Lessons Learned – EWG Map Navigation Experiment

Here we discuss the lessons taken from our initial work on designing the map navigation experiment.

Goals

As a first step in validating the EWG Evaluation Methodology Document, we have been engaged in designing and carrying out an experiment relating to task performance in a collaborative environment. Our goals in this work were:

1. to test the utility of the Evaluation Methodology document (draft 1) that we recently completed and to determine what revisions might be desirable;
2. to gain experience in experimental design for collaborative environments;
3. to debug the MITRE logging tool in the context of actual data collection in a collaborative setting;
4. to determine what additional tools and software were needed to make it easy to conduct evaluations in collaborative environments;
5. to gain sufficient experience to write a practical guide to evaluation of collaborative tools, to supplement the Methodology document.

Choice of Experiment

To this end, we designed a minimal but useful simple experiment. We discussed prototypical kinds of collaborative tasks (e.g., decision making, planning, information dissemination, etc.). We discussed practical issues of data collection in a laboratory environment (finding willing subjects knowledgeable enough about an artificial task to carry it out). And we discussed the kinds of things one might want to evaluate in a collaborative environment.

The upshot was that we chose a readily accessible problem domain: collaborative problem solving, using spatial materials - specifically a map. We carefully chose a scenario that would be readily familiar to a wide range of potential subjects: direction finding using a street map. And we designed the experiment to require information sharing and collaborative planning, by providing each participant with some private information (e.g., information marked on a non-shared map) about one-way streets or road blocks and traffic conditions.

We posed a question whose answer we were interested in, and where we were not entirely certain of the experimental outcome: How does the addition of audio communication to a textual and visual (whiteboard) communication environment affect

- task completion time,
- quality of solution and
- user satisfaction?

And we wished to look at some secondary questions, namely:

- How does the availability of audio affects participation?
- How does the availability of audio affect the use of the whiteboard?

The experiment makes it relatively straightforward to measure 1) time to complete the task, 2) quality of solution (length of route, whether it avoided all known obstacles, etc.) and 3) user satisfaction (via questionnaire). For the secondary questions, we are thinking about measures such as number of turns per participant length of turns per participant, mode of turn (whiteboard vs. typing or audio). Calculating these measures is more involved than the first three listed here, because some subjective assessment of the dialog is necessary.

The remainder of the experiment was designed to control for affects we were not interested in: the difficulty of different maps, a possible learning affect obtained by doing the same kind of task multiple times, the order of presentation, and differences among participants.

Lessons Learned

1. It is much easier to choose measures based on what hypotheses we are testing, rather than using the "layered" framework in the Methodology document.

2. The list of possible measures in the Methodology document was useful in giving us ideas of how to go about measuring things, particularly the secondary hypotheses.

3. Setting up DETAILED experimenter instructions is difficult but absolutely critical to getting data collection to run smoothly.

4. Doing MANY pilot runs before real data collection begins is necessary. Doing this, we have found (and fixed) bugs and awkwardnesses in the logger, developed new tools to aid in inspection of the data, and developed new hypotheses based on what we observed in the pilot data.

5. Looking at the data collected in the pilot stage is very important - because it turns out that we needed to log more data than we originally thought, e.g., time spent actually drawing lines in the Whiteboard.

6. We needed to learn a lot more about experimental design than we originally knew - and have been reading several books, including Paul Cohen's book on Empirical Methods for Artificial Intelligence.

7. Things that SEEMED easy to measure (time to complete task, number of Whiteboard turns) turned out to be more complicated than we thought. It was difficult to create a logged event associated with the start times of both participants, because they were not co-located (similarly for ending the task). There are several modes of drawing in the WhiteBoard (line segments vs. a continuous curve) that look different when logged (multiple events, one per line segment vs. one event for a long curve), so a subjective analysis of the data is necessary.

Results

As a result of this work, we plan to make some revisions in the Methodology document section on Measures and Metrics. We also plan to publish the materials relating to this experiment, as well as the results, to the IC&V community. Finally, we believe that it may be useful to write up an additional document, namely a guide to running experiments on collaborative technologies.