Learning about Voice Search for Spoken Dialogue Systems

Rebecca J. Passonneau¹, Susan L. Epstein²,³, Tiziana Ligorio², Joshua B. Gordon⁴, Pravin Bhutada⁴

¹Center for Computational Learning Systems, Columbia University
²Department of Computer Science, Hunter College of The City University of New York
³Department of Computer Science, The Graduate Center of The City University of New York
⁴Department of Computer Science, Columbia University
Outline

• Introduction: CheckItOut domain
  – Why voice search?

• Motivation
  – A single turn exchange
  – High accuracy to avoid re-prompting

• Experimental infrastructure
  – Wizard ablation method and architecture
  – Experimental design: 4200 book title requests

• Results: Learned models of individual wizards’ actions

• Conclusion
  – What we learned about voice search for SDS
  – Current and future work
• Andrew Heiskell Braille & Talking Book Library
  • Branch of New York City Public Library, and Library of Congress
  • One of first users of Kurzweil reading mach.
• Book transactions by phone
  • Patrons order books by telephone
  • Book orders sent/returned by U.S.P.O.
• CheckItOut dialog system
  • Based on 82 recorded patron/librarian calls
  • Replica of Heiskell Library catalogue (N=71,166)
  • Mockup of patron data for 5,028 active patrons
Voice search: query the backend catalogue with ASR string

• Minimal speech engineering
  – WSJ read speech acoustic models
  – Adaptation with ~12 hours of spontaneous speech
  – 0.49 WER in recent tests

• Take advantage of the domain knowledge to recover from poor WER, especially for book titles

<table>
<thead>
<tr>
<th>ROLL DWELL</th>
<th>Cromwell</th>
<th>0.67</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Robert Lowell</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Road to Wealth</td>
<td>0.50</td>
</tr>
</tbody>
</table>
High Accuracy Voice Search

• Minimize non-understandings/misunderstandings
  – User corrections in both contexts lead to poorer speech recognition (Litman et al., 2006)
  – Users seem to prefer system initiative with explicit confirmation (Litman & Pan, 1999)
  – Usability studies show a preference for mixed-initiative only in lab contexts; in real-world situations mixed-initiative is not sufficiently robust (Turunen et al., 2006)

• Wizard studies with simulated ASR, under high WER
  – High rate of misunderstandings (Williams & Young, 2004)
  – High rate of clarification requests (Rieser et al., 2005)
Challenges for SLU

- **Grammar**
  - 4,000 titles (cf. LREC 2010)
  - ~6,000 words in all sub-grammars (titles, authors, etc.)

- **Long utterances**: 9.1 words on average
  - Average title length: 4.5 words
  - Maximum title length: 40 words

- **Full database**: 71,600 titles

- **Confusability of**:
  - Between authors/titles
  - Among medium length titles
A Single Turn Exchange

• User requests books by title
  – Reads book synopses, orders the list of 20 books
  – Rates correctness of each wizard book offer
  – Rates wizard questions (e.g., answerable?)

• Wizard sees ASR, results of voice search
  – Can offer one of the voice search returns
  – **Or**, ask a question
  – **Or** give up

• Query: Ratcliffe-Obershemp string similarity
  – |Matching characters| / |Total characters|
  – Recursively find longest common subsequence

June 2-4, 2010 NAACL, Los Angeles
Wizard Ablation

• Wizard sees/manipulates modified system data
  – ASR in greyscale reflecting acoustic confidence
  – Three types of db return
    • Singleton list (matches in dark bold): RO $\geq 0.85$
    • Ambiguous list , 2-5 titles (matches in dark bold):
      $0.85 > RO \geq 0.55$
    • Noisy list, 6-10 titles (matches in greyscale bold):
      $0.55 > RO \geq 0.40$

• Machine learning methods to learn wizard actions
  – Linear regression
  – Logistic regression
  – Decision trees
Experimental Design

- 7 participants = 21 distinct pairs
- 20 titles per session
- Participants asked to maximize a session score
  - Winner awarded a prize
  - Wizard: +1 if correct, -1 if incorrect, 0.5 for good quest.
  - User: +0.5 for each correct title
- Two sessions per trial
  - Wizard/user rotate after first session
  - Rotation to encourage cooperation
- 5 trials per pair
- 5 x 2 x 20 x 21 = 4200 title cycles
User GUI

• Titles list
  – Green: correct offer
  – Red: incorrect offer
  – Yellow: in progress
• Responses to wizard questions
  – Can answer
  – Cannot answer
  – Undecided
  – Problem
Wizard GUI

- Display Types
  - Singleton
  - AmbiguousList
  - NoisyList

- Actions
  - Confident offer
  - Tentative offer
  - Question
  - Give up
Learned Models

• 60 initial features curated to 28 (cross-correlation)
  – GUI display type
  – Session features
  – Characteristics of or comparison of ASR and candidates and full DB
  – Recognition/NLU scores

• Models
  – Union of all wizards
  – Subset representing each wizard

• Supervised attribute selection reduced feature set to 8-12 features per decision tree
<table>
<thead>
<tr>
<th></th>
<th>Features</th>
<th></th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Display type</td>
<td>15</td>
<td>Avg. edit distance candidates</td>
</tr>
<tr>
<td>2</td>
<td>Requests to repeat</td>
<td>16</td>
<td>Num. ASR words in db</td>
</tr>
<tr>
<td>3</td>
<td>Title of 20</td>
<td>17</td>
<td>Num. db titles with ASR words</td>
</tr>
<tr>
<td>4</td>
<td>Titles correct</td>
<td>18</td>
<td>Ratio of feat. 9 to feat. 10</td>
</tr>
<tr>
<td>5</td>
<td>Recent titles correct</td>
<td>19</td>
<td>Acoustic model score</td>
</tr>
<tr>
<td>6</td>
<td>ASR length (words)</td>
<td>20</td>
<td>Helios confidence score</td>
</tr>
<tr>
<td>7</td>
<td>Avg. candidate length</td>
<td>21</td>
<td>Phoenix parse score</td>
</tr>
<tr>
<td>8</td>
<td>Avg. ASR word rarity</td>
<td>22</td>
<td>Language model score</td>
</tr>
<tr>
<td>9</td>
<td>Avg. edit distance</td>
<td>23</td>
<td>Num. frames in ASR</td>
</tr>
<tr>
<td>10</td>
<td>Avg. word matches</td>
<td>24</td>
<td>Avg. num. gaps in parse</td>
</tr>
<tr>
<td>11</td>
<td>Length longest match</td>
<td>25</td>
<td>Speaking rate in frames/word</td>
</tr>
<tr>
<td>12</td>
<td>Location longest match</td>
<td>26</td>
<td>Total number of parses</td>
</tr>
<tr>
<td>13</td>
<td>Max. gap size btw. matches</td>
<td>27</td>
<td>Num. words in parse</td>
</tr>
<tr>
<td>14</td>
<td>Number of candidates</td>
<td>28</td>
<td>Avg. words per parse slot</td>
</tr>
</tbody>
</table>
Distribution of Correct Actions

<table>
<thead>
<tr>
<th>Correct Action</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Return 1</td>
<td>2722</td>
<td>65.2445</td>
</tr>
<tr>
<td>Return 2</td>
<td>126</td>
<td>3.0201</td>
</tr>
<tr>
<td>Return 3</td>
<td>56</td>
<td>1.3423</td>
</tr>
<tr>
<td>Return 4</td>
<td>46</td>
<td>1.1026</td>
</tr>
<tr>
<td>Return 5</td>
<td>26</td>
<td>0.6232</td>
</tr>
<tr>
<td>Return 7</td>
<td>7</td>
<td>0.1678</td>
</tr>
<tr>
<td>Return 8</td>
<td>1</td>
<td>0.0002</td>
</tr>
<tr>
<td>Return 9</td>
<td>2</td>
<td>0.0005</td>
</tr>
<tr>
<td>Speak</td>
<td>Giveup</td>
<td>1186</td>
</tr>
<tr>
<td>Total</td>
<td>4172</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
## Correct Offers vs. Accuracy

<table>
<thead>
<tr>
<th>Particip.</th>
<th>Cycles</th>
<th>Session Score</th>
<th>Acc.</th>
<th>Offered Return 1</th>
<th>Correct Non-Offers</th>
</tr>
</thead>
<tbody>
<tr>
<td>W4</td>
<td>600</td>
<td>0.7585</td>
<td><strong>0.8550</strong></td>
<td>0.70</td>
<td><strong>0.64</strong></td>
</tr>
<tr>
<td>W5</td>
<td>600</td>
<td>0.7584</td>
<td><strong>0.8133</strong></td>
<td>0.76</td>
<td><strong>0.43</strong></td>
</tr>
<tr>
<td>W7</td>
<td>599</td>
<td>0.6971</td>
<td>0.7346</td>
<td>0.76</td>
<td>0.14</td>
</tr>
<tr>
<td>W1</td>
<td>593</td>
<td>0.6936</td>
<td>0.7319</td>
<td>0.79</td>
<td>0.16</td>
</tr>
<tr>
<td>W2</td>
<td>599</td>
<td>0.6703</td>
<td>0.7212</td>
<td>0.74</td>
<td>0.10</td>
</tr>
<tr>
<td>W3</td>
<td>581</td>
<td>0.6648</td>
<td>0.6954</td>
<td>0.81</td>
<td>0.20</td>
</tr>
<tr>
<td>W6</td>
<td>600</td>
<td>0.6103</td>
<td>0.6950</td>
<td>0.86</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Characteristics of Decision Trees

• Larger trees for more accurate wizards: 55 nodes for W4 [best], 7 nodes for W1 [worst]
• 5 features most often in top-level nodes of all trees
  – DisplayType
  – RecentSuccess
  – ContiguousWordMatch (averaged across candidates)
  – NumberOfCandidates
  – Helios confidence score
• Additional important features for W4
  – Number of frames in ASR
  – Acoustic Model Score
Conclusions

• Voice search can lead to high accuracy interpretations of book title requests

• Learning from embedded wizards makes it possible to model wizard actions using system features (e.g., AM score, speech rate, parse features, NLU confidence)

• Dialogue management can profit from more fine-grained representation of spoken language understanding results

• Machine learners should be selective about who to learn from (e.g., W4 and W5)
Current and Future Work

• Same methodology applied to full dialogues
• Focus on feature selection methods tailored to learning dialogue strategies
  – Replace filter method for feature selection with wrapper method
  – Combine heuristic selection with subset selection methods
• Assume DM has access to any level of representation Spoken Language Understanding