: 2019-03-01.

Appendix

id	train-1
hypothesis	DWUDZIESTEGO CZWARTEGO KWIETNIA BIEŻĄCEGO ROKU ROZMAWIALI O WIKIPEDII IN-TERNECIE WSPÓŁPRACY KLASYFIKOWANIU WIEDZY KSIĄŻKACH I WŁASNOŚCI **LEKTURA LNEJ**
reference	DWUDZIESTEGO CZWARTEGO KWIETNIA BIEŻĄCEGO ROKU ROZMAWIALI O WIKIPEDII IN-TERNECIE WSPÓŁPRACY KLASYFIKOWANIU WIEDZY KSIĄŻKACH I WŁASNOŚCI **INTELEK-TUALNEJ**
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 UTRUD-NIENIEM 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
id	train-3
hypothesis	DZIESIĄTEGO WRZEŚNIA DWA TYSIĄCE ÓSMEGO ROKU **LECH MAM BLADES** OGŁOSIŁ WYNIKI FINANSOWE TRZECIEGO KWARTAŁU WYNOSZĄCE TRZY I **DZIEWIEĆDZIESIĄ-TYCH** MILIARDA DOLARÓW STRAT
reference	DZIESIĄTEGO WRZEŚNIA DWA TYSIĄCE ÓSMEGO ROKU **LEHMAN BROTHERS** OGŁOSIŁ WYNIKI FINANSOWE TRZECIEGO KWARTAŁU WYNOSZĄCE TRZY I **DZIEWIEĆ DZIESIĄ-TYCH** MILIARDA DOLARÓW STRAT
source	https://pl.wikinews.org/w/index.php?curid=25282&actionaction=history
id	train-4
hypothesis	POCHÓD ROZPOCZĄŁ SIĘ NA PLACU SENATORSKIM ALKAMISTA SENAATINTORILLA A **PIERWSZE** W SZEREGU SZŁA SZKOŁA TAŃCA SAMBY SAMBIC TANSSIKOULU
reference	POCHÓD ROZPOCZĄŁ SIĘ NA PLACU SENATORSKIM ALKAMISTA SENAATINTORILLA A **PIERWSZA** W SZEREGU SZŁA SZKOŁA TAŃCA SAMBY SAMBIC TANSSIKOULU
source	https://pl.wikinews.org/w/index.php?curid=30303&actionaction=history
id	train-5
hypothesis	DZIESIĄTEGO PAŹDZIERNIKA W KATOWICKIM SPODKU ODBĘDZIE SIĘ DWUDZIESTA DZIEWIĄTA EDYCJA RAWA BLUES FESTIVAL NAJWIĘKSZEJ BLUESOWEJ IMPREZY TYPU IN-DOOR W EUROPIE
reference	DZIESIĄTEGO PAŹDZIERNIKA W KATOWICKIM SPODKU ODBĘDZIE SIĘ DWUDZIESTA DZIEWIĄTA EDYCJA RAWA BLUES FESTIVAL NAJWIĘKSZEJ BLUESOWEJ IMPREZY TYPU IN-DOOR W EUROPIE
source	https://pl.wikinews.org/w/index.php?curid=25476&actionaction=history
id	train-6
hypothesis	PRZEPROWADZONE W POŁOWIE GRUDNIA DWUSTRONNE ROZMOWY NIE PRZYNIOSŁY REZULTATU
reference	PRZEPROWADZONE W POŁOWIE GRUDNIA DWUSTRONNE ROZMOWY NIE PRZYNIOSŁY REZULTATU
source	https://pl.wikinews.org/w/index.php?curid=5050&actionaction=history

Table 2: Dataset samples