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
CheckItOut Domain

- 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
Why Voice Search?

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

| Roll Dwell     | Cromwell | 0.67   |
|----------------|----------|--------|
|                | Robert Lowell | 0.61   |
|                | Road to Wealth    | 0.50   |
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-Obershelf string similarity
  – |Matching characters| / |Total characters|
  – Recursively find longest common subsequence
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
Olympus/RavenClaw Architecture

Apollo
Audio and Interaction Manager

PocketSphinx
Automated Speech Recognition

Kalliope
Text-to-Speech Synthesizer

Phoenix
Semantic Parser
Natural Language Understanding

Helios
Confidence Annotator

RavenClaw
Dialogue Manager

Domain Reasoner

BE_1

BE_2
Olympus/RavenClaw Architecture
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
|   | Feature                                      |   | Feature                                         |
|---|---------------------------------------------|---|------------------------------------------------|
| 1 | Display type                                | 15| Avg. edit distance candidates                    |
| 2 | Requests to repeat                          | 16| Num. ASR words in db                             |
| 3 | Title of 20                                 | 17| Num. db titles with ASR words                    |
| 4 | Titles correct                              | 18| Ratio of feat. 9 to feat. 10                     |
| 5 | Recent titles correct                       | 19| Acoustic model score                             |
| 6 | ASR length (words)                          | 20| Helios confidence score                          |
| 7 | Avg. candidate length                       | 21| Phoenix parse score                              |
| 8 | Avg. ASR word rarity                        | 22| Language model score                             |
| 9 | Avg. edit distance                          | 23| Num. frames in ASR                               |
| 10| Avg. word matches                           | 24| Avg. num. gaps in parse                          |
| 11| Length longest match                        | 25| Speaking rate in frames/word                     |
| 12| Location longest match                      | 26| Total number of parses                           |
| 13| Max. gap size btw. matches                  | 27| Num. words in parse                              |
| 14| Number of candidates                        | 28| Avg. words per parse slot                        |
## Distribution of Correct Actions

| Correct Action       | N   | %      |
|----------------------|-----|--------|
| Return 1             | 2722| 65.2445|
| Return 2             | 126 | 3.0201 |
| Return 3             | 56  | 1.3423 |
| Return 4             | 46  | 1.1026 |
| Return 5             | 26  | 0.6232 |
| Return 7             | 7   | 0.1678 |
| Return 8             | 1   | 0.0002 |
| Return 9             | 2   | 0.0005 |
| Speak|Giveup |1186|28.4276|
| Total                | 4172|1.0000 |
## Correct Offers vs. Accuracy

| Particip. | Cycles | Session Score | Acc.  | Offered Return 1 | Correct Non-Offers |
|-----------|--------|---------------|-------|------------------|--------------------|
| W4        | 600    | 0.7585        | 0.8550| 0.70             | 0.64               |
| W5        | 600    | 0.7584        | 0.8133| 0.76             | 0.43               |
| W7        | 599    | 0.6971        | 0.7346| 0.76             | 0.14               |
| W1        | 593    | 0.6936        | 0.7319| 0.79             | 0.16               |
| W2        | 599    | 0.6703        | 0.7212| 0.74             | 0.10               |
| W3        | 581    | 0.6648        | 0.6954| 0.81             | 0.20               |
| W6        | 600    | 0.6103        | 0.6950| 0.86             | 0.03               |

June 2-4, 2010 NAACL, Los Angeles
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