Semi-Supervised Speech Recognition

What to do with 10,000 hours of unlabeled speech

Scott Novotney

07/15/10
The Two Equations that Matter to ASR

- Bayesian Definition of Automatic Speech Recognition

\[ \hat{W} = \arg \max_W P(A|W) P(W) \]

- Acoustic Model
- Language Model
The Two Equations that Matter to ASR

• Bayesian Definition of Automatic Speech Recognition

\[ \hat{W} = \arg \max_W P(A|W)P(W) \]

• Word Error Rate (WER) Metric

\[ WER = \frac{\text{sub} + \text{del} + \text{ins}}{\# \text{words}} \]
Motivation

• Speech recognition models hunger for data
  – ASR requires thousands of hours of transcribed audio
  – In-domain data needed to overcome mismatches like language, speaking style, acoustic channel, noise, etc…
Motivation

• Speech recognition models hunger for data
  – ASR requires thousands of hours of transcribed audio
  – In-domain data needed to overcome mismatches like language, speaking style, acoustic channel, noise, etc…

• Transcribing speech is expensive and time consuming.
Motivation

• Speech recognition models hunger for data
  – ASR requires thousands of hours of transcribed audio
  – In-domain data needed to overcome mismatches like language, speaking style, acoustic channel, noise, etc…

• Transcribing speech is expensive and time consuming

• Sufficient amounts of in-domain data may not be available
  – Conversational domains
  – Levantine Arabic
  – Sensitive data (voicemails, medical transcriptions, etc…)
Motivation

• Speech recognition models hunger for data
  – ASR requires thousands of hours of transcribed audio
  – In-domain data needed to overcome mismatches like language, speaking style, acoustic channel, noise, etc…

• Transcribing speech is expensive and time consuming.

• Sufficient amounts of in-domain data may not be available
  – Conversational domains
  – Levantine Arabic
  – Sensitive data (voicemails, medical transcriptions, etc…)

• But, we have one inexpensive and overflowing resource: unlabeled audio
We seek to:
- Minimize resource requirements such as:
  - Manual transcriptions
  - Language modeling text
  - Pronunciation lexicon
- While approaching supervised performance with current state of the art systems.

We take advantage of orders of magnitude more unlabeled audio than labeled audio.

We are not trying to improve supervised performance – that is a different problem.
Basic Method: Iterative Automatic Transcription

1. Build initial acoustic and language models from available manual data.

Labeled Audio

Acoustic Model

LM Text

Language Model

Unlabeled Audio

Automatic Transcriptions

Confidence Selection / Weighting

Adapted Transcriptions
1. Build initial acoustic and language models from available manual data.

2. Recognize large amount of untranscribed audio data.

- Labeled Audio
  - Acoustic Model
  - Language Model
  - Unlabeled Audio
  - Automatic Transcriptions
  - Confidence Selection / Weighting
  - Adapted Transcriptions

- LM Text
Basic Method: Iterative Automatic Transcription

1. Build initial acoustic and language models from available manual data.

2. Recognize large amount of untranscribed audio data.

3. Use word confidence from ASR system to select or weight observations.
Basic Method: Iterative Automatic Transcription

1. Build initial acoustic and language models from available manual data.

2. Recognize large amount of untranscribed audio data.

3. Use word confidence from ASR system to select or weight observations.

4. Train new models on adapted transcriptions and iterate. (twice usually enough)
2000 hours of ten minute conversations about an assigned topic between two strangers over the telephone.
English Fisher Corpus

• 2000 hours of ten minute conversations about an assigned topic between two strangers over the telephone.

• Rapid speech, phonetic reductions and varied speaking style make this a very hard task.
  – State of the art WER is ~18% (1 in 5 words wrong)
  – We’ll be dealing with WER in the 30%-70% range
• 2000 hours of ten minute conversations about an assigned topic between two strangers over the telephone.

• Rapid speech, phonetic reductions and varied speaking style make this a very hard task.
  – State of the art WER is ~18% (1 in 5 words wrong)
  – We’ll be dealing with WER in the 30%-70% range

• Expensive and time consuming to transcribe
  – $150 / hour of transcription
  – 50 hours of effort / hour of transcription
Semi-Supervised Training

Recovery

Absolute WER

<table>
<thead>
<tr>
<th>Labeled Audio</th>
<th>1</th>
<th>1</th>
<th>10</th>
<th>10</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>10</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>100k</td>
<td>1M BN</td>
<td>1B BN + Web data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment</td>
<td>Acoustic Model Self Training</td>
<td>Language Model Self Training</td>
<td>Combined Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Initial AM
Self-Trained AM
Supervised AM

Initial LM (after AM Self-Training)
Self-Trained LM
Supervised LM
Semi-Supervised Training

Absolute WER

Recovery

Labeled Audio
1

Unlabeled Audio
200

LM
100k

Experiment
Acoustic Model Self Training
## Semi-Supervised Training

### Graphs

- **Absolute WER**
- **Recovery**

### Table

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Labeled Audio</th>
<th>Unlabeled Audio</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 1 10 10 10 1 1 1 1 10 10</td>
<td>200 2000 200 2000 2000 200 2000 200 2000</td>
<td>100k 1M BN 1B BN + Web data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semi-Supervised Training</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Semi-Supervised Training

**Experiment:** Acoustic Model Self Training

<table>
<thead>
<tr>
<th>Labeled Audio</th>
<th>1</th>
<th>1</th>
<th>10</th>
<th>10</th>
<th>1</th>
<th>10</th>
<th>1</th>
<th>1</th>
<th>10</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>100k</td>
<td>1M BN</td>
<td>1B BN + Web data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Absolute WER**

**Recovery**
Semi-Supervised Training

<table>
<thead>
<tr>
<th>Labeled Audio</th>
<th>1</th>
<th>10</th>
<th>1</th>
<th>10</th>
<th>1</th>
<th>10</th>
<th>1</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>100k</td>
<td>1M BN</td>
<td>100k</td>
<td>1M BN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment</td>
<td>Acoustic Model Self Training</td>
<td>Language Model Self Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labeled Audio</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unlabeled Audio</td>
<td>2000</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>100k</td>
<td>1M BN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Experiment**

- Self-Trained AM / Initial LM
- Self-Trained AM / Initial LM
- Self-Trained AM / Initial LM + Supervised LM

**Recovery**

- Absolute WER
- Recovery

**Bar Graph**

- Self-Trained AM / Initial LM
- Self-Trained AM / Initial LM
- Self-Trained AM / Initial LM + Supervised LM

**Diagram**

- Language Model Self Training
Comparison Across Languages

- Successfully repeated acoustic model self-training with Spanish and Levantine
  - Conversational Telephone Speech
  - 1 hour initial model
  - 200 hour unlabeled audio
  - 100k word in-domain LM
Comparison Across Languages

• Successfully repeated acoustic model self-training with Spanish and Levantine
  – Conversational Telephone Speech
  – 1 hour initial model
  – 200 hour unlabeled audio
  – 100k word in-domain LM

• Adapted Modern Standard Arabic Broadcast News to Levantine Arabic
  – 10hrs of MSA equivalent to 1hr Lev AM
  – 1400hrs of MSA worse than 1hr Lev LM
  – 10hr MSA self-training has 40% WER
    Recovery with strong language model
Improving the Language Model

• With 2000 hours of speech, compare ngrams extracted from
  – Reference manual transcripts
  – Automatically decoded transcripts
Improving the Language Model

- With 2000 hours of speech, compare ngrams extracted from
  - Reference manual transcripts
  - Automatically decoded transcripts

- Divide up the V^3 trigrams into four categories
  - **Hit**: observed in the ASR output and in the reference
  - **Hallucination**: in ASR output, but not reference
  - **Miss**: not in ASR, but in reference
  - **Remaining**: everything else
Improving the Language Model

- With 2000 hours of speech, compare ngrams extracted from
  - Reference manual transcripts
  - Automatically decoded transcripts

- Divide up the V^3 trigrams into four categories
  - **Hit**: observed in the ASR output and in the reference
  - **Hallucination**: in ASR output, but not reference
  - **Miss**: not in ASR, but in reference
  - **Remaining**: everything else

- We are given Hit and Hallucination, our tasks are:
  - Deweight Hallucinations *(should be able to do it)*
  - Infer **Miss** from **Miss**+**Remaining** *(standard LM task)*
“To know your enemy…”
Oracle Experiments

• To understand what needs to happen, cheat
  – Assume all hallucinated n-grams are removed
To understand what needs to happen, cheat

- Assume all hallucinated n-grams are removed
- Then vary the *counts* of the n-grams seen in the ASR output and those unseen, but in the reference
To understand what needs to happen, cheat

- Assume all hallucinated n-grams are removed
- Then vary the *counts* of the n-grams seen in the ASR output and those unseen, but in the reference
  - **Single** – singleton counts
To understand what needs to happen, cheat

- Assume all hallucinated n-grams are removed
- Then vary the counts of the n-grams seen in the ASR output and those unseen, but in the reference
  - **Single** – singleton counts
  - **ASR** – counts from the decoded output
Oracle Experiments

• To understand what needs to happen, cheat
  – Assume all hallucinated n-grams are removed
  – Then vary the **counts** of the n-grams seen in the ASR output and those unseen, but in the reference
    – **Single** – singleton counts
    – **ASR** – counts from the decoded output
    – **Truth** – counts from reference
To understand what needs to happen, cheat
- Assume all hallucinated n-grams are removed
- Then vary the *counts* of the n-grams seen in the ASR output and those unseen, but in the reference
  - **Single** – singleton counts
  - **ASR** – counts from the decoded output
  - **Truth** – counts from reference

The ‘fair’ WER is 41%

<table>
<thead>
<tr>
<th>Unseen Types (C)</th>
<th>Seen Types (A)</th>
<th>ASR</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td>36.6</td>
<td>36.8</td>
</tr>
<tr>
<td>Single</td>
<td></td>
<td>34.1</td>
<td>33.7</td>
</tr>
<tr>
<td>Truth</td>
<td></td>
<td>33.2</td>
<td>32.0</td>
</tr>
</tbody>
</table>
To understand what needs to happen, cheat

- Assume all hallucinated n-grams are removed
- Then vary the *counts* of the n-grams seen in the ASR output and those unseen, but in the reference
  - Single – singleton counts
  - ASR – counts from the decoded output
  - Truth – counts from reference

The ‘fair’ WER is 41%
To understand what needs to happen, cheat
  – Assume all hallucinated n-grams are removed
  – Then vary the *counts* of the n-grams seen in the ASR output and those unseen, but in the reference
    – **Single** – singleton counts
    – **ASR** – counts from the decoded output
    – **Truth** – counts from reference

The ‘fair’ WER is 41%
• To understand what needs to happen, cheat
  – Assume all hallucinated n-grams are removed
  – Then vary the counts of the n-grams seen in the ASR output and those unseen, but in the reference
    – Single – singleton counts
    – ASR – counts from the decoded output
    – Truth – counts from reference

• The ‘fair’ WER is 41%

<table>
<thead>
<tr>
<th>Unseen Types (C)</th>
<th>Seen Types (A)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASR</td>
<td>Truth</td>
</tr>
<tr>
<td>None</td>
<td>36.6</td>
<td>36.8</td>
</tr>
<tr>
<td>Single</td>
<td>34.1</td>
<td>33.7</td>
</tr>
<tr>
<td>Truth</td>
<td>33.2</td>
<td>32.0</td>
</tr>
</tbody>
</table>
To understand what needs to happen, cheat
- Assume all hallucinated n-grams are removed
- Then vary the *counts* of the n-grams seen in the ASR output and those unseen, but in the reference
  - **Single** – singleton counts
  - **ASR** – counts from the decoded output
  - **Truth** – counts from reference

**The ‘fair’ WER is 41%**

<table>
<thead>
<tr>
<th></th>
<th>Seen Types (A)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASR</td>
<td>Truth</td>
</tr>
<tr>
<td><strong>Unseen Types (C)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>36.6</td>
<td>36.8</td>
</tr>
<tr>
<td>Single</td>
<td>34.1</td>
<td>33.7</td>
</tr>
<tr>
<td>Truth</td>
<td>33.2</td>
<td>32.0</td>
</tr>
</tbody>
</table>
Potential Solutions

- De-weighting hallucinated n-grams
  - Use contextual features
  - Adapt initial LM instead of retraining from scratch
Potential Solutions

• **De-weighting hallucinated n-grams**
  - Use contextual features to estimate counts
  - Adapt initial LM instead of retraining from scratch

• **Increasing weight for unseen n-grams**
  - Continuous language model like a neural net LM
  - Increased parameter sharing
Potential Solutions

• De-weighting hallucinated n-grams
  – Use contextual features to estimate counts
  – Adapt initial LM instead of retraining from scratch

• Increasing weight for unseen n-grams
  – Continuous language model like a neural net LM
  – Increased parameter sharing

• Move beyond n-grams
Potential Solutions

• De-weighting hallucinated n-grams
  – Use contextual features
  – Adapt initial LM instead of retraining from scratch

• Increasing weight for unseen n-grams
  – Continuous language model like a neural net LM
  – Increased parameter sharing

• Move beyond n-grams

• The goal is not to beat an n-gram LM on supervised corpora, but to *robustly* estimate parameters in the face of noise.
P( You | Thank) = 0.2

- [http://www.clsp.jhu.edu/people/snovotne/](http://www.clsp.jhu.edu/people/snovotne/)

- **Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription**
  Scott Novotney, Chris Callison-Burch (2010)
  *Proceedings of NAACL*, Los Angeles, USA

- **Analysis of Low-Resource Acoustic Model Self-Training**
  Scott Novotney, Richard Schwartz (2009)
  In *Proceedings of Interspeech*, Brighton, England

- **Unsupervised acoustic and language model training with small amounts of labelled data**
  Scott Novotney, Richard Schwartz, Jeff Ma (2009)
  In *Proceedings of ICASSP*, Taipai, Taiwan