Evaluation of multi-party virtual reality dialogue interaction

David R. Traum, Susan Robinson, Jens Stephan

USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292, USA
{traum, robinson, stephan}@ict.usc.edu

Abstract

We describe a dialogue evaluation plan for a multi-character virtual reality training simulation. A multi-component evaluation plan is presented, including user satisfaction, intended task completion, recognition rate, and a new annotation scheme for appropriateness. Preliminary results for formative tests are also presented.

1. Introduction

Evaluation of Dialogue systems is still a very difficult endeavor for a number of reasons, including lack of consensus on what a "good" or "better" dialogue system is, need for human involvement in testing/evaluation, and wide variety in tasks, domains, and goals for the systems. While there has been a lot of work in recent years on evaluation of dialogue systems (e.g., (Smith and Hipp, 1994; Danieli and Gerbino, 1995; Sikorski and Allen, 1996; Walker et al., 1997; Walker et al., 2002)), it is still often not possible to directly carry over one evaluation methodology to a new system and task, especially when the style of interaction, domain, task, and objectives are different.

In this paper we report on the evaluation plan and initial results for dialogue interaction as part of a larger multimodal story-based training simulation system (Swartout et al., 2004). The system is set in a life-sized virtual world, presented in a theatre with a 150 degree field of view screen and 3-D immersive sound. The virtual world includes graphical presentation of a scene including multiple animated characters, who can communicate with each other and the human trainee. We are planning multiple evaluations of different aspects of the interaction, including learning, story, and immersion, as well as usability. As part of the usability evaluation, we are evaluating the dialogue capability of the virtual agents. This evaluation also consists of multiple components, to try to cover different aspects of interaction. We have four main components to our dialogue interaction, each with several submetrics. The main components are user satisfaction, task success, recognition accuracy, and agent utterance appropriateness. In the next section, we briefly describe the domain and task in which the dialogue is embedded. In Section 3., we describe our evaluation plan, including a new coding scheme for appropriateness of agent behavior. Finally, in Section 4., we present some preliminary results and future directions for the evaluation.

2. The Mission Rehearsal Exercise System

The initial focus of the Virtual Reality system is on training leadership and decision-making for small group interaction. The trainee in the initial scenario is a US Army Lieutenant in charge of a platoon (about 30 soldiers) who must confront a dilemma in a peace-keeping scenario. His mission is to assist another platoon in a weapons inspection. However, en route, he discovers an accident between an army vehicle and a civilian car, with an injured boy lying prone on the ground. The Lieutenant must decide (with the assistance of the Sergeant, his 2nd in command) what should be done: to carry on the original mission, leaving the boy behind, to help the boy (at the possible risk to the other platoon), to split his forces and try to accomplish both, or some other course of action entirely. The Lieutenant can carry on face to face spoken conversations with the Sergeant and other people in view (both troops in the platoon and local people, including the injured boy's upset mother), and others on the radio (including his superiors and the other platoon). Figure 1 shows an example of some interaction in this domain.

Although the domain is task-oriented, as a training system, the main object of the interaction is not necessarily efficient task performance. Often more can be learned by confronting difficulties than by optimal performance. Thus, like in tutoring domains, the agents, even when they know exactly what they should do, should often give the user (trainee) a wide degree of initiative — even when this may lead to less efficient task performance. Likewise, user satisfaction does not seem to have the primary role that it has for some task oriented evaluations (Walker et al., 1997; Walker et al., 2002). Ultimately the real value of the system is how well it trains, and unsatisfactory agent behavior may lead the trainee to think things through, devising alternate approaches, and ultimately become better at decision-making. On the other hand, task performance and user satisfaction are still important – if the system is too hard to use or accomplish any tasks with, it will not play a role in learning (and no one will want to use it).

3. Evaluation plan

Given the complex goals for evaluation, we decided to measure several aspects of the dialogue interaction. User satisfaction is important as a subjective measure of how good the trainee thought the interlocutors were. Likewise task performance is also important, although the situation here is more complex than in some domains. Here we have two top-level tasks which may be in conflict – it may not be possible to solve both. Furthermore, the trainee may decide on other tasks based on assessment of the situation. There are also a number of sub-tasks to each main task. A third

45	base	Eagle 2 6 this is eagle base
46	base	medevac launching from operating base alicia time now
47	base	eta your location 0 3
48	lt	what should we do now
49	base	over
50	sgt	Thats not my responsibility
51	sgt	i dont know sir
52	lt	is the lz secure
53	sgt	thee L Z is secure sir
54	sgt	sir eagle one in celic need help
55	lt	send help to eagle one six
56	sgt	move what
57	lt	send second squad
58	sgt	sir that is a bad idea
59	lt	what should i do now
60	sgt	sir we already talked about this
61	sgt	i dont know sir
62	sgt	sir eagle one in celic need help
63	lt	send help to eagle one six
64	sgt	uh
65	sgt	sir we should get Fourth Squad to Celic
66	lt	send fourth squad
67	sgt	yes sir
68	sgt	Lopez
69	sgt	reconn forward along the route
70	4sldr	Fourth squad
71	4sldr	mount up

Figure 1: Example of MRE interaction

evaluation component is the recognition by system characters of what the user said. Since there is a pipelined recognition process (including speech recognition, semantic parsing, and pragmatic analysis), there are separate recognition rates for each component. Finally, we try to measure the response of agents. In some simple domains, the quality of the system response can be measured as "correct" or "incorrect". On the other hand, for "chatterbot" tests such as the Loebner competition, the ideal is to be indistinguishable from human responses (regardless of correctness). Our task is somewhere in the middle. Complete accuracy is not necessarily a goal (given that trainees must learn to deal with difficult communication conditions), nor is perfectly human response - what we try instead is to reach a compromise, and talk about the "appropriateness" of system character interventions both toward the domain and toward carrying on a natural conversation. In this section we describe each of these evaluation categories in more detail.

3.1. User Satisfaction

User satisfaction is obviously very important for a number of types of dialogue systems, since, to some degree, it will influence the future use of a system, especially if users have some choice. We follow the method in (Walker et al., 2002) of rated survey questions, although a slightly different set of questions is required. E.g., expected behavior and user expertise are not necessarily relevant, since training in unexpected circumstances is part of the task. Also, the interaction involves more than just information retrieval – a major part is negotiating and acting in the domain. Moreover, since the scenario has multiple characters – each with his or her own voice and body (except for distant radio characters), one can also rate satisfaction with each agent, as well as overall satisfaction. We are still piloting different questions, and have not reached a final consensus on the complete set, nor gathered enough samples of the same questions to make any meaningful comparisons.

3.2. Intended Task Completion

Task success is also very important, although less crucial in our domain than some – sometimes one can learn more from failure than from easy success. We measure the talk about tasks, using a modification of the *IU-coding* from (Nakatani and Traum, 1999). We are not yet trying to capture hierarchical intentional structure, but focusing on one level of granularity – the specific orders and questions that the LT gives to the Sgt and others. Each dialogue is annotated with a set of tasks that the participants bring up. We code each utterance as to which task(s) it is a part of, and we also note when a task has been accomplished. Inter-rater reliability was good (Kappa of 0.78 and 0.81 between two coders on two unseen dialogues). We also compute for each task whether it is in the task model of the agents or not (some tasks that the trainee would like to do are simply not possible because of the limited domain restrictions). We compute success rate both "subjectively" (as a ratio of all tasks the trainee attempted) and "objectively" (as a ratio of only those tasks that were in the domain model). It is also possible to compute efficiency measures for tasks, e.g., how much time it took to accomplish or how many utterances were part of the IU. In the dialogue fragment in Figure 1, there are no tasks mentioned that are out of the task model. Utterances 45-47 and 49 are related to calling a medevac (started before the fragment) and is successfully resolved. 48, 50-51 involve an unsuccessful attempt to find out what to do, which is continued in 59-61. 52-53 is a successful information exchange about the LZ's security. 54-58,62-71 is a successful (albeit extended) subdialogue about sending help to the Eagle-one platoon in Celic.

3.3. Recognition Rate

For Recognition of user contributions, the ultimate measure is whether the agents can classify the utterances as to who the addressee is, which dialogue acts are being performed, and which domain concepts (states, tasks, etc) are being referred to. We calculate an F-score measure over the utterances in the dialogue. We also calculate sub-measures, including speech recognition word error rate, and semantic parser slot-filler f-score.

3.4. Response Appropriateness

For our task of virtual reality simulation, one of the most important things is the naturalness of the interaction and it's contribution to the sense of immersion and being within the situation. Thus, unnatural interaction styles such as over-verification and strong system initiative with limited choice are seen as inappropriate, even though they might improve recognition results. Likewise, some utterances such as rejections and negotiation, while they might not lead effi-

ciently to task completion, may still be seen as very appropriate within the domain. We have thus developed an "appropriateness" coding scheme for rating agent interactions. We are marking appropriateness as seen from the trainees point of view. Our coding scheme consists of two subschemes, one for trainee utterances and one for all other utterances. Each trainee utterance is marked as to whether it receives a response or not. Some utterances, e.g., a simple acknowledgement, do not require further response, so we further marked trainee utterances with no response as to whether a response was expected. The coding for trainee (LT) utterances is shown in Figure 2. In our sample dialogue fragment in Figure 1, all of the LT's utterances are responded to, and are thus coded RES.

code	description
RES	gets response
NRA	no response appropriate
NRN	no response not appropriate

Figure 2: Trainee Appropriateness codes

Our coding scheme marks each agent utterance with one of six tags, marking a number of distinctions. The primary distinction is inappropriate vs appropriate. It is up to the judgement of the coder as to whether the utterance is appropriate or not. There are some special types of categories which, on the whole, are always conditionally appropriate. For instance, filled pauses and requests for repair – these are based on the understanding of the speaker and so hard for a listener to judge appropriateness (although many repetitions would certainly seem inappropriate). For this reason we code these two types of utterance separately and make appropriateness distinctions only on the remaining utterances. We also further classify the appropriate utterances into three categories, depending on how they relate to prior discourse. A direct (appropriate) response is coded separately from a new initiative (not directly related to what the LT or others have said before, but appropriate to move the situation forward toward solving the overall goals) and from a continuation of a prior system character utterance. The agent utterance coding is summarized in Figure 3. In our sample dialogue in Figure 1, utterances 53 and 67 are appropriate responses, while 50, 51, 58, 60, and 61 are judged inappropriate[1] 45-47, 49, and 68-71 are appropriate continuations. 64 is a filled pause, while 56 is a request for repair. 54 and 65 are agent initiatives.

Despite the subjective nature of appropriateness judgements, we were still able to achieve very high reliability with this coding scheme (Kappa of over 0.9 among four coders, including one who had no previous exposure to the scheme other than a coding manual).

We also assign a numerical score for each code, so that we can have a dialogue-wide measure of appropriateness. The scoring scheme is meant to capture the following intuitions: filled pauses are generally human-like and good for virtual agents to perform, but don't add a lot, they mainly prevent points from being taken away for non-responsiveness. Appropriate responses are very good, but even better are initiatives that push the interaction back on track rather than getting side-tracked into irrelevancy. Extended contributions are also "good" when appropriate, but not as important as new initiatives or responses. Repairs and clarifications are bad in their own right (especially too many in a row), but their use can still gain points by allowing a subsequent appropriate response. Inappropriate responses are seen as bad, but not as bad as no response. Our preliminary scoring is shown in figure 4. Future work includes trying to verify these scoring intuitions against human judgements of subdialogue sequences to see how robust the intuitions are in practice.

code	description
PF	filled pause
RR	request for repair
AP	appropriate response
INI	appropriate new initiative
CON	appropriate continuation
NAP	inappropriate response,initiative or continuation

Figure 3: System Agent Appropriateness codes

RES	NRA	NRN	PF	RR	AP	INI	CON	NAP
1	1	-2	0	-.5	2	3	.5	-1

Figure 4: Appropriateness Code scoring

4. Preliminary Results

We are currently in transition, in the MRE project, from a system that can be used to demonstrate advanced technologies to one that could actually be used by the target population for training. Thus the evaluations described here are currently being used formatively, with many changes to the system ongoing as the tests are being run. It is very important to use the target population rather than the general public for testing the system, as military cadets (people training to be Lieutenants) have a very different idea of the domain than people with no military training. Thus terms like "LZ" (a landing zone for a helicopter) and procedures like "secure the area" are used and understood by the cadets, while a general population, such as university students with no training, need more assistance.

We give here some comparisons in evaluations using the schemes in the previous section between tests run in March 2003 (with the system meant for demos) and a version of the system in December 2003 (which is not by any means the final system, but is a convenient benchmark). The March system included a finite-state grammar based recognizer, while the December one uses a bigram model, trained on data from previous tests (and wizard of oz tests). Also, the March tests used a purely user-initiative dialogue model, while the December one includes a mixed-initiative model in which the Sergeant can take the initiative, according to several parameters of the interaction.

[1]Note that "inappropriate" does not mean unnatural. The Sgt as second in command, is responsible for advising the LT and does have ideas about what to do, as can be seen elsewhere in the dialogue fragment. 58 is more controversial - in this case it seems inappropriate to reject the clarification without a reason for preferring to send fourth squad rather than second.

	IU coding					recognition rates				appropriateness	
session	# inits	resolved	oom	SUBJ	OBJ	ASR	NLU	SA	ADDR	total	avg
3-1	11	2	5	0.18	0.33	0.21	0.50	0.47	0.71	-25	-0.16
3-2	8	1	3	0.13	0.2	0.36	0.48	0.50	0.81	-8.5	0.12
3-3	16	4	6	0.25	0.33	0.30	0.60	0.35	0.75	-18	-0.07
12-1	12	2	6	0.17	0.33	0.50	0.38	0.61	0.82	16	0.16
12-2	10	5	3	0.5	0.71	0.62	0.58	0.65	0.65	83	0.47
12-3	7	6	0	0.86	0.86	0.74	0.73	0.86	1.00	93.5	1.02

Table 1: Dialogue evaluation comparison

Table 1 shows a summary of the results on three dialogues from each period. The first section shows the task completion results. The first column is the total number of task initiatives attempted in each dialogue. The second column shows the number that were successfully resolved, while the third column shows the number that were not part of the task model. The fourth and fifth column show the subjective and objective success rates, as defined in Section 3.2.. The next set of three columns shows the recognition rates for ASR, NLU, speech act recognition and addressee recognition for the same six dialogues. All numbers are F-scores to promote comparison across the areas. We are also coding task-based reference resolution, but do not have reportable numbers at this writing. The final two columns shows the response scoring for the six dialogues, including both a total score for the dialogue and a per utterance score, averaged over the whole dialogue.[2]

Figure 5 shows the distribution of appropriateness codes as a percentage of all coded utterances. For each code, the left bar shows a composite of the three March dialogues, while the right shows the distribution of the December dialogues. We can see that the three trainee codes all have higher percentages in March, meaning that the trainee spoke a higher percentage of the time for those dialogues. Note that the "not appropriate" responses are higher in March even with fewer proportionate total responses.

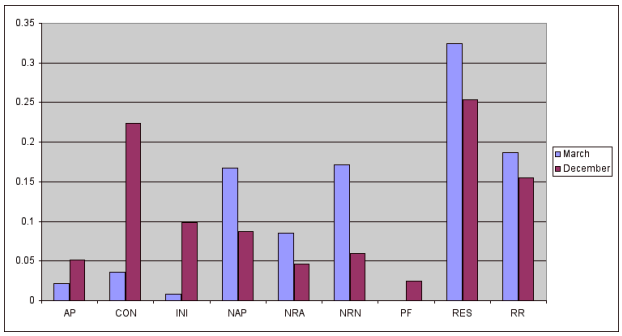

Figure 5: Appropriateness Code distribution

The coding schemes presented here do seem to do a good job of measuring different aspects of dialogue system performance for a training domain. Further testing is necessary both to develop a holistic view of what makes one dialogue better than another, and to compare different system strategies and their results.

[2]We are still not sure which is a more interesting measure. Since there are both plusses and minuses possible, it may be that a raw score is most representative of dialogue quality — a long good dialogue may be even better than a short good dialogue.

Acknowledgements

We would like to thank the many members of the MRE project team for help in this work. First, those who helped build parts of the system. Also Sheryl Kwak, Lori Weiss, Bryan Kramer, Dave Miraglia, Rob Groome, Jon Gratch, and Kate Labore for helping with the data collection, and Captain Roland Miraco and Sergeant Dan Johnson for helping find cadet trainees. Eduard Hovy, Shri Narayanan, Kevin Knight, and Anton Leuski have given useful advice on evaluation. The work described in this paper was supported by the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

5. References

Danieli, M. and E. Gerbino, 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Working Notes AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*..

Nakatani, Christine H. and David R. Traum, 1999. Coding discourse structure in dialogue (version 1.0). Technical Report UMIACS-TR-99-03, University of Maryland.

Sikorski, T. and J. F. Allen, 1996. A task-based evaluation of the TRAINS-95 dialogue system. In *ECAI Workshop on Dialogue Processing in Spoken Language Systems*.

Smith, Ronnie W. and D. Richard Hipp, 1994. *Spoken Natural Language Dialog Systems: A Practical Approach*. Oxford University Press.

Swartout, W, J. Gratch, R. W. Hill Jr., E. Hovy, R. Lindheim, S. Marsella, J. Rickel, and D. R. Traum, 2004. Simulation meets hollywood: Integrating graphics, sound, story and character for immersive simulation. In Oliviero Stock and Massimo Zancanaro (eds.), *Multimodal Intelligent Information Presentation*. Kluwer.

Walker, M., A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard, 2002. Darpa communicator: Cross-system results for the 2001 evalation. In Proceedings 7th International Conference on Spoken Language Processing (ICSLP-02).

Walker, M. A., D. J. Litman, C. A. Kamm, and A. Abella, 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings ACL-97*.