Maximum Entropy Techniques for min-WER Score Combination with Sausages

Kemal Sonmez
Summary Overview

• **Goal:** To improve lattice rescoring by including novel information sources with discriminatively trained weights

• **Approach:** Conditional probability model of the hypothesized word on a sausage edge being the true transcription
  - Exponential model conditioned on the context via a set of features
  - Maximum entropy (ME) estimation of the exponential model weights

• **Bottom line:** Not quite working yet, preliminary setup has so far not given a significant win (<0.1% abs)

• **Future Work:**
  - Discriminative framework for including side information in rescoring confusion networks, e.g. prosodic features --to be investigated further and many things in the pipeline to try
Talk Plan

• Rationale
  – Lattices and confusion networks
• Brief synopsis of prior work on discriminative score combination
• Approach
  – Min WER by ME estimation of conditional exponential model over confusion networks
• Experiments
• Preliminary Results
Rationale

• Lattice rescoring is an important part of information combination in ASR
• Rescoring by confusion networks allows minimization of WER directly
• Confusion network oracle error rates leave room for significant improvements
• Ideally, the scores need to be combined in a discriminative manner
• We develop a framework for rescoring of confusion networks based on a discriminatively estimated conditional model
Lattices to Sausages

- Lattice rescoring plays an important role in information combination in ASR
- Confusion networks are compacted lattices with nodes merged into ordered equivalence classes
- Word-level rather than sentence-level posteriors
- Minimize (an upper bound on) WER directly
RT03-dev sausages

• How much room is left in RT03-devset confusion networks?

<table>
<thead>
<tr>
<th>Max Depth in confusion network</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>top</td>
<td>25.8%</td>
</tr>
<tr>
<td>2</td>
<td>23.9%</td>
</tr>
<tr>
<td>3</td>
<td>23.0%</td>
</tr>
<tr>
<td>4</td>
<td>22.4%</td>
</tr>
<tr>
<td>5</td>
<td>22.0%</td>
</tr>
</tbody>
</table>
Some recent prior work

• Sentence Error Rate minimization
  – Yu, Waibel, ICASSP 2004

• Word Error Rate minimization
  – Mangu, Padmanabhan, ICASSP 2001

• Discriminative Model Combination
  – Beyerlein, ASRU 1997
Prior Work

- Sentence Error Rate Minimization by Conditional Exponential Models (Yu, Waibel, ICASSP 2004)
- Conditional exponential model of score combination estimated by ME
  \[ f_1(\text{obs}, \text{hyp}) = \log p_{AM}(\text{obs} | \text{hyp}) \]
- Set of feature functions:
  \[ f_2(\text{obs}, \text{hyp}) = \log p_{LM}(\text{hyp}) \]
  \[ f_3(\text{obs}, \text{hyp}) = [\# \text{words}(\text{hyp})] \]
  ...
- Similar to usual score combination, with a normalization term
  \[ \log P(\text{hyp} | \text{obs}) = \sum_i \lambda_i f_i(\text{obs}, \text{hyp}) - \log Z(\text{obs}) \]
- MMIE-like normalization computation
  \[ Z(\text{obs}) \approx \sum_{\text{hyp}(N-best)} \exp \left( \sum_i \lambda_i f_i(\text{obs}, \text{hyp}) \right) \]
Prior Work

- WER minimization via **error correction** over confusion networks (Mangu, Padmanabhan, ICASSP 2001)
- Transformation-based learning to train rules to distinguish hypotheses in a confusion network using additional information
  - *choose the 2nd candidate (‘-’) if 1st candidate is a short word with posterior < 0.46*
- 0.5% absolute improvement on WS97
Conditional Exponential Models of Word Error

• Probability that $w_e^i$, the word on edge $e$ of alignment is correct:

$$\log P(w_e^i = w_{ref}^i \mid context) = \sum_i \lambda_i f_i(context, w_e^i) - \log Z(context)$$

• Features to represent sausage context

$$f_1(context, w_e^i) = \log p_{AM}$$
$$f_2(context, w_e^i) = \log p_{LM}$$
$$f_3(context, w_e^i) = \log p_{DBN}$$
$$f_4(context, w_e^i) = [\# words(hyp)]$$

... 

• Weights estimated by ME
Sausage Context Features

- Normalized posterior rank
  - \( R_n = \frac{\text{rank}}{\text{(# edges in confusion network)}} \)
- Posterior
- Landmark Pronunciation Model scores
  - DBN scores
  - Discriminative pronunciation model scores
- AM and LM scores
- Duration
- Number of phones
- Relative confusion network position in the lattice
- Confusability
  - \( c(w) = \log(\text{# w in the training confusion network set}) \)
- Function word membership
- Delete feature
Experiments

• Selection of features
• Confidence smoothing
  – \text{conf\_score} = \frac{p(\text{top edge})}{p(\text{runner up edge})}
  – rerank edges only if \text{conf\_score} < \text{threshold}
• Two ways of dealing with \textit{--delete-} edges
  – Leave out sausages with deletes in the active depth
  – Include \textit{--delete-} edges in the training with binary delete features (f_{\text{delete}} = 1[w = \textit{--delete-}])
• Training edge depth into the confusion network:
  – True edge + top 2,3,4,5
Preliminary Results

• RT03 development set
  – sausages from 2000-best lists, aligned with references
  – divided into ME training (2000 sausages) and testing sets (930 sausages)

• Rescoring with ME trained posteriors
  – Test set performance:

<table>
<thead>
<tr>
<th>system</th>
<th>sub</th>
<th>del</th>
<th>ins</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.8</td>
<td>10.9</td>
<td>3.5</td>
<td>31.1</td>
</tr>
<tr>
<td>Rescored with top2</td>
<td>16.8</td>
<td>10.9</td>
<td>3.5</td>
<td>31.1</td>
</tr>
<tr>
<td>Conf-rescored with top2</td>
<td>16.7</td>
<td>11.0</td>
<td>3.4</td>
<td>31.1</td>
</tr>
</tbody>
</table>
Preliminary Results

• RT03 development set
  – sausages from lattices, aligned with references
  – divided into ME training (2000 sausages) and testing sets (930 sausages)

• Rescoring with ME trained posteriors
  – Test set performance:

<table>
<thead>
<tr>
<th>system</th>
<th>sub</th>
<th>del</th>
<th>ins</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>15.8</td>
<td>13.4</td>
<td>3.8</td>
<td>33.0</td>
</tr>
<tr>
<td>conf-rescored with sausage features</td>
<td>15.8</td>
<td>13.4</td>
<td>3.8</td>
<td>33.0</td>
</tr>
<tr>
<td>+ landmark (DBN) features</td>
<td>15.8</td>
<td>13.4</td>
<td>3.8</td>
<td>33.0</td>
</tr>
</tbody>
</table>
Summary and Future Work

- Sausage-based discriminative rescoring via ME
- Further work needed in assessing merits
  - as a score combination technique for landmark based pronunciation models as well as other side information
  - so far, results tentative and not conclusive
- Future Work:
  - New features from prosody
    - Stress accent levels
    - Energy and/or F₀ profiles
  - Many more things to try:
    - Interpolation of the exponential model with the original posterior
    - Confidence threshold informed by utterance and/or speaker characteristics (more in Emily’s talk)