Validating the Web-based Evaluation of NLG Systems

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Sara Dalzel-Job, Jon Oberlander and Johanna Moore
Overview

• The GIVE Challenge
• The Web-based Experimental Setting
• Evaluation Measures
• The Laboratory Experimental Setting
• Comparative Results
• Conclusions
NLG Evaluation Is Difficult

- Corpus-based evaluation suffers from the more-than-one-correct-answer problem
- Task-based evaluation is time-consuming and expensive
Move forward 2 steps, and then press the blue button.
The Three Worlds
The Web-based Evaluation Setup
Data Gathered

- **GIVE-1**: November 2008 to February 2009
- Five NLG systems
  - University of Texas at Austin; Union College, Schenectady, NY; Universidad Complutense de Madrid, University of Twente $\times 2$
- 1143 valid games by players in 48 countries
  - World 1: 374 games
  - World 2: 369 games
  - World 3: 400 games
GIVE-1 Participation

- **Gender:**
  - 80% male, 10% female, 10% unspecified

- **Source Country (by IP address):**
  - 37% US, 33% Germany, 17% China; 45 other countries

- **English Proficiency (self-reported):**
  - 62% 'expert', 34% native English speakers
Questionnaire

How clear were the directions?
- totally unclear
- n/a, 1, 2, 3, 4, 5
- very clear

How effective were the directions at helping you complete the task?
- not effective
- n/a, 1, 2, 3, 4, 5
- very effective

Did you feel the amount of information you were given was:
- N/A

What is your overall evaluation of the quality of the direction-giving system?
- very bad
- n/a, 1, 2, 3, 4, 5, 6, 7
- very good
Subjective Measures

1. overall evaluation of the quality of the direction-giving system
2. task difficulty
3. goal clarity
4. would you play this game again?
5. instruction clarity
6. instruction helpfulness
7. ease of understanding the system’s choice of wording
8. ease of interpreting referring expressions
9. ease of following navigation instructions
10. friendliness
11. informativity
12. timing
Objective Measures

- percentage of successfully completed games
- for the successfully completed games:
  - number of instructions generated by the NLG system
  - number of actions performed by the user
  - number of steps taken by the user
  - task completion time
Laboratory Experiment

• 91 participants
• Each played five games, one with each NLG system
• We use only the first game run in our comparison
• We used only World 1 (the easiest)
## Comparative Participation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Web-Based</th>
<th>Lab-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td># of participants</td>
<td>322</td>
<td>91</td>
</tr>
<tr>
<td>Gender</td>
<td>80% M, 10% F, 10% Unknown</td>
<td>31% M, 65% F, 4% Unknown</td>
</tr>
<tr>
<td>English proficiency</td>
<td>62% Expert, 34% native English</td>
<td>93% Expert, 81% native English</td>
</tr>
</tbody>
</table>
## Objective Measures

<table>
<thead>
<tr>
<th>Task</th>
<th>Success</th>
<th>Objective Measures</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>Indexes</strong></td>
<td><strong>Steps</strong></td>
<td><strong>Actions</strong></td>
<td><strong>Seconds</strong></td>
<td></td>
</tr>
<tr>
<td>Web</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>91%</td>
<td>A</td>
<td>83.4 B</td>
<td>99.8 A</td>
<td>9.4 A</td>
<td>123.9 A</td>
</tr>
<tr>
<td>M</td>
<td>76%</td>
<td>B</td>
<td>68.1 A</td>
<td>145.1 B</td>
<td>10.0 AB</td>
<td>195.4 BC</td>
</tr>
<tr>
<td>T</td>
<td>85%</td>
<td>AB</td>
<td>97.8 C</td>
<td>142.1 B</td>
<td>9.7 AB</td>
<td>174.4 B</td>
</tr>
<tr>
<td>U</td>
<td>93%</td>
<td>AB</td>
<td>99.8 C</td>
<td>142.6 B</td>
<td>10.3 B</td>
<td>194.0 BC</td>
</tr>
<tr>
<td>W</td>
<td>24%</td>
<td>C</td>
<td>159.7 D</td>
<td>256.0 C</td>
<td>9.6 AB</td>
<td>234.1 C</td>
</tr>
<tr>
<td>Lab</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>100%</td>
<td>A</td>
<td>78.2 AB</td>
<td>93.4 A</td>
<td>9.9 A</td>
<td>143.9 A</td>
</tr>
<tr>
<td>M</td>
<td>95%</td>
<td>A</td>
<td>66.3 A</td>
<td>141.8 B</td>
<td>10.5 A</td>
<td>211.8 B</td>
</tr>
<tr>
<td>T</td>
<td>93%</td>
<td>A</td>
<td>107.2 CD</td>
<td>134.6 B</td>
<td>9.6 A</td>
<td>205.6 B</td>
</tr>
<tr>
<td>U</td>
<td>100%</td>
<td>A</td>
<td>88.8 BC</td>
<td>128.8 B</td>
<td>9.8 A</td>
<td>195.1 AB</td>
</tr>
<tr>
<td>W</td>
<td>17%</td>
<td>B</td>
<td>134.5 D</td>
<td>213.5 C</td>
<td>10.0 A</td>
<td>252.5 B</td>
</tr>
</tbody>
</table>
### Subjective Measures

<table>
<thead>
<tr>
<th></th>
<th>Subjective Measures</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>overall</td>
<td>choice of words</td>
<td>referring expressions</td>
<td>timing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web</td>
<td>A  4.7 A</td>
<td>4.7 A</td>
<td>4.7 A</td>
<td>81%</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M  3.8 AB</td>
<td>3.8 B</td>
<td>4.0 B</td>
<td>70%</td>
<td>ABC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T  4.4 B</td>
<td>4.4 AB</td>
<td>4.3 AB</td>
<td>73%</td>
<td>AB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U  4.0 B</td>
<td>4.0 B</td>
<td>4.0 B</td>
<td>51%</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W  3.8 AB</td>
<td>3.8 B</td>
<td>4.2 AB</td>
<td>50%</td>
<td>BC</td>
<td></td>
</tr>
<tr>
<td>Lab</td>
<td>A  5.7 A</td>
<td>4.7 A</td>
<td>4.8 A</td>
<td>92%</td>
<td>A B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M  5.4 A</td>
<td>3.8 B</td>
<td>4.3 A</td>
<td>95%</td>
<td>A B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T  4.9 A</td>
<td>4.5 A B</td>
<td>4.4 A</td>
<td>64%</td>
<td>A B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U  5.7 A</td>
<td>4.7 A</td>
<td>4.3 A</td>
<td>100%</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W  5.0 A</td>
<td>4.5 AB</td>
<td>4.0 A</td>
<td>100%</td>
<td>B</td>
<td></td>
</tr>
</tbody>
</table>
Summary of Results

• 170 possible significant differences (17 measures × 10 pairs of systems)
  – Laboratory experiment found 6 that the Web-based experiment didn’t
  – Web-based experiment found 26 that the lab-based experiment didn’t

• All pairwise rankings are consistent across both evaluations
Differences

• Completion times in lab-based experiment higher
  – Gender distribution markedly different; and women took longer ← gender differences explain completion times?

• Success rates in lab-based experiment higher
  – Different language proficiencies ← explains lower task success rate on the web?

• Internet data skewed by tendency of unsuccessful participants not to fill in the questionnaire
  – Unsuccessful participants grade systems lower
Conclusions

• Evidence that web-based evaluation is safe!
• Consistent significance judgements in both settings
• More differences found as a consequence of more data
• Absolute values are likely due to demographic differences
  – Not a negative: online usage is arguably more reflective of ‘real life’ usage with laboratory artefacts