User-centered Evaluation of Interactive Question Answering Systems

Diane Kelly, University of North Carolina
Paul B. Kantor & Ying Sun, Rutgers University
Emile L. Morse & Jean Scholtz, NIST
Purposes of Work

- Develop user-centered evaluation methods and metrics for interactive QA systems
- Focus on portable evaluation methods and metrics, not tailored to specific systems or features
- Focus on evaluation of *end-product* and *totality of users’ experiences*
Goals of Presentation

• Describe and discuss key issues in the design of evaluations of interactive QA systems
• Potentially help others construct their own evaluations
• Encourage the discussion of standards and best practice
• NOT to tell you which system is the ‘best’
Evaluation Method

• Approach: 2-week workshop
• Hypotheses
• Participants
• Scenarios
• Corpus
• Systems (3 experimental; 1 baseline)
• Experimental Design
Data Collection

[1 Block = 1 System/2 Tasks/2 Days]

START BLOCK

System Tutorial

NASA TLX

Post-Scenario Questionnaire

Post-Session Questionnaire

Post-Session Interview

[Two Tasks]

NASA TLX Weighting

Post-System Questionnaire

Post-System Interview

Cross Evaluation & Focus Group

END BLOCK

• Observation
• System Logs
• Glass Box
• Spontaneous Self-report
Results

• Participants
  – Somewhat skewed toward managerial personnel; not many full-time analysts
  – Self-selection bias and snowball recruitment
  – High education is not necessarily a predictor of good analytic performance
  – Rating bias
Results

- **Scenarios**: judged reasonable by participants, but required extra-topical information to make them realistic
- **Corpus**: less than ideal, as analysts are used to using classified document
## Results

| Method               | Collection Cost | Analysis Cost | Value |
|----------------------|-----------------|---------------|-------|
| Observation          | $$              | $$            | ★     |
| System logs (SL)     | .               | $$            | ★★★   |
| Glass Box (GB)       | .               | $$$           | ★★★   |
| Self-report          | $               | $$            | ★     |
| NASA TLX             | $               | $             | ★     |
| Questionnaires       | $               | $             | ★★★   |
| Cross-Evaluation     | $$              | $             | ★★★★  |
| Interviews           | $$              | $$$           | ★★★★  |
| Focus Groups         | $               | $$            | ★★★   |
| Question | Description | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Difference |
|----------|-------------|--------|--------|--------|--------|------------|
| Q3:      | The [X] system helped me to better understand the scenario by the information that it provided. | 3.79 (.80) | 3.13 (1.06) | 2.79 (.89) | 2.53 (.83) | 1 > 2, 3, 4 |
| Q11:     | The [X] system stimulated my thinking about the scenario. | 3.79 (.70) | 3.20 (1.24) | 3.00 (.83) | 2.47 (.83) | 1 > 3, 4 |
| Q15:     | The [X] system expanded my understanding of the scenario. | 3.71 (.83) | 3.13 (.83) | 3.07 (1.21) | 2.60 (.51) | 1 > 4 |
| Q7:      | The [X] system helped me to think about the scenario in new ways. | 3.64 (.93) | 3.13 (.99) | 2.79 (.97) | 2.53 (.83) | 1 > 3, 4 |
System Effect on Factor_1 of Cross-Evaluation
Conclusions about Evaluation

• Participants, scenarios/tasks, corpus must be considered carefully and their effects measured when appropriate

• Questionnaires and Cross Evaluation were most effective at distinguishing between systems

• Interviews and Focus Groups most useful for providing feedback to developers and for identifying additional evaluation criteria and hypotheses
Users of interactive systems expect systems
– to exhibit behaviors which can be characterized as understanding what the user is looking for, what the user has done and what the user knows
– to track their actions over time, both with the system and with information
Example Hypotheses

• QA systems should assist analysts in exploring more paths/hypotheses.
• QA systems should help analysts recognize gaps in their thinking.
• QA systems should allow analysts to collect more data in less time.
• QA systems should enable analysts to produce higher quality reports.
### Description of Participants

|                        | Range            |
|------------------------|------------------|
| Age                    | 30-54            |
| Highest Degree         | H.S.-Ph.D.       |
| Years of Military Service | 2.5-31          |
| Years of Experience as Analyst | 0-23        |
| Computer Experience    | “medium”         |
**Scenario:** [country’s] Chemical Weapons Program
Before a [blank] is reestablished in [country], a current, thorough study of [country] chemical weapons program must be developed. Your task is to produce a report for [person] regarding general information on [country] and the production of chemical weapons. Provide information regarding [country] access to chemical weapons research, their current capabilities to use and deploy chemical weapons, reported stockpiles, potential development for the next few years, any assistance they have received for their chemical weapons program, and the impact that this information will have on the [country]. Please add any other related information to your report.

**Customer:** [customer]
**Role:** Country desk – [country]
**What they want:** General report on [country] and CW production
Characteristics of corpus in bytes

| Source | All Files | Documents | Images |
|--------|-----------|-----------|--------|
| CNS    | 40192     | 39932     | 945    |
| Other  | 261590    | 48035     | 188729 |
# Experimental Design

| Days  | 1 & 2  | 3 & 4  | 5 & 6  | 7 & 8  |
|-------|--------|--------|--------|--------|
| **Scenarios** | A, B | C, D | E, F | G, H |
| **System 1** | O₁   | O₂   | O₃   | O₄   |
|       | A₁    | A₂    | A₃    | A₄    |
|       | A₅    | A₆    | A₇    | A₈    |
| **System 2** | O₂   | O₁   | O₄   | O₃   |
|       | A₄    | A₃    | A₂    | A₁    |
|       | A₇    | A₈    | A₅    | A₆    |
| **System 3** | O₃   | O₄   | O₁   | O₂   |
|       | A₂    | A₁    | A₄    | A₃    |
|       | A₈    | A₇    | A₆    | A₅    |
| **System 4** | O₄   | O₃   | O₂   | O₁   |
|       | A₃    | A₄    | A₁    | A₂    |
|       | A₆    | A₅    | A₈    | A₇    |