Spoken Dialogue Evaluation for the Bell Labs Communicator System

Sungbok Lee, Egbert Ammicht, Eric Fosler-Lussier, Hong-Kwang Kuo, Alexandros Potamianos
Bell Labs, Lucent Technologies, NJ 07974, USA
sungbok, egbert, fosler, kuo, potam@research.bell-labs.com

ABSTRACT

Task-oriented dialogues collected by the Bell Labs Communicator system are analyzed based on objective dialogue metrics and user survey data on system usability. Analyses are presented that confirm and extend some of the previous results shown for other systems. An "error correction metric" is also introduced, and its effectiveness as an objective metric for system performance evaluation is investigated. Results indicate that both user satisfaction and task completion are better correlated with the error correction metric than with the recognition accuracy at the word or sentence level. Multiple regression analysis reveals that the most effective predictors of user satisfaction (among objective dialogue metrics tested in the study) are the number of user words in dialogue and the recognition accuracy at the "concept" level. These two metrics explain 27.2% of variation in user satisfaction. When perceived task completion judgment (a subjective measure that is the single best predictor of user satisfaction) is included in the regression, 45.6% of variation is explained. Adding more variables, such as dialogue duration, is found to have marginal or negative effects. Overall, recognition accuracy at the concept level (as opposed to plain word or sentence accuracy) is shown to be an important metric for dialogue evaluation. In this study, the dialogue system is also evaluated using subjective metrics at each dialogue turn to identify "hot-spots" in the dialogue. We also measure inter-labeler agreement with Communicator callers. Results indicate that it is difficult for independent human raters to reproduce the Likert ratings given initially by the Communicator callers. Thus, given the noisy nature of subjective metrics, an open question is how well objective metrics will ultimately be able to explain the individual variation in subjective evaluation.

General Terms

Dialogue evaluation, Communicator

Keywords

Dialogue metrics, error correction metric, system usability

HLT2002 '02 San Diego, California USA
.

1. INTRODUCTION

Evaluation of the quality of spoken dialogue systems continues to be a difficult task because of several factors: it is difficult to sum up the results of a human-conversation into one number expressing the quality of interaction. It is also hard to assign credit or blame to sub-parts of the system due to the complexity of inter-operating subparts of the system (e.g., the ASR engine, understanding module, spoken language generation, and world knowledge representation), particularly when only summary statistics are available for a particular conversation.

In the DARPA Communicator project, multiple sites have developed spoken dialogue systems that give information about travel reservations.[1] For each system, both objective and subjective metrics were collected for each dialogue; the objective measures were derived from system statistics such as duration of dialogue, number of user words, and recognition error, while the subjective measures were derived from user satisfaction surveys. While these two classes of variables are related, a linear regression framework [11] was only able to explain 38% of the variance in user satisfaction using the objective measures [9].

In the first part of this paper, we introduce an objective dialogue metric and expand on some of the analyses in [9], using the dialogues collected by the Bell Labs Communicator system [1, 6] during the DARPA 2001 Communicator Evaluation period. Since task completion, in addition to user satisfaction, is generally considered to be one of the most important factors in deploying a dialogue system, [2] we decided to investigate the effects of task completion on objective and subjective measures.

Dialogue-level metrics, however, do not really help system designers pinpoint "hot spots" in the dialogue where the system needs improvement. Thus, we initiated a second study where three naive subjects and three expert subjects evaluated the system responses in a given dialogue on a turn-by-turn basis. The findings presented in the second part of the paper suggest that although the averaged score of the turn-by-turn rating of each system response and the overall system usability rating at the end of a given dialogue are consistent for individuals, the inter-subject variation is large. Therefore, subjective judgments of dialogue quality may be difficult to reproduce due to the inter-subject variability.

2. BELL LABS COMMUNICATOR SYSTEM AND DIALOGUE COLLECTION

[1] See the Web pages: http://www.darpa.mil/ito/research/com/index.html, http://fofoca.mitre.org for detailed information about the DARPA Communicator project.

[2] Task completion was also found to be one of the significant factors in the multiple linear regression analysis of user satisfaction [9].

2.1 System Description

The Lucent Bell Labs Communicator system is a mixed-initiative spoken dialogue system applied to a travel reservation domain. The main components of the system are a semantic parser/interpreter, the context tracking/pragmatic analysis module, the dialogue manager, and the system response generation module, in addition to the ASR and TTS engines. Novel features of the system include a domain-independent understanding module with ambiguity resolution and pragmatic analysis that allows the user to provide values for specific attributes, and to directly or indirectly refer to attributes and/or values [1]. These features allow the user to change the system focus and to ask for information about existing attributes and values. It further provides a powerful error correction capability that can handle system errors without explicit confirmation, thus minimizing unnecessary interactions with user.

The system takes in spoken or typed natural language text, and derives candidate attribute-value (AV) pairs (e.g., "Fly to Atlanta in the morning" produces <TOCITY>=Atlanta and <TIME>=morning). We model two types of ambiguities in the system: *value ambiguities*, where the system is unsure of the value for a particular attribute, and *position ambiguities*, where the attribute corresponding to a particular value is ambiguous. We score candidate values based on supporting evidence for or against the candidate. Raw AV pairs may have any number of scores attached (e.g., acoustic confidences and distribution frequencies). The raw AV data are used as evidence; the scores (possibly from multiple dialogue turns) are combined to derive a single individual score p with range $[0, 1]$ for each AV candidate. Additional evidence is provided by pragmatic analysis and by matching a hypothesis to the current context.

The system carries out a pragmatic analysis of the user response to a particular system prompt. Individual candidates that may be in error are identified, and candidate scores are modified based on the derived evidence. For example, a "no" response to an explicit value ambiguity disambiguation question, (e.g., "Are you leaving from Atlanta or from New York?") is taken as strong evidence against both values and the pragmatic confidence for both ('departure.city','Atlanta') and ('departure.city','New York') is reduced.

The user can also talk about attributes directly, and request specific actions from the system. In particular, we allow *information requests*, e.g., "what is the departure city?" to ascertain the value for a specific attribute; *clear requests*, e.g., "clear the departure city," to force the removal of all candidate values for a given attribute; *freeze requests*, e.g., "freeze the departure city," to inhibit the system from further changing the value of a particular attribute; *change requests*, e.g., "change Atlanta to New York," "change the departure city to New York," or even "not Atlanta, New York!"

Other features of the system not discussed here are an agenda-driven dialogue manager, dynamic and adaptive electronic forms, and an adaptive initiative module [2]. The system uses the Galaxy-II hub architecture as outlined in [8]. The system is highly modular; most modules are domain-independent. Some of the design principles used in the Bell Labs Communicator system design are outlined in [5]. For more information on the Bell Labs and other Communicator systems see [1, 6, 9, 7, 12].

2.2 Dialogue Collection

During the DARPA Communicator 2001 evaluation period of about six months, 215 calls were made by 28 paid subjects. For each call the recognized user utterances and corresponding system responses were logged with associated dialogue events (such as the start and end time of user turns) so that objective dialogue metrics can be measured from the log files [10]. After the completion of each call, callers also judged whether the travel task was completed or not and answered a set of five user satisfaction survey questions on the system usability based on the NIST-derived Likert paradigm [3, 10].

Among the 215 dialogues, 139 dialogues for which the user survey data on system usability exist are analyzed in this study. Of 139 dialogues, in 115 the assigned task of finding a suitable travel itinerary was completed, while in 24 dialogues the task was not completed.

TABLE I: MEAN AND STANDARD DEVIATION OF OBJECTIVE DIALOGUE METRICS

	Task Comp.		No Comp.	
Metric	mean	std	mean	std
Dialogue Duration (secs)	547	428	429	423
# of User Turns	33.6	25.8	31.2	24.8
# of User Words	85.6	79.4	94.3	99.7
# of User Words per Turn	2.61	1.36	3.33	1.42
# of Concepts	35.2	27.9	28.7	20.2
# of Concepts per Turn	1.13	0.23	1.18	0.31
Word Error Rate	23.5	12.9	24.9	12.7
Sentence Error Rate	28.8	15.6	34.3	15.9
Concept Err. Rate (CER)	16.5	11.9	20.9	15.9
Sentence CER	15.5	11.2	18.9	15.9
Error Correction Metric	4.7	5.9	8.0	9.9

3. DIALOGUE EVALUATION

3.1 Measurement of Objective Dialogue Metrics

Objective dialogue metrics computed for each dialogue included dialogue duration, the number of user turns, the number of user words, the number of user words per turn, the number of user concepts, and the number of user concept per turn. Additionally, accuracy metrics, such as word and concept error rates, were considered at both the word and sentence level. Except for concept accuracy, each metric can be computed directly from corresponding log files. The concept accuracy error rates are computed based on the outputs of the semantic parser of recognized and transcribed user utterances.

An additional objective metric introduced in this study is the number of redundant occurrences of a given concept such as "city-name" or "date" in user utterances in a dialogue. The so-called "error correction metric" is computed by subtracting the default number of a concept required to complete the travel task from the total number of the occurrences of the concept. For example, two occurrences of two different city names are enough for a round trip. More than two occurrences of the "city" concept indicate that a user tried to correct system errors (e.g., "not Boston, Austin") or that the user abandoned the current dialogue and began a new dialogue by saying "start over". The metric can be computed using the outputs of the semantic parser of either transcribed user utterances or recognized utterances in a dialogue. The error-correction metric is computed in this study only for the two concepts, "city" and "date"[3], using human-transcribed user utterances.

Mean value and standard deviation of each objective dialogue metric are shown in Table I.

It is observed that task-completed dialogues have a longer dialogue duration and a larger number of user turns, although ANOVA

[3]The two concepts account for 61% of the total number of concepts in this set of dialogues.

analysis indicates that the differences are not significant. However, the number of user words is larger for task-incomplete dialogues. Moreover, the number of user words per turn, which is independent of dialogue duration or the number of user turns, is found to be significantly greater for the incomplete dialogues (F=5.94, p=0.01). The number of concepts per turn also shows the same tendency. These observations indicate that a user is more verbose when the system is in trouble (or that verbosity gets the user into more trouble!).

In [9], sentence accuracy was found along with task completion to be a significant predictor of user satisfaction. Here, we see that word error rates for task completion and non-completion do not differ significantly, but the concept error rate increases greatly for non-completion; this rise correlates with sentence accuracy (r=0.759, p=0.01), suggesting that in [9], sentence accuracy was substituting for concept accuracy. In fact, it can be shown that accuracy in concept level is more strongly correlated with user satisfaction than the plain word level or sentence level. This will be discussed in more detail later.

The error correction metric is found to be significantly less for task-completed dialogues (F=5.25, $p < 0.05$). Its correlation to user satisfaction is stronger (r=-0.403, p=0.01) when compared to that of sentence error rate (r=-0.336) or word error rate (r=-0.278). Its correlation to the task completion is also stronger (r=-0.198, p=0.01) than sentence error rate (r=-0.139) or word error rate (r=-0.032). These observations may signify its usefulness on system performance evaluation.[4]

3.2 Analysis of User Survey Data

Users' judgments on the system usability at the end of a dialogue session are based on the Likert paradigm [10] in which a subject expresses the degree of agreement with each of five statements related to the system performance. The degree of agreement is expressed using a numeric value ranging from 1 (disagree) to 5 (agree).

Table II displays the mean Likert scores for five user survey questions, as well as an averaged score, as a function of task completion. All subjective scores are higher when the task is completed; ANOVA analysis shows that differences between the two groups of dialogues are significant ($p < 0.01$) except for the 3rd question ("Know What to Say"). It seems that the second and third questions are less affected by task completion. The first, fourth, and fifth statements explain 94% of variance in user satisfaction.

Correlations between the system usability rating and the objective dialogue metrics were examined for task-completed dialogues. As a representative value of system usability, the sum of five user survey scores is used. All the correlations given below are significant (p=0.01). Interestingly, the best (although not strong) correlation is obtained for total number of user words (r=-0.549), followed by the number of user turns (r=-0.495) and the concept error rate at the sentence level (r=-0.458). However, the correlation between the number of user turns and the word error rate is relatively weak (r=0.229). This suggests that longer dialogue duration or the larger number of user turns is not necessarily due to poor recognition performance; users prefer shorter interactions.

Considering both task-completed and task-incomplete dialogues, the best correlation to user satisfaction is still obtained for the total number of user words(r=-0.458), followed by the concept error rate at the word level (r=-0.425), the concept error rate at the sentence level (r=-0.417), and the number of user turns (r=-0.411), as shown in Table III. It is worth noting that the error correction metric shows a higher correlation to the user satisfaction (r=-.405, p=.01) than the sentence error rate does (r=-.336, p=.01). In order to identify relatively important objective dialogue metrics for predicting user satisfaction, multiple linear regression analysis was performed. The number of user words in dialogue and the recognition accuracy at the "concept" level were found to be the best predictors of user satisfaction among objective dialogue metrics tested in the study. The number of user words alone can explain 19.9% of variance in the user satisfaction and the two variables combined can explain about 27.2% of the variation. The effectiveness of the number of user words in the prediction of user satisfaction results from the

[4] Although monthly change of user satisfaction during the period of evaluation is roughly inversely correlated with the error correction metric, the degree of correlation was not strong enough to make any conclusion against other competitive variables such as the number of user turns.

TABLE II: AVERAGE SUBJECTIVE EVALUATION SCORES WITH AND WITHOUT TASK-COMPLETION

	Task Comp.		No Comp.	
Question	mean	std	mean	std
Easy to Get Info	3.61	1.37	1.62	1.05
Easy to Understand	4.02	1.09	3.10	1.42
Know What to Say	3.56	1.27	2.97	1.47
Know What to Expect	3.55	1.29	1.98	1.30
Future Use	3.31	1.50	1.69	1.04
Overall	3.61	1.15	2.23	0.90

Summary of responses to five questions:

1. "In this conversation, it was easy to get the information that I wanted" (**Easy to Get Info**),
2. "I found the system easy to understand in this conversation" (**Easy to understand**),
3. "In this conversation, I knew what I could say or do at each point of the dialogue" (**Know What to Say**),
4. "The system worked the way I expected it to in this conversation" (**Know What to Expect**), and
5. "Based on my experience in this conversation using this system to get travel information, I would like to use this system again" (**Future Use**).

TABLE III: CORRELATION BETWEEN USER SATISFACTION AND OBJECTIVE METRICS

Rank	Objective Metric	Pearson Correlation
1	Number of User Words	-.458
2	Concept WER	-.425
3	Concept SER	-.417
4	Number of User Turns	-.411
5	Error Correction Metric	-.403
6	Total Number of Turns	-.370
7	Duration	-.359
8	Sentence Error Rate (SER)	-.336
9	Word Error Rate (WER)	-.278

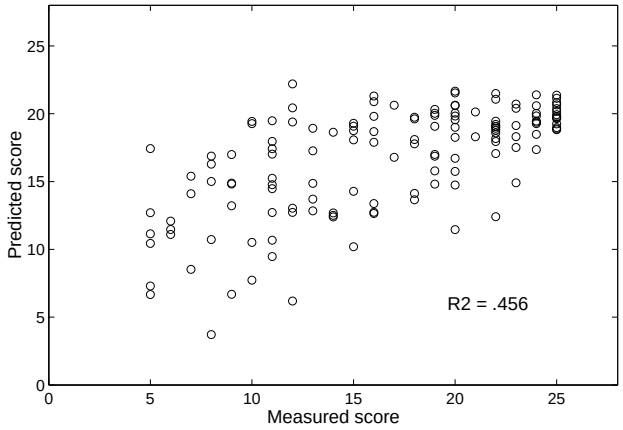

Figure 1: Distribution of the measured user satisfaction scores and predicted scores using the number of user words, concept word error rate, and task completion judgment

fact that the single variable is highly correlated with not only user satisfaction but also important objective metrics such as on-task dialogue duration, total number of turns, and recognition accuracy at the sentence levels.

When the perceived task completion judgment by users is added as another variable in the regression, about 45.6% of variation in the user satisfaction can be explained.[5] Adding other metrics such as dialogue duration show only marginal effects. A scatter plot of measured scores and predicted scores is shown in Figure 1. In [9] it was reported that 38% of variation in the user satisfaction data collected from 9 different sites can be explained by combining task completion judgment, sentence accuracy, ontask dialogue duration, and systemturn duration. The result of this study indicates that it may be possible to get more prediction power by substituting the sentence accuracy with the accuracy in the concept level and the system turn duration with the number of user words, respectively. A comparative study of the two user survey data sets may be needed to verify this assumption.

4. SYSTEM EVALUATION USING SUBJECTIVE METRICS AT EACH DIALOGUE TURN

We would like to locate "hot spots" in the spoken dialogue, i.e., system responses that are judged to be poor by the user. The possible causes of such problematic dialogue turns can be poor prompt wording, misrecognitions, or other semantic errors. In order to find these "hot spots", one must examine the dialogue turn-by-turn. However, it is not clear that post-hoc analyses can reproduce the subjective opinions of the original caller. Therefore, in these experiments, we also measured the inter-rater robustness of subjective metrics.

4.1 Method

Twenty dialogues (14 complete and 6 incomplete) were randomly selected from the 139 dialogues analyzed previously. Three naive subjects with no prior interaction with the Communicator system and three expert subjects, who have had experience designing and implementing one or more components of the Communicator system, were independently asked to evaluate each dialogue using the Likert questions of Table II. As opposed to the user survey, the questions were asked for every pair of the transcribed user utterance and the corresponding system response. Upon completion of a dialogue, the subjects then assigned the overall usability scores as the Communicator callers did. Most of subjects processed all twenty dialogues in a single session.

4.2 Analysis

In Figure 2, averaged scores of individual system response rating and overall rating assigned at the end of each dialogue are plotted for the two groups of dialogues. Each point is the averaged value across five question and dialogues in each group. Positive correlation and the linear tendency of data point distribution indicates that all subjects assign lower ratings to incomplete dialogues, indicating a general consensus on judging system usability. However, as is apparent from the ratings by subject 1 and subject 5, large inter-subject variability exists in terms of the magnitude of assigned values: Note that subject 1 used an extremely narrow range of scores.

The above observations suggest that subjects apply individual standards or criteria in assigning values to the system usability judgment. It thus may be difficult to reproduce human judgments in general. In fact, correlations of the overall scores among the subjects in the study including the Communicator callers are not strong and vary widely ($r_{\min} = 0.232$, $r_{\max} = 0.727$, p=0.01).

One of the major aims of the experiment was to isolate hot spots of the system determined as follows: the mean and standard deviation of the individual prompt scores are computed for each participant in the rating experiment. Outliers whose sum of scores are more than one standard deviation below the mean value are regarded as hot-spots determined by the subject. This procedure was repeated for each subject and hot spots which got four or five votes from 5 subjects were regarded as system's hot spots. Of 570 system responses presented, about 10.0% were regarded as the system's hot spots by the five subjects.

On average task-incomplete dialogues show a larger number of the hot spots (34.3 vs. 25.9) but it is not statistically significant. Examination of the system's hot spots also indicates that subject ratings are largely biased by speech recognition/understanding errors and vague system responses (e.g., "tell me what to change", "Thursday, July nineteenth is in the past. Would you like to continue on to another destination?"), which may or may not reflect system flaws.

5. CONCLUSION

We have presented some initial analyses of dialogues with our Communicator system, confirming and extending some of the results that have been previously shown for other systems. Specifically, we found that the error correction metric introduced in this study could be an effective objective metric for system performance evaluation. It is also better correlated with both user satisfaction and task completion than the recognition accuracy at the word or sentence level. More importantly, multiple regression analysis revealed that the total number of user words in dialogue and recognition accuracy at the concept level are the best predictors of user satisfaction among objective dialogue metrics tested in the study. The two variables explain about 27.2% of variation in user satisfaction. By adding perceived task completion judged by users as another variable in the regression, about 45.6% of variation in user satisfaction can be explained. Adding more variables such as dialogue duration was found to have marginal or negative effects.

We have also shown that it is difficult for independent human

[5] The task completion judgment alone can explain 23% of the variance in user satisfaction. However, it is not clear whether it can be regarded as an objective metric or not.

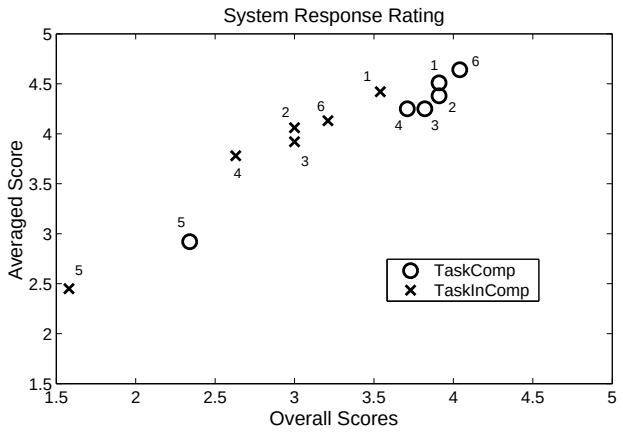

Figure 2: Averaged score vs. overall score obtained from the individual system response rating experiment.

raters to reproduce the Likert ratings given initially by the Communicator callers. Given the noisy nature of subjective metrics it is not surprising that there is little correlation between subjective and objective evaluation metrics, and between objective metrics and overall user satisfaction. The dependency of evaluation metrics on task completion and quality of service (e.g., fare price) makes dialogue evaluation an even harder task. From our experiments we conclude that objective measures (concept accuracy, error correction metric) can often be a better judge of the quality of the spoken dialogue interface than subjective measures. However, subjective measures and especially user satisfaction is the ultimate judge of service quality and should also be used (with care) when designing spoken dialogue systems.

6. ACKNOWLEDGMENTS

This work was funded by DARPA under the auspices of the Communicator project. The authors would like to express their sincere appreciation to NIST and the Communicator Evaluation Committee for designing and organizing the data collection and providing the user survey.

7. REFERENCES

[1] E. Ammicht, A. Potamianos, E. Fosler-Lussier, “Ambiguity Representation and Resolution in Spoken Dialogue Systems,” in *Eurospeech*, (Aalborg, Denmark), Sep. 2001.

[2] J. Chu-Carroll, “Form-based reasoning for mixed-initiative dialogue management in information-query systems,” in *Proc. Eurospeech*, (Budapest, Hungary), pp. 1519–1522, Sept. 1999.

[3] Larsen, L.B., ”Combining objective and subjective data in evaluation of spoken dialogues,” In ESCA Workshop on Interactive Dialogue in Multi-Modal Systems”, 1999.

[4] Boros, M. et al, ”Toward Understanding Spontaneous Speech: Word Accuracy vs. Semantic Accuracy,” Proc. of ICSLP, 1996

[5] A. Potamianos, H.-K. J. Kuo, A. N. Pargellis, A. Saad, and Q. Zhou, “Design principles and tools for multimodal dialog systems,” in *Proc. ESCA Workshop Interact. Dialog. Multi-Modal Syst.*, (Kloster Irsee, Germany), June 1999.

[6] A. Potamianos, E. Ammicht, and H.-K. Kuo, “Dialogue management in the Bell Labs communicator system,” in *ICSLP*, (Beijing, China), Oct. 2000.

[7] A. Rudnicky and W. Xu, “An agenda-based dialog management architecture for spoken language systems,” in *Proc. Workshop on Automatic Speech Recognition and Understanding*, (Keystone, Colorado), Dec. 1999.

[8] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, “Galaxy-II: A reference architecture for conversational system development,” in *Proc. ICSLP’98*, (Sydney, Australia), Dec. 1998.

[9] M. Walker *et al.*, “DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection,” in *Eurospeech*, (Aalborg, Denmark), Sep. 2001.

[10] M. Walker, L. Hirschman, and J. Aberdeen, “Evaluation for DRAPA Communicator Dialog Systems,” Proc. Language Resources and Evaluation Conference, LREC-2000.

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

[12] W. Ward and B. Pellom, “The CU Communicator system,” in *Proc. Workshop on Automatic Speech Recognition and Understanding*, (Keystone, Colorado), Dec. 1999.