The NICT ASR System for IWSLT 2012

H. Yamamoto, Y. Wu, C. Huang, X. Lu, P. R. Dixon, S. Matsuda, C. Hori and H. Kashioka

2012.12.06
Task of the ASR track

- Automatic transcription of TED Talks
  - TED Talks: A collection of lectures on www.ted.com
  - Spoken in English
  - Spontaneous style
  - Various topics: Technology, Entertainment, Design
  - Non-speech: laugh, applause, music
Our motivation

### IWSLT 2011

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT</td>
<td>15.3</td>
</tr>
<tr>
<td>KIT</td>
<td>17.1</td>
</tr>
<tr>
<td>LIUM</td>
<td>17.4</td>
</tr>
<tr>
<td>FBK</td>
<td>18.2</td>
</tr>
<tr>
<td>NICT</td>
<td>27.3</td>
</tr>
</tbody>
</table>

[Federico+2011]

### IWSLT 2012

- Catch up in straightforward way
- and tackle new problems
- 12 pt. behind!
- First challenge on English LVCSR

The NICT ASR System for IWSLT 2012
What we did

• Acoustic modeling
  – Spontaneous style speaking
  – Recording environment
  – Non-speech (noise, music)
  – Speaker switching
  – Non-native speaker

• Language modeling
  – Spontaneous style sentence
  – Variety of topics

Collect audio of the target condition
Train and combine two types of AMs
Important, but look small relatively
Extend RNNLM to use multiple features
Adapt N-gram to domain & topic
System Overview

• WFST-based two-pass decoding

Two Passes
1. Domain adapted LM
2. Topic adapted LM

Steps in each pass
1. Decoding w/ N-gram LM
2. System Combination
3. Rescoring w/ fRNN LM
AM: Training corpus

- TED Talks AM ← crawl movies and subtitles
- Prepare text-aligned speech segments
  - Take accurate time stamps, remove noisy parts

The NICT ASR System for IWSLT 2012

0 50 100 150
Corpus Size (hours)

WER (%) 25 30 35
Test on tst2011

204 hours
170 hours
1.6M words
1.8M words

Repeat for unaligned audio and text segments
Adapt at each iteration

SailAlign [Katsamanis+2011]
AM: Modeling

- Spontaneous speech AM
  - Cross-word triphone HMM
  - Two types of output pdf., GMM and SGMM

Kaldi [Povey+2011]

Training

- GMM ML
- GMM Boosted MMI
- Subspace GMM ML
  - 6.7K states
  - 35 mix (ave.)

Front-end

- Feature Extraction
- fMLLR (SAT)
- 9-frame x 13-dim MFCC
  - 40-dim by LDA, MLLT
- better than Δ+ΔΔ, HLDA

Show a different trend
 → System combination
LM: N-gram

- Data selection for LM adaptation
  - Corpora: TED\textit{(in-domain)}, Gigaword\textit{(out-of-domain)}
  - Cross-entropy difference metrics [Moore+2010]

The NICT ASR System for IWSLT 2012
LM: Factored RNNLM [Wu+2012]

• Incorporate multiple features into RNNLM

RNNLM [Mikolov+2010]

Word: difference between developed countries and developing countries

\[ P(\text{and} | h) \]

Recurrent Neural Network

Enable LM to consider long (entire) history
Incorporate multiple features into RNNLM

Factored RNNLM

Word: difference | between | developed
Lemma: difference | between | developed
Stem: differ | between | developed
Part-of-speech: NN | IN | JJ

Input “countries”

Surface

PoS

f1(t)

f2(t)

h(t)

h(t-1) → h(t)

Weight intensity

f1: word → h(t)

f2: Part-of-Speech → h(t)

Enable LM to utilize rich features
### Experimental Results

- **WER(%) of our transcriptions.**

<table>
<thead>
<tr>
<th>Step</th>
<th>1(^{st}) pass</th>
<th>2(^{nd}) pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1: GMM</td>
<td>12.3</td>
<td>11.8</td>
</tr>
<tr>
<td>Step 1: SGMM</td>
<td>12.9</td>
<td>12.5</td>
</tr>
<tr>
<td>Step 2: Comb.</td>
<td>12.0</td>
<td>11.5</td>
</tr>
<tr>
<td>Step 3: fRNN</td>
<td>10.9</td>
<td>10.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>1(^{st}) pass</th>
<th>2(^{nd}) pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>tst2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tst2012</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Each step contributed to reduce error
- Topic adaptation also worked
- After the submission, fRNNLM adaptation achieved 11.9%.
• **SprinTra decoder** [Dixon+2012]
  – On-the-fly WFST composition scheme.
  – Low computing resource, no degradation in WER.

<table>
<thead>
<tr>
<th></th>
<th>Computing Time</th>
<th>Memory Usage (max.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Building</td>
<td>Decoding</td>
</tr>
<tr>
<td>Kaldi</td>
<td>17.3h</td>
<td>3%</td>
</tr>
<tr>
<td>3g</td>
<td>0.5h</td>
<td>RTF:18.6</td>
</tr>
<tr>
<td>SprinTra</td>
<td>3%</td>
<td>48%</td>
</tr>
<tr>
<td>4g</td>
<td>RTF:8.9</td>
<td></td>
</tr>
</tbody>
</table>

(Xeon 2.67GHz)

The NICT ASR System for IWSLT 2012
• Decoding speed is also a key issue of ASR
  • To response quickly for online ASR (e.g. Closed captioning)
  • To process tons of data for offline ASR (e.g. Audio indexing)

![Graph showing WER vs. RTF](image)

- **RTF: Faster is better**
- **WER: Lattice has potential**
- **Small fall at RTF~1**

The NICT ASR System for IWSLT 2012
Thanks for your kind attention!
References


