Acoustic Model Self-Training with a Small Vocabulary

Scott Novotney, Richard Schwartz
BBN Technologies, Inc., 10 Moulton Street, Cambridge, MA 02138, USA
JHU HLT COE, 810 Wyman Park Drive, Baltimore, MD 21218, USA
{snovotney, schwartz}@bbn.com

Introduction
State of the art LVCSR requires hundreds to thousands hours of (expensive) manual transcriptions.

Can unlabeled audio reduce this cost?
(Lamel 02) Self-Training on English Broadcast News (TDT-2)
- 10 min. of transcripts, 135 hours of unlabeled audio
- 1 billion word in-domain LM
- 33% relative reduction in WER
(Ma 08) Self-Training on English CTS (Fisher corpus)
- 1hr of transcripts, 2000 hours of unlabeled audio
- 1 billion word out-of-domain LM
- 47% relative reduction in WER

“Whadya mean small vocabulary? Just go to the web!”
- Many languages are not written (and not just indigenous).
- Arabic dialects (Levantine, Gulf, etc.)
- Low-resource ASR derives the vocabulary from few hours of manual transcripts. High OOV is almost a given.
- Downstream tasks frequently add new words.
- Retraining with an expanded vocabulary isn’t cheap.
- Words are not treated equally.

Self Training Method
1. Build initial acoustic and language models from available manual data.
   - Acoustic Model
   - Language Model
   - Unlabeled Audio
2. Recognize untranscribed audio data with initial model.
   - Automatic Transcriptions
3. Use word confidences to select or weight observations.
   - Filtered Transcriptions
4. Train new models on adapted transcripts and iterate. (twice usually enough)

WER Recovery Metric
We gauge success as the relative reduction in WER recovered by self-training compared to models trained with manual transcripts of the same ‘unlabeled’ audio.

Absolute WER

Initial WER – Self Trained WER
Initial WER – Supervised WER

100% recovery means that the self-trained models performed as well as manually transcribed models.

Analyzing Self-Training
We measure WER Recovery as a function of:
- Labeled audio (1 or 10 hours)
- Unlabeled audio (200 or 2000 hours)
- Language modeling text (100k in-domain, 1M out of domain, 1B out of domain)
- Acoustic model or language model training

Trends
- Larger improvement for collecting ten times as much audio as transcribing.
- Self training most helpful with small initial models and large amounts of audio.
  - Best case: 80% WER recovery (24% out of 30%) with 1+2000hr AM and 1B BN LM.
  - Worst case: 42% WER recovery (5% out of 12%) with 10+200hr AM and 100k LM.
- Language modeling works, but provides much smaller benefit.
  - No parameter sharing; model memorizes the data.
  - No external knowledge source to correct errors.

Scenario 1: 5k train, 75k test
- Only labeled words can be recognized during decoding
- Unlabeled and OOT words perform very poorly.
- But unlabeled words benefit equally as labeled from self-training (4% absolute gain).

Scenario 2: 75k train, 5k test
- All words can now be recognized during self-training.
- Unlabeled words improve dramatically:
  - WER Recovery improves from 17% to 64%.
- Despite initial WER of 80%, for these words.
- OOT words benefit as well.

Detailed Word Analysis
Break down vocabulary words into three classes:
- Labeled: words in 10 hours of manual transcriptions.
- Unlabeled: words in 200hrs of audio but not in the 10hrs.
- OOT: words in neither. (Out Of Training)

Vocabulary
- Prevalence: 3:2:1
- Words can now be analyzed separately.
- Re-analyze the previous results with this new partition.

Small Vocabularies
Previous experiments used a large 75k word vocabulary.
- OOV rate of 0.14%
- All words in unlabeled audio unfairly appeared in dictionary
Now build initial models only from ten hours of manual transcriptions.
- Acoustic Model
- Language Model (100k tokens)
- Vocabulary (5k types, 4.5% OOV)
Repeat Self-Training with 200hrs and compare to using a full 75k vocab:
- Absolute WER increases by 2%
- WER Recovery nearly identical
- 41% for 5k versus 42% for 75k

OOV rate of 0.14%

Never all words in unlabeled audio unfairly appeared in the dictionary.

BBN Technologies, Inc., 10 Moulton Street, Cambridge, MA  02138, USA
JHU HLT COE, 810 Wyman Park Drive, Baltimore, MD 21218, USA
{snovotney, schwartz}@bbn.com

Vocabulary
Type Counts
Token Counts
Set
in Test Set
in Test Set
Labeled
Unlabeled
OOT
1,659
34,268
147
807
1,520
174

Re-training with an expanded vocabulary isn’t cheap.