Introduction – WOZ setup

• Use hidden, human component
• WOZ experimental protocol calls for holding “all other input and output … constant so that the only unknown variable is who does the internal processing” (Paek, 2001)
• WOZ systems appear automated to user
• Gather data for fully-automated system
Introduction – WOZ performance

• Assume user behavior is similar between the WOZ and automated (AUT) setups
• In one system, training with AUT data gave rise to better performance than training WOZ data (Drummond and Litman, 2011)
• System automation differences may have caused performance gap
• Differences in user behavior may weaken automated system performance
Introduction - goal

- Investigate differences in WOZ and AUT user behaviors
- Hypothesized that what users say and how they say it will differ between WOZ and AUT setups
Outline

• **Introduction**
• **Dialogue System**
• **Post-hoc Experiment**
• **Results**
• **Conclusions**
Dialogue System - ITSPOKE

- Our data comes from the Intelligent Tutoring Spoken Dialogue System (ITSPOKE)
- We draw from two prior experiments (one WOZ, one AUT) (Forbes-Riley and Litman(a), 2011; Forbes- Riley and Litman(b), 2011)
- Baseline, non-adaptive conditions of those experiments
- Users tutored in basic Newtonian physics
- Dialogues illustrated one or more basic physics concepts
Dialogue System – sample dialogue

- Tutor text is shown on a screen and read aloud via text-to-speech, and the user responds verbally to the tutor’s queries

<table>
<thead>
<tr>
<th>Tutor</th>
<th>So what are the forces acting on the packet after it’s dropped from the plane?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>um gravity then well air resistance is negligible just gravity</td>
</tr>
<tr>
<td>Tutor</td>
<td>Fine. So what’s the direction of the force of gravity on the packet?</td>
</tr>
<tr>
<td>Student</td>
<td>vertically down</td>
</tr>
</tbody>
</table>
Dialogue System - workflow

Front End

- Microphone
- Automatic Speech Recognition
- Text Display
- Speech Synthesizer

Dialogue Manager

- Human Wizard
- Correctness Evaluation
- Natural Language Understanding

Student

- Screen Display
- Tutor Audio

Next Tutor State

Curr Tutor State
Dialogue System – two user groups

- Setups varied by component for understanding and evaluating responses
  - One human, one automated
- Each student participated in only one setup
  - Students were not informed whether the system was fully automated
- Distinct student group responses constitute data
Outline

• Introduction
• Dialogue System
• Post-hoc Experiment
• Results
• Conclusions
Post-hoc Experiment

- Determine whether differences exist between WOZ and AUT responses
- Compared features of user turns to each question individually
- The table below shows the number of users and dialogue turns they took for each setup over 111 questions asked in both setups

<table>
<thead>
<tr>
<th>System</th>
<th>#Users</th>
<th>#Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOZ</td>
<td>21</td>
<td>1542</td>
</tr>
<tr>
<td>AUT</td>
<td>25</td>
<td>2034</td>
</tr>
</tbody>
</table>
Post-hoc Experiment - features

• Prosodic features: length of the pause before speech began, speech duration, pitch, and energy (RMS)
• Pitch and energy: maximum, minimum, mean, and standard deviation
• 10 total prosodic features
• Normalized each prosodic feature using same algorithm as live system
Post-hoc Experiment - features

• Lexical features: *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker et al., 2001)
  • *Tentative*($T$): “maybe”, “perhaps”, and “guess”
  • *Prepositions*($P$): “to”, “with”, and “above”

• Utterance “Maybe above” would receive feature vector:
  • $<0, \ldots, 0, T=50, 0, \ldots, 0, P=50, 0, \ldots, 0>$

• Used human transcriptions for all utterances
• 69 total LIWC lexical category features
Post-hoc Experiment

• Looked for response feature differences for each question in two ways:
• 1) A statistical comparison of features
• 2) Response classification via machine learning
Outline

• Introduction
• Dialogue System
• Post-hoc Experiment
• Results
  • Statistical Comparison of Features
  • Response Classification Experiments
• Conclusions
Statistical Comparison of Features

• For each question, all features between WOZ and AUT responses were compared
• Welch’s unpaired, two-tailed t-tests
Statistical Comparison of Features

- Possible that differences were inherent in WIZ/AUT student groups.
- Created control groups with evenly mixed, randomly selected WIZ/AUT students.
- We report only questions for which at least one feature differed between WOZ and AUT but not between these two control groups.
Statistical Comparison of Features

• The number of questions for which at least one feature differed statistically significantly ($p < 0.05$) between WOZ and AUT responses

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>#Questions</th>
<th>%Corpus by Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>42</td>
<td>46.22%</td>
</tr>
<tr>
<td>Lexical</td>
<td>33</td>
<td>35.46%</td>
</tr>
<tr>
<td>Either</td>
<td>61</td>
<td>66.86%</td>
</tr>
</tbody>
</table>
Statistical Comparison of Features

- 10/10 prosodic, 29/69 lexical features differed significantly (p < 0.05) for at least one question
- Features differing for at least 10% of the corpus:

<table>
<thead>
<tr>
<th>Feature</th>
<th>%Corpus</th>
<th>#Questions</th>
<th>#WOZ&gt;AUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>22.15%</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>RMS Min</td>
<td>16.86%</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Dictionary Words</td>
<td>15.13%</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>pronoun</td>
<td>12.56%</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>social</td>
<td>11.35%</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>funct</td>
<td>10.99%</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Six Letter Words</td>
<td>10.91%</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>
Statistical Comparison of Features

- Users used more words with the wizarded system.

<table>
<thead>
<tr>
<th>Feature</th>
<th>%Corpus</th>
<th>#Questions</th>
<th>#WOZ&gt;AUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary Words</td>
<td>15.13%</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>pronoun</td>
<td>12.56%</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>social</td>
<td>11.35%</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>funct</td>
<td>10.99%</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

- There exist features which differ for a substantial number of questions.
Statistical Comparison of Features

- A question for which the *Dictionary Words* feature was greater for WOZ responses:

<table>
<thead>
<tr>
<th>Tutor</th>
<th>So how do these two forces’ directions compare?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Most common responses</td>
</tr>
<tr>
<td>WOZ(9) AUT(2)</td>
<td>they are opposite</td>
</tr>
<tr>
<td>WOZ(3) AUT(8)</td>
<td>opposite</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• Dialogue System
• Post-hoc Experiment
• Results
  • Statistical Comparison of Features
  • Response Classification Experiments
• Conclusions
Response Classification Experiments

• Use classification models to distinguish WOZ/AUT setup
• J-48 model was trained and tested for each question
• Accuracy compared against a majority-class baseline
Response Classification Experiments

• 97 questions considered in total
• 21/97 outperformed the majority-class baseline
• 32.79% of the corpus by turns
Response Classification Experiments

- "Would you like to do another problem?"

Y

Duration <= 3.22

N

Word Count <= 1

WOZ (39:1)

AUT (134:39)

AUT (3:0)

WOZ (6:1)

Word Count <= 2
Response Classification Experiments

• This result is consistent with literature (Schechtman and Horowitz, 2003; Rosé and Torrey, 2005) that suggests that users interacting with automated systems will be more curt.
“Now let’s find the forces exerted on the car in the vertical direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?”
Outline

• Introduction
• Dialogue System
• Post-hoc Experiment
• Results
  • Statistical Comparison of Features
  • Response Classification Experiments
• Conclusions
Discussion

• There exist significant differences between user responses to a wizarded and an automatic dialogue system’s questions
• Contribution of the wizard was limited to speech recognition and correctness evaluation
Discussion

- Results suggest that user speech changes as a result of user confidence in the system’s accuracy.
- Relationship between user confidence and user speech may be analogous to observed differences in past experiments.
- These results suggest ways in which raw wizarded data may fall short of ideal for training an automated system.
Future Work - exploration

• Measure how the observed differences change over the course of the dialogue
• Use different methods of normalization for user speech values
Future Work - solutions

• Intentional wizard error could be introduced to frustrate the user; analogous to intentional errors produced in user simulation (Lee and Eskenazi, 2012)

• Generalizable statistical classification domain adaptation (Daumé and Marcu, 2006) and adaptation demonstrated to work well in NLP-specific domains (Jiang and Zhai, 2007)
DIFFERENCES IN USER RESPONSES TO A WIZARD-OF-OZ VERSUS AUTOMATED SYSTEM

Jesse Thomason and Diane Litman
University of Pittsburgh