Improved Data Selection for Domain Adaptation in ASR

Shannon Wotherspoon, William Hartmann, Matthew Snover, Owen Kimball
Raytheon BBN, Cambridge, MA, USA

Introduction

- We address domain adaptation in Automatic Speech Recognition (ASR) with semi-supervised training (SST).
- Data selection is critical with large domain shifts.
- We use min. phone error rate for oracle selection; we approximate min. PER with the avg. phone confidence of an utterance.
- With larger domain shifts, deletions and low lexical diversity are serious issues; we address these by using the phone rate of an utterance in our selection metric.
- We see up to 57% relative improvements over the baseline and good generalization across domain shift conditions.

Experimental Setup

- We examine a dialectal domain shift from UK to US English, as well as additional simulated domain shifts: 8 or 16kbps mp3 compression, additive noise with an SNR of 0 dB or -10 dB, and reverberation.
- Supervised data: UK English, 100hrs, Conversational Telephone Speech (CTS) & voicemail
- Unsupervised data: US English, 400hrs, CTS
- Test Set: US English, CTS

Acoustic Modeling

- Our ASR models are hybrid TDNN-Fs, trained using the supervised & unsupervised data for 1 epoch of LFMMI, followed by 1 epoch of sMBR with the unsupervised data.

Language Modeling

- We train trigram LMs using the supervised UK English data along with 62 million words of US English web data.

Data Selection and SST

Confidence

Baseline
- Average word-level CTM confidence
- A common selection method in traditional (non domain shift) SST
- confidence = the posterior probability of a word or phone

Minimum WER/PER

Oracle
- Utterances with the lowest Word or Phone Error Rate

PCPPS
- Proposed Method
- Average phone-level Cnet confidence, plus the phone rate (PPS) of the utterance
- PPS allows us to exclude high-deletion utterances
- Phone confidence is a closer approximation to acoustic error than word confidence

Selection Methods Across Conditions

- PCPPS selection matches or outperforms word confidence selection across all six data conditions
- The performance gap between the two is most prominent where the domain shift is the greatest, such as the 8 kbps compression and noise conditions
- However, even PCPPS selection is unable to match oracle minimum PER selection, indicating that there is still room for improvement in data selection methods

Lexical Diversity and Deletion Rates

- Word Confidence Selection is degenerate when performed iteratively
- PCPPS Selection, in contrast, improves with each iteration of semi-supervised training
- Word confidence selection is biased towards utterances with low lexical diversity
- By the 4th Iteration of SST, the Word Confidence selection has only 78 unique words

Iterative PCPPS Selection Across Conditions

- PCPPS selection generalizes well across all 6 conditions
- After 4 rounds of SST with PCPPS selection, we see 15-57% relative improvements over the out-of-domain baseline model
- The "True Transcript" bars follow our typical SST setup, but use 100% of our target-domain audio along with their true transcripts, representing a lower bound on SST WER