

DARPA COMMUNICATOR EVALUATION: PROGRESS FROM 2000 TO 2001

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

ABSTRACT

This paper describes the evaluation methodology and results of the DARPA Communicator spoken dialog system evaluation experiments in 2000 and 2001. Nine spoken dialog systems in the travel planning domain participated in the experiments resulting in a total corpus of 1904 dialogs. We describe and compare the experimental design of the 2000 and 2001 DARPA evaluations. We describe how we established a performance baseline in 2001 for complex tasks. We present our overall approach to data collection, the metrics collected, and the application of PARADISE to these data sets. We compare the results we achieved in 2000 for a number of core metrics with those for 2001. These results demonstrate large performance improvements from 2000 to 2001 and show that the Communicator program goal of conversational interaction for complex tasks has been achieved.

1. INTRODUCTION

The objective of the DARPA Communicator project is to support rapid development of multi-modal speech-enabled dialog systems with advanced conversational capabilities. Figure 1 illustrates the Communicator challenge problem; a system must support complex conversational interaction to complete this task within 10 minutes.

You are in Denver, Friday night at 8pm on the road to the airport after a great meeting. As a result of the meeting, you need to attend a group meeting in San Diego on Point Loma on Monday at 8:30, a meeting Tuesday morning at Miramar at 7:30, then one from 3-5 pm in Monterey; you need reservations (car, hotel, air).
You pull over to the side of the road and whip out your Communicator. Through spoken dialog (augmented with a display and pointing), you make the appropriate reservations, discover a conflict, and send an e-mail message (dictated) to inform the group of the changed schedule. Do this in 10 minutes.

Fig. 1. Darpa Communicator Challenge Problem

During the course of the Communicator program, we have been involved in developing methods for measuring progress towards the program goals and assessing advances in the component technologies required to achieve such goals. In previous work, we report on a data collection experiment with nine participating

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

Communicator systems in the travel planning domain [10]. During 2001, a second experiment, involving a subset of eight Communicator systems, was run. This paper describes the evaluation methodology and results of the experiments in 2000 and 2001. In a companion paper, we describe the 2001 evaluation and provide cross-system performance comparisons [11]. Section 2 compares the evaluation paradigm and experimental design for 2000 and 2001. Section 3 describes the application of the PARADISE framework [12], and compares the models we learned in 2000 and 2001. Section 4 compares the 2000 and 2001 results for metrics that PARADISE predicts are important and argues that this demonstrates that the Communicator program goals have been achieved. We sum up in Section 5.

2. EXPERIMENTAL DESIGN

Figure 2 summarizes advances in the evaluation paradigm from 2000 to 2001. The 2001 experiment was designed to test assumptions implicit in the 2000 experiment and improve upon that design. A major motivation was to test the systems in a more realistic setting. Eight systems from AT&T, BBN, University of Colorado, Carnegie Mellon University, IBM, Lucent Bell Labs, MIT, and SRI participated in the evaluation. In 2000, MITRE had also fielded a system to demonstrate plug-and-play of Communicator components. All systems accessed live data in the travel domain.

In 2000, we conducted a controlled experiment with recruited subjects who called each of the nine systems over three 3-day periods. The subjects conversed with the systems to complete 7 fixed and 2 open tasks. The fixed scenarios consisted of 3 domestic one way trips (DOW), 2 domestic round trips (DRT), and 2 international round trips (INT). The open tasks were planning an intended business trip and vacation. We used a textual tabular presentation for the fixed scenarios as an attempt to avoid “putting words in the user’s mouth”, although this was unsuccessful as discussed below. Figure 3 shows a sample scenario for an international round trip.

In 2001, the evaluation ran for 6 months with systems continuously accessible via a toll-free number. We conducted a within-system rather than a within-subject experiment in order to allow users to learn the interaction paradigm of their system and to allow systems to adapt to their users; we believed that continuous use of a particular system would better approximate actual conditions of use and provide data for analyzing the effects of expertise with a system. We collected more data and many dialogs for complex trips that demonstrate conversational interaction for complex tasks.

The 2001 experiment involved both SHORT and LONG subjects. All subjects were frequent travelers (people who make at least six trips per year). The SHORT users called their assigned system four times over the six month period when they had to

	2000	2001
Period	9 days	6 months
Time	Fixed Hours	Continuous
Design	Within subject	Within System
NumCalls	662	1242
Call Types	225 OneWay, 300 Round Trip, 137 Real	198 RoundTrip, 350 Complex, 694 Real
Scenarios	Text Tabular	Recorded Audio

Fig. 2. Evaluation Paradigm, 2000 to 2001

Your Preferences	
Preferred airline:	American Airlines
First Leg	
Starting location (home):	Phoenix, Arizona
Destination:	Seoul, Korea
Departure date:	Wednesday, October 11, 2000
Departure time:	Anytime
In Seoul	
Rental car:	No
Hotel:	No
Second Leg	
Return flight:	Yes, to Phoenix
Return date:	Sunday, October 15, 2000
Return time:	Anytime

Fig. 3. Task 7 Tabular Presentation from 2000

make travel plans. In addition to their real trips, the LONG group performed six fixed tasks. The aim of the fixed scenarios was to establish performance for complex tasks and allow comparisons with 2000. The fixed scenarios are described via their features in Figure 4. Scenarios 1 and 2 are simple round trips comparable to those in 2000 [10]. Scenarios 3 to 6 are COMPLEX tasks similar to the Challenge task in Figure 1; they require multiple legs (Multi-Leg), may require the user to fly on a particular airline (Airline?) and to make car and hotel arrangements (CarHotel?). Scenario 3 is shown in Figure 5. 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. To address concerns raised in 2000 about whether subjects seriously attempt to complete the tasks, subjects 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. 4. Parameters of Fixed Scenarios

In 2001, we experimented with a new method for scenario presentation. We were concerned with two aspects of the tabular pre-

sentation used in 2000. First, subjects had simply read off the table entries in response to system queries which resulted in unnatural responses such as *Wednesday, October 11th, 2000*. Second, we believed this presentation biased users against taking the initiative in the dialog by giving them the impression that their role is simply to provide the system with the values for the slots in the table. We also hypothesized that the problem of putting words in the user's mouth would only arise when the subject can actually read the description while simultaneously interacting with the system [8, 6]. Thus in 2001, users were provided with recorded speech descriptions of the tasks via an IVR system. See Figure 5. Users could listen repeatedly to the recording but were instructed to note the important points. In addition, we designed the verbal descriptions to vary common lexical items such as *fly, travel, go, depart*. Previous work suggests that normal comprehension and memory processes would leave subjects with an encoding of the meaning of these descriptions, but no memory for surface syntax [8, 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. 5. Scenario 3: A complex trip involving multiple legs, airline constraint, car and hotel arrangements

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 route in the last year. The user satisfaction survey is identical in 2000 and 2001, and is used to derive a quantitative satisfaction measure ranging from 5 to 25 [11]. 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.

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. 6. Metrics per Call.

A total of 662 dialogs were collected in 2000, and a total of 1242 were collected in 2001, with both logfiles and completed surveys. See Figure 2. Each dialog has: (1) logfiles generated with the logfile standard [1]; (2) user surveys; (3) metrics derived from the logfiles; (4) ASR and hand transcriptions of user utterances; (5) Task completion metrics. Metrics derived from this data and used in analysis are given in Figure 6.

3. GENERALIZATIONS VIA PARADISE FROM 2000 TO 2001

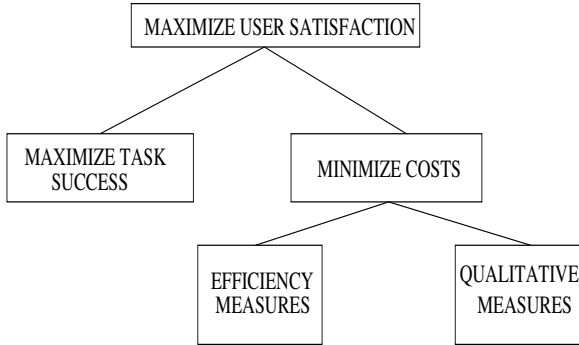


Fig. 7. PARADISE’s structure of objectives for spoken dialog performance

The PARADISE evaluation framework uses methods from decision theory [4] to combine a disparate set of performance measures (i.e., user satisfaction, task success, and dialog cost) into a single performance evaluation function [12, 5, 2]. The use of decision theory requires a specification of both the objectives of the decision problem and a set of measures (known as attributes) for operationalizing the objectives. The PARADISE model is based on the structure of objectives in Figure 7; it posits that the overall goal of a dialog system is to maximize usability, which is often measured by user satisfaction ratings [3, 9, 7]. The model further posits that two types of factors are potential contributors to user satisfaction (namely task success and dialog costs), and that two types of factors are potential contributors to costs: efficiency measures and dialog quality measures. Multivariate linear regression is used to derive a predictive model of user satisfaction as a function of the task success and dialog cost measures.

Figure 8 describes the linear model learned by applying PARADISE to the 2001 corpus which accounts for 27% of the variance in user satisfaction ($R = .52$). According to the model, task duration is the single largest contributor to user satisfaction ($-.44$), followed by the average number of words in system turns (.41) and task completion (CTC) (0.32). Overall improvements in task completion may have led task duration to be a more important predictor as compared with 2000. Sentence error rate (SERR), the number of overlaps between system and user turns, and the number of user words per turn are also significant predictors.

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

Fig. 8. PARADISE derived linear model for 2001.

Figure 9 shows the linear model learned via PARADISE on the 2000 corpus. This model accounts for 38% of the variance in user

Factor	Coefficient
Task Completion (ESC)	0.43
Sentence Accuracy (SACC)	0.21
Task Duration	-0.15
System Turn Duration	0.14

Fig. 9. PARADISE derived linear model for 2000.

satisfaction. The metrics were calculated slightly differently in 2000 (Sentence Accuracy SACC rather than SERR), but note that the actual factors that were predicted as important in 2000 are also the 4 most important factors in the model learned for 2001. The difference in the model fits from 2000 to 2001 indicates more variability in the 2001 dialogs. This is not surprising. In 2001, there was a greater difference in task complexity across tasks and there were two different populations of users, SHORT and LONG. In addition, subjects used one system over six months rather than calling all systems over a few days.

4. ACCOMPLISHMENTS FROM 2000 TO 2001

This section summarizes the accomplishments of the Communicator program by comparing progress in 2001 with baselines established in the 2000 data collection experiment for the three metrics predicted by PARADISE to be the most important (Task completion, Sentence accuracy or error, and Task Duration). We assume that the open trips from 2000 are comparable to the real trips in 2001. Figure 10 shows the means for user satisfaction for different types of trips. As described above, the 2000 data included domestic one-way (DOW) trips that were not tested in 2001. The figure shows increases in satisfaction for domestic (DRT) and international round trips (IRT) and the performance baselines established for the two types of complex trips, multileg trips that required car and hotel (MultiCH) and those that also required satisfying an airline constraint (MultiAirCH). As the figure shows, the complex tasks are more difficult and result in lower user satisfaction.

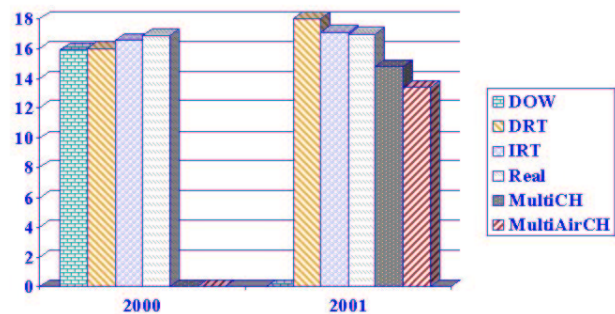


Fig. 10. Mean User Satisfaction from 2000 to 2001 by Trip Type

The systems as a whole also improved on every single dimension that PARADISE predicts to be a strong contributor to user satisfaction. Figure 11 details the large increases in rates of task completion for domestic (DRT) and international round trips (IRT) and for the real trips (Real). The figure shows that just completing the airline portion of the multileg trips (Multi) was more difficult, and that satisfying additional constraints such as an airline constraint or car and hotel arrangements added to the task complexity.

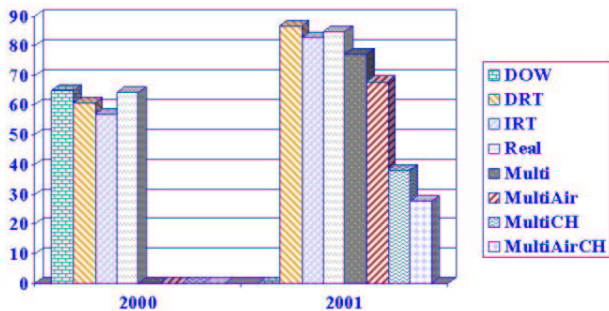


Fig. 11. Task Completion from 2000 to 2001 by Trip Type

Figure 12 shows that we maintained task durations for simple trips, but complex trips take more time. However, the mean durations for complex trips are below the 10 minutes specified in Figure 1, demonstrating that the program goals were achieved.

Figure 13 details the large improvements in ASR accuracy from 2000 to 2001 for tasks that were not completed (CTC=0), tasks where an airline booking was completed (CTC=1) and tasks where airline and car and hotel bookings were completed (CTC=2). The improvement is even more notable since in 2001: (1) more systems supported voice barge-in; (2) the user population included non-native speakers; and (3) subjects could use any type of phone.

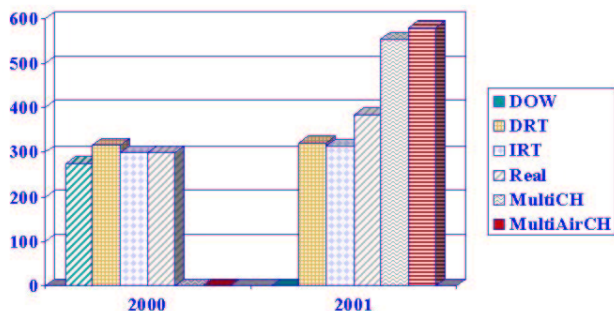


Fig. 12. Task Durations (secs) from 2000 to 2001 by Trip Type

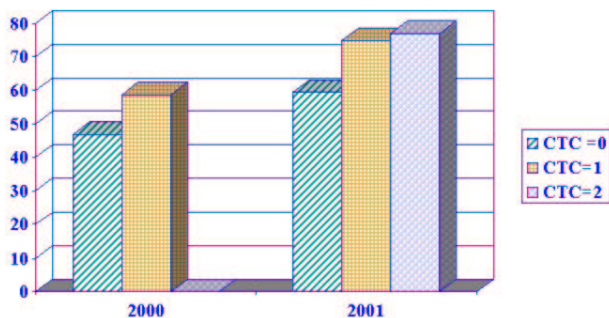


Fig. 13. ASR Performance, 2000 to 2001, by completion status

5. DISCUSSION

This paper compares the experimental paradigm and results of the 2000 and 2001 evaluations for DARPA Communicator. We show

large performance improvements from 2000 to 2001 for significant metrics and establish performance baselines for complex tasks. The rates of completion and the task durations in Figures 11 and 12 show that, in the large, the program goals have been achieved. These results provide a baseline for future work to demonstrate further improvements in conversational capability for complex tasks.

6. 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] C.A. Kamm. User interfaces for voice applications. In Roe and Wilpon, eds, *Voice Communication between Humans and Machines*, pp. 422–442. National Academy Press, 1995.
- [4] R. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley, 1976.
- [5] L. Lamel and S. Rosset. Considerations in the design and evaluation of spoken language dialog systems. In *ICSLP*, 2000.
- [6] W. J. M. Levelt and G. Kempen. Semantic and syntactic aspects of remembering sentences: A review of some recent continental research. In A. Kennedy and A. Wilkes, editors, *Studies in Long Term Memory*. John Wiley, 1975.
- [7] J. Polifroni, L. Hirschman, S. Seneff, and V. Zue. Experiments in evaluating interactive spoken language systems. In *Proc. of the DARPA Speech and NL Workshop*, pp. 28–33, 1992.
- [8] J. D. Sachs. *Recognition memory for syntactic and semantic aspects of connected discourse*. PhD thesis, University of California Berkeley, 1967.
- [9] E. Shriberg, E. Wade, and P. Price. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In *Proc. of the DARPA Speech and NL Workshop*, pp. 49–54, 1992.
- [10] 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. Rudnicki, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. DARPA Communicator dialog travel planning systems: The June 2000 data collection. In *EUROSPEECH 2001*.
- [11] M. Walker, A. Rudnicki, 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. Darpa Communicator: Cross-system results for the 2001 evaluation. In *ICSLP 2002*.
- [12] 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.