The LIMSI RT03 BN Systems

J.L. Gauvain, L. Lamel, G. Adda, L. Chen, H. Schwenk

RT03 meeting
Boston, MA
May 19, 2003
TALK OUTLINE

- LIMSI 2003 BN system overview
- Development set design
- BN English system
- BN Mandarin system
- Conclusions
BN SYSTEM OVERVIEW
(English & Mandarin)

● Same partitioning as ’98 BN system
  – Iterative maximum likelihood segmentation/clustering procedure using GMMs and agglomerative clustering

● Updated acoustic and language models
  – 4 sets of tied state triphones (31k contexts, 11.5k states), 16 Gaussians per mixture
  – MMI training
  – 65k vocabulary, 4-gram LM
  – Use of TDT4 audio data with closed-captions for training

● Revised decoding strategy (same as dryrun03 system)
  – 2 step decoding
STT ENGLISH DEVELOPMENT SET

- No appropriate BN dev data available
- Selected 6 TDT shows from the second half of January 2001
  - 20010117_2000_2100_PRI_TWD
  - 20010120_1830_1900_ABC_WNT
  - 20010122_2100_2200_MSN_NBW (no captions available)
  - 20010125_1830_1900_NBC_NNW
  - 20010128_1400_1430_CNN_HDL
  - 20010131_2000_2100_VOA ENG
- Selection criteria: representative WER and date
- Normalized closed-captions aligned with recognizer hypothesis
- Manual correction for scoring shared with BBN, CUED and SRI
- Verification marked commercials segments to ignore during scoring
ACOUSTIC MODELS

- PLP-like frontend, cepstral mean and variance normalization (by segment cluster)
- Triphone models (31k contexts, 16 Gaussian mixtures)
- Separate cross-word/word-internal statistics
- Tied states with decision tree
- Training data: \( \sim 150 \) hours (1995, 1996, and 1997 Hub4 data) + \( \sim 90 \) of selected TDT4 data
- Telephone and wideband models
- Gender-dependent models from SI seed models with MMI training
TRAINING TEXTS

- Old newspapers and newswires (1994-1999, 1.37G words)
- Recent newspapers and newswires (01/2000-31/01/2001, 54M words)
- Manual transcripts of the HUB4 acoustic training data, old dev and eval sets (1.9M words)
- TDT2 and TDT3 captions and transcripts (1998, 9.6M words)
- TDT4 captions and transcripts (10/2000-15/01/2001, 2.2M words)
- CNN data from CNN archive (01/2000-15/01/2001, 12M words)
- **Wordlist**: selected using cutoffs for each source
  Minimize OOV on dev03 data
  Lexical coverage \( \sim 99.5\% \) on dev03
LANGUAGE MODELS

- 65233 words including compound words (300) and acronyms (1000)
- Language models: 2-gram, 3-gram and 4-gram
  - Development LMs trained all sources predating Jan 15, 2001
  - Interpolation coefficients minimize perplexity on Dev03
  - Eval LMs trained on all sources predating Feb 1, 2001
  - **RT03 LM:** 21M bigrams, 44M trigrams, 34M fourgrams

<table>
<thead>
<tr>
<th>LM</th>
<th>PRI</th>
<th>ABC</th>
<th>MSN</th>
<th>NBC</th>
<th>CNN</th>
<th>VOA</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT02</td>
<td>10.1</td>
<td>12.3</td>
<td>11.1</td>
<td>11.8</td>
<td>18.6</td>
<td>17.8</td>
<td>13.6</td>
</tr>
<tr>
<td>RT03</td>
<td>9.5</td>
<td>11.8</td>
<td>10.0</td>
<td>10.6</td>
<td>17.7</td>
<td>16.5</td>
<td>12.6</td>
</tr>
</tbody>
</table>
DECODING STRATEGY

- Initial hypothesis generation with 3-gram LM, small cross-word position-dependent, gender-specific AMs (total 1.4xRT)
- Lattice rescoring with 4-gram
- MLLR adaptation and word lattice generation (2 global regression classes) with 2-gram LM and large cross-word position-dependent, gender-specific AMs
- Lattice expansion with 4-gram LM
- Consensus decoding with pronunciation probabilities
BN ENGLISH PROGRESS ON DEV03

<table>
<thead>
<tr>
<th>System</th>
<th>PRI</th>
<th>ABC</th>
<th>MSN</th>
<th>NBC</th>
<th>CNN</th>
<th>VOA</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT02 (10x)</td>
<td>11.9</td>
<td>13.4</td>
<td>11.1</td>
<td>12.9</td>
<td>19.0</td>
<td>18.4</td>
<td>14.5</td>
</tr>
<tr>
<td>RT03 (10x)</td>
<td>8.6</td>
<td>11.0</td>
<td>9.6</td>
<td>10.0</td>
<td>16.7</td>
<td>14.8</td>
<td>11.8</td>
</tr>
<tr>
<td>BBN+LIMSI (17x)</td>
<td>8.2</td>
<td>9.0</td>
<td>8.3</td>
<td>8.9</td>
<td>14.7</td>
<td>12.5</td>
<td>10.3</td>
</tr>
<tr>
<td>BBN⊗LIMSI (9.2x)</td>
<td>8.0</td>
<td>9.2</td>
<td>7.9</td>
<td>9.0</td>
<td>14.9</td>
<td>12.9</td>
<td>10.3</td>
</tr>
</tbody>
</table>
BBN-LIMSI INTEGRATED SYSTEM (9.2xRT)

speech

LIMSI decode
5.8xRT

dev03: 12.3
eval03: 11.8

output 1

BBN adapted decode
3.4xRT

dev03: 11.1
eval03: 10.5

output 2

Combine

dev03: 10.3
eval03: 9.9

final output
MANDARIN BROADCAST NEWS SYSTEM

L. Lamel, L. Chen, J.L. Gauvain
BN MANDARIN - OVERVIEW

- Same basic system as for English BN STT
- Modified audio partitioner for CBS/CTS (speech-in-noise GMM)
- Wideband & narrowband acoustic models
- Gender-specific, position-dependent triphones
- Lightly supervised acoustic model training
- 4-gram LM
- 57k wordlist includes all characters
- 2 pass decoding (1.4xRT + 8.4xRT)
ACOUSTIC MODEL TRAINING

- Hub4 Mandarin data from LDC (27 hours)
- 120 hours from TDT4 corpus
- Light acoustic model training: transcripts generated automatically with
  - AMs trained on LDC data
  - Source-specific LMs trained on TDT4 captions for Mainland sources (CNR, CTV and VOA) and the CBS Taiwan source
- CER about 7% on 4 CBS shows
ACOUSTIC MODELS

- Wideband models trained on Hub4-Mandarin and TDT4 Mainland sources (CNR, CTV, VOA)
- Narrowband models trained on narrowband version of above and TDT4 CBS data and 20 CBS shows (6 hours) with manual segmentations
- Gender-specific models
- Pass 1: 5500 contexts, 5500 tied-states, 16 Gaussians
- Pass 2: 21k contexts, 11500 tied-states, 16 Gaussians
LANGUAGE MODEL TRAINING

• Text sources available from LDC
  – TDT2,3,4 Mandarin transcripts (10.2M characters)
  – People Daily newspaper 1991-1996 (85M characters)
  – China Radio transcripts 1994-1996 (87M characters)
  – Xinhua news 1994-1996 (22M characters)
  – Acoustic training transcripts (0.43M characters)

• Text sources shared by BBN
  – Central Daily News text 1997-2000 (61M characters)
  – CTS transcripts 1997-2000 (14M characters)
LEXICON

- 57707 words (including all characters)
- Essentially no OOVs
- 59152 phone transcriptions (2% alternate pronunciations)
- 61 phones including silence, fillers and breath
- 24 consonants
- 11 vowels, with 3 tones for each vowel (rising, flat and falling)
• Source specific language models (CBS, CNR, CTV, CTS VOA)
• Text segmentation using maximum match method
• Component LMs trained on each text source and each audio source
• Mixture weights chosen to minimize perplexity on Mandarin Dev03 data (shared by BBN)
• Weight of the audio transcript component set to 0.1.
• Minimum Discrimination Information adaptation for Taiwan sources (CBS, CTS) using the TDT4 CTS (0.66M chars) and CBS (0.46M chars) closed captions as adaptive data
• RT03 dev LMs trained on data through mid-Dec (predating Dev03 epoch)
• RT03 eval LMs trained on all data through Jan’03
# LANGUAGE MODELS - CHARACTER PERPLEXITY

<table>
<thead>
<tr>
<th>Show</th>
<th>TDT LM</th>
<th>Source LMs</th>
<th>MDI-adapt</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV_MAN</td>
<td>191</td>
<td>167</td>
<td>-</td>
</tr>
<tr>
<td>CNR_MAN</td>
<td>248</td>
<td>204</td>
<td>-</td>
</tr>
<tr>
<td>VOA_MAN</td>
<td>274</td>
<td>249</td>
<td>-</td>
</tr>
<tr>
<td>CBS_MAN</td>
<td>508</td>
<td>412</td>
<td>390</td>
</tr>
<tr>
<td>CTS_MAN</td>
<td>623</td>
<td>495</td>
<td>460</td>
</tr>
<tr>
<td>Avg.</td>
<td>351</td>
<td>282</td>
<td>-</td>
</tr>
</tbody>
</table>
## DEV’03 RESULTS

<table>
<thead>
<tr>
<th>Show</th>
<th>Initial 3-pass SI</th>
<th>2-pass decoding</th>
<th>Common TDT4 LM</th>
<th>Source LMs + addl texts SI</th>
<th>GD+wb/nb</th>
<th>TDT4 AMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV_MAN</td>
<td>17.3</td>
<td></td>
<td>11.5*</td>
<td>13.4</td>
<td>12.8</td>
<td>9.7</td>
</tr>
<tr>
<td>CNR_MAN</td>
<td>16.2</td>
<td></td>
<td>14.1</td>
<td>11.6</td>
<td>10.9</td>
<td>9.8</td>
</tr>
<tr>
<td>VOA_MAN</td>
<td>15.0</td>
<td></td>
<td>12.9</td>
<td>12.5</td>
<td>11.9</td>
<td>10.8</td>
</tr>
<tr>
<td>CBS_MAN</td>
<td>43.2</td>
<td></td>
<td>34.0</td>
<td>30.4</td>
<td>29.5</td>
<td>24.1</td>
</tr>
<tr>
<td>CTS_MAN</td>
<td>75.9</td>
<td></td>
<td>72.2</td>
<td>65.6</td>
<td>59.4</td>
<td>52.8</td>
</tr>
<tr>
<td>Avg.</td>
<td>34.5</td>
<td></td>
<td>30.2*</td>
<td>28.0</td>
<td>25.8</td>
<td>22.6</td>
</tr>
</tbody>
</table>

* unfair LM for the CTS sources due to a naming reversal in captions
## EVAL’03 RESULTS

<table>
<thead>
<tr>
<th>Show</th>
<th>Dev03</th>
<th>Eval03</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV_MAN</td>
<td>9.7</td>
<td>8.0</td>
</tr>
<tr>
<td>CNR_MAN</td>
<td>9.8</td>
<td>6.1</td>
</tr>
<tr>
<td>VOA_MAN</td>
<td>10.8</td>
<td>11.6</td>
</tr>
<tr>
<td>CBS_MAN</td>
<td>24.1</td>
<td>24.5</td>
</tr>
<tr>
<td>CTS_MAN</td>
<td>52.8</td>
<td>54.8</td>
</tr>
<tr>
<td>Avg.</td>
<td>22.6</td>
<td>21.7</td>
</tr>
</tbody>
</table>
CONCLUSIONS

- Updated BN systems for English and Mandarin
  - Improved acoustic models using additional TDT4 data
  - Improved language models (additional texts, improved smoothing)
  - WER reduction of 18% for English and 35% for Mandarin
  - CBS and CTS data are much more challenging than Mainland data

- Design of dev03 set for English
- Dev03 data are good indicators of eval performance