Development of the MIT ASR System for the 2016 Arabic Multi-Genre Broadcast Challenge

Tuka AlHanai, Wei-Ning, and James Glass

Spoken Language Technologies Workshop
Tuesday 18th October 2016
Multi-Genre Broadcast (MGB) Challenge

• Started in 2015 with ASRU for English (BBC) broadcast news data (~1,600 hours)

• Evaluate:
  • Speech-to-text transcription of broadcast television
  • Alignment of audio to transcript
Arabic MGB Data

• 10 years of Al-Jazeera News Channel programming (2005-2015)
• 1,200 hours transcribed audio
• 8M words (200K Vocab)
• 375K utterances
• 10 hour development set
• 10 hour unseen test set for evaluation

Extra Text

• 120M Words (1.4M Vocab)
• 1.75% Out-of-Vocabulary (OOV) rate
Motivation

- 300 million speakers: diverse set of dialects
- Speech technologies need to accommodate this diversity
Pipeline

- Lexical Model
- Language Model
- Acoustic Model
- Recognizer
- Audio
- Transcription
- 3gram
- DNN
- DNN - mbr
- CNN
- TDNN
- (H/G)-LSTM
- 4gram
- RNNLM
- Rescoring
- GMM

Diagram of a pipeline for speech recognition, showing the flow from audio input through to transcription.
Methods

Acoustic Modeling:
• Feed-forward Neural Networks (DNN)
• Time-Delay Neural Networks (TDNN)
• Convolutional Neural Networks (CNN)
• Recurrent Neural Networks (RNN)
  • Long-short Term Memory (LSTM)
  • Highway-LSTM (H-LSTM)
  • Grid-LSTM (G-LSTM)
• Various Objective Functions
  • Cross-Entropy (CE)
  • Minimum Phone Error (MPE)
  • Minimum Bayes Risk (MBR)
  • Lattice-Free Maximum Mutual Information (LF-MMI)
Methods

Toolkits
• Kaldi Speech Recognition
• CTNK
• SRILM

Features
• 39-dim MFCC + LDA + MLLT + fMLLR (GMM-HMM)
• 30/80 Mel-filterbanks + pitch (DNN)

Language Modeling
• 3-gram with Kneser-Ney Smoothing
• 4-gram rescoring with MGB + Extra Text
• RNN
  • 1000 hidden units + Hierarchical Softmax
  • 300 hidden units + Noise Contrastive Error (NCE) Criterion
Methods

Model Combination
• Lattice combination and hypothesis scoring using Minimum Bayes Risk (MBR)

Evaluation
• Word Error Rate (WER)
• Significance Testing using Matched Pair Sentence Segment Word Error (MAPSSSWE)
Main Contribution

• Applied a range of Neural Network topologies under a single setup.
  • Feed-forward, CNN, LSTM
  • Newer: TDNN, LF-MMI Criterion, Highway-LSTM
• One of the first applications of Grid-LSTM to speech.
Time-Delay Neural Network

LSTM Models

LSTM

Highway-LSTM

Grid-LSTM
## Results (Dev)

<table>
<thead>
<tr>
<th>Model</th>
<th>Topology</th>
<th>Alignment</th>
<th>WER (%)</th>
<th>( p &lt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN CE</td>
<td>5x1024</td>
<td>GMM</td>
<td>28.1</td>
<td>-</td>
</tr>
<tr>
<td>CNN</td>
<td>4x2000</td>
<td>GMM</td>
<td>28.1</td>
<td>0.734</td>
</tr>
<tr>
<td>TDNN</td>
<td>6x3000</td>
<td>GMM</td>
<td>25.8</td>
<td>0.001</td>
</tr>
<tr>
<td>DNN MPE</td>
<td>5x1024</td>
<td>CE</td>
<td>24.7</td>
<td>0.001</td>
</tr>
<tr>
<td>TDNN LF-MMI</td>
<td>7x625</td>
<td>GMM</td>
<td>23.4</td>
<td>0.001</td>
</tr>
<tr>
<td>LSTM</td>
<td>3x1024</td>
<td>CE</td>
<td>22.7</td>
<td>0.001</td>
</tr>
<tr>
<td>H-LSTM 3L</td>
<td>3x1024</td>
<td>CE</td>
<td>22.6</td>
<td>0.250</td>
</tr>
<tr>
<td>H-LSTM 5L</td>
<td>5x1024</td>
<td>CE</td>
<td>22.4</td>
<td>0.184</td>
</tr>
<tr>
<td>G-LSTM 3L</td>
<td>3x1024</td>
<td>CE</td>
<td>21.7</td>
<td>0.001</td>
</tr>
<tr>
<td>G-LSTM 5L</td>
<td>5x1024</td>
<td>CE</td>
<td>21.5</td>
<td>0.070</td>
</tr>
<tr>
<td>G-LSTM 3L sMBR</td>
<td>3x1024</td>
<td>CE</td>
<td>19.5</td>
<td>0.001</td>
</tr>
<tr>
<td>G-LSTM 5L sMBR</td>
<td>5x1024</td>
<td>CE</td>
<td>19.2</td>
<td>0.034</td>
</tr>
<tr>
<td><strong>Top 2 Combined</strong></td>
<td><strong>G-LSTM sMBR (3L + 5L)</strong></td>
<td><strong>CE</strong></td>
<td><strong>18.3</strong></td>
<td><strong>0.001</strong></td>
</tr>
</tbody>
</table>
# Final Results (Test)

<table>
<thead>
<tr>
<th>Team</th>
<th>WER (%)</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eqra</td>
<td>56.8</td>
<td>58.5</td>
</tr>
<tr>
<td>Univ of Seville</td>
<td>55.0</td>
<td>57</td>
</tr>
<tr>
<td>Cairo University</td>
<td>43.3</td>
<td>45.6</td>
</tr>
<tr>
<td>NHK</td>
<td>34.7</td>
<td>39.1</td>
</tr>
<tr>
<td>China National Digital Switching</td>
<td>24.4</td>
<td>29.5</td>
</tr>
<tr>
<td>MIT</td>
<td>23.7</td>
<td>26.2</td>
</tr>
<tr>
<td>LIUM</td>
<td>23.0</td>
<td>25.5</td>
</tr>
<tr>
<td>QCRI</td>
<td>21.1</td>
<td>23.7</td>
</tr>
</tbody>
</table>

### Model Details:
- **G-LSTM sMBR (3L + 5L) + 4-gram LM**
- **LF-MMI TDNN + DNN\(^{BN}\)**
- **LF-MMI BLSTM + 4-gram RNNLM**
Conclusions

- Models that capture temporal context are superior – LSTM
- TDNN outperformed DNN CE
  - Captures wider temporal context

Areas for improvement

- Parameter Tuning
- RNNLM