DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection

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, S. Whittaker

AT&T Labs, MITRE, University of Louisiana, SRI International, NIST, MITRE, NIST,AT&T Labs, AT&T Labs, IBM, University of Colorado, MIT, Lucent Bell Labs, AT&T Labs, Carnegie Mellon University, NIST, MIT, BBN Technologies, AT&T Labs

walker@research.att.com

Abstract

This paper describes results of an experiment with 9 different DARPA Communicator Systems who participated in the June 2000 data collection. All systems supported travel planning and utilized some form of mixed-initiative interaction. However they varied in several critical dimensions: (1) They targeted different back-end databases for travel information; (2) The used different modules for ASR, NLU, TTS and dialog management. We describe the experimental design, the approach to data collection, the metrics collected, and results comparing the systems.

1. Introduction

The objective of the DARPA Communicator project is to support rapid, cost-effective development of multi-modal speech-enabled dialog systems with advanced conversational capabilities. In order to make this a reality, it is important to be able to evaluate the contribution of various techniques to users' willingness and ability to use a spoken dialog system [15]. In June of 2000, we conducted an exploratory experiment with 9 participating communicator systems. All systems supported travel planning and utilized some form of mixed-initiative interaction. However the systems varied in several critical dimensions: (1) They targeted different back-end databases for travel information; (2) System modules such as ASR, NLU, TTS and dialog management were typically different across systems.

The experiment was designed by the Evaluation Subcommittee composed of representatives of each Communicator site and NIST. A logfile standard was developed by MITRE and used by all systems to collect a set of core metrics for making cross-site comparisons[9]. These are described on the Evaluation Committees WebPage [10]. We also collected user satisfaction metrics via a web-based survey. The results, to be discussed in more detail below, show that user satisfaction differed considerably across the 9 systems. Subsequent modeling of user satisfaction applying the PARADISE framework [16] gave us some insight into why each system was more or less satisfactory. While other metrics were also significant predictors of user satisfaction, the four metrics of task completion, task duration, recognition accuracy and mean system turn duration accounted for 38% of the variance in user-satisfaction. Section 2 explains the experimental design and Section 3 presents the results. Section 4 discusses future plans.

2. Experimental Design and Setup

Nine different Communicator travel planning systems participated in the data collection, one from each of AT&T Labs, BBN Technologies, Carnegie Mellon University, University of Colorado, IBM, Lucent Bell Labs, MITRE, SRI International. Here we report results anonymously by a random number between 1 and 9 assigned to each site.

We ran a controlled experiment in which a set of realistic subjects from the target population of frequent travelers interacted with each of the 9 Communicator spoken dialog systems. We recruited 72 native U.S. English speakers to call all 9 systems over 3 periods of 3 days to plan travel tasks according to a set of 9 realistic scenarios. Subjects carried out the scenarios in a fixed order. The goal was to have 8 dialogs per task per system, but since not all subjects called all systems, the resulting corpus consists of 662 dialogs.

The task scenarios consisted of 7 *fixed* and 2 *open* scenarios. The 7 *fixed* scenarios were designed to vary task complexity, where task complexity for this purpose was defined simply as the number of constraints that the user had to communicate to the system. These were presented to the user in tabular format. The *open* scenarios were defined by the user. After completing 7 pre-defined tasks with 7 of the systems, the users were asked to use the remaining two systems to *plan a recent or intended business trip* and *plan a vacation*. By asking the users to define their own tasks, the *open* scenarios were intended to approximate the conditions under which these systems would be used in the field [1], although as we discuss below this intention was not achieved.

The dialogs were recorded in full at NIST by connecting each call through a central call router. Each site provided a standard logfile, as well as transcriptions and recordings user utterances. At the end of each call, users gave subjective feedback via a web survey.

The Communicator data collection was designed to make it possible to apply the PARADISE evaluation framework which integrates and unifies previous approaches to evaluation [16,2,10,3,4]. This framework posits that maximizing user satisfaction is the system's overall objective and that task success and various interaction costs calculated as metrics can be used as predictors of user satisfaction.

Metrics collected per call consisted of objective metrics extracted from the logging and subjective metrics collected via a survey. The survey was used to calculate User Satisfaction by asking the user to specify the degree to which they agreed with the set of statements below on a 5 point Likert scale [5,7,14].

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

The values of the responses were then summed, giving a per dialog measure ranging from 5 to 25. In addition, a ternary definition of Task Completion was annotated by hand for each call. We distinguish between exact scenario completion (ESC), other scenario completion (OTHER) and no scenario completion (NOCOMP). This metric arose because some callers completed an itinerary other than the one assigned. This may have resulted from caller's inattentiveness, e.g. she didn't correct the system when it misunderstood. In this case, the system could be viewed as having done the best it could with the information provided. This argues for defining Task Completion as ESC + OTHER. However, examination of the dialogs suggests that sometimes the OTHER category arose as a rational reaction to repeated recognition error. The fact that 85% of the surveys included comments also supports the conclusion that users were generally attempting to complete the described scenarios. Thus we decided to distinguish cases where users completed the assigned task, completed some other task, or the call ended without itinerary completion. In the analysis below, we present results for both exact scenario completion (ESC only) and ANY scenario completion (ESC + OTHER). Because we were concerned with user behavior in this experimental setup, we also separately hand tagged each dialog for user behavior. Descriptions are provided below. The set of metrics were:

- **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**: Word Accuracy, Sentence Accuracy, Mean Response latency, Response latency variance
- **Task Success Metrics**: Perceived task completion, Exact Scenario Completion, Any Scenario Completion
- **User Satisfaction**: Sum of TTS performance, Task ease, User expertise, Expected behavior, Future use.

3. Experimental Results

The experiment resulted in 662 dialogs with dialogs per system numbering between 60 and 79. Variation in the number of dialogs per system and task resulted from problems with system stability and the stability and load on the central call router. Thus, although the design was a within-subjects design, only 49 of the subjects actually called all 9 systems. Here, we report an analysis of all the data.

User Behavior: We labeled each dialog with one of 6 types of user behavior. Percentages per behavior are below. A Goal-Directed user (71%) is completely focused on the task and never exhibits any of the following behaviors. An Initially-Inattentive user (1.3%) took some seconds to respond to the system, either not responding or answering wrongly. The False-Acceptance users (8.8%) failed to correct a system misunderstanding. The Wrong-Information users (2.4%) provided information inconsistent with a *fixed* scenario. The Scenario Switch category (4.3%) were *open* tasks where the user changed plans during the dialog (often in response to repeated recognition error). The Unknown case (10.7%) covers those dialogs where no logfile was generated (i.e., the system either crashed or prematurely ended the call).

User Satisfaction: We initially examined differences in user satisfaction across the 9 systems as shown in the box plot in Figure 1. The box plot indicates the full range of values for user satisfaction, and the interquartile range as a box within that. The median of the distribution is shown by a horizontal line within the box. A one-way ANOVA for user satisfaction by site using the modified Bonferroni statistic shows that the user satisfaction metric distinguishes four groups of performers with sites 3,2,1,4 in the top group, sites 4,5,9,6 in a second group, and sites 8 and 7 defining a third and a fourth group.

We also examined the relationship between the individual components of user satisfaction, namely Task Ease, TTS Performance, User Expertise, Expected Behavior and Future Use and the cumulative user satisfaction measure. In contrast to previous work, we found that all of the components contributed similarly to the overall measure. The correlation between User Satisfaction and Task Ease was 0 .9, TTS Performance was .72, User Expertise was .83, Expected Behavior was .91 and Future Use was .91. This suggests that if only one question could be asked that the Future Use question could stand in for the rest. However, we also examined whether there were significant differences across systems in any of these components. As one might expect there were significant differences in all of these components, however the pattern for each component tended in the main to mirror the overall pattern shown in Figure 1

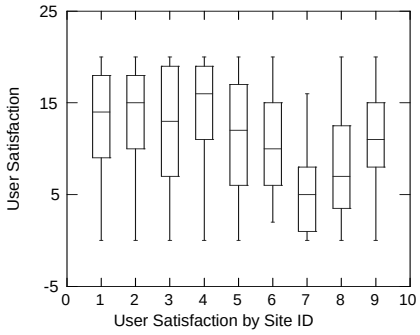

Figure 1: User Satisfaction by Site ID.

We then applied PARADISE to develop models of user satisfaction and examined differences across sites for metrics that were significant predictors of user satisfaction. In order to provide a baseline performance model, we initially derived a model using a set of core metrics typically available for any dialog corpus, namely task completion, task duration and sentence accuracy. Models of user satisfaction based on these

core metrics account for 35% of the variance in user satisfaction. A 2-tailed t-test shows that these predictors were significant at the p=.0001 level.

The finding that task completion and recognition performance are significant predictors duplicates previous results [16]. The fact that task duration is also a significant predictor may simply indicate larger differences in task duration in this corpus. When all of the metrics available from the Communicator logfile standard are utilized, the best model fits can be obtained by the addition of only one other metric, namely System Turn Duration. The model accounts for 38% of the variance in user satisfaction[1] . The learned model is that User Sat is the sum of:

.43 * ESC1 - 1.5 * TaskDur + .21 Sacc + .14* SysTurnDur (1)

where ESC1 is the ternary task completion metric, TaskDur is the time on task, Sacc is Sentence Accuracy and SysTurnDur is the average time for each system turn. We then examined how the metrics significant for predicting user satisfaction distinguished among sites .

Task Completion: We first examined Task Completion by Scenario and by Site. We examined Task Completion by Scenario in order to determine whether the experimental manipulation of task complexity had indeed made some tasks more difficult, and whether there were differences in completion between the *open* tasks (scenarios 8,9) and the *fixe*d tasks (scenarios 1 to 7).

A one-way ANOVA for ESC and for ANY by session showed significant differences between sessions for the ESC metric (p = .01), but not for the ANY metric. One reasons for this is that, contrary to our expectations, users more readily modified their travel plans for the *open* tasks, i.e. if the system couldn't understand Denpasar airport in Bali, and thought the user wanted to fly to St. Petersburg in Russia, the users changed their vacation plans. The fact that there were no differences for ANY completion by session also suggests that experience with the systems did not improve the users' ability to complete their tasks. This may have been because users called each system only once. See the discussion below.

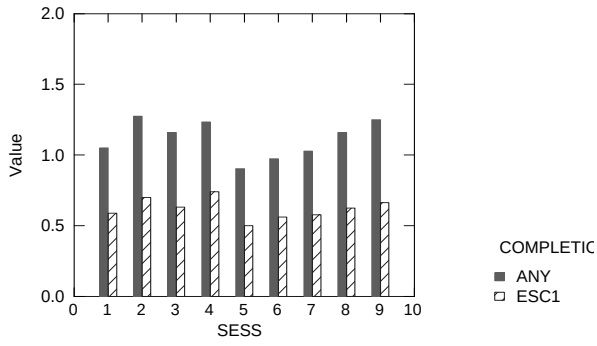

Figure 2: Completion by Session ID. Each user did all tasks in the same sequence, but order of systems varied.

One-way ANOVAs for ESC and ANY by site indicate significant differences in task completion (F> 13.9, p < .001). A one-way ANOVA for ESC by site using the modified Bonferroni statistic for multiple comparisons defines three groups of performers, with sites 2,3,4,1,5 in the top group, sites 5,6,9 in a second group and sites 8,7 in the lowest group. A one-way ANOVA for ANY Scenario Completion by site using the modified Bonferroni statistic defines the same three groups. Figure 3 shows task completion performance .

Task Duration:. A one-way ANOVA for Task Duration by site using the modified Bonferroni statistic for multiple comparisons indicates significant differences in Task Duration (F=10.8, p = .0001), and distinguishes three groups of performers with site 3 in the top group (shortest durations), sites 1, 2, 4, 7 in a second group and sites 5, 6, 8, 9 in a third group. However the interesting case for Task Duration is for calls in which an itinerary is completed, since some failed tasks were due to system crashes early in the dialog. A box plot in Figure 4 indicates the performance of each site for Task Duration for the ANY task completion subset. A one-way ANOVA for Task Duration by site for this subset also indicates significant differences (F= 10.9, p < .0001)

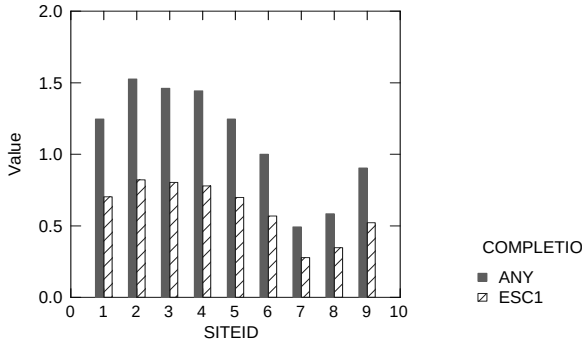

Figure 3: Completion by Site ID

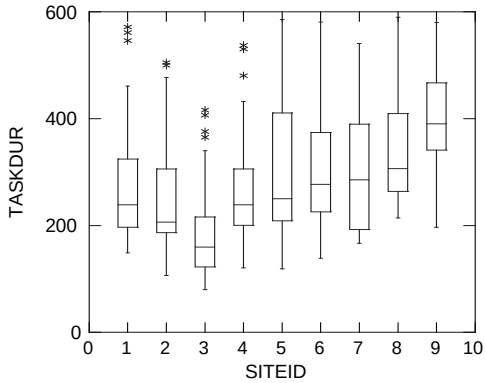

Figure 4: Task Duration for Completed Tasks by Site ID.

Sentence Accuracy: A one-way ANOVA for Sentence Accuracy by site using the modified Bonferroni statistic showed significant differences between sites (F= 40.5, p < .0001) and two groups of performers (1, 2, 4, 8, 9 and 3, 5, 6, 7). Some systems did not support voice barge-in, and this correlated with higher accuracy. However, there was also a strong interaction between gender and sentence accuracy by site; recognition performance at some sites was much better

[1] Tree models using the full set of metrics account for 36% of the variance in user satisfaction

for female speakers, at others better for males, and for some there was no difference. See the box plot in Figure 5. Furthermore, although the experimental design attempted to balance for gender, subjects were added as users failed to call. In the end, the user population was 64% female and 36% male, causing problems for sites with poor recognition performance for female speakers.

System Turn Duration: The PARADISE model above indicates that System Turn Duration is *positively* correlated with satisfaction. However because flight presentation utterances were longer than other system turns, this may simply indicate the presentation of potential itineraries.

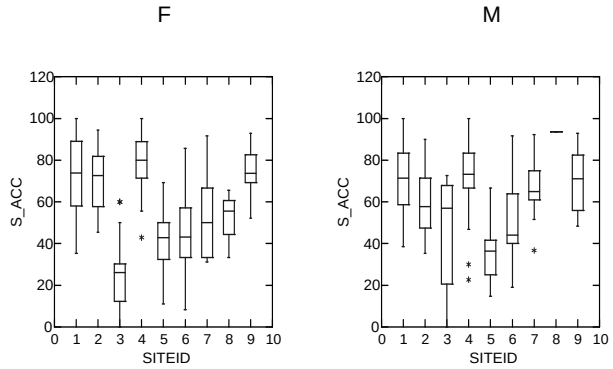

Figure 5: Sentence Accuracy Females vs. Males by Site ID.

User Words per Turn: Finally, even though User Words per Turn was not a significant predictor of user satisfaction, we examined this metric as an indicator of user initiative. A one-way ANOVA by site revealed that there were significant differences among sites in the amount of initiative that users took. In particular site 5 was the only site in which at least half the dialogs had an average user words per turn greater than 4. Further examination of the dialogs from that site suggests that this may be due to the use of more open prompts, both at the beginning of the dialog, e.g. *Tell me about your travel plans*, and at other phases of the dialog. For example, when system 5 was having trouble understanding the user, it would make open-ended suggestions such as *Try asking for flights between two major cities* rather than using directive prompts such as *Please tell me your destination*

4. Discussion and Future Work

Our analysis identified several issues with the 2000 data collection. The first issue was the within-subjects design. We thought this would allow us to learn about comparisons across systems, but we believe this design may result in using behavior reflecting the *least common denominator*; as users called one system after another, they accommodated their behavior to the least flexible system. A second issue was the tabular presentation of the *fixed* scenarios; users took very little initiative and this presentation format may lead them to believe a conversation is simply filling in the slots in the table. A third issue was that users doing the *open* scenarios were more likely to change their task midstream (20% vs. 5%); thus these scenarios **did not** approximate users planning real trips.

We expect to address these problems in several ways. First, the second data collection scheduled to begin in April 2001 is a longitudinal experiment (6 months) where users repeatedly use the same system. This should better approximate the real conditions of use and users should be able to learn how to use the systems as well as providing system designers an opportunity to explore algorithms for system adaptation to users. Second, all users are frequent travelers who call their system to plan real trips. There will be both SHORT and LONG users. The LONG users will perform 4 fixed learning scenarios in the beginning of the data collection; this will provide data for adaptation algorithms and will create an *expert* population. Third, we hope to use audio presentation of the learning tasks to address the problems of tabular presentation while avoiding the problem of putting words into the user's mouth. The experimental design is described in more detail on the Evaluation Committee web page [9].

In related work we have developed additional qualitative metrics based on dialog act tags for comparing Communicator systems[17] and found that dialog act metrics improve models of user satisfaction. We plan to utilize these metrics in the 2001 data collection.

Acknowledgments

This work was supported by DARPA Grant MDA 972 99 3 0003.

References

[1] Baggia, P., G. Castagneri, and M. Danieli, 1998. Field trials of the Italian ARISE train timetable system. In IVTTA.

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

[3] Hirschman, L., M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann, 1993. Multi-site data collection and evaluation in spoken language understanding. In Proc. of the Human Language Technology Workshop.

[4] Hirschman, Lynette, 2000. Evaluating spoken language interaction: Experiences from the DARPA spoken language program 1990--1995. In S. Luperfoy (ed.), Spoken Language Discourse . Cambridge, Mass.: MIT Press.

[5] Jack, M.A., J. C. Foster, and F. W. Stentiford, 1992. Intelligent dialogs in automated telephone services. In Int. Conf. on Spoken Language Processing,

[6] Larsen, L. B, 1999. Combining objective and subjective data in evaluation of spoken dialogs. In ESCA Workshop on Interactive Dialog in Multi-Modal Systems .

[7] Love, S. R. T Dutton, J. C. Foster, M. A. Jack, and F.W. M. Stentiford, 1994. Identifying salient usability attributes for automated telephone services. In Int. Conf. on Spoken Language Processing,

[8] Polifroni, J. and S. Seneff, 2000. Galaxy-II as an architecture for spoken dialog evaluation. In 2nd International Conference on Language Resources and Evaluation (LREC)

[9] DARPA COMMUNICATOR Evaluation Committee WebSite. http://www.research.att.com/~walker/eval/eval.html

[10] Aberdeen, J. http://fofoca.mitre.org/logstandard.

[11] Rudnicky, A. I., 1993. Factors affecting choice of speech over keyboard and mouse in a simple data-retrieval task. In EUROSPEECH93.

[12] Sanderman, A., J. Sturm, E. den Os, L. Boves, and A. Cremers, 1998. Evaluation of the Dutch train timetable information system developed in the ARISE project. In IVTTA.

[13] Shriberg, E., E. Wade, and P. Price, 1992. Human-machine problem solving using spoken language systems

(SLS): Factors affecting performance and user satisfaction. In Proc. of the DARPA Speech and NL Workshop
[14] Sparck-Jones, K. and J. R. Galliers, 1996. Evaluating Natural Language Processing Systems . Springer.
[15] Stallard, D. (2000) Talk'n'Travel: A Conversational System for Air Travel Planning. In Proceedings Applied Natural Language Processing Conference, Seattle, Washington.
[16] Walker, M. A., C. A. Kamm, and D. J. Litman, 2000. Towards developing general models of usability with PARADISE. Natural Language Engineering: Special Issue on Best Practice in Spoken Dialog Systems
[17] Walker, M. A., and R. Passonneau, 2001. DATE: A Dialog Act Tagging Scheme for Evaluation of Spoken Dialog Systems. Human Language Technology Conference. San Diego, March 2001.