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

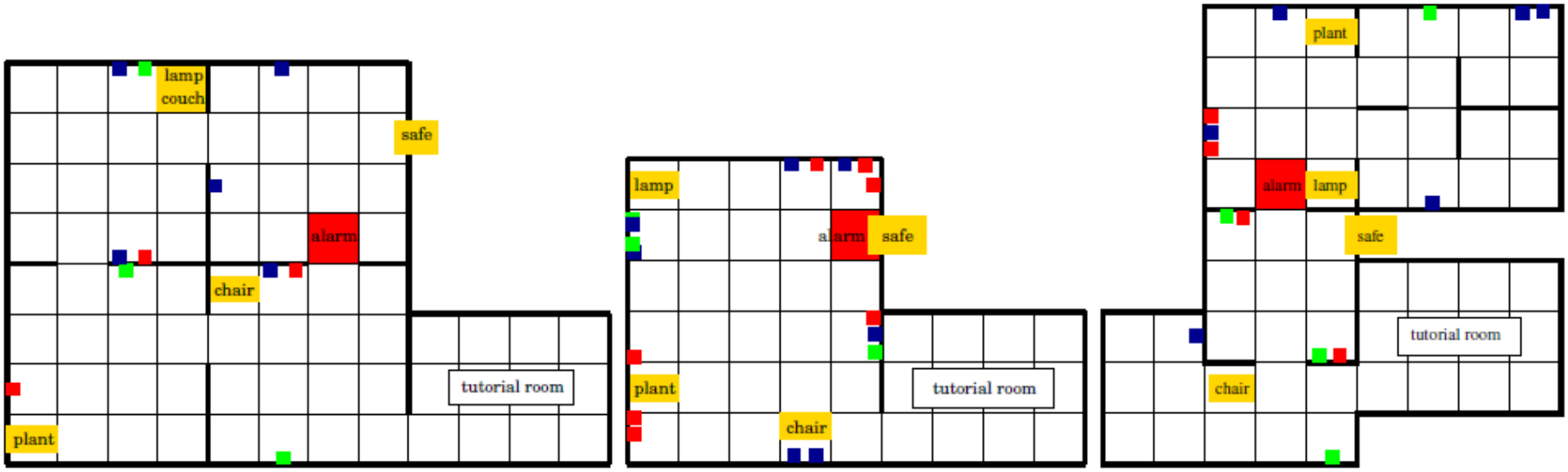
- The GIVE Challenge
- The Web-based Experimental Setting
- Evaluation Measures
- The Laboratory Experimental Setting
- Comparative Results
- Conclusions

NLG Evaluation Is Difficult

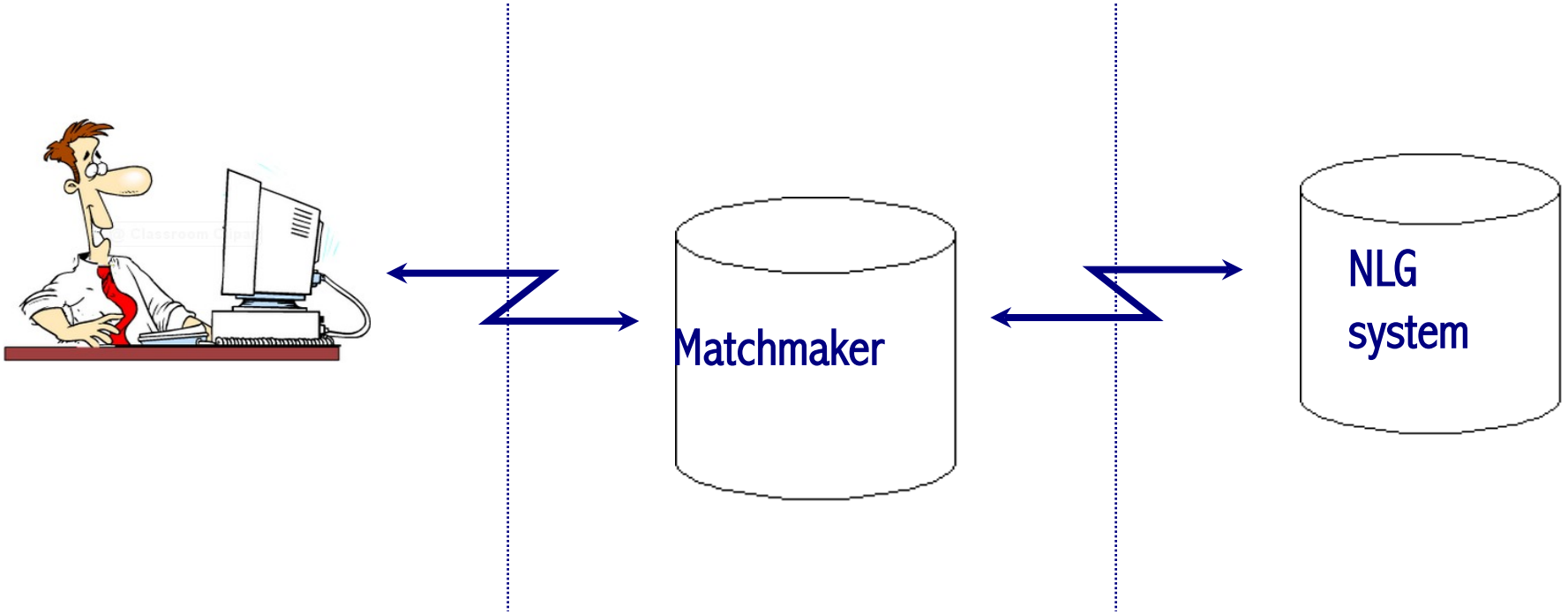
- Corpus-based evaluation suffers from the more-than-one-correct-answer problem
- Task-based evaluation is time-consuming and expensive

GIVE:
Generating Instructions in Virtual Environments

The Three Worlds



The Web-based Evaluation Setup



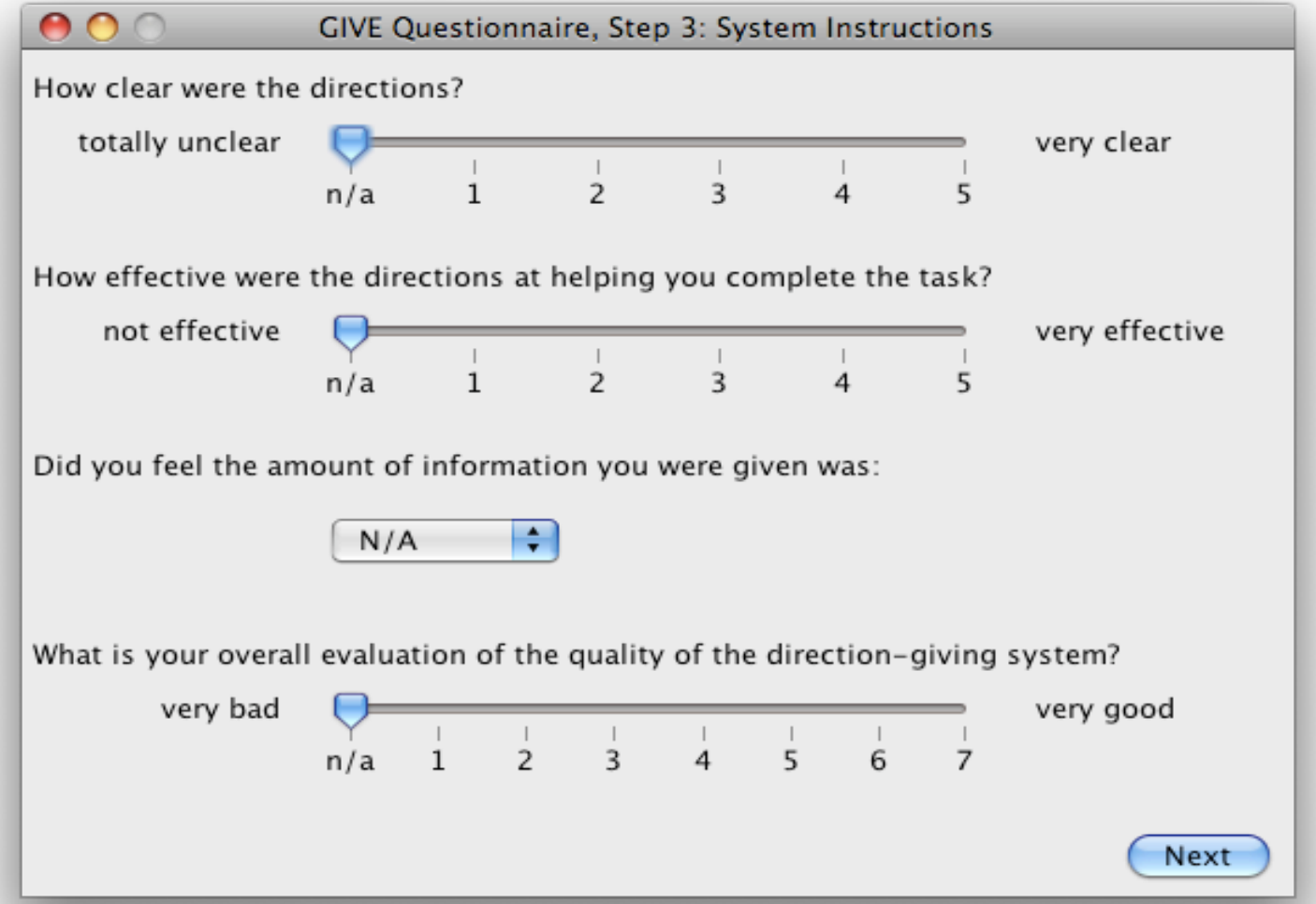
Data Gathered

- GIVE-1: November 2008 to February 2009
- Five NLG systems
  - University of Texas at Austin; Union College, Schenectady, NY; Universidad Complutense de Madrid, Univiersity of Twente × 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



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

Parameter	Web-Based	Lab-Based
# of participants	322	91
Gender	80% M, 10% F, 10% Unknown	31% M, 65% F, 4% Unknown
English proficiency	62% Expert, 34% native English	93% Expert, 81% native English

Objective Measures

		Objective Measures									
		task success		instructions		steps		actions		seconds	
Web	A	91%	A	83.4	B	99.8	A	9.4	A	123.9	A
	M	76%	B	68.1	A	145.1	B	10.0	AB	195.4	BC
	T	85%	AB	97.8	C	142.1	B	9.7	AB	174.4	B
	U	93%	AB	99.8	C	142.6	B	10.3	B	194.0	BC
	W	24%	C	159.7	D	256.0	C	9.6	AB	234.1	C
Lab	A	100%	A	78.2	AB	93.4	A	9.9	A	143.9	A
	M	95%	A	66.3	A	141.8	B	10.5	A	211.8	B
	T	93%	A	107.2	CD	134.6	B	9.6	A	205.6	B
	U	100%	A	88.8	BC	128.8	B	9.8	A	195.1	AB
	W	17%	B	134.5	D	213.5	C	10.0	A	252.5	B

Subjective Measures

		Subjective Measures							
		overall		choice of words		referring expressions		timing	
Web	A	4.7	A	4.7	A	4.7	A	81%	A
	M	3.8	AB	3.8	B	4.0	B	70%	ABC
	T	4.4	B	4.4	AB	4.3	AB	73%	AB
	U	4.0	B	4.0	B	4.0	B	51%	C
	W	3.8	AB	3.8	B	4.2	AB	50%	BC
Lab	A	5.7	A	4.7	A	4.8	A	92%	A B
	M	5.4	A	3.8	B	4.3	A	95%	A B
	T	4.9	A	4.5	A B	4.4	A	64%	A B
	U	5.7	A	4.7	A	4.3	A	100%	A
	W	5.0	A	4.5	A B	4.0	A	100%	B

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