Hypothesis Posterior Student-Teacher Training

Jeremy H. M. Wong and Mark J. F. Gales
Department of Engineering, University of Cambridge
jhmr2@cam.ac.uk, mffg@eng.cam.ac.uk

1 INTRODUCTION

• Ensemble methods
  – improve ASR performance
  – are computationally expensive to decode.
• Student-Teacher (S-T) training
  – trains single student model to emulate teacher ensemble.
  – Existing methods only transfer frame posterior information.
• This work incorporates sequence discriminative criteria into S-T training by:
  – sequence discriminative training of the teacher ensemble
  – further sequence discriminative training of the student model after frame-level S-T training
  – a proposed hypothesis-level S-T criterion.

2 TEACHER ENSEMBLE

• Diversity obtained by
  – different DNN random initialisations.
• Teachers can be trained using the following criteria:
  – Cross-Entropy (CE)
    \[ F_{CE} = -\sum \sum \delta(s_t, s'_t) \log P(s_t | o_t, \Phi_{t}) + \eta \sum \log P(h_t | o_t, \Phi_{t}) \]
  – Maximum Mutual Information (MMI)
    \[ F_{MMI} = -\sum \delta(h_t, h'_t) \log P(h_t | o_t, \Phi_{t}) \]
  – state-level Minimum Bayes Risk (sMBR)
    \[ F_{sMBR} = \sum L(h_t, h'_t) P(h_t | o_t, \Phi_{t}) \]

3 INFORMATION PROPAGATION

• Frame posteriors
  – Existing method.
  – Minimise KL-divergence between frame posteriors.
  – Interpolate with hard alignments.
  \[ C_{CE} = -\sum \sum \delta(s_t, s'_t) \log P(s_t | o_t, \Phi_{t}) + \lambda \sum \alpha_o P(s_t | o_t, \Phi_{t}) \log P(s_t | o_t, \Theta) \]
  – Setting \( \lambda = 0 \) reduces to CE.
• Hypothesis posteriors
  – Novel approach.
  – Minimise KL-divergence between hypothesis posteriors.
  – Interpolate with manual transcriptions.
  \[ C_{MMI} = -\sum \sum \delta(h_t, h'_t) \log P(h_t | o_t, \Phi_{t}) + \eta \sum \beta_h P(h_t | o_t, \Phi_{t}) \log P(h_t | o_t, \Theta) \]
  – Setting \( \eta = 0 \) reduces to the MMI criterion.

4 EXPERIMENTS

4.1 TEACHER ENSEMBLE TRAINING CRITERION

• Training ensemble with different criteria, in Tok Pisin

<table>
<thead>
<tr>
<th>Ensemble criterion</th>
<th>Single system WER (%)</th>
<th>Combined WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>51.4 51.3 51.8 0.1</td>
<td>50.5</td>
</tr>
<tr>
<td>MMI</td>
<td>49.3 49.1 49.4 0.1</td>
<td>48.4</td>
</tr>
<tr>
<td>sMBR</td>
<td>48.2 48.1 48.4 0.1</td>
<td>47.0</td>
</tr>
</tbody>
</table>

– Training teachers with sequence discriminative criteria improves combined ensemble performance.
– Frame-level S-T training with sequence-trained teachers, in Tok Pisin

4.2 REFINEMENT OF THE STUDENT MODEL

<table>
<thead>
<tr>
<th>Training</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame level S-T</td>
<td>47.7 5.07</td>
</tr>
<tr>
<td>frame level S-T + MMI</td>
<td>47.6 5.09</td>
</tr>
<tr>
<td>frame level S-T + sMBR</td>
<td>47.2 4.94</td>
</tr>
</tbody>
</table>

• Student is initialised using frame-level S-T training with the sMBR-trained teacher ensemble.
• For WSJ,
  – mean single sMBR system WER = 5.09 %
  – combined ensemble WER = 4.84 %.
• Further sMBR training of student improves performance.
• Further MMI training does not give significant gains, as the teacher ensemble has been sMBR-trained.

4.3 PROPAGATING HYPOTHESIS POSTERIOR INFORMATION

<table>
<thead>
<tr>
<th>Training</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypothesis level S-T</td>
<td>50.0 4.85</td>
</tr>
<tr>
<td>hypothesis level S-T + MMI</td>
<td>49.7 4.91</td>
</tr>
<tr>
<td>hypothesis level S-T + sMBR</td>
<td>47.4 4.94</td>
</tr>
</tbody>
</table>

• Hypothesis-level S-T training improves the student performance beyond frame-level S-T training, even with further MMI training.

5 CONCLUSIONS

• Sequence discriminative training of the teacher ensemble improves the resulting student performance.
• Further sequence discriminative training after frame-level S-T training brings additional gains.
• Proposed hypothesis-level S-T training yields gains over frame-level S-T training, even with further sequence discriminative training.

ACKNOWLEDGEMENT

This work is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense (DoD), ARPA, contract number D17PC00002. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARP, DoD/ARL, or the U.S. Government. The authors would like to thank the LORELEI team for providing the KWS infrastructure and those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARP, DoD/ARL, or the U.S. Government.