Experiments with lightly supervised discriminative training on TDT data.

H.Y. Chan and P.C. Woodland

Cambridge University Engineering Department

5th Sept 2003
Experiments with lightly supervised discriminative training on TDT data

Overview

• Compare with other sites' approaches

• Lightly supervised acoustic model training

• Investigate interactions between discriminative training and lightly supervised acoustic model training

• Only closed-caption transcripts are available for TDT data
  - TDT4 (300h raw data)
  - TDT2 (450h raw data)
  - 144 hours of accurately transcribed data

• Improve the HTK English Broadcast News system by adding large amount of TDT data
Experiments with lightly supervised discriminative training on TDT data

Lightly supervised training

- Filter: based on sentence confidence score

  - Closed-caption filtering (LIMSI/BBN approach)

- Construct individual LM from each text source and perform interpolation

  - Bias the interpolated LM to the closed-caption sources

- Use all TDT data for training

  - Language model includes closed-captions

  - Use reasonably fast and accurate transcription system

  - Recognize the TDT data to get the training transcriptions

lightly supervised training
Experiments with lightly supervised discriminative training on TDT data

**Training and Testing Data Sets**

- **Acoustic training data**
  - **bnac**
  - **144 hours broadcast news acoustic with accurate transcriptions**
  - **TDT2**, **TDT3**, **TDT4**
  - **Feb 2001 (6 shows, 3h data)**
  - **Feb 2001**
  - **Oct 2000 - Jan 2001 (448 shows, 300h raw data)**
  - **Feb 1998 - June 1998 (902 shows, 450h raw data)**

- **Text corpora:**
  - **TDT2 CC** (closed captions)
  - **TDT3 CC**
  - **TDT4 CC**
  - **Marketplace**
  - **PSM broadcast newstranscriptions**
  - **CNN, commercial newswire, all before and after Jan 2001**

- **Test sets**
  - **dev03 17th Jan 2001 - 31 Jan 2001 (6 shows, 3h data)**
  - **eval03 Feb 2001 (6 shows, 3h data)**
Experiments with lightly supervised discriminative training on TDT data

TDT data transcription: Recognition LM

- Interpolated weight for TDT4 model is 0.90
- OOV rate is 0.44%
- Trigram perplexity is 53.2, fourgram perplexity is 25.6

Transcribing TDT4 - minimize perplexity of dev03

- Interpolated weight for TDT2 model is 0.92
- OOV rate is 0.68%
- Trigram perplexity is 44.5, fourgram perplexity is 21.3

Transcribing TDT2 - minimize perplexity of 10h accurate transcription set

- Interpolated word LM (14K unknown words)
- 0.85% OOV rate on TDT4 closed-captions (14K unknown words)
- 0.76% OOV rate on TDT2 closed-captions (23K unknown words)

59K word-list of CU-HTK 2003 10xRT system

Interpolated word LM (16, 48): one model for each LM data set

Cambridge University Engineering Department EARS Meeting Sept 2003
Experiments with lightly supervised discriminative training on TDT data

**TDT datatranscription: Decoding**

- **Gender labelling**
  - Clustering
  - SPron
  - HLDA
  - MPE

- **Alignment**
  - CNC
  - P1, P2 of CU-HTK 2003 10xRT system

- **Automatic segmentation**
  - GI
  - MPE triphones, HLDA, 59k, tgint, fgint
  - LSLR, MLLR, 1 speech transform

- **Recognition of TD T data**
  - WB/NB
  - GIWB

- **WER on 10h TDT2 data (P2+CN)**
  - 12.4% •
  - P1, P2 of CU-HTK 2003 10xRT system LM:
    - 12.4%
  - - biased TD T2 LM: 9.3%

- **WER on dev03 (P2+CN)**
  - 12.4% •
  - - biased TD T4 LM: 8.3%

- **WER is 10.3% •**
  - - biased TD T2 LM: 9.3%
  - - biased TD T4 LM: 8.3%

- **TD T2 closed-captions WER is 10.3% •**
  - - biased TD T2 LM: 9.3%
  - - biased TD T4 LM: 8.3%

- **5xRT ~**
  - - Confusion network, re-alignment system

- **WER on 10h TD T2 data (P2+CN)**
  - 12.4% •
  - - CU-HTK 2003 10xRT system LM:
    - 12.4%
  - - P1, P2 of CU-HTK 2003 10xRT system

- **WER on dev03 (P2+CN)**
  - 12.4% •
  - - biased TD T4 LM: 8.3%

- **WER is 10.3% •**
  - - biased TD T2 LM: 9.3%
  - - biased TD T4 LM: 8.3%

- **TD T2 closed-captions WER is 10.3% •**
  - - biased TD T2 LM: 9.3%
  - - biased TD T4 LM: 8.3%
Experiments with lightly supervised discriminative training on TDT data

Acoustic model training

- Wide-band model training data
  - Recognized TDT 4: 255h (234h wb, 21h nb)
  - Recognized TDT 2: 420h (370h wb, 50h nb)

Narrow-band model training data: 144h base only (NB analysis)

- MLE, HLDA, MMI, MPE acoustic modelling

• Cross-word triphone, ~ 7000 tied states, 16 Gaussian mixture components

• Discriminative training
  - I-smoothing for both MMI and MPE
  - Heavily pruned bigram for both word lattices and phone-marked lattices
  - Denominator: re-recognize with MLE + HLDA model
  - Numerator: accurate base transcriptions / recognized TDT transcriptions

Cambridge University Engineering Department EARS Meeting Sept 2003
Experiments with lightly supervised discriminative training on TDT data
Further adding TDT2 on bnac+TDT4 only gives very small improvement

TDT4 data is more useful than TDT2 data

Adding TDT2/TDT4 to bnac improve the performance

Using 370h wb TDT2 alone outperforms bnac with 144 hours of data

<table>
<thead>
<tr>
<th></th>
<th>13.6</th>
<th>13.7</th>
<th>14.2</th>
<th>14.4</th>
<th>14.2</th>
<th>14.9</th>
<th>15.3</th>
<th>15.5</th>
<th>15.5</th>
<th>MPE+HLD</th>
<th>MMI+HLD</th>
<th>MLE+HLD</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+370h TDT2

bnac+230h TDT2

bnac+250h TDT4

bnac+420h TDT2

bnac+370h wb TDT2

370h wb TDT2

bnac (144h)

Wide-band data

Undepated single pass decoding WER - dev03
Experiments with lightly supervised discriminative training on TDT data

Wide-band data

<table>
<thead>
<tr>
<th>Model</th>
<th>+230h TDT2</th>
<th>+230h wb TDT2</th>
<th>+230h wb TDT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPE</td>
<td>17.1</td>
<td>17.1</td>
<td>17.1</td>
</tr>
<tr>
<td>MLE</td>
<td>17.1</td>
<td>17.1</td>
<td>17.1</td>
</tr>
<tr>
<td>MPE+HLDA</td>
<td>15.1</td>
<td>15.1</td>
<td>15.1</td>
</tr>
<tr>
<td>MLE+HLDA</td>
<td>15.1</td>
<td>15.1</td>
<td>15.1</td>
</tr>
<tr>
<td>bnac</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
</tr>
<tr>
<td>bnac+420h TDT4</td>
<td>13.4</td>
<td>13.4</td>
<td>13.4</td>
</tr>
<tr>
<td>bnac+255h TDT4</td>
<td>13.4</td>
<td>13.4</td>
<td>13.4</td>
</tr>
<tr>
<td>bnac+370h TDT2</td>
<td>13.4</td>
<td>13.4</td>
<td>13.4</td>
</tr>
<tr>
<td>bnac+370h wb TDT2</td>
<td>13.4</td>
<td>13.4</td>
<td>13.4</td>
</tr>
</tbody>
</table>

MPE always outperforms MMI

Adding narrow-band TDT2 data (wide-band analysis) harms the model 0.1-0.3%

More gains are obtained in discriminative stage than maximum likelihood

Undeadapted single pass decoding WER - Eval03

Experiments with lightly supervised discriminative training on TDT data
Experiments with lightly supervised discriminative training on TDT data

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>dev03 P1</th>
<th>dev03 P2</th>
<th>eval03 P1</th>
<th>eval03 P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>bnac</td>
<td>16.2</td>
<td>15.8</td>
<td>15.5</td>
<td>14.5</td>
</tr>
<tr>
<td>370h wb TDT2</td>
<td>12.5</td>
<td>12.3</td>
<td>12.0</td>
<td>11.4</td>
</tr>
<tr>
<td>bnac+370h wb TDT2</td>
<td>14.8</td>
<td>11.9</td>
<td>14.2</td>
<td>13.6</td>
</tr>
<tr>
<td>bnac+420h TDT2</td>
<td>15.1</td>
<td>11.4</td>
<td>14.0</td>
<td>11.4</td>
</tr>
<tr>
<td>bnac+255h TDT4</td>
<td>11.5</td>
<td>10.7</td>
<td>13.6</td>
<td>10.7</td>
</tr>
</tbody>
</table>

- More WER reduction in P1 (GI, unadapted, tight beam-widths) than in P2
- Adding 255h TDT4 to bnac
  - 1.1% (dev03) and 0.8% (eval03) WER reduction in P2 output
  - 10.7 WER on eval03, the same as full CU-HTK 2003 10xRT evaluation system

- Adding TDT2 to bnac obtain much less gain, 0.6% (dev03) and 0.2% (eval03) WER reduction in P2 output
Experiments with lightly supervised discriminative training on TDT data

Data selection experiments with TDT4

- Remove the last two weeks of TDT4 data which cover the time period of dev03, 230h TDT4 data is remaining (new baseline)

- Con £© confidence measure filtering
  - CC mismatch: only retain segments which match worst with CC
  - CC match: only retain segments which match best with CC
  - Align the recognized transcriptions with the CC on a whole show basis

- Closed-captions filtering

- Dataselection experiments with TDT4
  - Dataselection experiments with TDT4

- Closed-captions filtering
Experiments with lightly supervised discriminative training on TDT data

- Using all data is the best for MPE
- Confidence measure filtering doesn’t appear to reduce WER
- Very small difference in performance on CC match and CC mismatch!!

<table>
<thead>
<tr>
<th>Wide-band data</th>
<th>MPE+HLD A</th>
<th>MLE+HLD A</th>
</tr>
</thead>
<tbody>
<tr>
<td>bnac+230h TDT4</td>
<td>17.8</td>
<td>15.0</td>
</tr>
<tr>
<td>bnac+213h TDT4 (0.85 CM)</td>
<td>17.1</td>
<td></td>
</tr>
<tr>
<td>bnac+115h TDT4 (CC mismatch)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bnac+115h TDT4 (90% CC match)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bnac+80h TDT4 (94% CC match)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bnac</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data selection: unadapted single pass decoding - dev03
Experiments with lightly supervised discriminative training on TDT data

Data selection: unadapted single pass decoding - eval03

<table>
<thead>
<tr>
<th>bnac</th>
<th>MPE+HLDA</th>
<th>MLE+HLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similar pattern as in the results on dev03

bnac+115h TDT4 (90% CC match) - reduces 0.6% WER for both MLE and MPE over bnac baseline

bnac+230h TDT4 - reduces 0.5% and 1.0% WER for MLE and MPE respectively

bnac+213h TDT4 (0.85 CM) - bnac baseline

bnac+115h TDT4 (CC mismatch) - bnac baseline

bnac+80h TDT4 (94% CC match) - bnac baseline

Wide-band data
Experiments with lightly supervised discriminative training on TDT data

Dataselection: CU-HTK P1-P2 System - dev03/eval03

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>dev</th>
<th>eval03</th>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>bac</td>
<td>10.9</td>
<td>13.6</td>
<td>11.8</td>
<td>14.5</td>
</tr>
<tr>
<td>bac+30h TDT4</td>
<td></td>
<td></td>
<td>14.5</td>
<td></td>
</tr>
<tr>
<td>bac+15h TDT4</td>
<td></td>
<td>11.9</td>
<td>11.2</td>
<td>15.2</td>
</tr>
<tr>
<td>bac+15h TDT4 90% CC match</td>
<td></td>
<td></td>
<td>11.1</td>
<td>15.1</td>
</tr>
<tr>
<td>bac+15h TDT4 94% CC match</td>
<td></td>
<td></td>
<td>11.0</td>
<td>15.1</td>
</tr>
<tr>
<td>bac+80h TDT4</td>
<td></td>
<td>11.0</td>
<td>11.2</td>
<td>14.2</td>
</tr>
<tr>
<td>bac+80h TDT4 94% CC match</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Adding 230h TDT4 to bac gives P2 results only slightly worse than adding 115h TDT4 to bac
- No difference in performance between CC match and CC mismatch
- Much better performance than all CC filtering in P1 output
- Adding 230h TDT4 to bac gives P2 resultsonly slightly worse than adding 230h TDT4 to bac

CC filtering

- No difference in performance between CC match and CC mismatch
- Much better performance than all CC filtering in P1 output
- 0.8% (dev03) and 0.6% (eval03) WER reduction in P2 output
Conclusion

Successfully apply MPE and MMI for lightly supervised discriminative training on TDT data.

MPE outperforms MMI in both supervised and lightly supervised training.

By adding 255h TDT4 data to 144h broadcast news acoustic training data, on eval03

- 0.8% absolute WER rate decrease on the 5xRT PL-P2 CU-HTK system.
- 10.7% WER, which is the same as the CU-HTK 2003 10xRT Broadcast News Evaluation system.

Closed-captions filtering and sentence based confidence measure filtering don’t appear useful for improving recognition accuracy for MPE.

The best MPE result comes from the model trained with all data.

News Evaluation system

Cambridge University Engineering Department EARS Meeting Sept 2003
Experiments with lightly supervised discriminative training procedures

Conclusion

By adding 230h TDT4 data, on dev03

- 0.8% absolute WER rate decrease in the 5xRT PI-P2 CU-HTK system
- 11.8% WER, compared with 11.6% WER of the CU-HTK 2003 10xRT Broadcast News Evaluation system

Future work

- Improve lightly supervised discriminative training procedures
- Use more complex models for more data
- Adding outdated TDT2 on banc+TDT4 doesn’t give much further WER reduction with current set-up

By adding 230h TDT4 data, on dev03

Cambridge University Engineering Department EARS Meeting Sept 2003