Browsing Oral History

Phil Green, Oscar Saz, Jonathan Kilgour, George Powell, Trudy Pankhurst Green, Mark Gales, Pierre Lanchantin, Yanmin Qian
Oral History

- Captures the personal and historical experiences of individuals
- Many collections:
  - English Heritage
  - Heritage Lottery (funded 2000)
  - British Library
- Memories, reminiscences
- Long Interviews
- No central database..
- Utility limited by retrieval problems.. *The Deep Dark Secret of oral history ...that nobody spends much time listening to or watching recorded and collected interview documents’* (Frisch (2008)}
Oral History Collections

• Topic-focussed (but topic may be very wide)
• Data back to the 1990s (at least)
• Variable recording quality
• Metadata
  • Typically summary of an interview
  • Occasionally a transcription
• OOV problems .. Use the summary
Oral History recordings .. WWII experiences
• around 25 interviewees
• ~50 hours
Using Speech Technology to Browse Oral History

• ASR Transcription as a browsing tool
• Link Audio, summary and ASR
• Browsing Tool for Duty Calls (‘System1’)
  • Transcribed with a Sheffield MGB system
  • Speech activity detection based on DNNs
  • 4 independent speaker adapted systems (2 DNN-HMM and 2 DNN-GMM-HMM) trained on 700 hours of BBC broadcasts
  • Lexicon of 50,000 words
  • 4-gram language model and RNN rescoring, trained on 700 million words
• Demo on http://brodsworthhall.azurewebsites.net
• 3 interviews .. with permission
Summary

Search Summary

SummaryForGarnett-Sheila_04-08-2010.docx

00:12:48 Description of Woodlands
00:05:9 Role of Miners' Welfare
00:27:7 Humbolt? Family
00:08:6 Opening Percy Jackson Grammar School delayed because of outbreak of war
00:24:58 How family came to new estate
00:09:0 Father's job at the colliery
Recognisers for Duty Calls

System 1
- Sheffield MGB system
- Speech activity detection based on DNNs
- 4 independent speaker adapted systems (2 DNN-HMM and 2 DNN-GMM-HMM) trained on 700 hours of BBC broadcasts
- Lexicon of 50,000 words
- 4-gram language model and RNN rescoring, trained on 700 million words

System 2
- SI Hybrid system,
- sequence trained 700hrs trained on MGB data.
- LM data from MGB systems (includes subtitle data).
- IBM KWS
Duty Calls: Trial Corpus

Ground Truth: 8 interviews manually transcribed

<table>
<thead>
<tr>
<th></th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>33'39”</td>
<td>1h13'</td>
<td>16'29”</td>
<td>47'54”</td>
<td>1h43'</td>
<td>12'47”</td>
<td>39'42”</td>
<td>21'38”</td>
<td>5h49'</td>
</tr>
<tr>
<td>Segments</td>
<td>666</td>
<td>1,965</td>
<td>259</td>
<td>1,151</td>
<td>2,203</td>
<td>335</td>
<td>949</td>
<td>392</td>
<td>7,919</td>
</tr>
<tr>
<td>Words</td>
<td>4,605</td>
<td>12,399</td>
<td>1,992</td>
<td>6,647</td>
<td>18,608</td>
<td>2,041</td>
<td>5,843</td>
<td>2,563</td>
<td>54,698</td>
</tr>
</tbody>
</table>
Duty Calls: using the Summaries

7 interviews contain a summary

<table>
<thead>
<tr>
<th>Entity</th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>544</td>
<td>0</td>
<td>663</td>
<td>1,600</td>
<td>891</td>
<td>505</td>
<td>370</td>
<td>853</td>
<td>5,426</td>
</tr>
</tbody>
</table>

A set of relevant named entities (e.g. “Hooton Pagnell”) were manually identified from the summaries, and matched to spoken occurrences in the audio.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>41</td>
<td>0</td>
<td>23</td>
<td>23</td>
<td>74</td>
<td>5</td>
<td>29</td>
<td>45</td>
<td>240</td>
</tr>
</tbody>
</table>

Many of the named entities are not in the vocabulary (~40-50%). When they are, they are underrepresented in the LM training data (either n-gram or RNN).
Duty Calls: finding named entities

Evaluation of WER and Recall of named entities using two different systems
WER over 30%, recall 35%-40% with a high precision (100%-94%)

<table>
<thead>
<tr>
<th>System 1</th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>31.2%</td>
<td>34.0%</td>
<td>19.0%</td>
<td>43.9%</td>
<td>35.2%</td>
<td>46.2%</td>
<td>27.5%</td>
<td>33.6%</td>
<td>34.4%</td>
</tr>
<tr>
<td>Recall</td>
<td>24.4%</td>
<td>N/A</td>
<td>73.9%</td>
<td>26.1%</td>
<td>36.5%</td>
<td>20.0%</td>
<td>51.7%</td>
<td>20.0%</td>
<td>35.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System 2</th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>37.9%</td>
<td>35.8%</td>
<td>24.0%</td>
<td>45.6%</td>
<td>33.9%</td>
<td>51.1%</td>
<td>33.0%</td>
<td>40.4%</td>
<td>36.3%</td>
</tr>
<tr>
<td>Recall</td>
<td>24.4%</td>
<td>N/A</td>
<td>73.9%</td>
<td>39.1%</td>
<td>50.0%</td>
<td>0.0%</td>
<td>55.2%</td>
<td>20.0%</td>
<td>40.8%</td>
</tr>
</tbody>
</table>
Duty Calls: adapting the Language Model from the Summary

Lexicon and language model adaptation using the interview summary:

- A new lexicon is created for each summary adding words from the summary
- A new 4-gram is trained interpolating the baseline LM with the summary n-gram
- RNN is fine tuned using the summary

Small WER improvement, but recall increased to >60%

<table>
<thead>
<tr>
<th>System 1</th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>31.6%</td>
<td>N/A</td>
<td>17.3%</td>
<td>42.3%</td>
<td>35.1%</td>
<td>42.2%</td>
<td>28.6%</td>
<td>31.4%</td>
<td>34.1%</td>
</tr>
<tr>
<td>Recall</td>
<td>48.8%</td>
<td>N/A</td>
<td>87.0%</td>
<td>52.2%</td>
<td>68.9%</td>
<td>40.0%</td>
<td>69.0%</td>
<td>62.2%</td>
<td>63.8%</td>
</tr>
</tbody>
</table>

Precision 92%
LM Adaptation Across Summaries

Lexicon and language model adaptation using a pool of all summaries:

<table>
<thead>
<tr>
<th>System 1</th>
<th>Sheila Coates</th>
<th>Jean Covell</th>
<th>Joyce Durdy</th>
<th>Ernest Egginton</th>
<th>David Gilling</th>
<th>Sheila Miles</th>
<th>Jean Paton</th>
<th>Madge Rouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>30.7%</td>
<td>34.8%</td>
<td>16.2%</td>
<td>42.7%</td>
<td>34.3%</td>
<td>42.4%</td>
<td>27.7%</td>
<td>27.7%</td>
<td>33.6%</td>
</tr>
<tr>
<td>Recall</td>
<td>51.2%</td>
<td>N/A</td>
<td>91.3%</td>
<td>39.1%</td>
<td>68.9%</td>
<td>20.0%</td>
<td>58.6%</td>
<td>82.2%</td>
<td>65.4%</td>
</tr>
<tr>
<td>Precision</td>
<td>92%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Precision 92%
Duty Calls: Keyword Spotting from Lattices

A different operating point can be chosen from the DET curves (keyword spotting based on lattices)

- Recall up to 40-50%
Duty Calls: adding pronunciations

Recall can be increased to 60% by generating pronunciations for the missing words and searching for them in lattices.
Conclusions

• ASR can be used as the basis for good browsing tools for Oral History
• Using summaries to tune language models is highly effective
• Keyword Spotting from lattices is also effective

Future Work

• Adapt acoustic models for individual speakers
• Links between interviews and between collections
• Add value to the summaries