DARPA COMMUNICATOR: CROSS-SYSTEM RESULTS FOR THE 2001 EVALUATION

*Marilyn A. Walker, Alex Rudnicky, Rashmi Prasad, John Aberdeen, Elizabeth Owen Bratt, John Garofolo, Helen Hastie, Audrey Le, Bryan Pellom, Alex Potamianos, Rebecca Passonneau, Salim Roukos, Greg Sanders, Stephanie Seneff, Dave Stallard**

ABSTRACT

This paper describes the evaluation methodology and results of the 2001 DARPA Communicator evaluation. The experiment spanned 6 months of 2001 and involved eight DARPA Communicator systems in the travel planning domain. It resulted in a corpus of 1242 dialogs which include many more dialogues for complex tasks than the 2000 evaluation. We describe the experimental design, the approach to data collection, and the results. We compare the results by the type of travel plan and by system. The results demonstrate some large differences across sites and show that the complex trips are clearly more difficult.

1. INTRODUCTION

The objective of the DARPA Communicator program is to support rapid development of multi-modal speech-enabled dialog systems with advanced conversational capabilities. To track progress toward these goals, we have been developing methods for measuring the performance of Communicator systems. In previous work, we report an evaluation experiment with nine participating Communicator systems in the travel planning domain [5, 8]. During 2001, we carried out a second evaluation, involving systems from AT&T, BBN, University of Colorado, Carnegie Mellon University, IBM, Lucent Bell Labs, MIT, and SRI. Our goal was to develop an evaluation paradigm that (1) supports continuous, lightweight, data collection for heterogeneous systems; (2) allows sites to develop and test new metrics; (3) supports comparisons with the 2000 evaluation data; and (4) supports the application of the PARADISE framework [7]. We also wanted to establish a performance baseline for complex tasks.

This paper describes the evaluation methodology and results of the 2001 evaluation. A companion paper compares these results with those for 2000 and shows that there was a large performance improvement in every core metric from 2000 to 2001 [6]. Section 2 describes the evaluation paradigm and experimental design. Section 3 reports the results, the application of PARADISE, and differences across sites for the measures identified by PARADISE as important. We sum up in Section 4.

2. EXPERIMENTAL DESIGN

The experimental design in 2001 was motivated in part by a desire to improve upon the design for 2000. A major motivation was to test the systems in a more realistic setting. The evaluation took place over 6 months with systems continuously accessible via a toll-free number. This required systems to be stable; it was not possible for developers to continously monitor their systems for six months.

Subjects were recruited by NIST and assigned to a particular system. We believed that continuous use of a particular system would best approximate actual conditions of use and provide us data for analyzing the effects of expertise with a system. Half of the sites provided a site-specific web page for users to enroll with the system before making any calls. These pages often included instructions about how to interact with the system, detailed meta-dialog commands for making corrections, or an example dialog interaction, illustrating how users could take the initiative.

There were two groups of subjects: SHORT and LONG subjects. Both groups called their assigned system in order to plan travel for real tasks. All subjects were recruited from a population of frequent travelers (people who make at least 6 trips per year). The SHORT users called their system four times over the six month period when they had to make travel plans. In addition to their real trips, the LONG group performed six fixed tasks. The fixed scenarios are described via their features in Figure 1. The LONG users completed scenarios 1 to 4 to familiarize themselves with the system before doing their real trips. Then after planning four real trips, they completed scenarios 5 and 6. As a way to motivate subjects to seriously attempt each task, they received a bonus for completing an itinerary.

Scenario	TripType	Destination	Airline?	CarHotel?
1	Round	Domestic	No	No
2	Round	International	Yes	No
3	MultiLeg	Domestic	Yes	Yes
4	MultiLeg	International	No	Yes
5	MultiLeg	Domestic	Yes	Yes
6	MultiLeg	International	No	Yes

Fig. 1. Parameters of Fixed Scenarios

The fixed scenarios were intended to establish performance for complex tasks and support comparisons with the 2000 evaluation. See Figure 1. Scenarios 1 and 2 are simple round trips comparable to those in 2000 [5]. Scenarios 3 to 6 are COMPLEX tasks; they require multiple legs (MultiLeg), may require flying on a particular airline (Airline?) and making car and hotel arrangements (CarHotel?). Scenario 3 is in Figure 2. A system with advanced conversational capabilities should complete the COMPLEX tasks within

*AT&T Labs, Carnegie Mellon University, AT&T Labs, MITRE, SRI, NIST, AT&T Labs, NIST, University of Colorado, Lucent Bell Labs, AT&T Labs, IBM, NIST, MIT, BBN Technologies. This research was funded by DARPA contract MDA972-99-3-0003.

10 minutes [6]. We experimented with a new method for scenario presentation using recorded speech descriptions of the tasks via an IVR system to avoid biasing users' syntax. See [6].

You live in Boston Massachusetts, and you want to fly to Detroit Michigan for a meeting. After the meeting, you want to fly to San Francisco to visit a friend. You will fly from Boston to Detroit on November 2nd, arriving in Detroit around 2 PM. You will fly from Detroit to San Francisco on November 6th, leaving Detroit in the late morning. You will fly back to Boston on November 11th, leaving San Francisco in the afternoon. You will fly on Northwest Airlines. While in Detroit you need a compact rental car and a single room in a downtown hotel.

Fig. 2. Scenario 3: A complex trip involving multiple legs, airline constraint, car and hotel arrangements.

We also collected a user profile for each subject with information on subject gender, native language, dialect region, experience with dialog systems, normal mode of booking travel, frequency of business and personal trips, number of miles and areas traveled to, and the number and status of frequent flyer accounts.

Task Ease: In this conversation, it was easy to get the information that I wanted.
TTS Performance: I found the system easy to understand in this conversation.
User Expertise: In this conversation, I knew what I could say or do at each point of the dialog.
Expected Behavior: The system worked the way I expected it to in this conversation.
Future Use: Based on my experience in this conversation using this system to get travel information, I would like to use this system regularly.

Fig. 3. User Survey assessing User Satisfaction

After each call, the subjects completed a survey probing their user satisfaction, task requirements, perception of task completion, type of phone used and how many times the subject had traveled this itinerary in the last year. Users stated their degree of agreement with each statement in the user satisfaction survey in Figure 3 on a scale of 1 to 5. These responses were summed resulting in a user satisfaction metric ranging from 5 to 25. The task requirements and user perception of task completion were used to derive a ternary task completion measure CTC (corrected task completion); a 2 indicates that both the airline itinerary and any car and hotel arrangements were completed; a 1 indicates that only an airline itinerary was completed; a 0 indicates that no task was completed.

The corpus contains 1242 dialogs with both logfiles and completed surveys. There were 198 RoundTrip, 350 Complex, and 694 Real calls. As mentioned above, the dialogs for complex trips were intended to demonstrate advanced conversational interaction for complex tasks. Each dialog has: (1) logfiles generated using the Communicator logfile standard [1]; (2) user surveys completed at the end of each call (1052 with user comments); (3) the user profile; (4) metrics derived from the logfiles; (5) ASR and hand transcriptions of user utterances; (6) Task success metrics. The metrics that we extracted from these sources for each dialog are in Figure 4.

Dialog Efficiency Metrics: Total elapsed time, Time on task, System turns, User turns, Turns on task, Time per turn for each system module
Dialog Quality Metrics: Sentence error rate, Word error rate, Number of Overlaps in System/User turns
Task Success Metrics: Ternary measure of perceived task completion

Fig. 4. Metrics per Call.

3. EXPERIMENTAL RESULTS

As mentioned above, one of the main goals in 2001 was to establish performance for COMPLEX tasks such as Scenarios 3 to 6 in Figure 1. Figure 5 shows user satisfaction and Figure 6 shows task durations by the type of trip. These results are shown as box plots: a box plot shows the full range of the distribution of values, with the box itself marking the interquartile range, and the median of the distribution as a line within that box. One way ANOVAs for user satisfaction and task duration by type of trip indicates significant differences with multileg trips resulting in lower user satisfaction and taking significantly longer (df=2,F= 41.1, p = 0.00; df=2,F=7.12, p= 0.001). Real and round trips have similar values for both user satisfaction and task duration, although users indicated that 210 out of the 694 real trips required car and hotel arrangements.

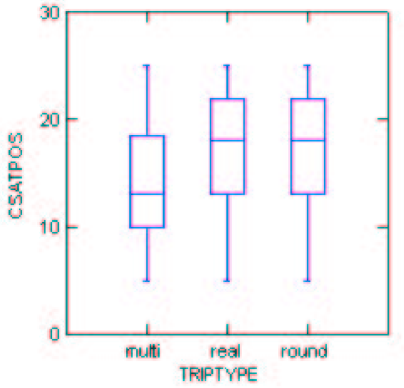

Fig. 5. User Satisfaction by TripType

3.1. Application of PARADISE

The PARADISE evaluation framework has been applied in other work [7, 2]. The framework posits that user satisfaction is the overall objective to be maximized and that task success and various interaction costs can be used as predictors of user satisfaction. Figure 7 describes the linear model learned by applying PARADISE to the corpus which accounts for 27% of the variance in user satisfaction (R = .52). According to the model, task duration is the largest contributor to user satisfaction (-.44), followed by the mean words in system turns (.41) and task completion (CTC). Sentence error rate (SERR), the number of overlaps between system and

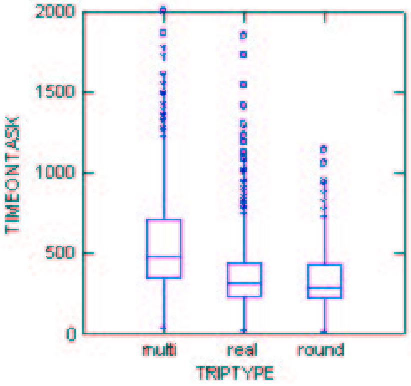

Fig. 6. Task Duration by TripType

Factor	Coefficient
Task Duration	-0.44
System Words Per Turn	0.41
Task Completion (CTC)	0.32
Sentence Error Rate (SERR)	-0.17
Number of Overlaps	-0.15
User Words Per Turn	0.04

Fig. 7. PARADISE derived linear model for 2001.

user turns, and the number of user words per turn are also significant predictors. More of the variance in user satisfaction can be predicted using tree models with dialogue act metrics [3].

3.2. Cross System Comparisons

We restrict the cross-system comparisons to the four largest magnitude predictors of user satisfaction in Figure 7, i.e. Task Duration, System Words per Turn, Task Completion and Sentence Error rate. Systems are identified by random numbers from 1 to 9; the assignments are the same as in 2000.

Figure 8 shows a box plot of user satisfaction by site. A one-way ANOVA for user satisfaction by site (F=8.235, p=.0001) using the modified Bonferroni statistic indicates several groups with overlapping performance. Site 1 is statistically worse than site 6 (p = 0.001) but statistically indistinguishable from all the other sites. Site 3 is statistically worse than site 6 (p = .04) and statistically better than site 9 (p = .01) but indistinguishable from sites 1,2,4,5 and 8. Sites 2,4,5 and 8 are all statistically better than site 9 (p < .001), but indistinguishable from one another and from sites 6 and sites 1 and 3.

Figure 9 shows a box plot of task duration by site. A one-way ANOVA for task duration by site (F=1.76, p=.09) using the modified Bonferroni statistic indicates only a trend toward statistical significance, due to the large variance in task duration. The large amount of variance is due to the variation in task type between multileg (complex), real, and round trips and also due to the fact that within these task types, sites 5 and 6 did not implement the functionality for cars and hotels (because they were not able to obtain a realistic database with the required information). Thus task durations for multileg trips for those sites are essentially the same as durations for real and round trips.

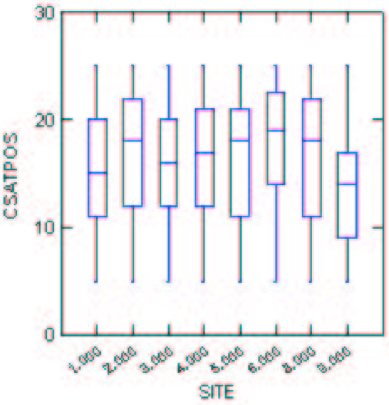

Fig. 8. User Satisfaction by Site

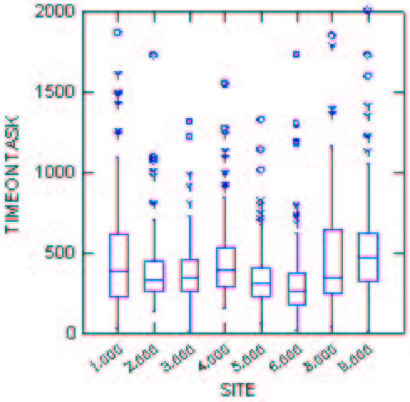

Fig. 9. Task Duration by Site in Seconds

As in 2000, system turn length is a positive predictor of user satisfaction. Our hypothesis is that this is because instructions and itinerary presentations tend to be longer. This is supported by the fact that removing this term from the PARADISE model makes task completion the most important predictor of user satisfaction, but reduces the model fit from .27 to .19. Figure 10 shows the mean system words per turn by site.

Figure 11 shows the percentages of task completion for each value of task completion by site. The figure shows large variations in completion for COMPLEX tasks, and indicates, as mentioned above, that some sites (5,6) chose not to implement this functionality. A one-way ANOVA for task completion by site (F=17.96, p=.000) using the modified Bonferroni statistic indicates several groups of performers. Sites (2,3,4,8) are the top performing group; their performance is indistinguishable from one another, although there is a trend (p = .09) for Site 8 to be statistically worse. Sites

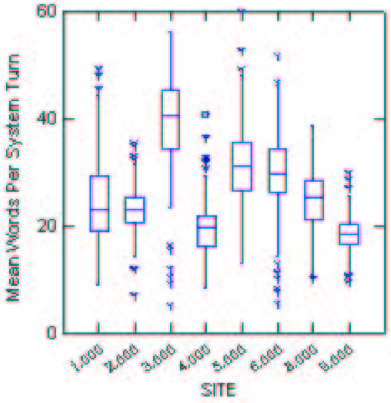

Fig. 10. System Words per Turn by Site

(1,5,6,9) define a lower performing group that are also statistically indistinguishable from one another. The companion paper presents results for task completion by type of trip [6].

Site	CTC=0	CTC = 1	CTC= 2
1	27.46%	55.56%	16.96%
2	13.70%	36.30%	50.00%
3	12.14%	46.43%	41.43%
4	13.29%	44.30%	42.41%
5	10.98%	89.02%	0.00%
6	13.45%	86.55%	0.00%
8	21.43%	42.86%	35.71%
9	36.81%	36.81%	26.39%
ALL	18.56%	55.92%	25.53%

Fig. 11. Cross Site Task Completion as Percentages: CTC=0 is no completion, CTC=1 is completion of the itinerary only and CTC=2 is completion of airline, car and hotel arrangments.

Figure 12 shows the means for Sentence Error by site and by gender. A one-way ANOVA for Sentence Error by site using the modified Bonferroni statistic shows significant differences between sites (F=23.07, $p < .0001$) and four overlapping groups of performers. Sites (2,4,5) form a top group, followed by groups for sites (6,8,9), (6,1) and (1,3). There is also a strong interaction between gender and sentence error by site (F = 2.9, p = 0.004); recognition performance at some sites is much better for female speakers, at others better for males, and for some there is no difference; see Figure 12. Most systems had the capability for voice-bargein, but Site 8 only had voice bargein available in some dialog states and Sites 2 and 6 turned it off for the data collection.

4. DISCUSSION

This paper presents the results of the 2001 evaluation for the DARPA Communicator program. The results establish performance baselines for real and complex tasks. The differential performance for real and complex tasks illustrate an issue in evaluation design. The complex tasks stress the system's capabilities for advanced conversational interaction as specified in the program goals, but tasks of this complexity occur rarely in the dialogues about real trips, i.e. these may not be *representative* tasks for this domain. The results indicate that the Communicator program goals of demonstrating a baseline functionality for conversational interaction for complex tasks has been achieved.

Site	Female	Male
1	34.6 %	34.3 %
2	20.0 %	11.2%
3	42.1 %	38.1 %
4	22.2 %	11.6%
5	23.5 %	31.3 %
6	34.4 %	27.3%
8	27.3 %	29.0 %
9	30.1 %	19.9 %
ALL	28.9 %	26.0 %
N	832	332

Fig. 12. Sentence Error rate by Site by Gender

5. REFERENCES

[1] J. Aberdeen. Darpa communicator logfile standard, 2000. http://fofoca.mitre.org/logstandard.

[2] H. Bonneau-Maynard, L. Devillers, and S. Rosset. Predictive performance of dialog systems. In *Language Resources and Evaluation Conference*, 2000.

[3] H. Wright Hastie, R. Prasad, and M. Walker. What's the trouble: automatically identifying problematic dialogues in DARPA communicator dialogue systems. In *Proc. of ACL*, 2002.

[4] P. Price, L. Hirschman, E. Shriberg, and E. Wade. Subject-based evaluation measures for interactive spoken language systems. In *Proc. of the DARPA Speech and NL Workshop*, pages 34–39, 1992.

[5] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. DARPA Communicator dialog travel planning systems: The June 2000 data collection. In *EUROSPEECH 2001*, 2001.

[6] M. Walker, A. Rudnicky, J. Aberdeen , E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, R. Prasad, S. Roukos, G. Sanders, S. Seneff and D. Stallard. DARPA Communicator Evaluation: Progress from 2000 to 2001. In *ICSLP 2002*. 2002.

[7] M. A. Walker, C. A. Kamm, and D J. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

[8] M. A. Walker, R. Passonneau, and J. E. Boland. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proc. of the Meeting of the Association of Computational Lingustics, ACL 2001*, 2001.