Robust and Flexible Mixed-Initiative Dialogue for Telephone Services

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

Abstract

In this work, we present an experimental analysis of a Dialogue System for the automatization of simple telephone services. Starting from the evaluation of a preliminar version of the system we[1] conclude the necessity to desing a robust and flexible system suitable to have to have different dialogue control strategies depending on the characteristics of the user and the performance of the speech recognition module. Experimental results following the PARADISE framework show an important improvement both in terms of task success and dialogue cost for the proposed system.

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 sytems 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 behaviours. 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 (Walker et al., 1998) to analyze both task success and agent dialogue behaviour related to subjective user satisfaction.

[1] ♮ Dep. SSR ETSIT-UPM Spain

2 ROBUST AND FLEXIBLE SYSTEM

Following the classification of Dialogue Systems proposed by Allen (Allen, 1997), 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 dialogue 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 dialogue acts.

Therefore, from this preliminar evaluation of the system we found that in order to increase its permormance 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. We designed four complementary strategies to improve its performance:

1. To estimate the performance of the speech recognition module. This was done from a count on the number of corrections during previous interactions with the same user.

2. To classify each user as belonging to group A or B that will be described later in the Experimental Results section. This was done combining a normalized average number of utterances per task and the amount of information in each utterance, especially at some particular dialogue points (for example when answering to the question of our previous example).

3. To include a control module that from the results of steps 1 and 2 defines two different kinds of control management allowing a flexible mixed-initiative strategy: more user initiative for Group A users and high recognition rates, and more restictive strategies for Group B users and/or low recognition performance.

All of these strategies have been included in our system as it is depicted in Figure 1.

3 EXPERIMENTAL RESULTS

In order to test the improvements over our original system (described in (Alvarez et al., 1996)) we designed a simulated evaluation environment where the performance of the Speech Recognition Module (recognition rate) was artificially controlled.

A Wizard of Oz simulation environment was designed to obtain different levels of recognition performance for a vocabulary of 1170 words: 96.4% word recognition rate for high performance and 80% for low performance. A pre-defined single fixed mixed-initiative strategy was used in all the cases.

We used an annotated data base composed of 50 dialogues with 50 different novice users and 6 different simple telephone tasks in each dialogue: 25 dialogues were simulated using 94.6% recognition rate and 25 with 80%. Performance results were obtained using the PARADISE evaluation framework (Walker et al., 1998), determining the contributions of task success and dialogue cost to user satisfaction. Therefore 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 $N(x) = \frac{x+\bar{x}}{\sigma_x}$ score

	Both Group		High ASR	
	Lo ASR	Hi ASR	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 ASR. And separately for each Group A and B, only in high ASR 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 (Walker et al., 1998). The ANOVA for Kappa, the cost measure and user satisfaction demostrated 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 suscessfully completed tasks and an important increase in the average number of utterances.

However we also found that in high ASR situation the task success measure of Kappa was surprisingly low.

A closer inspection of the dialogues in Table 1 revealed that this low performance under high ASR situations was due to the presence of two groups of users. A first group, Group A, showed a "fluent" interaction with the system, similar to the one supposed by the mixed-initiative strategy (for example, as an 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 the other group of users, Group B, exibited a very restrictive interaction with the system (for example, a short answer "yes" for the same question).

As a conclusion of this first evaluation we found that in order to increase the permormance of our baseline system, 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.

Therefore we designed an adaptive strategy to adapt our dialogue manager to Group A or B of users and to High and Low ASR situations. The adaptation was done based on linear discrimination, as it is ilustrated in Figure 2, using both the average number of turns and recognition errors from the two first tasks in each dialogue.

	Low ASR	High ASR	
	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 high ASR situations and for both in low ASR.

Table 2 shows mean results for each Group A and B of users for High ASR performance, and for all users in Low ASR situations. These results show a more stable behaviour of the system, that is, less difference in performance between users of Group A and Group B and, although to a lower extend, between high and low recognition rates.

4 CONCLUSIONS

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

References

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

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

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

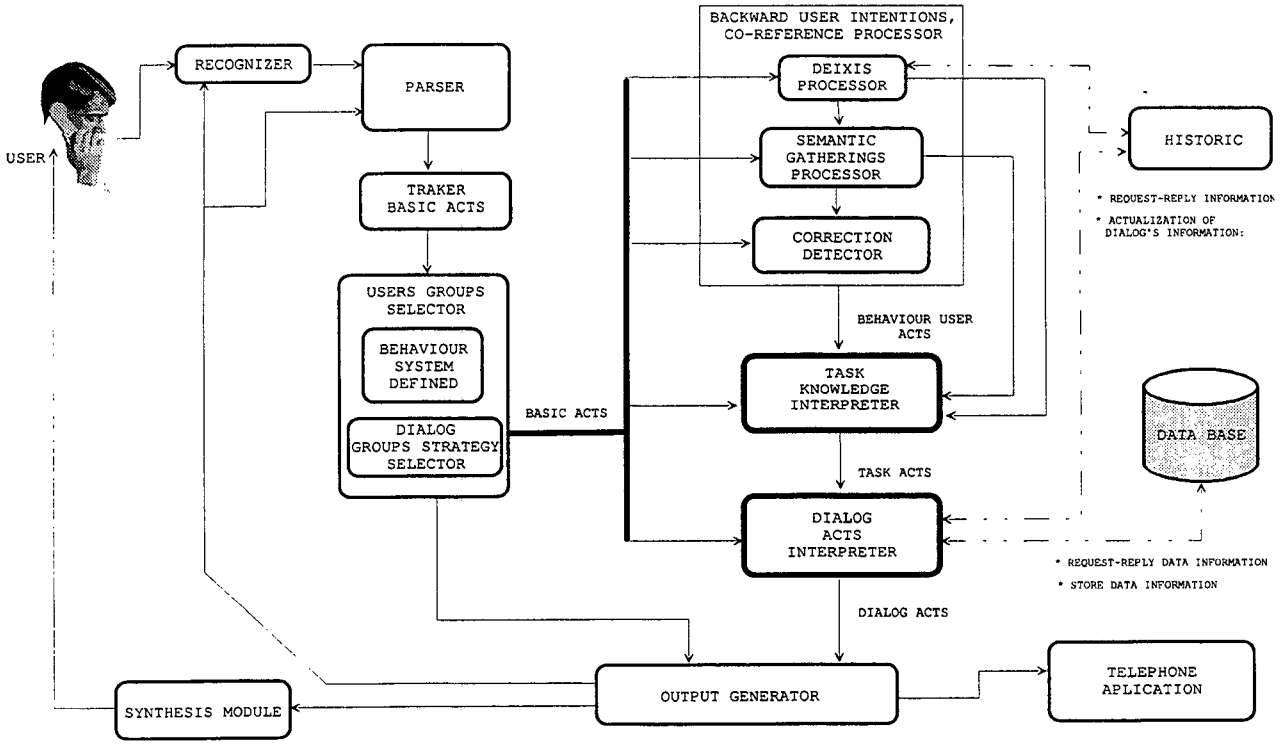

Figure 1: Modules of Robust and Flexible Mixed-Iniciative Dialogue

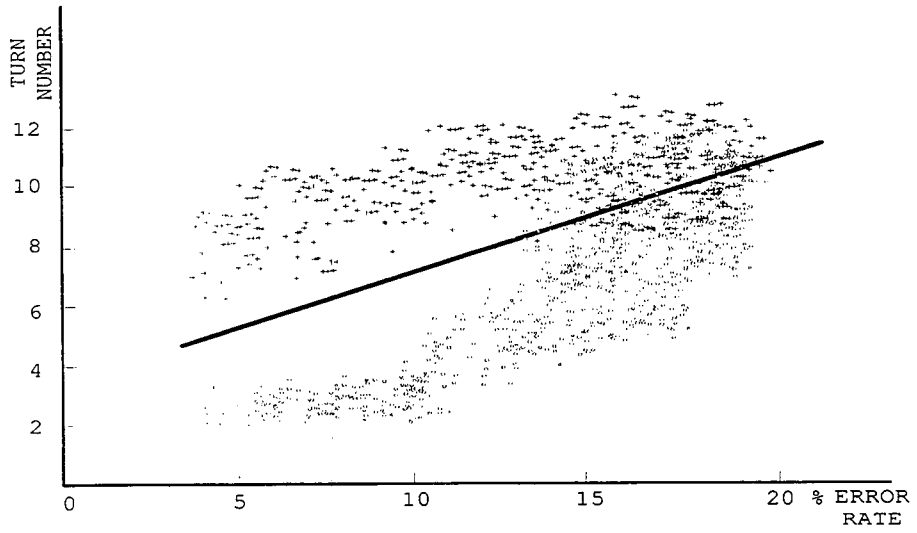

Figure 2: User clasification