User Modeling For Spoken Dialogue System Evaluation

Wieland Eckert, Esther Levin, Roberto Pieraccini
180 Park Avenue
Florham Park, NJ, 07932

Abstract - Automatic speech dialogue systems are becoming common. In order to assess their performance, a large sample of real dialogues has to be collected and evaluated. This process is expensive, labor intensive, and prone to errors. To alleviate this situation we propose a user simulation to conduct dialogues with the system under investigation. Using stochastic modeling of real users we can both debug and evaluate a speech dialogue system while it is still in the lab, thus substantially reducing the amount of field testing with real users.

1 Introduction

Recent literature shows an increasing number of speech dialogue systems being implemented and used in the field [1, 2, 3, 4, 5, 6]. However, the development of such dialogue systems (especially the dialogue manager) is still considered art rather than an engineering task. While the optimality criteria of a low level component like the speech recognizer are obvious (i.e. generate an accurate transliteration of a speech signal) there exists no "ideal behavior" for the high level dialogue control module. Therefore all evaluation is still left to human judgment and personal taste. Another obstacle for the design of dialogue control strategies is the need for large dialogue corpora. Only with a sufficiently large sample of dialogues can any evaluation be performed. Unfortunately, these corpora can not be static (as e.g. the speech signal & transliteration corpora used for recognizer development) since different dialogue strategies result in very different dialogues.

Any evaluation and optimization on the dialogue level would greatly profit from a standardized automatic evaluation methodology where the adverse effects of subjective judgment could be minimized [7, 8, 9, 10]. In this paper we propose using a simulated user for an automatic assessment of a spoken dialogue system. With user simulations we gain the following advantages:

- automatic evaluation of a large number of dialogues can be performed without expensive manual investigation,
- due to reduced manual work the results are more consistent and less prone to errors,

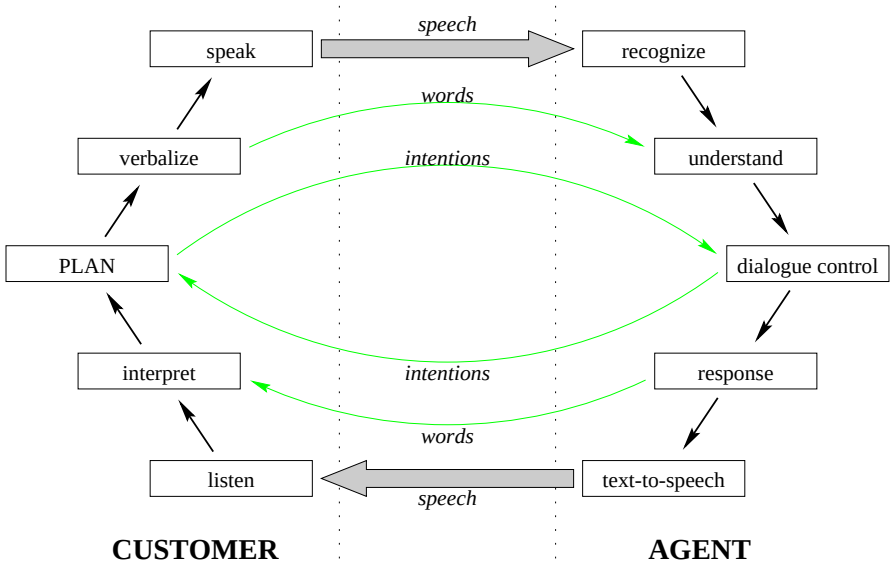

Figure 1: Conversational interaction between user and agent.

- characteristics of different user populations can be modeled easily,
- the same user model can be used for comparative evaluation of different (competing) dialogue systems, and
- the simulations provide a mechanism for employing automatic optimization techniques for dialogue strategies, such as reinforcement learning (described in a companion paper in these proceedings [11]).

A substantial amount of guesswork can be eliminated by employing standard procedures. In the following sections we describe our approach toward automatic evaluation of dialogue systems. First we show the different possible levels of interactions, then we show an outline of our stochastic user simulation, and finally we report experiments and results in the ATIS domain.

2 Intention Based Interactions

Speech dialogue can be separated into several layers of interactions. Independent modules perform the transfers between these layers. Figure 1 shows this separation and the modules associated with each transformation step. Note that the dialogue is depicted as a closed loop system where every module has a specific transfer function.

For the matter of speech dialogue between a customer and an agent we consider three different levels of interaction: speech signals, word sequences,

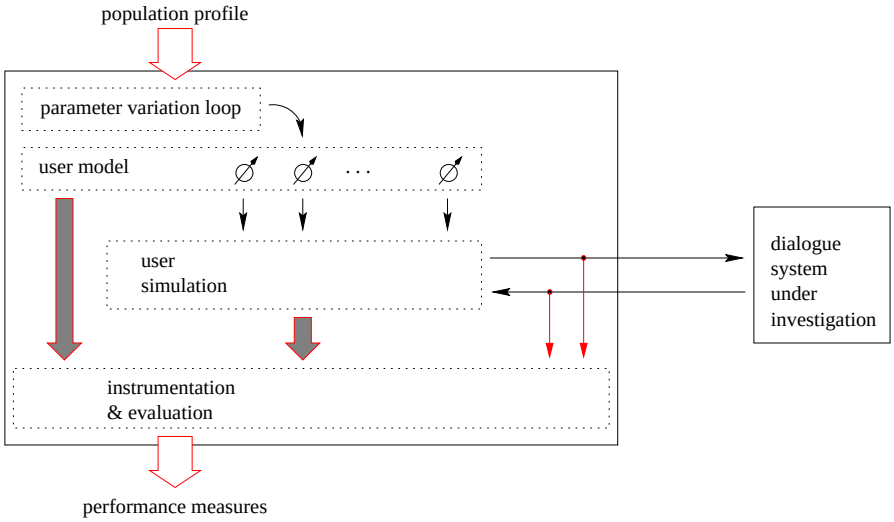

Figure 2: Structure of the user-system interaction utilizing different user profiles.

and intentions. Speech is carried by acoustic signals that can be transported, for instance, via telephone lines. Word sequences represent the next level reducing the variability of speech production to clean text, e.g. written sentences. Intentions can not be observed directly but can be described in terms of speech acts [12] or dialogue acts [13]. In a simplified view they represent the actual information whereas the other two levels can be seen as transport mechanisms only.

We have implemented an automatic agent that simulates the customer of Figure 1 and connected it to an already existing dialogue system. In our current version of the simulated interactive user we started with the intentional level of interaction. This means that there is no room for misunderstanding (which can eventually be modeled as imperfections in the transmission channel). Interaction on the level of intentions requires intimate knowledge of the semantic representation scheme used in the dialogue system. Subsequent transition toward the word sequence level or even the level of speech signals will make our approach more portable since these interfaces are inherently easier.

We added an evaluation environment that controls sessions between the user model and the dialogue system, that instruments the user model, counts, and does rudimentary evaluation of dialogues. It is depicted in Figure 2.

In our first implementation we use interactions based on intentions. However, the evaluation environment is capable of managing dialogues using all kinds of interactions. We plan to go a step further and evaluate dialogues that are based on exchanging word sequences. On the world level we can, for

instance, simulate the effect of different word recognition rates or even characteristics of typical misrecognitions made by a particular recognizer. This adds another dimension to our evaluation space.

3 Stochastic Modeling of User Behavior

Interaction between humans and machines is characterized by a particular degree of freedom of interaction: humans can always take the initiative and change the direction of a goal directed dialogue. We assume that this variability can be modeled by probability distributions. Furthermore we assume that for a fixed population of users these distributions are conditional only on the previous system responses and a state description of the user. For the moment we consider only the single preceding system response. Thus a user response is modeled by the conditional probability

$$p_{ij} = p(U_t = I_i | S_{t-1} = I_j) \tag{1}$$

of replying with a response I_i when receiving the stimulus I_j from the system (I_i and I_j denote sets or sequences of intentions). We assume that the process is time invariant, i.e. these probabilities do not depend on the absolute value of t.

User initiative is modeled by the probability $p(U_t = I_i | S_{t-1} = \epsilon)$ of presenting I_i in response to an open ended question (i.e. there was no request for a particular value). Using this technique we obtain a homogeneous representation mechanism. Obviously, all probabilities can be estimated easily from an annotated corpus of human–machine dialogues.

For the moment we restrict our model to decisions based on a history of length one, i.e. the user response depends only on the preceding system turn. Using the same mechanism we can incorporate longer dialogue contexts by calculating the probabilities

$$p_{ij\,kl\ldots} = p(U_t = I_i | S_{t-1} = I_j, U_{t-2} = I_k, S_{t-3} = I_l, \ldots) \tag{2}$$

which is closely related to estimating language models for recognition tasks. Because of sparse training data we limited our current user model to a memory depth of one turn.

Additional parameters in our user modeling approach deal with conversational customs, like *After a dialogue lasts X turns a user just hangs up unhappily* or *In the initial utterance the user gives Y pieces of information without being asked for them.* Other parameters are the probability for an overinformative response, or the probability to go into a subdialogue. Again, these parameters are modeled by (normal) densities which are specified for a population.

While investigating the transcriptions in the ATIS corpus we realized that the corpus is not well suited to estimate these conditional probabilities. In the ATIS corpus all information is usually given in the initial user utterance (class

User Model	Informal Description
reference	a first try to construct a "reasonable" behavior; all probabilities selected according to common sense
patient	a reference user with nearly infinite patience that would hang up the phone after 99 turns; all dialogues are expected to lead to success
submissive	questions will be answered, but no additional information will be volunteered
experienced	much more over-informative than the reference user, gives information on her own, with slightly higher patience

Table 1: Characteristics of our experimental user models.

Criteria		User Model			
		ref	*pat*	*subm*	*exp*
avg. length of dialogue (turns)	μ	4.25	31.24	8.87	5.10
	σ	2.32	42.94	7.05	3.33
dialogue success rate		0.65	0.70	0.73	0.56

Table 2: Overview of our evaluation results. The four different user models are described in Table 1.

A) and there are rarely any followup utterances necessary (class B). Thus we had to handcraft some of the probabilities but have been inspired by the characteristics of the corpus and by common sense[1]. We have defined four sets of user characteristics resembling different populations. It seems reasonable to explore the system's capabilities in different specialized situations in order to obtain more detailed analyses of different cases.

4 Experiments & Results

We have implemented a user model for the airline traffic information task (ATIS). This simulated user was connected to our AMICA system [5]. In order to get a feeling for different user interactions we implemented several parameter sets which resemble different populations of users. An outline of their characteristics is shown in Table 1. We ran simulations of 10,000 dialogues for each model and compared the evaluation results. Table 2 shows a summary of our evaluation. For this evaluation a dialogue is considered successful when the user is given information about a flight connection.

The most prominent finding is that it is actually possible to identify dialogue shortcomings by investigating compressed statistics only. For instance by looking at the distributions of the dialogue length we found several bugs in our dialogue system that degraded the performance substantially. By examin-

[1] We will have a better foundation for these estimates when actual dialogue corpora are available.

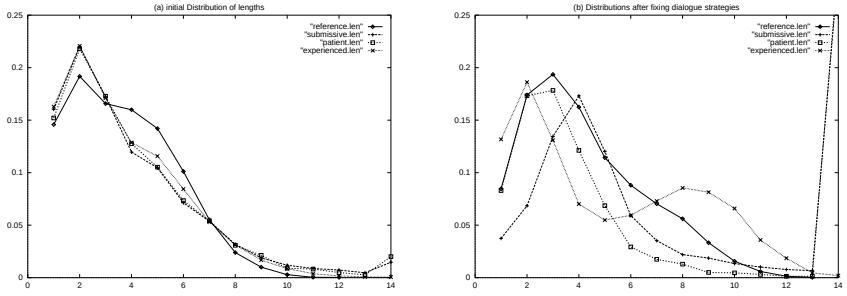

Figure 3: Distribution of the length of the dialogues (in turns).

ing the dialogues of the "patient" user we also found that a different dialogue strategy is adequate: it is advised that the system should close down the connection (after announcing unsuccessful completion) when there is no progress in the overall dialogue for many turns. After integrating these changes into our AMICA system we observed a better overall performance.

The plots in Figure 3 (a) show the distribution of the dialogue length after running the above described experiments. We observe an unusually high number of very short dialogues of three user turns or less. These were caused by a severe bug in the system's dialogue strategy. After fixing this bug we obtained the plots shown in Figure 3 (b) which do not show these extremely short dialogues.

Another irregularity which is still present in Figure 3 (b) is the large amount of long dialogues (14 user turns and more) for the *patient* and *submissive* user models. Examining an example dialogue showed that these dialogues follow a stereotypical interaction where the system tells the user that no information for a particular (impossible) connection is available, but in this population model the user keeps insisting on that flight. A modified dialogue strategy where the system terminates the dialogue is perfectly feasible.

Even though we have not yet altered the dialogue strategy to terminate such unsuccessful dialogues, the impact of this kind of strategy modification will be huge. A comparison of the two plots in Figure 3 shows substantial differences between them after small bug fix. Changes in strategy will have a much larger effect.

It is especially noteworthy that these deficits in dialogue strategies have been found by examining the plots only. Our approach of automatically conducted dialogues between a simulated user and an existing dialogue system enables us to find these shortcomings much more easily than by manual examination of all dialogues. Furthermore, we can test different strategies and decide on their utility.

Additionally, we realized that the parameters of the user profile do not need to be highly accurate. Due to the stochastic process that is superimposed on the values of our model, slight variations of model parameters typically lead to just slight variations of the evaluation result. This property is very

convenient since many problems of estimating model parameters (sparse data, smoothing, etc.) disappear.

5 Further Work

There is work underway to employ automatic learning techniques for the design and optimization of dialogue systems. Clearly, it is necessary to run a large number of dialogues for a learning system and for practical reasons these dialogues can not be collected with human users. Simulation of user interaction is indispensable for this endeavor [11].

Furthermore, we will be engaged in the development and application of automatic evaluation criteria [10]. We want to provide quantitative measures for the "quality" of dialogues with as few manual operations as possible.

6 Summary

Automatic evaluation of speech dialogue systems is accomplished by conducting dialogues with a simulated user. User behavior is specified in terms of conditional probabilities resulting in a stochastic user model. Our approach permits off-line test and evaluation in an automated fashion, thus reducing the necessary number of dialogues to be conducted with real users. Automatic evaluation enables an objective comparison of different dialogue systems, or different strategies in the same system. Cumbersome manual work is greatly reduced by applying our automatic evaluation procedure. However, we believe that tests with human users are still vital for verifying the simulation models.

We have applied this framework to an existing dialogue system for the ATIS task. We estimated the parameters of our user profiles by analyzing the available ATIS corpus and completed our user model manually where the corpus did not provide any examples. Our evaluation environment conducted 10 000 dialogues between the user simulation and the existing dialogue system. By analyzing the statistical data we found several shortcomings and bugs in the dialogue system, both in simple oversights and more serious difficulties in the dialogue strategy. Automatic evaluation has proven to be a useful approach for the assessment of a dialogue system.

References

[1] Norman M. Fraser and Paul Dalsgaard. Spoken dialogue systems: A European perspective. In ISSD 96 [14], pages 25–36.

[2] A. L. Gorin, B. A. Parker, R. M. Sachs, and J. G. Wilpon. How may I help you? In IVTTA 96 [15], pages 57–60.

[3] L. F. Lamel, J. L. Gauvain, S. K. Bennacef, L. Devillers, S. Foukia, J. J. Gangolf, and S. Rosset. Field trials of a telephone service for rail travel information. In IVTTA 96 [15], pages 111–116.

[4] A. Kellner, B. Rueber, and F. Seide. A voice-controlled automatic telephone switchboard and directory information system. In IVTTA 96 [15], pages 117–120.

[5] Roberto Pieraccini, Esther Levin, and Wieland Eckert. AMICA: The AT&T Mixed Initiative Conversational Architecture. In *Proc. European Conf. on Speech Communication and Technology*, pages 1875–1878, Rhodes, Greece, September 1997.

[6] M. D. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget, and J. Simonin. Effective human-computer cooperative spoken dialogue: The AGS demonstrator. In ISSD 96 [14], pages 169–172.

[7] Andrew Simpson and Norman Fraser. Black Box and Glass Box Evaluation of the SUNDIAL System. In *Proc. European Conf. on Speech Communication and Technology*, pages 1423–1426, Berlin, Germany, September 1993.

[8] Morena Danieli and Elisabetta Gerbino. Metrics for Evaluating Dialogue Strategies in a Spoken Language System. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39, 1995.

[9] L. Hirschman and H. Thompson. Overview of evaluation in speech and natural language processing. In R. Cole, editor, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge, 1996. to appear.

[10] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proc. Conf. of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain, 1997.

[11] Esther Levin, Roberto Pieraccini, and Wieland Eckert. A Stochastic Model of Computer-Human Interaction for Learning Dialogue Strategies. In *IEEE ASRU Workshop*, Santa Barbara, 1997. (this issue).

[12] J. R. Searle. *Speech acts. An essay in the philospphy of language.* University Press, Cambridge, 1969.

[13] H. C. Bunt. Rules for the interpretation, evaluation and generation of dialogue acts. In *IPO annual progress report 16*, pages 99–107. Technische Universiteit, Eindhoven, 1981.

[14] Acoustical Scociety of Japan. *Proceedings International Symposium on Spoken Dialogues*, Philadelphia, October 1996.

[15] IEEE Communication Scociety. *Proceedings of the IEEE Third Workshop on Interactive Voice Technology for Telecommunications Applications*, Basking Ridge, September 1996.