EVALUATION OF DIALOG STRATEGIES FOR A TOURIST INFORMATION RETRIEVAL SYSTEM*

L. Devillers, H. Bonneau-Maynard

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

ABSTRACT

In this paper, we describe the evaluation of the dialog management and response generation strategies being developed for retrieval of touristic information, selected as a common domain for the ARC AUPELF-B2 action. A large number of spoken dialog systems have been reported in the literature making use of different strategies for dialog management. Comparing and evaluating different strategies is a difficult task, which often remains unexplored, because in most cases evaluation approaches require a unified database structure and efficient integration of data from several disparate sources and forms. To avoid this problem, we implemented two dialog strategy versions within the same general platform. We investigate qualitative and quantitative criteria for evaluation of these dialog control strategies: in particular, by testing the efficiency of our system with and without automatic mechanisms for guiding the user via suggestive prompts. An evaluation phase has been carried out to assess the utility of guiding the user with 32 subjects. We report performance comparisons for naive and experienced subjects and also describe how experimental results can be placed in the PARADISE framework for evaluating dialog systems. The experiments show that user guidance is appropriate for novices and appreciated by all users.

1. INTRODUCTION

Recent advances in the speech and language processing have led to the development of human-machine dialog systems for a variety of applications. While some common metrics are used to measure speech recognition performance and measures have been proposed for natural understanding language, evaluation of dialog strategies is less straightforward. A critical point with system development is the lack of common framework for comparing performances of dialog systems. The main difficulty is to find a paradigm for dialog system evaluation which is independent of the database and the task model. The PARADISE framework [1] seems promising as a method for comparing systems performing different tasks by normalizing for task complexity.
Several projects concerned with spoken dialog system development and evaluation are underway, such as the DISC, AUPELF, DARPA Communicator and on a more general level ELSE projects. The aim of the ARC AUPELF-UREF B2 action [2] is to evaluate French spoken language systems on a common domain task of touristic information.

This paper aims to investigate qualitative and quantitative criteria for evaluation of dialog control strategies. We describe an evaluation study that has been carried out with 16 naive and 16 experienced subjects to assess the utility of guiding the user. Each speaker used two versions of the system: with and without guiding prompts. An analysis of these dialogs is presented, as well as a summary of the users' subjective and objective assessments. We also apply the PARADISE task success measure to our experimental results.

* This work was partially financed by the ARC AUPELF-B2 action.

2. SYSTEM DESIGN

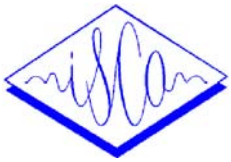
The tourist information system is built upon the MASK [3] and RAILTEL/ARISE [4] rail travel information dialog systems developed at LIMSI. PARIS-SITI (*Système d'Informations Touristiques Interactif*) is a French language information retrieval system, that allows users to obtain information (such as 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. In this study, we focus on hotels and restaurants located in the district of Saint-Lazare station in Paris. The database contains information on approximately a hundred different objects.

The system is composed of a 2000-word speaker-independent continuous speech recognizer, and components for natural langage understanding, dialog management and response generation. The speech recognizer uses acoustic models from MASK and a bigram language model estimated on the transcriptions of 5200 utterances recorded previously. 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. In order to have flexibility in testing different response strategies we used the LIMSI text-to-speech synthetizer.

3. DIALOG STRATEGIES

A variety of spoken dialog systems have been developed making use of different strategies for dialog management (c.f. [4], [5]). Some papers report comparisons between different strategies such as system-initiative or mixed-initiative strategies [6] and explicit or implicit confirmation strategies [1].

To independently test the effects of the suggestive prompts in the dialog, we compared two systems differing in their responses strategies: one uses an automatic mechanism for guiding the user



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

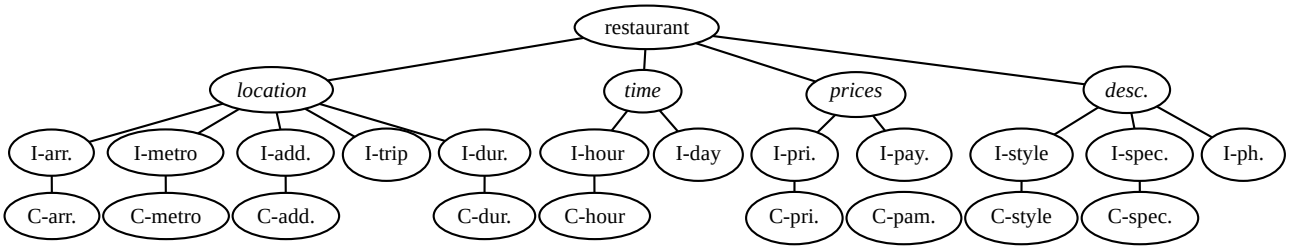


Figure 1: Hierarchical domain representation used by the dialog manager for the restaurant object. The first level corresponds to the object level, the second to classes level. I-* is the information level and C-* the constraint level.

via non-directive prompts (or system suggestions); the other does not use this mechanism. Both of the systems use a mixed-initiative dialog with an implicit confirmation strategy where the user is free to ask any question at any time. The suggestive prompt strategy helps the subject to use valid vocabulary items without the constraining aspect of system-initiative strategies which ask directive questions.

We view the aim of a tourist information retrieval system as providing the user with information which will allow him to take a decision. From this point of view, the dialog strategy should serve not only keep the user within the boundaries of the system, but also help him to discover the different possibilities of the system and the contents of the database. The dialog strategy is more open than typically needed for simple train timetable retrieval in that the user does not necessarely have an a priori idea of what the system is able to provide.

Our generation strategy differs from other approaches in that clarification dialogs are determined by the domain model which is hierarchically represented along with the generation and dialog histories. The hierarchic representation of the domain is derived from the database, where each database object contains attributes. We distinguish 4 classes of attributes: *location*, *hour*, *price* and *description*. The proximity of objects is given with respect to the closest metro station, which is the reference commonly used by Parisians. Figure 1 shows how these attributes are hierarchically represented for the restaurant object. Each level in the hierarchy can be associated with a precision degree of the user query. 4 precision levels are distinguished: object, class, information and constraint. The generation frame integrating the user's query contains the identification of the query as well as its level in this hierarchy. For example, the generation frame for the query "*I'd like seafood and I'm looking for a restaurant*" contains the following attributes:

```
object-level = restaurant
class-level = description, location
information-level = speciality, address
C-speciality = seafood
```

The dialog manager maintains a history frame which keeps track of all the types of information and their level asked for by the user since the beginning of the dialog.

The dialog manager chooses which prompt to generate as a function of the levels instantiated in the generation frame. The strategy illustrated in Figure 2 is described in more detail in [7]. When a class-level is reached, the dialog manager will encourage the speaker to provide additional information in the same class, according to the object hierarchy. When the speaker has asked for all of the information in the same class, or when the generation frame contains more than one `class-level`, the system orients the user to ask for information in other classes. The suggestion prompts rules have been improved after observing dialogs from experienced users interacting with the system. The two alternative dialog strategies are illustrated in dialogs D1 (Figure 2) and D2 (Figure 3).

U: please I want a hotel
S: Here is the hotel list, do you have a location preference ? Do you want prices or descriptions of these hotels ?
U: I'm looking for a hotel near the Galeries Lafayette **P-1 (location)**
- the user followed the system suggestion by adding a location constraint
S: Here is the list of the hotels near the Galeries Lafayette. You can have more information about subway trip and duration from the Saint-Lazare station.
U: How much time does it take to go to these hotels ? **P-1 (location)**
- the user followed the system suggestion by adding a new location constraint
S: Here is the trip and its duration from Saint-Lazare station for the hotels near the Galeries-Lafayette. You can ask for prices, categories, services or descriptions of these hotels
- When all usefull location constraints are given by the user, the system suggests that he ask for information in the other classes
U: I'll take the Adria hotel **P-0**
- the user did not follow the system suggestion
S: Here is hotel Adria's address, you can ask for prices, categories, services or a description of this hotel
U: I want prices please **P-1(prices)**
- the user followed the system suggestion and asked for prices
S: Here are hotel Adria's prices, you can ask for its services or a description
- location and prices have been visited, suggestion for the 2 remaining classes
U: The services please **P-1(services)**
S: Here are hotel Adria's services.

Figure 2: Dialog example D1 with system suggestions. S: system prompt, U: userquery. Prompt labelling P-0: user did not follow suggestion and P-1: user did follow suggestion. Comments are given in italics.

Figure 2 shows an example of a dialog where a novice user mostly follows the prompts of the system. He obtains information on location, prices, services, trip and duration. Figure 3 is a dialog of an experienced user without suggestive prompts. This user easily obtained information selecting fewer constraints than the previous example. Based on this observation we hypothesize that experienced users would benefit less from the suggestions than would novice users.

U: I'd like the prices of the hotels euh near the "Galeries-Lafayette" please
S: Here are the prices of the hotels near the "Galeries-Lafayette"
U: I want the least expensive
S: Here is the least expensive hotel near the "Galeries-Lafayette"

Figure 3: Excerpt taken from dialog D2 of an experienced user, without suggestive prompts. S: system prompt, U: userquery.

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 4: Prototype scenarios used in the experiment. 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.

4. EVALUATION AND SYSTEM PERFORMANCE

4.1. Methodology

Experiments were carried out with 32 subjects interacting with the system: 16 experienced users (PhD students or researchers in the speech and natural language area but not necessarily familiar with this dialog system) and 16 novice users.

To independently test the effects of the prompting mechanisms, half of the subjects (8 novices and 8 experienced users) used first the system with prompts (SP) and the other half used first the other system (S). All subjects were familiar with computers. In order to measure the user task learning, each subject tested both systems with the 4 scenarios listed in Figure 4. The scenarios were chosen after some preliminary experiments conducted with scenarios containing sufficient information to select only one hotel or restaurant from the database. An example scenario of this type *(that would not be the case with a realistic Paris tourist database)* is "Find a calm hotel near the Saint-Lazare station (no more far than 5 minutes away), with a double-room and a price less than 200 francs". Such precise scenarios do not correspond to the reality of tourist information task, and are also not a good way to test the efficiency of the suggestive prompts as users tended to paraphrase the written scenario. In real situations, people usually have an idea of just few of the constraints such as location, price or description: Each scenario sets only one type of constraint. These contraints select a subset of the database (3 hotels for A, 5 for B, 9 restaurants for C and 4 for D). User's were asked to select only one of these objects, using their own constraints.

4.2. Measuring Dialog Costs

The first 4 scenarios per subject are used to evaluate the impact of the dialog strategy and to compare both systems. Each utterance of each dialog was labelled in terms of understanding success. When the understanding is complete (all semantic slots are correctly instanciated), the system response is labelled (C-1). If an error occurs, we distinguish the case of a recognition error (CR-0) or an understanding component error (C-0). Out of domain utterances were labelled (O-0). For the SP system, we also labelled as a failure (P-0), as shown in Figure 2, the utterances of the speaker which did not follow the preceding prompt and as a success (P-1) if the user asked for at least one of the suggested items. Results

	E-S	N-S	E-SP	N-SP
Recognition error rate	24.9	25.0	28.4	26.9
(C-0) rate	9.4	10.4	4.3	5.6
(CR-0) rate	19.3	21.2	25.0	24.5
Success prompt rate	-	-	51.0	66.3
Mean #user turns	6.6	7.8	6.5	8.3
Mean #information	6.7	6.9	6.8	7.6

Table 1: Recognition error rate corresponds to exact string match with the literal transcription. Understanding error rates, success prompts rate (#P-1/(#P-1+#P-0)), average number of utterances and mean of different informations. The total number of labelled dialogs is 128, with 935 total user utterances. *(E-S, N-S means experienced, novice with the system S, E-SP, N-SP experienced, novice with the system SP)*

Users	1st	2nd
Novice	SP= 7.8 S= 6.0	S= 7.7 SP= 7.2
Experienced	SP= 7.4 S= 5.6	S= 6.9 SP= 6.6

Table 2: Experienced and novice users satisfaction - average of the global marks given at the end of the session for the first and second system tested

are summarized in Table 1.

The global understanding error rate (CR-0 + C-0) is approximately the same for both systems. When the recognition is correct, the understanding rate (C-0) of the SP system is better than that of the S system.

In order to compare the efficiency of SP and S systems, we labelled every dialog in terms of the different information obtained by the user. The results, in terms of the average number of different information items per scenario are shown in lower part of Table 1. As expected, there is no essentially difference for experienced users with both systems. These users follow only half of the prompts, with a prompt success rate of 51%. The results for novice speakers are quite encouraging. Novice users have a tendency to follow the prompts (66.3%) resulting in an increase in the mean number of different constraints obtained with the system SP (7.6) compared to the system S (6.9). There were very few dialogs in which the subjects did not follow any prompts (1% of the experienced users and 0.5% of the novices).

4.3. User Satisfaction

At the end of a recording session each user completed a questionnaire concerning the usability of the system and the helpfulness of the system prompts. Users rated both systems (in the range 0 to 10) (see Table 2) and were asked to describe the differences between them. Satisfaction marks for all the users after using first system, showed that the SP system is prefered. Users who first interacted with the prompted system (SP), generally did not notice any difference between the systems. Novice users gave the same mark for both systems whereas experienced users prefered the SP system. All users who first interacted with the system without guiding prompts (S), noticed the difference between both systems and clearly prefered the SP system. Very encouraging remarks on the interest of the prompts were given (by example: *"the second system (SP) is more helpful to reach the goal by proposing different information. With the first system (S) the user has no idea of the system limits"*).

SP	O1	O2	C1	C2	C3	C4	N1	N2	N3	N4	S	O1	O2	C1	C2	C3	C4	N1	N2	N3	N4
O1	32										**O1**	32									
O2		32									**O2**		32								
C1			16								**C1**			13							
C2				14							**C2**				12						
C3					16						**C3**					11					
C4						16					**C4**						16				
oth.				2							**oth.**			3	4	5					
N1							16	1			**N1**							15			
N2								14			**N2**								11		
N3									16		**N3**									11	
N4										15	**N4**										16
oth.								1		1	**oth.**							1	5	5	
t_i	32	32	16	16	16	16	16	16	16	16	**t_i**	32	32	16	16	16	16	16	16	16	16

Figure 6: Confusion matrices for system SP and system S *(O_i: mean objects, C_i constraints of the 4 scenarios and N_i are the 4 groups of the object names matching with the scenarios).*

attribute	actual value
object	hotel
constraint	near the Galeries Lafayette
name	N1 = {Adria, Ambassador, de Beauharnais}

Figure 5: AVM instantiation for scenario A (see Figure 4)

	S		SP	
Kappa	0.869		0.967	
	E_S	N_S	E_SP	N_SP
Kappa	0.923	0.803	0.988	0.954

Table 3: Kappa measures of the success of the dialog of novice and experienced users using S or SP systems (only for the first system used).

4.4. Task success evaluation with PARADISE

The PARADISE framework [1] has been proposed as a means to evaluate spoken dialog systems enabling comparisons across tasks. The paradigm aims to separate how a system uses dialog strategies from what a system achieves in terms of task requirements. Notably, it uses the Kappa statistic which normalizes for task complexity as a measure of success. We have compared the task success of the SP and S systems. To be consistent with the PARADISE evaluation framework, each scenario is represented by its Attribute Value Matrix (AVM), as shown in Figure 5.

We have built a confusion matrix for both systems (see Figure 6) using data from 16 subjects, consisting of the 4 dialogs from the fisrt system used by subject. Given a confusion matrix M, success at achieving the information requirements of the task is measured with the Kappa coefficient (equation 1), where P(A) is the proportion of time that the AVMs for the actual set of dialogs agree with the AVMs for the scenarios keys, P(E) is the proportion of time that the AVMs for the dialogs and the keys are expected to agree by chance, t_i is the sum of the counts in column i of M and T is the sum of all counts in $M(t_1 + ... + t_n)$, here T=192. For both systems, P(E) = 0.083.

$$k = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

$$P(E) = \sum_{i=1}^{n} (\frac{t_i}{T})^2 \qquad P(A) = \frac{\sum_{i=1}^{n} M(i,i)}{T} \quad (2)$$

The Kappa measure is 0.967 for the speakers who interacted with the SP system (for a total of 64 tests of novice and experienced users), compared with 0.869 for the speakers that interacted with the S system. This difference in task success suggests that system SP is more successful than S in achieving the task goals. In Table 3, we compare the Kappa measures for novice and experienced users using each system. The largest difference is observed for the novice users, 0.803 for the system S against 0.954 for the system SP. The Kappa measures are in agreement with the user satisfaction, both indicating that the system with suggestive prompts gives better results, especially for the novice users.

5. CONCLUSION

An evaluation of two dialogs strategies has been carried out using objective dialog measures such as speech recognition and understanding error rates and a subjective measure of user satisfaction. This subjective measure has been compared with an objective measure of overall system performance (the Kappa measure). All measures indicate that user guidance is appropriate for novice users and appreciated by all the users. These measures are being correlated with objective measures of system performance in an effort to determine in which dialog contexts user guidance is appropriate.

6. REFERENCES

1. M. Walker, D. Litman, C. Kamm, and A. Abella, "Paradise: a general framework for evaluating spoken dialog agents", ACL/EACL 1997.
2. J. Mariani "The Aupelf-Uref Evaluation-Based Language Engineering Action and Related Projects", LREC, May 1998.
3. J.L. Gauvain, S. Bennacef, L. Devillers, L. Lamel, R. Rosset: "Spoken Language component of the MASK Kiosk" in K. Varghese, S. Pfleger(Eds.) "Human Comfort and security of information systems", Springer-Verlag, 1997.
4. S. Bennacef, L. Devillers, S. Rosset, L. Lamel, "Dialog in the RAILTEL telephone-based system", ICSLP-96 and ISSD-96, October 1996.
5. M. Denecke, A. Waibel, "Dialogue strategies guiding users to their communicative goals", ESCA Eurospeech Rhodes, September 1997.
6. M. Walker, D. Hindle, J. Fromer, G. Di Fabbrizio, C. Mestel, "Evaluating competing agent strategies for a voice Email agent", ESCA Eurospeech Rhodes, September 1997.
7. H. Bonneau-Maynard, L. Devillers. "Dialog Strategies in a tourist information spoken dialog system", SPECOM 98, St-Petersbourg, October 1998.