Dialogue Tuning
Power-ups

Presented by:
Dominique Boucher
CTO, Nu Echo

Tuesday, May 24th, 2016
Why tuning?

Improperly tuned speech apps tuned have a negative impact on your business

Lower completion rates
Customer dissatisfaction
Still a manual activity

Requires experience and expertise

Requires the mining of thousands of utterances, in search of tuning opportunities
Still a manual activity

What if we could decrease this time from hours to minutes?
Company

Solution Providers & Consultants

Services

- Telephony infrastructure & contact center system selection
- System architecture, implementation and deployment
- System performance testing and quality assurance
- Speech recognition tuning and development
- Solutions design and or customization
Dialogue tuning

What is speech tuning? What are we trying to achieve?

- Optimizing recognition accuracy and setting confidence thresholds

Focus is mainly on speech tuning

- Properly setting thresholds also optimizes dialogue
- Gives hints on dialogue optimization through call analysis
A disciplined approach

Clear goals

A rigorous tuning process

Tools and platform supporting the process

Analytics and visualization tools to quickly discover tuning opportunities
The tuning process

1. Recognition Results
2. Score Results
3. Semantic References
4. Generate Semantic References
5. Transcribed Utterances
6. Transcribe
7. Field Utterances
8. Modify Grammars/Dictionaries
9. Identify Improvements
10. Perform Speech Recognition

nuecho
The tuning platform

Goals of an effective tuning/benchmarking environment:

- Experiments are inexpensive to run
- Experiments are repeatable
- Simple method for experiment comparison
- Lab results reproduce ones from the field
- Effective visualization and analysis tools!
The importance of sound metrics

The “In-Grammar” example - two possible definitions

**Grammar-centric**
Transcription covered by the grammar

**User-centric**
Utterance contains a valid response to the application's prompt

Only the user-centric definition makes sense!
Tagged utterances

Manual annotations

- Transcription
- Reference interpretation
- OOG flag

Automatic annotations

- Initial transcription
- Recognition results
- Call info (time, date)
- Application attributes (dialogue state, previous state, ...)
- Grammars
- Audio attributes
- User attributes (age, gender, sentiment, ...)
- Others
# Tagged utterances

Various data types

<table>
<thead>
<tr>
<th>Type</th>
<th>Example tags</th>
<th>Example values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free-form text</td>
<td>Transcription</td>
<td>“Forty nine ninety five”</td>
</tr>
<tr>
<td></td>
<td>Dialogue state name</td>
<td>“confirm-departure-date”</td>
</tr>
<tr>
<td>Enumerations</td>
<td>Scoring category</td>
<td>“correct” / “incorrect” / “oog”</td>
</tr>
<tr>
<td></td>
<td>Age category (teen / adult / elderly)</td>
<td>“teen” / “adult” / “elder”</td>
</tr>
<tr>
<td></td>
<td>Weekday</td>
<td>“Monday” / “Tuesday” / … / “Sunday”</td>
</tr>
<tr>
<td>JSON values</td>
<td>[Reference] Semantic Value</td>
<td><code>{ “dollarAmount”: 49, “cents”: 95 }</code></td>
</tr>
<tr>
<td>Boolean</td>
<td>Accepted</td>
<td>true</td>
</tr>
<tr>
<td>Numbers</td>
<td>Confidence score</td>
<td>0.61</td>
</tr>
</tbody>
</table>
What we do with those tags?

**First approach**
- Partition data using tags
- Create reports
- Identify tuning opportunities

**Challenges**
- Lots of data, many dimensions!
- Time consuming!
The power-ups
The power-ups

Essential tools to detect dialogue/speech problems

- **Visualization**
  - Help understand problems in their context
  - Find trends

- **Automatic problem detection**
  - Automate the reporting process
The power-ups

Dialogue Visualization

- Shows callers behavior
- Hotspot indicators

Mean Confidence score < 0.55

Miscommunication errors > 30%
The power-ups

Dialogue Visualization

- State statistics
- Path statistics
# The power-ups

**Single/multiple tag partitions reports**

- Automate the selection of partition types and the reporting process
- Guide speech scientists in best tuning opportunities

<table>
<thead>
<tr>
<th>Single tag partitions report</th>
<th>Multiple tag partitions report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect tag values associated with poor performance metrics</td>
<td>Detect abnormal variations among partitions based on some context</td>
</tr>
</tbody>
</table>

nuecho 😄
Tag partitions reports: A real-world example

Tuning project for a large utility in Canada

Call analysis & tuning of hotspots

- Pronunciation problems for some acronyms
- Pronunciation problems for numbers multiples of 100 in French

Several days of call analysis to detect problems
Tag partitions reports: A real-world example

Running the single tag partitions report

<table>
<thead>
<tr>
<th>Tag Name</th>
<th>Tag Value</th>
<th>Tag Count</th>
<th>AC/ing</th>
<th>(AI+Aoog)/all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Value (Reference)</td>
<td>{institution=&quot;TD&quot;}</td>
<td>370</td>
<td>56.76%</td>
<td>2.43%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{institution=&quot;CIBC&quot;}</td>
<td>173</td>
<td>57.23%</td>
<td>6.94%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;50000&quot;}</td>
<td>90</td>
<td>67.78%</td>
<td>7.78%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;40000&quot;}</td>
<td>63</td>
<td>41.27%</td>
<td>20.63%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;70000&quot;}</td>
<td>54</td>
<td>57.41%</td>
<td>11.11%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;10000&quot;}</td>
<td>50</td>
<td>46.00%</td>
<td>16.00%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;60000&quot;}</td>
<td>50</td>
<td>44.00%</td>
<td>24.00%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;20000&quot;}</td>
<td>50</td>
<td>66.00%</td>
<td>12.00%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;30000&quot;}</td>
<td>48</td>
<td>47.92%</td>
<td>20.83%</td>
</tr>
<tr>
<td>Semantic Value (Reference)</td>
<td>{result=&quot;80000&quot;}</td>
<td>47</td>
<td>46.81%</td>
<td>14.89%</td>
</tr>
</tbody>
</table>

8,000 utterances
Count >= 30
Recog. Rate < 80%

Setup time: 2 mins!
Multiple tag partitions report

Compare partitions based on some context

- “Reference Semantic Value” vs “Dialogue State”
- “Reference Semantic Value” vs “Hour of day”
- Etc.
## Tag partitions reports: A real-world example

### Running the multiple tag partitions report

<table>
<thead>
<tr>
<th>Primary Tag Values</th>
<th>Tag Count</th>
<th>AC/ing</th>
<th>(AC/ing)</th>
<th>(AI+Aoog)/all</th>
<th>[(AI+Aoog)/all]</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;track=&quot;</td>
<td>2302</td>
<td>56.39%</td>
<td>[-75.40%, -69.80%]</td>
<td>39.27%</td>
<td>[-64.85%, -58.05%]</td>
</tr>
<tr>
<td>&quot;info=&quot;</td>
<td>266</td>
<td>90.23%</td>
<td>[-11.42%, 3.94%]</td>
<td>9.02%</td>
<td>[-8.99%, 5.56%]</td>
</tr>
<tr>
<td>&quot;ship=&quot;</td>
<td>360</td>
<td>90.00%</td>
<td>[-10.70%, 3.44%]</td>
<td>9.17%</td>
<td>[-11.80%, 2.12%]</td>
</tr>
<tr>
<td>&quot;track_a_package&quot;</td>
<td>571</td>
<td>88.09%</td>
<td>[-10.58%, 1.99%]</td>
<td>1.75%</td>
<td>NA</td>
</tr>
<tr>
<td>&quot;representative&quot;</td>
<td>995</td>
<td>87.64%</td>
<td>[-9.91%, 6.62%]</td>
<td>0.90%</td>
<td>NA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary Tag Values</th>
<th>Tag Count</th>
<th>AC/ing</th>
<th>(AC/ing)</th>
<th>(AI+Aoog)/all</th>
<th>[(AI+Aoog)/all]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue State</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPSRequestTrackingNumber</td>
<td>1488</td>
<td>82.06%</td>
<td>[69.80%, 75.40%]</td>
<td>17.54%</td>
<td>[58.05%, 64.85%]</td>
</tr>
<tr>
<td>UPSCollectTrackingNumber</td>
<td>814</td>
<td>9.46%</td>
<td>[-75.40%, -69.80%]</td>
<td>78.99%</td>
<td>[-64.85%, -58.05%]</td>
</tr>
</tbody>
</table>
### Tag partitions reports: A real-world example

#### Running the multiple tag partitions report

<table>
<thead>
<tr>
<th>Hour of Day</th>
<th>Tag Count</th>
<th>AC/ing</th>
<th>[AC/ing]</th>
<th>(Al+Aoog)/all</th>
<th>[(Al+Aoog)/all]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance Hour of Day: 10</td>
<td>169</td>
<td>50.00%</td>
<td>[-17.60%, -2.01%]</td>
<td>46.75%</td>
<td>[-18.32%, -2.78%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 11</td>
<td>164</td>
<td>52.44%</td>
<td>[-14.76%, 1.04%]</td>
<td>43.90%</td>
<td>[-15.33%, 0.36%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 12</td>
<td>188</td>
<td>54.79%</td>
<td>[-11.78%, 3.00%]</td>
<td>40.96%</td>
<td>[-11.68%, 2.92%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 13</td>
<td>191</td>
<td>53.68%</td>
<td>[-13.25%, 1.46%]</td>
<td>39.79%</td>
<td>[-10.33%, 4.10%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 14</td>
<td>194</td>
<td>47.94%</td>
<td>[-19.14%, -4.52%]</td>
<td>43.81%</td>
<td>[-14.74%, -0.23%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 15</td>
<td>201</td>
<td>52.24%</td>
<td>[-14.92%, -0.53%]</td>
<td>43.78%</td>
<td>[-14.61%, -0.34%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 16</td>
<td>181</td>
<td>59.12%</td>
<td>[-7.15%, 7.73%]</td>
<td>36.46%</td>
<td>[-6.80%, 7.76%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 17</td>
<td>167</td>
<td>60.00%</td>
<td>[-6.59%, 8.80%]</td>
<td>37.72%</td>
<td>[-8.48%, 6.73%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 18</td>
<td>153</td>
<td>62.50%</td>
<td>[-4.49%, 11.40%]</td>
<td>34.64%</td>
<td>[-5.37%, 10.21%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 19</td>
<td>252</td>
<td>66.53%</td>
<td>[2.07%, 14.44%]</td>
<td>31.35%</td>
<td>[0.12%, 12.25%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 20</td>
<td>164</td>
<td>65.85%</td>
<td>[-0.03%, 15.03%]</td>
<td>29.27%</td>
<td>[0.94%, 15.41%]</td>
</tr>
<tr>
<td>Utterance Hour of Day: 21</td>
<td>115</td>
<td>71.30%</td>
<td>[4.56%, 21.56%]</td>
<td>26.09%</td>
<td>[3.09%, 19.60%]</td>
</tr>
</tbody>
</table>
Tag partitions reports: A real-world example

Running the multiple tag partitions report
Other tags that can be leveraged

- Age/Gender detection
- Emotion detection
- Automatic Hotspot detection
In Summary...

- Reduced time in finding the issues allows you to invest in problem resolution.
- IMPORTANT for tight schedules, fixed budgets, other
- RESULT: More problems resolved and more efficient applications
When should you tune your app?

- Just after first deployment
- After conversion to a new speech engine (or new version)
- Regularly!
Thank you!
Questions & Answers

www.nuecho.com