Automatic Speech Recognition: Conditional Random Fields for ASR

Jeremy Morris
Eric Fosler-Lussier
Ray Slyh
9/19/2008

Overview

- Traditional Automatic Speech Recognition (ASR) systems built using components
  - Feature extraction
  - Acoustic models
  - Lexicon
  - Language model

Overview

- AFRL SCREAM Lab and OSU SLaTe Lab have both investigated using phonetic features for ASR
  - Based on linguistic phonetic attributes of speech
  - Language independent model of speech – phonetic attributes cross languages
  - Acoustic models built using these features have the potential to be language independent
  - Both SCREAM and SLaTe Labs have investigated these features as replacements for acoustic features

Overview – ASR System

- Frequency Analysis
- Acoustic Modeling
- Pronunciation Modeling
- Language Modeling

Objective

- Investigate new methods for integrating phonetic features into ASR systems
  - Phonetic features can provide a level of language independence that acoustic features do not
  - Phonetic features can provide a method for leveraging linguistic theory and observations in our ASR systems

Approach

- Conditional Random Fields
  - We make use of a discriminative statistical model known as a Conditional Random Field (CRF)
  - Traditional ASR systems use a generative statistical model known as a Hidden Markov Model (HMM)
    - Any HMM can be transformed into a CRF, though the reverse is not true
    - As a model, a CRF has fewer independence assumptions than a corresponding HMM, potentially giving us a better model for our data
Approach

- Two methods of integration
  - Use the CRF model to generate inputs for a traditional ASR HMM-based system
  - Use the CRF model as a replacement for an HMM in a new type of ASR system
- We have performed some investigation into both of these approaches

Approach 1 - CRANDEM

- Use the CRF to generate inputs to an HMM-based system
  - This approach parallels Tandem ASR systems
    - Use neural networks to generate inputs to a standard HMM-based ASR system
    - CRF-Tandem -> CRANDEM
  - CRFs are used to generate local posterior probabilities for each frame of input speech data
    - These posteriors are then used as inputs to an HMM

Progress 1 - CRANDEM

- Pilot system 1
  - Train on phonetically transcribed speech data (TIMIT corpus)
  - Test to see if phone recognition accuracy shows improvement over baseline
  - Two experiments – phone classes and phonological feature classes

<table>
<thead>
<tr>
<th>Experiment (61 phone classes)</th>
<th>Phn. Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM Reference baseline (PLP)</td>
<td>68.1%</td>
</tr>
<tr>
<td>Tandem baseline</td>
<td>70.8%</td>
</tr>
<tr>
<td>CRF baseline</td>
<td>71.6%</td>
</tr>
<tr>
<td>CRANDEM</td>
<td>72.4%</td>
</tr>
</tbody>
</table>

(Fosler-Lussier & Morris, ICASSP 2008)

Progress 1 – Pilot System

<table>
<thead>
<tr>
<th>Experiment (44 phonetic feature classes)</th>
<th>Phn. Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM Reference baseline (PLP)</td>
<td>68.1%</td>
</tr>
<tr>
<td>Tandem baseline</td>
<td>71.2%</td>
</tr>
<tr>
<td>CRF baseline*</td>
<td>71.6%</td>
</tr>
<tr>
<td>CRANDEM</td>
<td>72.4%</td>
</tr>
</tbody>
</table>

(Fosler-Lussier & Morris, ICASSP 2008)

* Best score for CRF is currently 74.5% (Heinz, Fosler-Lussier & Brew, submitted)

Progress 1 - CRANDEM

- CRANDEM system
  - Use results from pilot system to train system on word transcribed speech data (WSJ corpus)
  - Test to see if word recognition accuracy shows improvement over baseline
Progress 1 – CRANDEM System

<table>
<thead>
<tr>
<th>Experiment</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>(54 phone classes)</td>
<td></td>
</tr>
<tr>
<td>HMM Reference baseline (MFCC)</td>
<td>9.15%</td>
</tr>
<tr>
<td>Tandem baseline (MLP+MFCC)</td>
<td>7.79%</td>
</tr>
<tr>
<td>CRANDEM (CRF+MFCC)</td>
<td>11.01%</td>
</tr>
</tbody>
</table>

Approach 2 – CRF ASR

- Use the CRF as a replacement for the HMM
  - CRFs generate local acoustic probabilities as a lattice of possibilities
  - This lattice is composed with a probabilistic word network (language model) to recognize word strings

Progress 1

- Pilot CRANDEM systems showed:
  - Improvement in phone accuracy
  - Degradation in word accuracy
- More tuning may be required
  - Initial CRANDEM results had 21.32% WER
  - Tuning has dropped this to 11% so far
- Error may be due to different assumptions between CRF and HMMs
  - CRFs very “sure” of themselves, even when they are wrong
  - CRF outputs do not fit a Gaussian distribution, which the HMMs assume

Progress 2

- Pilot system
  - Train a CRF over phonetically transcribed speech corpus (TIMIT)
  - Use this to derive training data from word transcribed, restricted vocabulary speech corpus (TI-DIGITS)
  - Perform training and testing over restricted vocabulary

Progress 2 – TI-DIGITS recognition

<table>
<thead>
<tr>
<th>Experiment</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>(PLPs only)</td>
<td></td>
</tr>
<tr>
<td>HMM baseline (32 mix)</td>
<td>0.18%</td>
</tr>
<tr>
<td>HMM (1 mix)</td>
<td>1.55%</td>
</tr>
<tr>
<td>CRF (PLPs only)</td>
<td>1.19%</td>
</tr>
</tbody>
</table>

Progress 2 – TI-DIGITS recognition

<table>
<thead>
<tr>
<th>Experiment</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>(PLPs + phone classes)</td>
<td></td>
</tr>
<tr>
<td>HMM baseline (32 mix)</td>
<td>0.25%</td>
</tr>
<tr>
<td>CRF (PLPs only)</td>
<td>1.01%</td>
</tr>
</tbody>
</table>
Progress 2 – TI-DIGITS recognition

- Results show CRF performing within reach of state-of-the-art for this task
  - Some more tuning may be necessary – certain parameters of the CRF have not been exploited to full potential yet
  - HMMs have had over thirty years of exploration in how to tune parameters for speech recognition – the space for CRFs is just beginning to be explored
  - Our initial results were much worse – over 10% WER – but experimentation has brought them down to almost 1%

Progress Summary

- Phone recognition
  - Over the last year, we have improved CRF phone recognition results on TIMIT to near best-reported HMM result (75% accuracy)
- Large Vocabulary Recognition
  - Code framework now in place for Large Vocabulary recognition experiments
  - Challenges have been identified, and ideas are being examined for tackling these challenges

Future Work

- CRF model has not yet been fully explored for word recognition with these experiments
  - More tuning possibly required for both CRANDEM experiments and direct recognition experiments
  - Direct word recognition experiments need to be performed on larger vocabulary corpus (WSJ)
  - More analysis on CRANDEM results
- Results suggest the need for pronunciation modeling for the direct CRF model
  - These experiments have been begun, but no results to report as yet

Thank You