FLEXIBLE MIXED-INITIATIVE DIALOGUE FOR TELEPHONE SERVICES

Relaño Gil, José ♮ and ***Tapias, Daniel*** and ***Villar, Juan M.*** ♮
Gancedo, Maria C. ♮ and *Hernández, Luis A.* ♮
Speech Technology Group, Telefónica Investigación y Desarrollo, S.A.
C. Emilio Vargas, 6 28043 - Madrid (Spain)
♮ Dep. SSR ETSIT-UPM Spain. Tel:34.1.549500. Fax:34.1.3367350.
e-mail:jrelanio@gaps.ssr.upm.es

10.21437/Eurospeech.1999-302

Abstract

In this work, we present an experimental analysis of a Dialogue System for the automatization of simple telephone services. A first evaluation of a preliminary version of the system was done based on the Speech Recognizer error rate and on the identification of two groups of users, that we refer to as group A and B. From this evaluation we conclude the necessity to design a robust and flexible system suitable to have different dialogue control strategies depending on the characteristics of the user and the performance of the speech recognition module. A system adaptation procedure combining a normalized average number of utterances per task, the amount of information in some particular utterances, and an estimate of the recognition error rate. Our adaptation procedure defines two different control management strategies for a flexible mixed-initiative strategy: fixed for high recognition errors and Group B users, and mixed for low error rates and A users. Experimental results following the PARADISE framework showed an important improvement.

1 INTRODUCTION

In this contribution we present some improvements on the design of a Dialogue Management System for the automatization of simple telephone tasks in a PABX environment (automatic name dialing, voice messaging, ...). From the point of view of its functionality, our system is a very simple one because there is no need of advanced Plan Recognition strategies or General Problem Solving methods. However we think that even for these kind of dialogue systems there is still a long way to demonstrate their usability in real situations by the "general public".

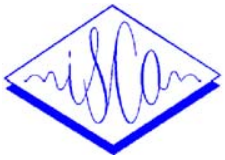
In our work we will concentrate on systems designed for the telephone line and for a wide range of potential users. Therefore our evaluations will be done taking into account different levels of speech recognition performance and user behaviors. In particular we will propose and evaluate strategies directed to increase the robustness against recognition errors and flexibility to deal with a wide range of users. We will use the PARADISE evaluation framework [4] to analyze both task success and agent dialogue behavior related to subjective user satisfaction.

2 BASELINE SYSTEM

Following the classification of Dialogue Systems proposed by Allen [1], our baseline dialogue system could be described as a system with topic-based performance capabilities, adaptive single task, a minimal pair clarification/correction dialogue manager and fixed mixed-initiative.

One of the most important objectives of our dialogue manager has been the implementation of a collaborative dialogue model. So the system has to be able to understand all the user actions, in whatever order they appear, and even if the focus of the dialog has been changed by the user. In order to achieve this, we organize the information in an information tree, controlled by a task knowledge interpreter and we let the data to participate in driving the dialogue. However, to control a mixed-initiative strategy we use three separate sources of information: the user data, the world knowledge embedded in the task structure and the general dialog acts.

The basic architecture of the system is composed of three major modules, a Semantic Parser, a Dialogue Manager and different data structures that we called Repository of Information. Each of these principal items includes the possibility of configuration through different mechanisms: a set of semantic networks inside the Parser, local task rules and general rules of dialogue within Dialogue Manager, and automata and local understanding rules within an information storage Tree into the Repository. Moreover, we have defined an historic of dialogue inside the Repository. So our objective is a system configurable in relation to the data structure of the task, dialogue strategies, and decisions of the output strategies, depending on the knowledge of the environment, shaping different behaviors and adapting itself to several



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

users.

The parser is a modified version of the Phoenix system of C.M.U. [3] This is a flexible frame-based parser implemented through Recursive Transition Networks (RTN). In our system we have modified the semantic parser in order to allow word insertions at any part of the RTN. This ability is very useful to make the system robust from word insertions, either from the recognizer or from the speaker who usually includes non-keyword expressions in spontaneous speech.

The dialogue manager uses the repository module like a super-structure for grouping the two major sources of information of the system: the task information, and the dialogue history. The task information is stored in a tree structure, and the dialogue history in a list of dialogue events.

Based on the architecture and control procedures previously described our system evolves in a mixed-initiative environment resolving pre-designed confirmation strategies. We consider that it is very important to control the confirmations leading by dialogue, which means, in our approach a confirmation have to be made only when strictly necesary. For obtaining it, we propose three confirmation kinds, depending on the system status: explicit, implicit and "confirmation on wait" the system communicates the user the action required and it waits a moment for a negative answer, if it does not happen the system takes the silent like a confirmation.

3 FIRST EVALUATION RESULTS

In order to test and improve the original system (described in [2]) we designed a simulated evaluation environment where the performance of the Speech Recognition Module (recognition rate) was artificially controlled. As we already point out in the Introduction our system is intended to be used by a variety of users through the telephone line. Therefore we must take into account that state-of-the-art speech recognition systems are still very dependent on the noise and different characteristics of different telephone lines and also on the different voice characteristics of different users. It is well known that speaker independent ASR system always show bad recognition rates for certain populations of speakers, even though new speaker adaptation techniques try to minimize this problem.

A Wizard of Oz environment was designed to obtain a dialogue database simulating two different levels of recognition performance for a vocabulary of 1170 words: 96.4% word recognition rate for high performance and 80% for low performance. The same pre-defined single fixed mixed-initiative strategy of our system was used by the Wizard in all the cases. Fifty different dialogues were obtained for 50 different novice users. The similar instructions on the use of the system were given to each novice user. Each dialogue consisted of six different telephone tasks: 3 were simulated using 94.6% recognition rate and 3 with 80%. Performance results, presented in Table 1, were obtained using the PARADISE evaluation framework [4], determining the contributions of task success and dialogue cost to user satisfaction. As task success measure me obtained the Kappa coefficient while dialogue cost measures were based on the number of users turns. In this case it is important to point out that as each tested dialogue is composed of a set of six different tasks which have quantify different number of turns, the number of turns for each task was normalized to it's $N(x) = \frac{x+\overline{x}}{\sigma_x}$ score

	Both Group		Low ER	
	High ER	Low ER	Gr. A	Gr. B
κ	0.68	0.81	1	0.61
User Turn	7.3	5.4	4.2	6.9
Satisf	26.4	30.1	35.4	25.2

Table 1: Shows means results for both group in low and high ER (Error Rate). And separately for each Group A and B, only in Low ER situation

User satisfaction in Table 1 was obtained as a cumulative satisfaction score for each dialogue by summing the scores of a set of questions similar to those proposed in [4]. The ANOVA for Kappa, the cost measure and user satisfaction demonstrated a significant effect of ASR performance. As it could be predicted, we found that in all cases a low recognition rate corresponds to a dramatical decrease in the absolute number of successfully completed tasks and an important increase in the average number of utterances. Our major conclusions are:

1. There is an obvious correlation between low performance and low recognition rates. In all cases a low recognition rate corresponds to a dramatical decrease in the absolute number of successfully completed tasks and an important increase in the average number of utterances.

2. A less obvious result was to discover that in high recognition situations the performance for a group of 54% of the users (Group A), both in terms of number of completed tasks and average number of utterances, was clearly higher that for the rest of the users (46% Group B). A closer inspection of the dialogues revealed that this difference in performance was mainly due to a clear mismatch between the actual user dialogue strategy and the fixed strategy of the system. That is, users belonging to Group A showed a "fluent" interaction with the system similar to the one supposed to establish the mixed-initiative strategy (for example, as an

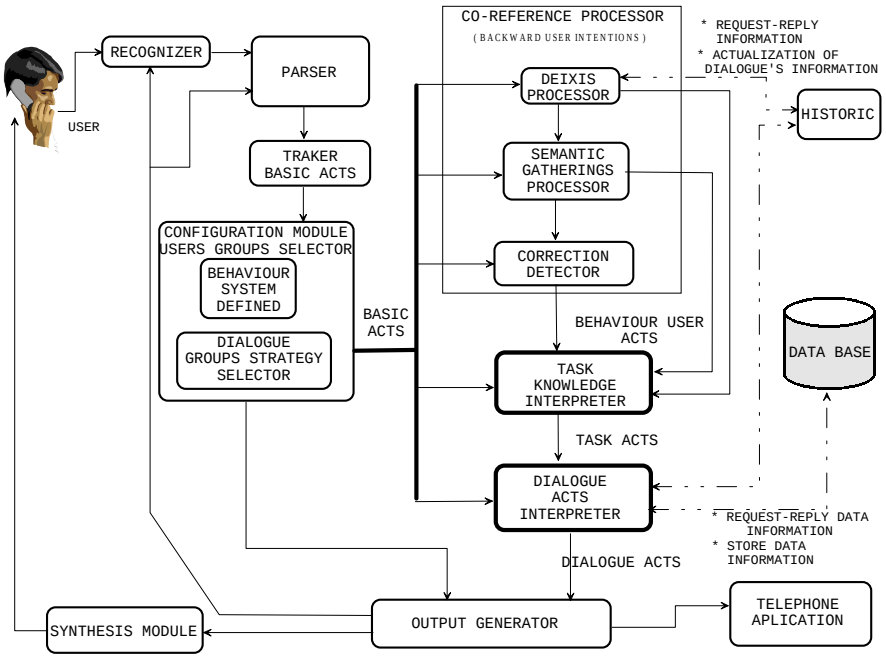


Figure 1: Modules of Robust and Flexible Mixed-Iniciative Dialogue

answer to the question of the system "do you want to do any other task?", these users could answer something like "yes, I would like to send a message to John Smith"). While users in Group B exhibited a very restrictive interaction with the system (for example, a short answer "yes" for the same question).

3. We also noticed that a closer analysis of the important reduction in performance for B users was not only due to a poor recognition rate. In almost a 50% of the cases (of Group B) we observed that the increase in average number of utterances was due to the use of anaphorical and eliptical expressions. And we must say that our baseline system did not include any kind of treatment for these cases. Therefore we can say that in almost a 23% of the 50 speakers an extensive use of anaphorical references was used even for a simple task as our limited number of telephonic functions. This fact points out the importance of dealing with these phenomena in real situations.

4. Finally, we observed that looking at the general behavior of the system, in an important number of utterances produced during confirmation or correction phases of the dialogue (approximately in a 15% of these situations) , users produce highly redundant expressions that are very difficult to be processed by our parser; for example: "yes, ofcourse, yes, yes I want to make a call" or "no, no, thats not what I want, I do not want to make a call". In most of these cases the parser recognition rate decreases.

Therefore, from this preliminary evaluation of the system we found that in order to increase its performance two major points should be addressed: a) robustness against recognition and parser errors, and b) more flexibility to be able to deal with different user models.

4 FLEXIBILITY AND ROBUSTNESS

Flexibility to deal with different user behaviors and robustness against different Speech Recognition error rates was tested throught the design of an adaptation strategy of our dialog manager. In particular we designed strategies to adapt our dialogue manager to Group A or B of users and to the simulated High and Low ER situations. Adaptation was based on the combination of a jointly characterization of user behaviour and recognition errors according to the following procedure:

1. To estimate the performance of the Speech Recognition module. This was done in the Co-reference Processor, see Figure 1, from a count on the number of corrections during previous interactions with the same user. This processor is activated when an important number of assertory, negative expressions or words that mean co-references are detected at the output of the parser. The Co-reference Processor is based on a set of compactation rules related to both the present speech act and the dialogue state. In order to have a proper estimation of the number of user corrections, special

rules have been designed to deal with the observed highly redundant structures during confirmation and correction phases. (We are also considering the use of confidence scores directly from the speech recognizer, but we do not have any experimental results at this moment).

2. To classify each user as belonging to group A or B. This was done from a count of user turns in the Configuration Module (Figure 1), combining a normalized average number of utterances per task.

3. To decide the dialogue strategy to be followed by the system. From the estimation of recognition performance (Co-reference Processor) and normalized average of user turns (Configuration Module), the dialogue management adaptation is made through Linear Discrimination Analysis (LDA). As it can be seen in Figure 2, the two-dimensional space representation of normalized average number of turns and recognition error rate defines two different regions, corresponding to Group A and B users, that can be linearly separated for low recognition error rates. While during high error rates these two regions are merged due to the low performance of the system for both kind of users. Then mixed-initiative was decided for Group A and low error rates, and system fixed-initiative for Group B and/or high error rates.

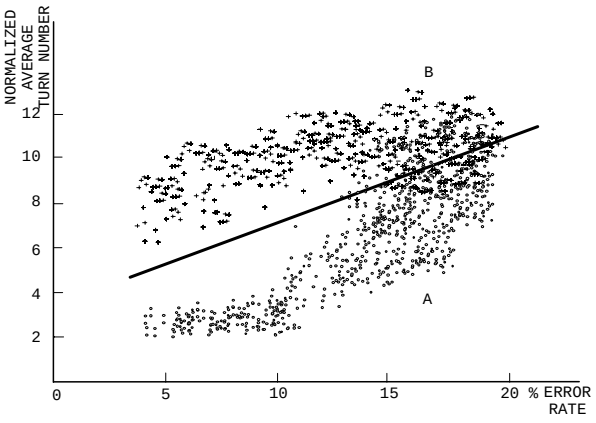


Figure 2: User classification

As a result of the final decission, the fixed or mixed initiative is implemented by means of a dynamic change of the dialog control rules that implements the Dialogue Acts Interpreter module (Figure 1).

5 FINAL EXPERIMENTAL RESULTS

Experimental results to test the improvements over our baseline system used the same evaluation environment described in Section 3 but with the proposed dynamic adaptation initiative. Table 2 shows mean results for each Group A and B of users for low ASR error rate, and for all users in Low ASR situations. In this case, although no detailed results are given, due to the low performance of the ASR system, the low performance of the dialogue system is quite similar for both Group A and B users. However, compared to Table 1, for high error rates only a moderate improvements are obtained with a low impact in the level of user satisfaction. Finally for low error rates the switch between fixed and mixed initiative provides a more stable behavior of the system, that is, less difference in performance between users of Group A and Group B.

	High ER	Low ER	
	Both Gr.	Gr. A	Gr. B
κ	0.71	1	0.83
User Turn	7.2	5.3	6.1
Satisfaction	26.9	32.1	29.4

Table 2: Shows means results for each Group in Low ER situations and for both in High ER.

The main conclusion of the work is the necessity to design adaptive dialogue management strategies to make the system robust against recognition performance and different user behaviors.

6 ACKNOWLEDGEMENTS

The authors from ETSIT-UPM are grateful to Telefónica I+D and CICYT (TIC-96-0956-C04-03) for their financial support.

References

[1] James Allen. *Tutorial: Dialogue Modeling.* ACL/ERACL Workshop on Spoken Dialogue System, Madrid, Spain, 1997.

[2] J. Alvarez, J. Caminero, C. Crespo, and D. Tapias. *The Natural Language Processing Module for a Voice Asisted Operator at Telefónica I+D.* ICSLP '96, Philadelphia, USA, 1996.

[3] S. Issar and W. Ward. *CMU's Robust Spoken Understanding System.* Proceding in Eurospeech '93, Berlin, 1993.

[4] M. Walker, D. Litman, C. Kamm, and A. Abella. *Evaluating spoken dialog agents with PARADISE: Two case studies.* Computer speech and language, 1998.