Predictive performance of dialog systems

H. Bonneau-Maynard, L. Devillers, S. Rosset

LIMSI-CNRS, BP 133 91403 Orsay cedex, FRANCE {hbm,devil,rosset}@limsi.fr

Abstract

This paper relates some of our experiments on the possibility of predictive performance measures of dialog systems. Experimenting dialog systems is often a very high cost procedure due to the necessity to carry out user trials. Obviously it is advantageous when evaluation can be carried out automatically. It would be helpfull if for each application we were able to measure the system performances by an objective cost function. This performance function can be used for making predictions about a future evolution of the systems without user interaction. Using the PARADISE paradigm, a performance function derived from the relative contribution of various factors is first obtained for one system developed at LIMSI: PARIS-SITI (kiosk for tourist information retrieval in Paris). A second experiment with PARIS-SITI with a new test population confirms that the most important predictors of user satisfaction are understanding accuracy, recognition accuracy and number of user repetitions. Futhermore, similar spoken dialog features appear as important features for the Arise system (train timetable telephone information system). We also explore different ways of measuring user satisfaction. We then discuss the introduction of subjective factors in the predictive coefficients.

1. Introduction

Evaluation of spoken dialog systems is currently a very active research area. While there are commonly used measures and methodologies for evaluating speech recognizers, the evaluation of spoken language systems is considerably more complicated due its interactive nature and the human perception of the performances. Experimenting dialog systems is often a very high costly procedure due to the necessity to carry out user experiments. Obviously it is advantageous when evaluation can be carried out automatically. It seems very promising to assume that for each application we can measure the system performances by an objective cost function. This performance function could be used for making predictions about a future evolution of the systems without user interaction.

This paper relates our experiments with predicting user satisfaction of dialog systems for two systems developed at LIMSI: ARISE (Lamel et al., 1999) and PARIS-SITI (Devillers and Bonneau-Maynard, 1998).

PARIS-SITI and ARISE are data retrieval dialog systems based on the same basic architecture derived from MASK system (Lamel et al., 1998). ARISE is a train timetable telephone information system. PARIS-SITI (*Système d'Informations Touristiques Interactif*) is a spoken dialog tourist information system kiosk which allows users to obtain information such as prices, payment procedures, opening hours, address, trip, descriptions and services offered, for a variety of objects (hotels, restaurants, cinemas, department stores, museums and monuments) in Paris.

PARADISE (Walker et al., 1997; Walker et al., 1998) is an evaluation paradigm in which a combined performance metric for dialog systems is derived as a weighted linear combination of a task-based success measure and dialog costs. The PARADISE model posits that user satisfaction is the top-level objective. Using the PARADISE paradigm, performance functions, derived from the relative contribution of various factors, are obtained for PARIS-SITI and ARISE. Identical metrics for dialog costs are used for both systems.

A preliminary experiment with PARIS-SITI shows that the most important predictors of user satisfaction are understanding accuracy, recognition accuracy and number of user repetitions. A validation of this observation is first explored by repeating the experiment with a new test population on ARISE and PARIS-SITI.

Since user satisfaction is a major parameter for this paradigm, we explore and compare different ways of measuring it. For example we compare directly asking the user to rate the system with deriving the user satisfaction from a questionnaire on the different modules of the system. We also discuss about the introduction of subjective factors and the generalisation across systems.

2. System Description

The tourist information system PARIS-SITI (Devillers and Bonneau-Maynard, 1998) and rail travel information dialog ARISE (Lamel et al., 1999) are built upon the LIMSI dialog systems architecture (Lamel, 1998).

PARIS-SITI is a French language information retrieval system, that allows users to obtain information in Paris. In this study, we focus on hotels and restaurants located in the district of Saint-Lazare station in Paris. PARIS-SITI uses a generation strategy in which clarification dialogs are determined by the domain model which is hierarchically represented along with the generation and dialog histories.

ARISE is a French language train timetable telephone information system, which uses a 2 level mixed-initiative strategy, where the user has maximum freedom when all is going well and the system takes the initiative if problems are detected (Rosset et al., 1998). For these experiments we used a version of the ARISE system which also offers to the user the ability to interrupt the system (barge-in capability).

Both systems are composed of a speaker-independent real-time continuous speech recognizer, and components for natural langage understanding, dialog management, database access and response generation.

Statistical models are used at the acoustic and word levels. Acoustic modeling makes use of context independent continuous density hidden Markov models (HMM) with Gaussian mixture. Speaker independence is achieved by using acoustic models which have been trained on speech data from a large number of speakers. Bigram backoff language models are estimated on the orthographic transcriptions of the training set spoken queries. Word classes for cities, dates and numbers providing more robust estimates of the n-gram probabilities are used for ARISE. Both recognizers have a medium vocabulary size (about 2000 words). The PARIS-SITI vocabulary contains approximately a hundred different objects (including 24 hôtels and 26 restaurants) and the ARISE vocabulary contains 600 different station names.

The speech recognizer transforms the input signal into the most probable sequence of words and then forwards it to the natural langage understanding component which carries out a caseframe analysis and generates a semantic frame representation. If enough information is present in the semantic frame the dialog manager generates a database query. The retrieved information, in the form of a generation frame, is formatted into a natural langage response by the response generator (taking into account the dialog context) and vocal feedback is provided to the user (along with a visual display of the different objects already selected for the PARIS-SITI system).

3. PARADISE Paradigm

PARADISE paradigm was proposed by AT&T (Walker et al., 1997; Walker et al., 1998). Using methods from decision theory, PARADISE allows a disparate set of performance measures to be combined into a single performance evaluation function. The PARADISE model posits that performance can be correlated with a meaningful external criterion such as usability, and thus that the overall goal of a spoken dialog agent is to maximize an objective relative to usability. Using multivariate regression it is possible to estimate a performance equation. Regression is used to find the weights in the performance equation, thereby quantifing the relative contribution of the performance parameters.

4. Experimental methods

All of our experiments used a similar setup. The experiments were carried out with 18 subjects interacting with PARIS-SITI and ARISE. Each user first carried out the 4 scenarios listed in Figure 1 with PARIS-SITI and then the 3 scenarios listed in Figure 2 with the ARISE system. The scenarios used for ARISE are very precise, and contain sufficient information to select a very few number of possible trains corresponding to the constraints. Users were asked to make a reservation for one train at the end of the dialog, corresponding to typical train reservation task. Therefore the task success may be estimated as a function of whether or not, the user was able to performe the reservation. In contrast, each scenario for PARIS-SITI specifies only one type of constraint. With these constraints a subset of the database objects can be selected (this case corresponds better to the actual tourist information task where people may just have an approximate idea of what they are looking for). Users were asked to select only one of these objects, using their own constraints. Therefore the task success for PARIS-SITI dialogs may be measured as a function of whether or not at the end of the dialog, the user selected an object corresponding to the contraints fixed in the scenario.

Scenario	Constraint
A- Find a hotel near the *Galeries-Lafayette*	location
B- You are looking for a luxurious hotel	description
C- You are looking for a restaurant open late	hour
D- You want to eat seafood	speciality

Figure 1: Prototype scenarios used to test PARIS-SITI system. These scenarios were presented to subjects in a picture form with keywords corresponding to the contraints (ex. scenario B: hotel luxurious) so as to minimally influence the vocabulary used by the subjects.

Scenario
A- Lyon Paris, 23 June, 8 o'clock the morning, direct.
B- Grenoble Paris, next Monday, evening, arrival time ?
C- Paris Carcassonne, Christmas, arrival time around 18 hours, Dining car.

Figure 2: Prototype scenarios used to test ARISE system.

These experiments resulted in 72 dialogues with PARIS-SITI and 54 dialogues with the ARISE. The dialogues produced from the two tasks are quite different (Table 1). The mean number of user interactions per dialogue is much greater for ARISE (16.2) than for PARIS-SITI (8.2) whereas the average user interaction length is much greater for PARIS-SITI (9.5) than for ARISE (5.5).

	PARIS-SITI	ARISE
#utt	589	871
#utt per dial	8.2	16.2
#words per utt	9.5	5.5
SAT	6.0	5.9
QUEST	6.6	6.5
TC	87.1	87.5

Table 1: Test dialog characteristics and global performance measures: total number of user interactions (#utt), mean number of user interactions per dialogue, mean user utterance lenght, mean satisfaction mark, questionnaire evaluation of user satisfaction, and percentage of dialog completed.

4.1. User satisfaction measure

Since user satisfaction is a major parameter for the PARADISE model, we have explored and compared different ways of measuring it. At the end of each dialog, the subjects were asked to complete a questionnaire which is the same for PARIS-SITI and ARISE. Before completing the questionnaire the subject was asked to first give a satisfaction mark (SAT) about the overall system performance (see Table 1) by marking a cross on a scale which was afterwards transformed into a score ranging from 0 to 10 (0 corresponding to very bad satisfaction). The questionnaire addressed the users on specific aspects of the dialog as listed in Table 2. For each aspect affirmation (shown in bold) the user must state his/her disagreement or agreement

The system was easy to understood in this conversation
- TTS performance
The system understood what I said
- Understanding and recognition performance
I have obtained the information I asked for
- Contextual understanding performance
The pace of interaction was appropriate
- Interaction pace
At each point of the dialog I knew what to say
- System help
The system clearly explained what he understood
- Generation performance
The system suggestions or questions helped me
- Generation strategy performance

Table 2: Questionnaire submitted to the user at the end of each dialog, inspired from (Walker et al., 1998). The assertion given to the user is shown in italics and the underlying system aspect whose performance is tested in shown in bold.

on a 5 level scale: not agree at all, not agree, approximately agree, agree or very much agree, with the affirmation. Each response was mapped to an integer from 0 to 4. We then sum the user satisfaction scores for each dialog to derive an estimator (QUEST). At the end of each dialog, subjects also reported whether they believed they had completed the task. We took this measure for evaluating task completion (TC).

4.2. Dialog cost measures

Each dialog was annotated with a set of dialog cost measures. Table 3 gives the mean values for both tasks. The different measures are: recognition word accuracy (ASR), literal understanding accuracy (LU), contextual understanding accuracy (CU), generation error rate (G0), history management error rate (H0), user repetitions rate (UR), and task completion (TC) rate. This measures are calculated for each dialog. LU, CU, G0, H0, UR are normalized by the number of utterances in the dialog. The literal understanding accuracy (LU) is obtained by running the understanding module on the exact transcription of the user utterance, whereas the contextual understanding accuracy (CU) is obtained from the dialog observation, taking into acount the possible recognition and history managment errors. UR corresponds to the percentage of times that the user had to repeat an information because of a misunderstanding of the system. Each predictor factor x is normalized to its Z score:

$$\mathcal{N}(x) = \frac{(x - \overline{x})}{\sigma_x}$$

where σ_x is the standard deviation for x.

5. Experiments with two different sets of subjects

In a previous experiment (Devillers and Bonneau-Maynard, 1998), which aimed to evaluate the guiding prompt strategy of PARIS-SITI, a set of 16 subjects were recorded , using the same 4 scenarios listed in Figure 1. The resulted 64 dialogs were evaluated in terms of recognition accuracy (corresponding to the ASR factor), contextual understanding accuracy (corresponding to the CU factor), number of user turns, number of different information obtained during the dialog, number of user repetitions per dialog (corresponding to UR factor). After each dialog, the subjects were asked to rate the systems in the range 0 to 10 (this corresponds to the SAT factor).

	PARIS-SITI	ARISE
ASR	75.3	74.0
LU	81.9	79.4
CU	73.9	58.0
G0	6.0	1.5
H0	12.0	1.6
UR	5.3	20.0

Table 3: Mean dialog cost values for PARIS-SITI and ARISE including per dialog : mean recognition accuracy (ASR), mean Literal Understanding Accuracy (LU), mean Contextual Understanding accuracy (CU), mean Generation error rate (G0), mean history management error rate (H0) and mean User Repetitions (UR).

We took this opportunity of having data from two sets of subjects using the same system (PARIS-SITI) for different experiments to validate the idea that user satisfaction can be predicted with a similar combination of the most important dialog cost features. Note that between the two experiments PARIS-SITI was slightly changed: the training set used to estimate the language model was increased with 2000 transcriptions and the understanding component has been rewritten.

We applied the PARADISE paradigm on both sets of dialogs. The factor to be predicted by the linear regression is SAT. A correct prediction (p=0.003) for the first set of subjects was obtained with the predictive factors: CU, ASR and UR, as shown in equation 1. These factors together explain 39% of the variance of the user satisfaction.

$$\mathrm{SAT} = 0.45 * \mathrm{CU} + 0.12 * \mathrm{ASR} - 0.133 * \mathrm{UR} \quad (1)$$

When using the same prediction factors for the second set of subjects, equation 2 was obtained (p<0.0001), explaining 30% of the variance of the user satisfaction.

$$\mathrm{SAT} = 0.34 * \mathrm{CU} + 0.11 * \mathrm{ASR} - 0.28 * \mathrm{UR} \quad (2)$$

Both equations show a similar behavior. It can be observed that the weights associated with the recognition performances are equivalent. The contextual understanding accuracy is still the most important predictor. The number of user repetitions is more important for the second set of subjects.

We tested the use of the same predictors for the ARISE system with the second set of subjects. The set of factors shown in equation 3 was obtained, explaining 44% of the variance of the user satisfaction with a very statistically significant level (p<0.00001),

$$\mathrm{SAT} = 0.45 * \mathrm{CU} + 0.15 * \mathrm{ASR} - 0.21 * \mathrm{UR} \quad (3)$$

So we can conclude that the weights of these three important factors are pretty stable, across different sets of subjects or different tasks (such as PARIS-SITI and ARISE).

6. Comparing different ways of measuring user satisfaction

We have compared different ways of measuring user satisfaction. For the first one (SAT), the user was asked to give a satisfaction mark about the dialog immediately at the end of the dialog. The second one (QUEST) was derived from a questionnaire by summing the responses of the user of questions about the system performance for each dialog. We also derived a third estimation of the user satisfaction (QComb) by a weighted combination of the QUEST score, the SAT score and the fact that the user obtained the object that he wants for PARIS-SITI and that he obtained the reservation that he wanted for ARISE, namely task completion.

Equations 4,5,6 show the weights obtained for PARIS-SITI, with the three different user satisfaction measures and with the same predictor coefficients.

$$\begin{aligned} \mathrm{SAT} = 0.26 * \mathrm{CU} + 0.08 * \mathrm{ASR} - 0.27 * \mathrm{UR} \\ + 0.27 * \mathrm{LU} - 0.09 * \mathrm{G0} - 0.04 * \mathrm{H0} \end{aligned} \tag{4}$$

$$\begin{aligned} \mathrm{QUEST} = 0.33 * \mathrm{CU} + 0.16 * \mathrm{ASR} - 0.12 * \mathrm{UR} \\ + 0.32 * \mathrm{LU} - 0.07 * \mathrm{G0} + 0.04 * \mathrm{H0} \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{QComb} = 0.29 * \mathrm{CU} + 0.13 * \mathrm{ASR} - 0.20 * \mathrm{UR} \\ + 0.34 * \mathrm{LU} - 0.09 * \mathrm{G0} + 0.06 * \mathrm{H0} \end{aligned} \tag{6}$$

Several observations can be made from these equations. Firstly the number of history management errors and generation errors are not significant predictors. This could have been hypothetized because the number of these errors is very low (see Table 3).

Secondly, for PARIS-SITI the best explanation of the variance of the user satisfaction (44.3%) with a strong statisticall significance level ($p < 0.00001$), is obtained with QComb factor, which includes the questionnaire estimator, the SAT mark and the task completion filling of the user (TC). This way of estimating user satisfaction will be kept for our future experiments. The degree of explanation is 37% for SAT alone and 42% for QUEST.

Thirdly, the literal understanding performance is a significant factor. Its weight is quite equivalent to the weight of the contextual understanding performance. This is interesting because the literal understanding performance can be automatically estimated given a reference set of utterance transcriptions along with the corresponding reference semantic frames.

We performed the same test with the ARISE measures. The results are quite different. The best explanation (43%) of the variance of the user satisfaction with the predictors (CU, ASR, UR, LU) is obtained when user satisfaction is measured by the satisfaction mark (SAT). The resulting equation is given in equation 7:

$$\begin{aligned} \mathrm{SAT} = 0.43 * \mathrm{CU} + 0.15 * \mathrm{ASR} \\ - 0.21 * \mathrm{UR} + 0.02 * \mathrm{LU} \end{aligned} \tag{7}$$

The explanation is 34% with the QUEST measure and 42% with the QComb measure. Therefore we can see that for ARISE evaluation, the users are able to give a relieable mark to the system performances, and that the questionnaire does not contain more information. This may be explained by the fact that the scenarios used for ARISE are more precise, and that the subjects are more familiar in the train reservation task (most subjects travel by train more than 3 times a year), so they have a good idea of what the system is supposed to perform.

Secondly we can observe that the weight of LU is very low. The number of user repetitions for ARISE is relatively high (20%) due in part to recognition errors on station names (there are 600 stations) which are obviously very important for the understanding process. We hypothetize that the literal understanding performance is a less reliable predictor for a system which has a such ASR error rate on important task words. This is also illustrated by the difference observered between the global LU (79.4%) and CU (58.0%) performances of ARISE.

7. Introducing subjective measurements in the predictor coefficients

In the equations given above, the predictors factors all consist of objective measures. In the last two experiments, we introduce subjective factors.

7.1. Introduction of task completion in predictors

The set of performance measures used by Walker et al includes task success (evaluated by KAPPA coefficient or task completion) along with dialogue costs. The task completion factor (TC) was then integrated in the prediction factors for PARIS-SITI evaluation. The QUEST factor was used for the user satisfaction estimation (we did not used the QComb factor because it already includes the TC factor).

$$\begin{aligned} \mathrm{QUEST} = 0.31 * \mathrm{CU} + 0.16 * \mathrm{ASR} - 0.05 * \mathrm{UR} \\ + 0.25 * \mathrm{LU} + 0.28 * \mathrm{TC} \end{aligned} \tag{8}$$

The percentage of the QUEST variance explanation increases from 41% to 48% ($p < 0.00001$). This shows that task completion is an important factor in user satisfaction. However the introduction of this factor masks in a certain way, the importance of the other relative contributors (such as the number of repetitions for example).

7.2. Introduction of HU in predictors

At the end of the ARISE questionnaire, the user was asked to respond to the following question: “In real life, would you have hung up before the end of the conversation?”. The user answer to this question was transformed in a HU coefficient in a range from 0 to 4 (0 surely I do, 4 surely I would not).

Adding the HU factor in the predictor coefficients gives equation 9.

$$\begin{aligned} \mathrm{SAT} = 0.20 * \mathrm{CU} + 0.07 * \mathrm{ASR} - 0.18 * \mathrm{UR} \\ + \mathrm{O}.28 * \mathrm{TC} + 0.33 * \mathrm{HU} \end{aligned} \tag{9}$$

The percentage of user satisfaction variance explanation with this equation is very high (60.5%, p <0.000001), showing that HU and task completion factors are very good predictors of user satisfaction.

8. Discussion

Several experiments were performed to study the possibility of predicting user satisfaction from dialog cost and task completion measurements. It was observed that the weights between three important dialog cost factors (ASR accuracy, contextual understanding and number of user repetitions) are quite comparable while changing subject sets on the PARIS-SITI task, or even changing the application domain (tourist information and train time-table information) for the same set of subjects. Our experiment about different ways of measuring user satisfaction indicates that a combination of user satisfaction marks, user questionnaire responses and task completion is the most accurate. Some differences between the two tasks were observed when introducing the literal understanding factor. The reliability of this factor clearly depends on the ASR accuracy of important words for the understanding modules (station names for ARISE for example). Results show that task completion is a good predictor, but that it may mask the importance of other contributors.

9. References

L. Devillers and H. Bonneau-Maynard. 1998. Evaluation of dialog strategies for a tourist information retrieval system. In *ICSLP*.

L. Lamel, S. Bennacef, J.L. Gauvain, Hervé Dartigues, and Jean-Noël Temem. 1998. User evaluation of the mask kiosk. In *International Conference on Speech and Language Processing*, Sydney.

L. Lamel, S. Rosset J.L. Gauvain, and S. Bennacef. 1999. The limsi arise system for train travel information. In *ICASSP*.

L. Lamel. 1998. Spoken language dialog system development and evaluation at limsi. In *ICSLP*.

S. Rosset, S. Bennacef, and L. Lamel. 1998. Design strategies for spoken language dialog systems. In *Eurospeech*.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1997. Paradise: a general framework for evaluating spoken dialog agents. In *ACL/EACL*.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language*.