GERALD PENN, University of Toronto

Predicting Levenshtein Edit Sequences for Fine-Grained Estimation of Automatic Speech Recognition Error

The predominant method for scoring the quality of automatic speech recognition (ASR) transcripts when ground-truth labels are not available is to predict the word error rate (WER) from the corresponding audio segment. We propose WAV2LEV, a novel paradigm for WER estimation which predicts the underlying sequences of Levenshtein edit operations (substitutions, deletions, insertions and matches) from which the WER can be computed. This approach offers more fine-grained token-level error estimation in comparison to previous work without compromising on performance for WER estimation. To support this investigation, we present Mini-CNoiSY (Miniature Clean-Noisy Speech from YouTube), a bespoke 353 hour noisy speech corpus which ensures confidence in ground-truth labeling and captures a diverse range of noise artifacts which degrade ASR performance. Our results show that WAV2LEV achieves near state-of-the-art performance for the task of WER estimation with a root mean square error (RMSE) of 0.1706 and a Pearson correlation coefficient (PCC) of 87.42%, while generating predictions of ASR error that are more informative and fine-grained than that of direct WER estimators.