Dialogue and Prompting Strategies Evaluation in the DEMON System

Carine-Alexia Lavelle, Martine de Calmès, Guy Pérennou

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
118, route de Narbonne,
31062 Toulouse, France
lavelle@irit.fr

Abstract

In order to improve usability and efficiency of dialogue systems a major issue is of better adapting dialogue systems to intended users. This requires a good knowledge of users' behaviour when interacting with a dialogue system. With this regard we based evaluations of dialogue and prompting strategies performed on our system on how they influence users answers.

In this paper we will describe the measure we used to evaluate the effect of the size of the welcome prompt and a set measures we defined to evaluate three different confirmation strategies. We will then describe five criteria we used to evaluate system's question complexity and their effect on users' answers. The overall aim is to design a set of metrics that could be used to automatically decide which of the possible prompts at a given state in a dialogue should be uttered.

1. Introduction

When developing a spoken dialogue inquiry system for everyday life information, such as train or air travel schedules, weather forecast, movie theatres or restaurant facilities around, usability and efficiency appear as two important issues.

Theses systems, providing casual information to the general public, are by essence destined to be used by occasional and inexperienced users. Within this context the design of dialogue appear to be a crucial element in the provision of a usable service. The dialogue management should be so that users, and in particular first time users, get sufficient guidance about what they may say and do, accommodate user behaviour and allow graceful error recovery (Peckham, 1995); Thomson & Wisowaty, 1999). The overall usability of a spoken dialogue system therefore dependent upon the ability of dialogue designers to produce clearly understandable prompts.

But understandable prompts are not enough. For prompts to be easily understandable it may be argued that it is advisable to make them as informative as possible and to ask for no more than one information at a time. This kind of dialogue strategy lead to very long dialogues annoying for users.

A efficient dialogue and prompting strategy should lead the dialogue toward the accomplishment of the goal, here providing users with some information, as quickly as possible. With this regard it may be advisable to ask for the most possible information at a time.

These goals appear to be at odds. The problem is then which among all the possible questions at a given state in a dialogue should be uttered.

The dialogue and prompting strategies evaluation metrics we are presenting in this paper were designed with this idea in mind. We based our evaluations of prompts and dialogue moves in our system DEMON on how they were handled by users. More precisely, on whether the users manage to answer in a way that put the dialogue towards its goal.

We were firstly interested in observing how it helps users to provide them with guidance on how to interact with the system in the welcome prompt.

As erroneous confirmations appear as an important problem in dialogue systems (van Haren et al., 1998), we have also been interested in comparing three different confirmation strategies. The comparison was base on how they helped the users with correcting badly understood and confirming correctly understood information.

The last point we have been working on was to build a set of metrics to measure complexity of system's questions and to evaluate how different sources of complexity may influence users' answers. The complexity of questions was defined considering cognitive load.

First we will shortly introduce our system DEMON, its developing and evaluation frame and the different kinds of prompts it uses. We will then present the particular points we have been interested in dialogue, the measures we developed for evaluation and some results obtained for complexity criteria.

2. The DEMON System

DEMON is a spontaneous speech dialogue system allowing mixed initiative. It has been developed in the frame of A.R.I.S.E. LE3-4229 project (Automatic Railway Inquiry Systems for Europe) (ARISE, 1996). This project consists in parallel development of Dutch, Italian and French demonstrators allowing telephonic access to train timetable and in their comparison in order to share experience from experimentation with various users and various technical improvements. See (Blasband, 1998).

DEMON (Pérennou, 1997) has been developed on a Philips platform including: telephonic interface ensuring calls management, PHICOS software ensuring speech recognition and DR software ensuring understanding, dialogue control, messages synthesis and database interface. Philips system is described in (Aust et al., 1995). Currently DEMON gives actual information from S.N.C.F.'s (French Railway National Service) RIHO (train timetable) database for the 600 most frequently asked for train stations. Dialog strategy in DEMON is primarily a slot filling strategy. In order to give a time schedule information the system requires four pieces of information: *departure station*, *arrival station*, *day of departure or arrival* and preferred *time of departure or arrival*. When all this information has been provided, DEMON prompts the users with closest connections. User may then ask for a later or earlier connection, return travel or other connections.

3. System Prompts in DEMON

In DEMON tow types of system questions are used: direct questions and confirmation questions.

Direct questions such as (1) are uniquely used to obtain from the user an information so far unknown to the system.

(1) DEMON - "Quel jour désirez-vous partir?"
DEMON - "On which day do you want to leave?"

Confirmation questions are used to make the user confirm or deny what was understood by the system. They may take tow different forms: either explicit or implicit. An explicit confirmation question, illustrated by (3) as an answer to (2), explicitly ask for confirmation that some information was correctly understood. Alternative implicit confirmation question (3bis) includes the request for confirmation in a direct question. An answer to the direct question induces implicit confirmation of included information.

(2) USER - "Je veux partir avant 8 heures"
USER - "I want to leave before 8 o'clock"
(3a)DEMON - "Désirez-vous partir avant 8 heures?"
DEMON - "Do you want to leave before 8 o'clock"
(3b) DEMON – "Vous partez avant 8 heures. Quel jour?"
DEMON - "You are leaving before 8 o'clock. Which day?"

4. Giving Advice in the Welcome Prompt

The first point we have been interested in was the effect of giving advice on how to use the system in the welcome prompt. Three advice messages were tested.

A first approach for evaluating the effect of advice is to observe the users behaviour during the whole dialogue. To differentiate the possible effects of the different advice, different combinations of advice were used in different experiments. Advice in the welcome prompt did not seem to be of any help for users. Actually, guidance given in the welcome prompt seems not to be remembered by users as showed survey performed during the experiments (Lavelle & al., 1998). Or when they are, it appears that a user answer to a system's prompt is too much based on communication reflex to include some theoretic information.

A second approach for evaluating this prompt was based on its length. Actually, depending on the number of advice given the prompt was longer or shorter. The measure for evaluating the effect of the length of this was based on first users' sentence: naturalness and amount in information provided.

Complete welcome prompt is as follow. Elements in brackets appear or not depending on system version.

(4) DEMON_ *"Ici serveur vocal experimental de renseignements sur les horaires. Quelques conseils. [Parlez de maniere concise et naturelle.] [Repondez simplement à votre tour aux questions posées.] [N'hésitez pas à corriger le system en cas d'erreur.]*
Veuillez indiquer les villes de départ et d'arrivée du trajet souhaité."
DEMON - "Experimental vocal server for train schedule information speaking. [A few advise.] [Speak naturally and concisely.] [Answer simply at your turn to asked questions.] [Do not hesitate to correct the system in error case.]
Please indicate departure and arrival stations for required travel. »

The expected answer to this Prompt (users' first utterances) is a statement including at least departure and arrival stations and possibly indication of travel date and time.

Therefore we identify three main situations:

- Users may give the expected information in a natural sentence. We consider a sentence as natural if it includes a conjugated verb.
- Users may give the expected information in a straightforward manner. Only relevant information is given with possibly useful prepositions. Utterances are like :"Toulouse Paris demain" (*"Toulouse Paris tomorrow"*).
- Users may not answer or answer not relevant sentences. We classify those answers as lost users

There was no evidence that the welcome prompt length has any influence on users' first utterances. The same prompt (prompt without advice) was used in two different experiments that show very different results. Even, corpora obtained from experiments with a welcome prompt indicating people they may speak naturally are among those having the poorest natural sentence rates.

5. Comparing Confirmation Strategies

We have also been interested in comparing different confirmation strategies (lavelle, 1999a). Three strategies were experimented in three different version of the system.

5.1. Strategies

5.1.1. DEMON_0: Explicit and Implicit Single Sentence Questions

In this first version of the system implicit confirmations were used as largely as possible. They were used each time there was at least one parameter to confirm and they included every parameter to confirm at the stage they were uttered. They were made of one single sentence.

For example, if Monday the 5th of January 1998 as to be confirmed as departure day and departure time as not been provided yet, the question was:

(5)"A quel heure désirez vous partir le lundi 5 janvier 1998?"

"At what time do you which to leave on Monday the first of January?"
Explicit confirmation questions were used only if there was no parameter left to fulfil.

5.1.2. DEMON_1: Explicit and Implicit Two Sentence Questions

In this second version, implicit questions were redesigned to match the order of mental operation necessary to answer the question. They were therefore divided in to sentences: first a statement of parameters to confirm, then a question on parameters still to fulfil.
The question for the situation described above was then:
(6)"Vous partez lundi 5 janvier 1998. A quelle heure?"
"You are leaving on Monday the 5th of January 1998. At what time?"
Application cases for both explicit and implicit questions remained unchanged.

5.1.3. DEMON_2: Explicit and Semi-implicit Questions

In this last version we introduced semi-implicit questions to replace implicit ones. Those carry the same information as implicit questions: a question on new information and a request for confirmation but, as do explicit questions, they clearly state they can be denied.
In the situation described above the confirmation question would now be:
(7)"Vous partez lundi 5 janvier 1998. En cas d'erreur corrigez moi, sinon, indiquez l'heure de votre départ."
"You are leaving on Monday the 5th of January 1998. In error case, correct me, otherwise, indicate your departure time."
Application cases for explicit and semi-implicit questions were also redefined. Semi-implicit questions only include confirmation request for concept of the same semantic type. In this applications we consider two semantic types: place for departure and arrival train station and time for departure or arrival day and time.

5.2. Measures and Results

As we are dealing here with confirmation strategies they have to be compared with regard to how they help the user with confirming correctly understood information and correcting badly understood information. We therefore decided to evaluate the dialogue at two levels: whole dialogue evaluation and system's question/user's answer pairs.

5.2.1. Dialogue Evaluation

Two metrics were used to evaluate the dialogue as a whole.
First, was based on task complexion. Three values used to characterise task complexion: 1) *success* if the user has obtained the information he or she asked for, 2) *failures* if the dialogue ends before he or she got any information at all, 3) and what we call *irrelevant answer,* if the system provide the user with information for some travel which does not match the one he or she asked for. This last dialogue result value highlights situation where erroneous confirmation as caused dialogue system's task complexion failure.
The second was a measure for efficiency: the dialogue length in terms of number of dialogue turns.

5.2.2. Measures for System's Question/user's Answer Pairs

These metrics were designed to evaluate how system question were understandable or actually answerable for users and their efficiency in the correction process. Two measures were used to evaluate understandability of questions: the *refutation rate,* percentage of denied confirmation over the number of confirmations including badly understood parameters, and the *no answer rate,* percentage of question that were not answer on the first time and had therefore to be repeated. Evaluation of efficiency is based on the observation that if the user immediately answer a confirmation request holding information to be corrected with the correct information we save an extra dialogue turn asking for the correct information. We so decided to use *direct correction rate,* percentage of this immediate correction over the number of denied confirmation as our measure for question's efficiency.

5.2.3. Results

From our observations it appears that refutation rates were better with tow sentence implicit questions than with one sentence implicit questions and that refutation rate for semi-implicit questions equals the one for explicit questions. Dialogue success rate seems to be tight with refutation and direct correction rates. Finally, semi-implicit confirmations, by increasing refutation and direct correction rates, have helped to increase dialogue success rate and shorten successful dialogues.

6. Measuring System's Question Complexity

The last point we have been working on was observation of users' answer depending on complexity of system's questions (Lavelle, 1999b). We considered five criteria that we thought might imply complexity in questions. They were chosen according to experience from reviewing corpora and defined so that they can be evaluated automatically during the course of the dialogue. The four first criteria are based on information concepts involved in questions and question focus. In DEMON we have four information concepts: <departure station>, <arrival station>, <day of departure or arrival> and <time of departure or arrival>. We consider a concept as involved in a question if the question includes a confirmation request for a value of that concept. Those first criteria are therefore only relevant for confirmation questions.

6.1. Complexity Criteria

The number of concepts involved in a question. As having all four concepts in one system's utterance is quite rare in our application, we consider three values for this criterion: one concept, two concepts and over two concepts.

The semantic distance between the different concepts involved in a question. In the DEMON system we consider tow semantic types for information concepts:

<place> for information concepts <departure station> and <arrival station> and <time> for concepts <day> and <time>. We consider two values for this criterion: All concepts involved are of the same semantic type (noted USemConY in figures) and not all concepts involved are of the same semantic type (noted USemConN in figures).

The semantic distance between the concepts involved in a question and the focus of that question. Here again we consider two values for this criterion: semantic unity between involved concepts and question focus (noted UsemQueY) or no unity (noted UsemQueN). In the implicit confirmation question (3b), involved concept <departure time> is of the same semantic type, <time>, than the question focus <departure day>. We should note here that this criterion is essentially relevant for implicit confirmation questions. In explicit confirmation questions, such as (3a), involved concepts are also the focus of the question. Semantic unity for involved concept and question focus is therefore always true for those questions.

The formulation of concepts: We are interested here with the lexical end syntactical continuity between user's formulation of an information concept value and DEMON's formulation of that same value. It should be noted here that DEMON does not intend to reproduce users' formulations but uses formulations that its comprehension module better understands. This criterion can therefore only be evaluated when the user pronounces in first place a value for a concept. This criterion can take two values: FormY if the formulation used by DEMON is the one used by the user and FomN otherwise. Considering we are on Monday the 19th of June 2000, the exchange (8-9) is an example of the first case. The value presented by the system for the concept <day of departure> matches the value demanded by the user, but the formulation used does not match the one used by the caller.

(8) USER – "Je veux partir dimanche 25 juin."
USER – "I want to leave on Sunday the 25th of June."
(9) DEMON – "Voulez-vous partir dimanche prochain?"
DEMON – "Do you want to leave next Sunday?"

The contextual relevance of questions: This criterion was design to evaluate if the system's question is relevant considering the last system-user exchange. The positive case is noted ContextY in figures; the negative case is noted ContextN Evaluating such a criterion appear highly subjective in the first place. As our measures were intended to be automatically evaluated during the course of the dialogue we had to define a computable measure. A question from the system is considered to be contextually relevant if either its focus or its involved information concepts match the focus of the previous system's question. With this definition, considering the exchange (10-11), the question (12a) is relevant whereas the question (12b) is not.

(10) DEMON - "Vous partez pour Grenoble. Indiquez votre ville de depart."
DEMON – '"You are going to Grenoble State your departure town."
(11) USER – "Je pars de Toulouse le dimanche 18 juin."
UQSER – "I am leaving from Toulouse on Sunday the 18th of June."
(12a) DEMON – "Vous partez le dimanche 18 juin de Toulouse. A quelle heure?"
DEMON "You are leaving on Sunday the 18th of June from Toulouse. A what time?"
(12b) DEMON – "Avez-vous dit le dimanche 18 juin?"
DEMON – "Did you say on Sunday the 18th of June?"

6.2. Measure

Here again we aimed at evaluating effects of these different types of complexity on users answers. The measure we used here is an extension of the refutation rate used to evaluate confirmation questions. As every system's questions were to be considered we rather used *correct answer rate,* where an answer is considered to be correct if, for a confirmation question, it denies incorrect values for involved concept, and otherwise confirms proposed values; or, for a direct question, provides values for requested concepts. When the user does not answer a direct question or does not deny a wrongly understood value for some information concept, the answer is considered *not correct*. If the user does not answer at all or if is answer is not relevant in the context of dialogue we consider answers to be not classifiable.

6.3. Experiments and Corpora

Observations for evaluation of the three different confirmation strategies have been performed on 3 corpora (see table 1), obtained from 3 different experiments. Observations for evaluating the effect of question complexity were performed only on the tow last corpora. In each experiment a few users hang up before the end of the welcome prompt. Those calls have not been considered in corpora. These experiments were realised in collaboration with S.N.C.F. They organised the tests selected the callers and asked them to call the system and fill appreciation forms. They used appreciation forms to evaluate the users' acceptation of the system (Gitton & Temem, 1997). Our evaluations are based on written transcriptions of recorded calls.

Corpus	Number of call	Callers
November_97 (DEMON_0)	41	French railways employees
February_98 (DEMON_1)	88	French railways users
November_98 (DEMON_2)	200	French railways users

Table 1: Summary of experiments

6.4. Results

For both studied corpora the percentage of correct answer over all answer is 83%. The rate of incorrect answer is 10% and the rate of not classifiable answers is 7%. We present below observations confirmations question. We will only present correct answer rates.

Number of concepts involved in a prompt. If we look at table 2 and 3, we notice that questions involving more than two concepts to be confirmed seem to be difficult to handle for users. This is verified for both explicit and implicit confirmations. But, it must be noted than in DEMON, having more than tow concept in a question implied that they are not of the same semantic type.

For explicit confirmations the number of concepts involved does not seem to have any influence on users answers. We must note here that in DEMON explicit confirmations involving one concept are used in cases of misunderstanding or ambiguity.

	1 concept	2 concept	More than 2 concepts
Explicit questions	76%	87%	77%
Implicit questions	90%	87%	74%

Table 2: Correct answer rates in corpus February_98 depending on number of concepts involved

	1 concept	2 concept
Explicit questions	78%	76%
Implicit questions	84%	89%

Table 3: Correct answer rates in corpus November_98 depending on number of concepts involved.

Semantic unity of information concepts involved in prompts. Because of low occurrence number we could not calculated this rate for corpus November_98. Results for corpus February_98, presented in table 4, show that semantic unity seems to simplify handling of questions.

Semantic unity of concepts involved and focus of system's prompts. This measure can only be evaluated for implicit confirmations. From table 5, we may suppose that no semantic unity between involved concepts and question focus lead to more answerable questions. It is possible that when the concepts involved are of one semantic type and the focus is of the other, this create a contrast that highlights the concept to be confirmed.

	UsemConY	UsemConN
Explicit questions	90%	77%
Implicit questions	83%	72%

Table 4 : Correct answer rates in corpus February_98 depending on semantic unity of concepts involved.

	UsemQueY	UsemQueN
Implicit questions (February_98)	75%	86%
Semi-implicit questions (November_98)	84%	88%

Table 5: Correct answer rates for implicit questions in corpus February_98 and November_98 depending on semantic unity of concepts involved and focus.

Formulation of concepts. Results presented in tables 6 and 7 tend to show that whether the system re-use the user's words or not for describing references does not make any difference. Using formulation that the system better understand to influence users seems therefore to be a safe approach.

Contextual relevance of system's prompts. Table 8 and 9 show for both corpora a significant difference for correct answer rates for contextually relevant and not contextually relevant prompts. We may notice that disambiguation in DEMON are performed previous to any other action. As a result not contextually relevant prompts as we defined them tend to occur in cases of recognition errors. This is of course not the easiest situation to deal with for users. But anyway, changing the focus of the question when it is not necessary should be avoided. And dealing with ambiguity first does not seem to be a good idea with regard to usability of systems.

	FormY	FormN
Explicit questions	87%	89%
Implicit questions	80%	81%

Table 6 : Correct answer rates in corpus February_98 depending on formulation of concepts involved.

	FormY	FormN
Explicit questions	83%	72%
Implicit questions	84%	88%

Table 7 : Correct answer rates in corpus February_98 depending on formulation of concepts involved.

	ContextY	ContextN
Explicit questions	89%	69%
Implicit questions	86%	59%

Table 8 : Correct answer rates in corpus February_98 depending on contextual relevance.

	ContextY	ContextN
Explicit questions	79%	69%
Implicit questions	87%	60%

Table 9: Correct answer rates in corpus February_98 depending on contextual relevance.

7. Conclusion

If we look at evaluations presented in this paper with the idea of deciding which, among all the possible questions, should be asked at any state of a dialogue, several observations may be noted.

First results from those various experiments was that welcome prompt was probably not the right place for advice. From evaluation of semi-implicit confirmations we see that adding advice or guide comments inside the dialogue, at the time it is needed, was a lot more efficient. We also observed the importance of the order of concepts and question focus in implicit confirmation questions. When they are ordered to match the required mental operations, users seem to handle these questions more easily. From our work on complexity criteria it appears that contextual relevance and semantic proximity of involved concepts seem to have the strongest influence on users answers. Not contextually relevant question should be avoided as much as possible. This excludes the dialogue strategy that consists in disambiguating first and caring on the dialogue afterward. Confirmation questions should involve only concept of the same semantic type and apparently no more than two concepts.

In the general case, formulation of concepts and semantic unity between concept involved and focus of questions does not seem to have a strong influence in the complexity of prompts. It may be interesting to investigate their influence in cases were correction is required. Since, this situation is always more difficult to handle for users they be more vulnerable.

It would also be interesting to evaluate cross-effect between the different criteria. This would help to decide which weight should be given to each criterion. Unfortunately for results to be significant this would require bigger corpora those that we have.

We are presently developing a dialogue system similar to DEMON, where prompting strategy is based on observation showed her. We hope this will allow us to better evaluate the influence of criteria and guidelines defined here.

Acknowledgements

The authors acknowledge the invaluable contribution of *Philips Research Labs Aachen* for providing the Dialogue workstation and used during ARISE Project and SNCF for their help in the test experiments.

8. References

A.R.I.S.E. LE3-4229 Project Programme (1996).

Aust H, Oerder M, Seide F, Steinbiss V. (1995). The Philips Automatic Train Timetable Information System. Speech Com., 17, pp. 239--250

Blasband, M., Speech Recognition in Practice the ARISE Project (Automatic Railways Inquiry Systems for Europe). (1998). la lettre de l'I.A.,134-135-136, 207--210

Gitton S., Temem J.N. (1997). Specification and Evaluation of French Automatic Telephone Information systems in the ARISE Project. In proceedings of WCCR.

Lavelle, C.A., de Calmès , M., Pérennou, G.. (1998). A Study of Users' Behaviors in Different States of a Spontaneous Oral Dialog with an Automatic Inquiry System. In proceedings of the IEEE fourth workshop on Interactive Voice Technology for Telecommunications Applications, pp 118--123.

Lavelle, C.A., Pérennou, G. (1999). Spoken Dialogue Systems: Toward a complexity Measure for System's Questions. In proceedings of the Workshop on Speech and Computer, pp 79--82.

Lavelle,C.A., de Calmès, M., Pérennou, G.. (1999) Confirmation Strategies to Improve Correction Rates in a Telephonic Inquiry Dialogue System, In proceeding of the sixth European Conference on Speech Communication and Technology, pp 1999--2002.

Peckham J. (1995). Conversational Interaction: Breaking the Usability Barrier, In proceedings of ESCA workshop on Spoken Dialogue Systems, pp 1-8.

Pérennou, G. de Calmès, M. Lavelle, A. Tronel, R. (1998). Un système de dialogue oral spontané pour l'accès téléphonique aux informations d'horaire de train. la lettre de l'I.A.,134-135-136, 207--210

Springer, S., Basson, S., Kalyanswamy, A., Man, E., Yashcin, D. (1995). The Money Talks Interactive Speech Technology Assessment: a Report from the Field. In proceedings of the fourth European Conference on Speech Communication and Technology, pp 1939--1942.

Thomson, D.L., Wisowaty, J.J.. (1999). User Confusion in Natural Language Services. In proceedings of the ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems.

van Haaren, L., Blasband, M., Gerristen, M., van Schijndel, M. (1998). Evaluating Quality of Spoken Dialogue Systems: Comparing a technology-focused and a User-focused Approach,. In proceedings of the first International Conference on Language & Evaluation, pp 655--660.