COMBINATION OF END-TO-END AND HYBRID MODELS FOR SPEECH RECOGNITION

Jeremy Wong, Yashesh Gaur, Rui Zhao, Liang Lu, Eric Sun, Jinyu Li, and Yifan Gong

Microsoft Speech and Language Group
Data

• Training:
  • 75K hours from variety of Microsoft applications.

• Testing:
  • Average of 13 application scenarios (Cortana, far-field, ....).
  • Total 1.8M words, 260K utterances.
Model architectures

Hybrid model

\[ P(\omega_{1:L} | \mathbf{O}_{1:T}) \propto P(\omega_{1:L}) \sum_{s_{1:T} \in \omega_{1:L}} \prod_{t=1}^{T} \frac{P(s_t | \mathbf{o}_t)}{P(s_t)} P(s_t | s_{t-1}) \]

- Language model
  \[ P(\omega_{1:L}) = \prod_{l=1}^{L} P(\omega_l | \omega_{l-n+1:l-1}) \]
- Makes conditional independence assumptions.
- Uses external lexicon and language model.
Model architectures

**LAS model**

\[
P(\tau_{1:j}|O_{1:T}) = \prod_{j=1}^{J} P(\tau_j|\tau_{1:j-1}, O_{1:T})
\]

- No conditional independence assumption.
- All components jointly trained.
- Not frame-synchronous.

**RNN-T model**

\[
P(\tau_{1:j}|O_{1:T}) = \sum_{s_{1:T+j} \in \mathcal{B}(\tau_{1:T})} \prod_{k=1}^{T+j} P(s_k|s_{1:k-1}, O_{1:T})
\]

- No conditional independence assumption.
- All components jointly trained.
- Frame-synchronous.
Hypothesis-level model combination

• The models may behave differently and predict diverse error patterns.
• Combine the hypotheses together to correct each other’s errors.
• Use MBR combination decoding.

\[ \omega^* = \arg\min_{\omega'} \sum_{m=1}^{M} \lambda_m \sum_{\omega \in \mathbb{N}} L(\omega, \omega') \frac{P_{km}^m(\omega | O_{1:T})}{\sum_{\tilde{\omega} \in \mathbb{N}} P_{km}^m(\tilde{\omega} | O_{1:T})} \]

• Only hypothesis posteriors are needed, not per-word scores.
• Performance depends on the accuracy of the hypothesis posteriors.
Bias toward short hypotheses

• LAS and RNN-T produce hypothesis posteriors that are biased toward short sequences.
• Alleviate using length normalisation.

\[ \tilde{P}(\tau_{1:j}|O_{1:T}) \propto P^1(\tau_{1:j}|O_{1:T}) \]

<table>
<thead>
<tr>
<th>Length norm</th>
<th>LAS WER (%)</th>
<th>Insertion (%)</th>
<th>Deletion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>10.40</td>
<td>0.79</td>
<td>4.82</td>
</tr>
<tr>
<td>yes</td>
<td>7.90</td>
<td>1.32</td>
<td>1.38</td>
</tr>
</tbody>
</table>
Model architectures

Hybrid model

\[
P(\omega_{1:L} | \mathbf{o}_{1:T}) \propto P^\gamma(\omega_{1:L}) \sum_{s_{1:T} \in \omega_{1:L}} \prod_{t=1}^{T} \frac{P(s_t | \mathbf{o}_t)}{P(s_t)} P(s_t | s_{t-1})
\]

- Language model

\[
P(\omega_{1:L}) = \prod_{l=1}^{L} P(\omega_l | \omega_{l-n+1:l-1})
\]

- Makes conditional independence assumptions.
- Uses external lexicon and language model.
Model architectures

**LAS model**

\[ P(\tau_{1:j}|O_{1:T}) = \prod_{j=1}^{J} P(\tau_j|\tau_{1:j-1}, O_{1:T}) \]

- No conditional independence assumption.
- All components jointly trained.
- Not frame-synchronous.

**RNN-T model**

\[ P(\tau_{1:j}|O_{1:T}) = \sum_{s_{1:T+j} \in B(\tau_{1:T},T)} \prod_{k=1}^{T+J} P(s_k|s_{1:k-1}, O_{1:T}) \]

- No conditional independence assumption.
- All components jointly trained.
- Frame-synchronous.
Bias toward short hypotheses

- LAS and RNN-T produce hypothesis posteriors that are biased toward short sequences.
- Alleviate using length normalisation.

\[
\tilde{P}(\tau_{1:j} | o_{1:T}) \propto P^1(\tau_{1:j} | o_{1:T})
\]

<table>
<thead>
<tr>
<th>Length norm</th>
<th>LAS WER (%)</th>
<th>Insertion (%)</th>
<th>Deletion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>10.40</td>
<td>0.79</td>
<td>4.82</td>
</tr>
<tr>
<td>yes</td>
<td>7.90</td>
<td>1.32</td>
<td>1.38</td>
</tr>
</tbody>
</table>
MBR training

- Can also alleviate bias by using discriminative training.
- Conditional maximum likelihood implicitly minimises alternative hypotheses through softmax.

\[ F_{CML} = - \log P(\omega^{\text{ref}} | O_{1:T}) \]

- Minimum Bayes’ risk explicitly minimises alternative hypotheses within criterion.

\[ F_{MBR} = \sum_{\omega \in \mathbb{N}} L(\omega, \omega^{\text{ref}}) \frac{P(\omega | O_{1:T})}{\sum_{\omega' \in \mathbb{N}} P(\omega' | O_{1:T})} \]

- Length normalisation can be used inside MBR criterion.

\[ F_{MBR-LN} = \sum_{\omega \in \mathbb{N}} L(\omega, \omega^{\text{ref}}) \frac{\frac{1}{P[\omega]} (\omega | O_{1:T})}{\sum_{\omega' \in \mathbb{N}} \frac{1}{P[\omega']} (\omega' | O_{1:T})} \]
## MBR training

<table>
<thead>
<tr>
<th>Training</th>
<th>Decoding length norm</th>
<th>LAS WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{\text{CML}}$</td>
<td>no</td>
<td>10.40</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>7.90</td>
</tr>
<tr>
<td>$F_{\text{MBR}}$</td>
<td>no</td>
<td>8.95</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>7.92</td>
</tr>
<tr>
<td>$F_{\text{MBR-LN}}$</td>
<td>no</td>
<td>9.29</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>7.85</td>
</tr>
</tbody>
</table>

- MBR training reduces bias toward short hypotheses.
MBR decoding of end-to-end NN model

• Decoding process:

  N-best ➔ Length norm ➔ Posterior scale ➔ N-best to lattice ➔ Determinise ➔ Lattice ➔ MBR decode ➔ Text

• Treat length-normalised scores as hypothesis posteriors.

• N-best to lattice conversion example:

  a brown cat 0.7
  the bound cat 0.3

  text graph:
  a, 0.7
  brown, 1
  cat, 1
  the, 0.3
  bound, 1
  cat, 1
## MBR decoding of end-to-end NN model

<table>
<thead>
<tr>
<th>Model</th>
<th>1-best WER (%)</th>
<th>MBR WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>8.03</td>
<td>8.01</td>
</tr>
<tr>
<td>LAS</td>
<td>7.85</td>
<td>8.42</td>
</tr>
<tr>
<td>RNN-T</td>
<td>8.16</td>
<td>8.16</td>
</tr>
</tbody>
</table>

- N-best list size = 16.
- No significant gain from MBR decoding.
Model combination

- Hypothesis-level MBR combination.

<table>
<thead>
<tr>
<th>Models</th>
<th>WER (%)</th>
<th>Relative WERR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>8.03</td>
<td>-</td>
</tr>
<tr>
<td>LAS</td>
<td>7.85</td>
<td>-</td>
</tr>
<tr>
<td>RNN-T</td>
<td>8.16</td>
<td>-</td>
</tr>
<tr>
<td>Hybrid + LAS</td>
<td>7.32</td>
<td>6.8</td>
</tr>
<tr>
<td>Hybrid + RNN-T</td>
<td>7.26</td>
<td>9.6</td>
</tr>
<tr>
<td>LAS + RNN-T</td>
<td>7.62</td>
<td>2.9</td>
</tr>
<tr>
<td>Hybrid + LAS + RNN-T</td>
<td>6.89</td>
<td>12.2</td>
</tr>
</tbody>
</table>

- Combination between different model architectures yields significant gains.
Model combination

- Compare combination methods for hybrid + LAS + RNN-T.

<table>
<thead>
<tr>
<th>Combination method</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-best of merged N-best</td>
<td>7.59</td>
</tr>
<tr>
<td>ROVER</td>
<td>7.33</td>
</tr>
<tr>
<td>MBR</td>
<td>6.89</td>
</tr>
</tbody>
</table>

- MBR combination performs the best.
Conclusion

• Propose hypothesis-level combination between hybrid and end-to-end NN models.
• Length normalisation and MBR training can reduce bias toward short hypotheses.