Automatic Evaluation of Spoken Dialogue Systems

Wieland Eckert, Esther Levin, Roberto Pieraccini
AT&T Labs Research
180 Park Avenue
Florham Park, NJ, 07932
E-mail: eckert@research.att.com

Abstract

We advocate an objective evaluation methodology for the automated evaluation of spoken dialogue systems that eliminates manual interaction and reduces annotation errors and personal bias. The evaluation is performed by observing interactions between the system and a simulated user. We argue that user simulation is an inexpensive and feasible method for optimizing a dialogue system in the lab. Using a simulated user we can conduct dialogues until the performance measure reaches a predetermined confidence level. A simulated user not only exercises the dialogue system and points out defects, it also helps predict the success of a modified dialogue strategy.

1 Introduction

An increasing number of speech dialogue systems are being deployed (Fraser and Dalsgaard 1996; Gorin et al. 1996; Lamel et al. 1996; Kellner et al. 1996; Pieraccini et al. 1997; Sadek et al. 1996). These systems deal with information retrieval tasks using unconstrained speech. However, there are no commonly agreed methods to evaluate a system, or to compare the performance of competing systems for the same application. Several approaches have been proposed (Simpson and Fraser 1993; Danieli and Gerbino 1995; Hirschman and Thompson 1996; Walker et al. 1997), but the current literature shows that none of them has been commonly adopted by system developers. Still, everyone reports performance measures using different metrics. Recent activities of the EAGLES groups (Gibbon et al. 1998) have not resolved this problem.

We tackle some of these difficulties by proposing an *automatic evaluation framework* that uses a *simulated user* to interact with the dialogue system. Using a simulated user that interacts with the system, we obtain results that are unbiased, accurate, and inexpensive. Evaluation is done automatically and therefore quickly, and with less effort than manual evaluation. Comparison of different dialogue systems (or different strategies within the same system) is simple and conclusive.

2 Dialogue Processing as a Feedback System

We adopt the view that a dialogue partner can be modeled as a feedback system. It is quite common to decompose this dialogue system into several modules which resemble the different fields of research involved. Typically, we deal with speech recognition, natural language understanding, dialogue control, language generation, and text to speech synthesis modules. Figure 1 illustrates this approach of specialized modules. Furthermore, we assume that both participants are treated equally, i.e. *both* are described by interacting modules. Feedback to one dialogue partner is provided by the other participant. Each of them can be seen as a system that follows their particular goal and receives feedback which, in turn, influences the goals and strategies.

An idealized interaction would involve an error-free communication channel: Speech recognition would be exactly the inverse operation of speech production and information is neither lost nor

speak · speech · recognize

words

verbalize · understand

intentions

PLAN · dialogue control

interpret · response

intentions

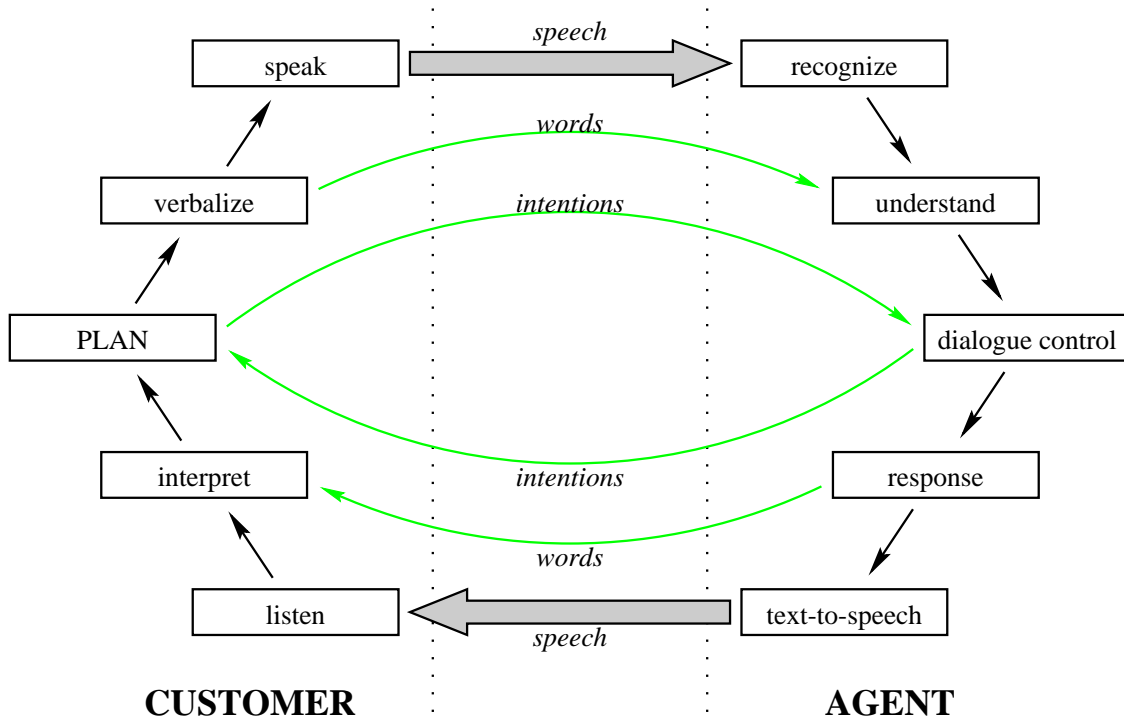listen · words · text-to-speech

speech

CUSTOMER · **AGENT**

Figure 1: Dialogue interaction can be modeled using several different levels: speech signals, sentences, and intentions.

added. This reduction changes the view of dialogue interactions from exchanging acoustic speech signals to communicating via text (word sequences). In yet another step of simplification we may want to assume ideal understanding capabilities. Then we can model dialogues by means of exchanging information on the level of intentions (also known as *dialogue acts*). We began our research on this level of exchanging intentions. Having understood the benefits and limitations of an idealized interaction we started to extend our framework to deal with the understanding module as well. An example dialogue is shown in Figure 2. Corresponding intentional descriptions[1] in Figure 3 are only given to illustrate the data exchange mechanism.

It is clear, that *both* dialogue partners are

- **responsive**, they consider the current input for generating their output;

- **adaptive**, they change their goals (global strategies) or their methods (local strategies) according to the dialogue history;

- **goal driven**, a customer wants some information, an agent wants a happy customer;

- **cost conscious**, they use some measure of the "cost" or "reward" of different alternatives, usually described by the overall satisfaction of the speaker.

The assessment of the quality of a dialogue system needs to eliminate as many variabilities as possible. A most general approach is to only observe the external behavior (black box evaluation) without detailed knowledge about internal proceedings. The only method to gain insight about a system without knowing internals is by exploration (probing the behavior), i.e. conducting dialogues with the system. White box evaluation would need to describe and evaluate internal processing. Details about internal procedures are not really helpful for comparative evaluation of different dialogue systems. Furthermore, any comparison between different dialogue systems

---

[1] In this paper we refrain from taking a position about "good" or "bad" formalisms or semantic representations.

System:	This is the AT&T CHRONUS Air Travel Information System. How may I help you?
User:	I need a flight from Philadelphia with T W A.
System:	Can you please tell me your destination airport?
User:	To Oakland
System:	There are no flights from Philadelphia to Oakland served by Trans World Airlines. Instead there are flights served by American Airlines, Delta Air Lines, United Airlines and US Air. Do you have any preference?
User:	either one is fine
System:	Oakland is served by 2 airports: Metropolitan Oakland International and San Francisco International. Please choose one.
User:	to Oakland
System:	Oakland is served by 2 airports: Metropolitan Oakland International and San Francisco International. Please choose one.
User:	Oakland airport
System:	Flight UA355/UA673 leaves at 5 53 PM and arrives at 10 05 PM. Flight UA131/UA551 leaves at 9 05 AM and arrives at 1 39 PM. Other questions?
User:	no thanks

Figure 2: Example dialogue in the domain of flight information inquiries.

must also consider the user populations that were used to evaluate these systems. Objective comparison is possible when the same population is used for both systems. Thus, any systematic bias is minimized.

3 Automatic Evaluation

We advocate an objective description of a dialogue system's capabilities. Ideally, one would like to specify a *characteristic function* or *transfer function* of this system (cf. control theory). While we usually don't know the transfer function of a complex dialogue system, we can get a glimpse of it by exploration. This is exactly what a corpus of dialogues shows: the results of applying the system's transfer function to a distribution of inputs. An external observation mechanism can record and extract features of the unknown transfer function.

Since evaluation is based on interactions between the system and users, we have to make sure that these interactions are typical of the projected use of the system, i.e. that the actual operating point of the system is observed. As in all experimental setups, we face the dilemma of considering the user's ability to "learn the ropes", i.e. to adapt to the system's strategies. Some evaluation methodologies ignore the user's adaptation by considering only novice or expert users. Using an automatic evaluation approach we can specify characteristics of a *typical* user population that contain various different behavior patterns, ensuring a proper mix of novice, intermediate, and expert users.

Figure 4 shows a block diagram of our evaluation environment. We observe the information exchange between an existing dialogue system and a simulated user. According to a population profile a particular set of values of simulation parameters is picked and used to drive the user simulation. This simulation is terminated when a predetermined number of dialogues is conducted. Another useful termination criterion is the stability of the evaluation results, e.g. when the performance measure reaches a particular confidence level.

Definition of a single, global *performance* figure is not trivial since it may vary depending on the customer's needs. We propose to use a compound quality measure $q(d)$ of a dialogue d that assigns a single value to a dialogue. This compound measure might consist of several components that are weighted according to the customer's demands. For our experiments we used a compound

```
S:    ANNOUNCE: OPEN
U:    SUBJECT: FLIGHT
      ORIGIN_CITY: PPHL
      AIRLINE: TW
S:    REQUEST_CONSTRAINT: DESTINATION
U:    DEST_CITY: OOAK
S:    ANNOUNCE: FAIL
      RELAXED_CONSTRAINT: AIRLINE
      CHOICE_TOKEN: AIRLINE
      ORIGIN_CITY: PPHL
      DEST_CITY: OOAK
      AIRLINE: TW
      DATA: AIRLINE: {AA | DL | UA | US}
U:    RELEASE: AIRLINE
S:    CHOICE_TOKEN: DEST_AIRP
      DEST_CITY: OOAK
      NDATA: 2
      DATA: DEST_AIRP: {OAK | SFO}
U:    DEST_CITY: OOAK
S:    CHOICE_TOKEN: DEST_AIRP
      DEST_CITY: OOAK
      NDATA: 2
      DATA: DEST_AIRP: {OAK | SFO}
U:    DEST_AIRP: OAK
S:    DATA: ORIGIN_TIME:1753 DEST_TIME:2205 FLIGHT_NUMBER:UA355/UA673
      DATA: ORIGIN_TIME:905 DEST_TIME:1339 FLIGHT_NUMBER:UA131/UA551
      YESNO: QUESTION
U:    DUMMY: NO
```

Figure 3: Intentions of the dialogue in Figure 2. The representation consists of (sequences of) attribute value pairs that describe the information content conveyed in the utterance.

measure that is a weighted sum of simple components c_i,

$$(1) \qquad q(d) = \sum_i k_i c_i(d)$$

where the components c_i specify elementary cost functions that can be determined easily for a particular dialogue. Examples for costs are the dialogue length in turns, cost for database access, or penalty for not replying with an answer. While it is still not obvious how to specify the weights k_i, a reasonable approach shown in (Walker et al. 1997) is to select predictor coefficients that have found most predictive for an overall user satisfaction.

The overall performance measure $perf(D, U)$ of a dialogue system D interacting with a user population U is then the expected value of the quality function $q(d)$. It is calculated by weighting the performance for each possible dialogue d with the probability that it occurs, and can be approximated by evaluating this integral over a small, representative corpus C.

$$(2) \qquad perf(D, U) \quad = \quad \mathcal{E}_d\{q(d)\}$$

$$(3) \qquad \qquad\qquad = \quad \int_d P(x)\, q(x)\, dx$$

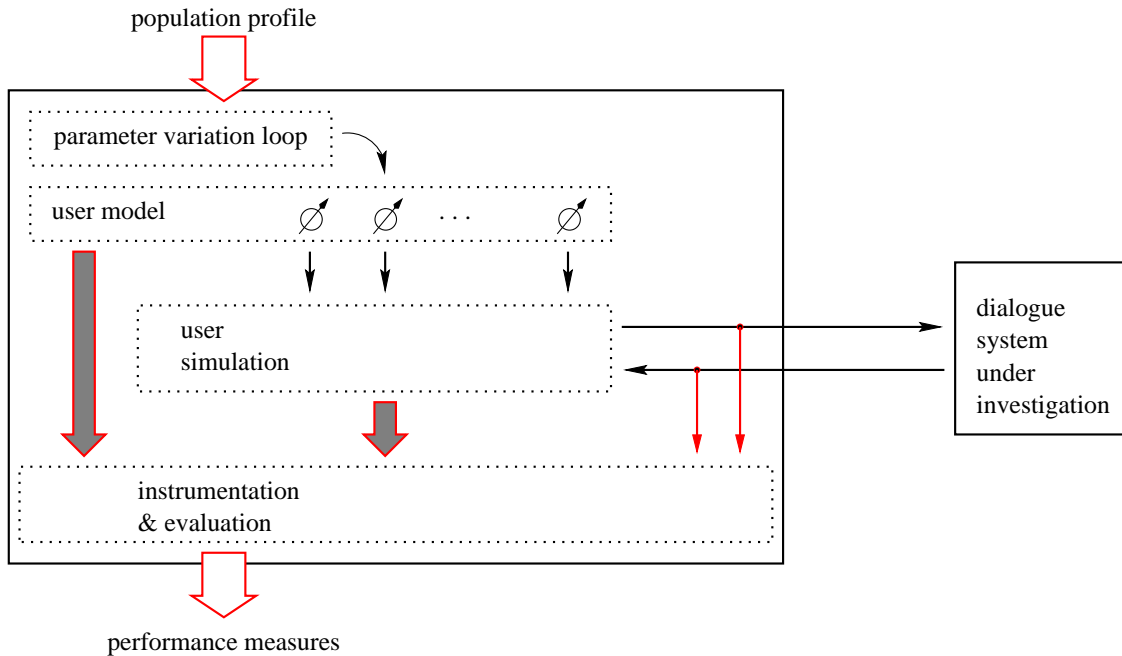$$(4) \qquad \qquad\qquad \approx \quad \sum_{d \in C} \frac{1}{|C|}\, q(d)$$

Figure 4: Block diagram of our evaluation environment.

All corpus based evaluation methodologies have to use the latter approximation and they have to rely on C being representative. In fact, the approximation $P(d) \approx 1/|C|$ is coarse and can only hold for large corpora with duplicated entries. Otherwise one has the added inconvenience of estimating $P(d)$.

Note, that a dialogue system does not allow for arbitrary dialogues, it follows a strategy. In general, a dialogue system has the property of stabilizing a dialogue, i.e. after any unexpected or inintelligible user response the system has a strategy dictating how to continue in order to reach the dialogue goal. The same holds for the user's response. Both are caused by the fact that the dialogue partners want to reach their respective goals. Cooperative dialogue partners have even fewer opportunities to diverge since their goals are correlated. This means that for cooperative behavior (of both user and system), the approximation (4) will converge much faster than for non-cooperative behavior. In other words: the evaluation of stereotypical (i.e. uninteresting, dumb) dialogues is easier than the assessment of diverse (i.e. interesting, intelligent) dialogue behavior.

4 User Simulation

The automatic evaluation approach outlined in the previous section can be applied to any existing corpus of dialogues. However, serious limitations are, that

- a corpus must exist, i.e. data was collected during user trials;

- the corpus must be representative, i.e. trials with users of the projected population, not the designers of the system;

- the corpus must be large enough to obtain significant results.

Additionally, for every slight change in the system's dialogue strategy a new data collection is necessary, because even with only a different prompt users may respond in a different way, resulting in a big change of the dialogue continuation.

We present a new approach for dealing with these problems: We employ a *simulated user* that interacts with the dialogue system. Simulation overcomes all the above mentioned limitations of

a static corpus. We can conduct an arbitrary number of dialogues, and for a modified dialogue system we can rerun the simulation easily. We can imagine having an infinite corpus of dialogues that is generated on demand.

There is no doubt that adequate simulation can accurately represent all the properties of the underlying sample the simulation model is based on.[2] In order to generate "interesting" dialogues we want to have some variety, not stereotypical behavior, i.e. we need nondeterministic models. Some generalization is obtained when dialogues are described in terms of their statistical properties, which also ensures consistent dialogue behavior with some predetermined number of outliers. This is the reason why we do not want to employ rule-based simulation models. Rule-based approaches are, in general, used to model the ideal behavior of a system. In this context we need, however, a model that is capable of representing variations from ideal behavior. Actually, we need a model that describes *real users* as accurately as possible in order to generate reliable performance figures.

The main problem is to find a sound, stochastic representation of dialogue behavior. Dialogues can be described at several levels of abstraction. Most commonly used are the levels of speech (acoustic signals, utterances), syntax (lexical, word sequences, sentences), and intentions (semantics, dialogue acts). These levels coincide with the modularization of most current dialogue systems containing a speech recognizer, natural language parser, and a dialogue manager component, cf. Figure 1. If we adopt a stochastic simulation of user responses, we might want to factor the resulting probabilities into

$$(5) \qquad \begin{aligned} P(utterance) \quad &= \quad P(acoustics|syntax, semantics) \\ &\quad \cdot\, P(syntax|semantics) \\ &\quad \cdot\, P(semantics) \end{aligned}$$

for any utterance, assuming independence of these description levels. Furthermore, we have to introduce conditions on the dialogue context

$$(6) \qquad P(utt_{user,t}|utt_{system,t-1}, utt_{user,t-2}, utt_{system,t-3}, \ldots)$$

that represent the *dialogue strategy*. This representation becomes tractable when we make assumptions about conditional independence of representational levels, and limitations of dialogue context to a length of n turns

$$(7) \qquad P(utt_{user,t}|utt_{system,t-1}, \ldots, utt_{\ldots,t-n})$$

The length n of previous utterances considered for the current response might be considered the user's short term memory.

In our implementation we limit the context length to $n = 1$, i.e. we make a user response dependent only on the previous system utterance. Using the descriptional level of intentions, we have to determine the bigrams

$$(8) \qquad p_{ij} = P(utt_{user,t} = I_i|utt_{system,t-1} = I_j)$$

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. This bigram model can be extended canonically to cover longer context lengths.

Additional parameters in our user modeling approach deal with conversational customs, like After a dialogue lasts more than 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 of an over-informative response, or the probability of going into a sub-dialogue. Again, these parameters are modeled by (normal) densities which are specified for a population.

---

[2] Proof: one can *store* the whole sample and show *exact* behavior.

Using stochastically generated simulated user interactions we can run a large number of dialogues and reach results that are significant according to a predetermined confidence level. Due to the nature of stochastic modeling we can be sure that all results will eventually converge, and we can run experiments until a predefined confidence level is reached.

Note, that in equation (3) we consider all possible dialogues. Since the dialogue system and the user both constrain the set of dialogues, we can rewrite this equations and calculate the performance of a dialogue system by considering all possible user populations, i.e. we calculate the expected value

$$(9) \qquad perf_D = \mathcal{E}_U\{perf(D, U)\}$$

of the system D conducting a dialogue d with any user u of any population U. Without loss of generality, we can actually consider each user having only a single dialogue with the system when we consider a simple renaming of any subsequent dialogue of the same individual. This assumption collapses all possible user populations into a single, representative population. User's learning experiences are not hurt since this is already represented in the statistics of the population. For the matter of evaluation a dialogue system it does not matter if a subsequent dialogue is conducted with the same, now more experienced individual, or with a different one who is better accustomed to this task in the first place.

5 Obtaining Simulation Models

Ideally, one would extract the probabilities of these stochastic models from a corpus of human-machine dialogues. Learning algorithms that are well understood, e.g. for the extraction of language models for speech recognizers, can also be applied to estimate the parameters of our user models. However, we face the problem that a given corpus might be too small for reliable estimates of the model parameters. Either we have to work with bad estimates, or reduce the model complexity. Simpler models can be obtained, for instance, by reducing the context length n, or by making additional independence assumptions. The trade off is either having a sophisticated model using unreliable parameters, or a simplified model with highly reliable parameters. This dilemma is well known.

The worst case for obtaining simulation models is having no dialogue corpus at all. This typically happens when a new dialogue application is envisaged and no previous studies are available. In this case, a simple model with unreliable parameters can be built manually by guessing typical user reactions. Even though this model would shift the operating point of the system substantially, a properly designed dialogue system would try to "help this inexperienced user" and respond accordingly, thus stabilizing the dialogue. We see the whole process of evaluating a dialogue system as an incremental process going through several iterations, starting with a manually generated model based on "common sense" when a dialogue corpus is unavailable or too limited.

In our case we have the official ATIS corpus at hand. While investigating this corpus we faced the following problems:

- While the corpus is large enough to estimate language models (i.e. working on word sequences), on the intentional level it is sparse.

- In the ATIS corpus all information is usually given in the initial user utterance (class A) and there are rarely any followup utterances necessary (class B). As a consequence there is a large imbalance of training material for initial and follow-up utterances.

- The characteristics of the ATIS corpus differ from real user dialogues. ATIS dialogues are quite short and contain a lot of information (kind of complete inquiries) in each utterance, whereas our experiments show longer dialogues with less information per utterance.

We concluded that the official ATIS corpus is based on a very different dialogue strategy, and therefore of limited use. Class A sentences are a rich source for estimating parameters for initial utterances, but a large class of probability densities is not represented in this corpus at all, for

instance yes/no-questions or simply requesting the departure time from the user. Thus we had to hand craft some of the probabilities but have been inspired by the characteristics of the corpus and by common sense.[3] While we are not satisfied with this kind of guesswork, we see it as a bootstrap process for obtaining better models. Standard machine learning textbooks (Sutton and Barto 1998, Section 9.2) show how to obtain an improved model of the user while interacting.

Even with only rough estimates of response probabilities we found that these simulated dialogues are perfectly reasonable, i.e. we can easily explain the behavior and draw parallels to real user's behavior. We attribute that to both partners having the property of following their goals. A minor or even a major deviation has no catastrophic effects on the dialogue, although it is still visible in the performance results.

In equation (9) we calculate the performance of a dialogue system dealing with a variety of users. As a dual problem one could calculate the performance of a user population that is exposed to a variety of dialogue systems

$$(10) \qquad perf_U = \mathcal{E}_D \left\{ perf(D, U) \right\}$$

A simulation model S would have to show the same performance as the underlying population U of real users. Thus, a metric for the quality of a simulation model is the mean square error.

$$(11) \qquad E(S, U) = \mathcal{E}_D \left\{ perf(D, S) - perf(D, U) \right\}^2$$

A proper model design and refinement technique would try to minimize this kind of model approximation error. If this error approaches 0, it means that simulated dialogues and interactions with real users are indistinguishable. Another point of view is to consider (10) as the degree of goal correlation between the system and the user. Collaborating agents usually reach a higher performance measure than opposing agents. If we assume collaborating agents, a drop in performance within a dialogue may be caused by misrecognition or misunderstanding. Therefore, a dialogue system may compare a user's actual response with the predicted response to adapt the system's strategy in order to reach better overall performance.

6 Utility

There are several reasons for employing automatic evaluation methods and interactions with simulated users:

- We want an *objective* methodology that eliminates any bias of manual evaluation.

- Automatic evaluation is inexpensive and quick. Large amounts of data can be processed more quickly and reliably than with human intervention.

- Automatic evaluation using simulated users enables comparative evaluation of different dialogue systems, or different strategies of the same system.

- Improvements of a dialogue system can be performed in the lab. Costly trials with real users can be reduced.

The first two items are quite obvious since they follow directly from the automation of labor expensive and tedious manual work.

While there is considerable literature about evaluation of a particular dialogue system, comparative evaluation of two or more systems has not been tackled, yet. Even for the same application (e.g. ATIS) there was no comparison possible since researchers adopted different dialogue strategies, experimented with different user populations, and reported different performance metrics. Using a simulation environment like ours, one can connect to a variety of dialogue systems and run experiments with a consistent population of simulated users.

---

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

This framework allows for documenting the improvement of a dialogue system in the lab. In the following we outline the reasoning behind this method. Let's assume we have an automatic dialogue system D_1 and we collect a corpus $C_U(D_1)$ of dialogues with real users. Based on that corpus a simulation model $S(C_U)$ can be built. Now that we have a user simulation tool and an automatic evaluation environment, we are able to optimize the existing dialogue system according to some given objective performance measure. The resulting system D_2 is better on the simulation model, i.e. $perf(D_2, S) > perf(D_1, S)$. This optimized dialogue system can be exposed to real users and a test corpus $C_U(D_2)$ is collected. We claim that the simulation is useful, iff the performance is not worse than that of the previous system, i.e. $perf(D_2, C_U(D_2)) \geq perf(D_1, C_U(D_1))$. Even though the simulation model is not necessarily *accurate* in the sense that it represents the user population *exactly*, it is still useful for optimizing an existing dialogue system. Most likely this process will be iterated several times, leading to better dialogue systems as well as better simulation models.

Keep in mind that the whole optimization can be done in the lab and only for testing purposes we do have to validate the system using interactions with real users. On the other hand, we can easily test different dialogue strategies in the lab and get an idea about their estimated performance in the field.

7   Our Implementation

We started our experiments in the ATIS domain (Eckert et al. 1997) and on the intentional level, i.e. a simulated user dealing with a dialogue manager in terms of intentions. These experiments essentially exercise the dialogue manager's ability to complete a dialogue, even with unorthodox user behavior. It assumes perfect recognition and perfect understanding. The majority of the dialogues went well, but a number of problems in the system's dialogue strategy were found and fixed. These problems occurred when the stochastic simulation encountered a situation that was not foreseen and never exercised by the developers. Thus, simulation did help debugging and optimizing the dialogue manager.

One defect in our dialogue strategy was, for example, that the system was too polite and never hung up on the user. The system would repeat infinitely that it did not find a solution for the user's particular request, but suggested a relaxation of some parameter, e.g. airline, instead. It is a perfectly reasonable strategy to repeat this information only n times and give up on a stubborn user. We found these defects by looking at the compressed performance statistics of a simulation run. This particular defect showed its footprint in an unusual large amount of dialogues with very many turns. After looking at one of these long dialogues we explained the defect and fixed it easily. We performed more experiments with different user populations (Eckert et al. 1997) in order to exercise the AMICA dialogue manager (Pieraccini et al. 1997). These simulations provided us with valuable hints for improving the strategies of the system.

Next, we went to the syntactic level and sent word sequences to the natural language understanding component which passed on its results to the dialogue manager. Thus, we test the robustness of NLU in combination with dialogue management, assuming perfect speech recognition. An example dialogue of our simulation experiments is shown in Figure 5. Again, we found several shortcomings of our NLU component that did not show up in previous tests with expert users. Even though we only used a very simple, template based text generation method to transform the user's intentions into word sequences, we found a few instances of simulated hesitations and interruptions where the NLU component could not extract all intentions of the user. As a result, these dialogues tend to take more turns than those assuming perfect understanding. Ideally, one would expect identical performance when transmitting intentions or when communicating in word sequences. The task of a robust speech understanding system is to minimize this difference as far as possible.

As a logical consequence one may think of using synthesized speech for the simulated user in order to exercise the complete speech dialogue system with its speech recognizer. While this approach is definitely beneficial to the understanding of interacting using complex media, we think that there are still many open research issues to be solved before we benefit from this kind

of simulation results.

System:	This is the AT&T CHRONUS Air Travel Information System. How may I help you?
User:	Can you get me a flight from Denver to Pittsburgh Airport?
System:	There are too many flights that match your request. I need additional information. Flights from Denver to Greater Pittsburgh International are served by 2 airlines: Delta Air Lines and US Air. Do you have any preference?
User:	The cheapest
System:	At what time do you want to leave?
User:	The earliest flight
System:	Flight DL296/DL842 leaves at 0 55 AM and arrives at 10 AM. Other questions?
User:	`hangup`

Figure 5: Actual Example dialogue that was conducted using our simulated user.

8   Discussion

Evaluation of a (speech) dialogue system is like exploring the unknown. Using a black box approach, we don't have an analytical description of the exact behavior of the dialogue system. Additionally, we can only gather statistical data of our user population. Therefore, we can not calculate the performance of the system using an analytical approach, but a stochastic approach to evaluation dialogue systems is necessary. We have to probe the dialogue system. In order to control as many parameters as possible, we employ a simulated user with stochastically well defined behavior. User simulation and automatic evaluation is the key to enable comparative evaluation of dialogue systems.

Simulation using stochastic models is a well known field. Our original work is the application of stochastic user simulation to evaluate dialogue systems, and the automatic evaluation of the simulation results. Stochastic modeling is an assumption that is made in our approach, but these models adequately represent the underlying corpus.

One of the first evaluation methodologies for speech dialogue systems was described in (Simpson and Fraser 1993). While it intended to enable comparative studies of different dialogue systems, It did not address the problem of diverse user populations. Any experimental result is prone to some bias caused by the user sample picked. In order to normalize results one would have to determine the prior distributions of user characteristics, and either factor out their influence or normalize to some standards. Using a simulated user we provide every examined dialogue system with the same user characteristics, repeatedly. Additionally, experiments with a different user population can be run easily.

Recent work in the field of machine learning techniques led to the development of an automatic, learning dialogue agent (Levin et al. 1997). This learning approach utilizes a cost measure for any action taken and finds an optimal strategy how to interact with a user. While the idea is fascinating, the drawback is that a very large number of dialogues is necessary to learn a good strategy. Obviously, one would not take advantage of human dialogue partners (especially if one might loose customers), so a simulation approach is necessary. We were able to show in (Levin et al. 1997) that user simulation is a useful strategy for learning dialogue strategies.

9   Summary & Further Work

In this paper we presented a new approach toward the automatic evaluation of speech dialogue systems. This approach alleviates two of the dominant difficulties: (1) manual work is greatly reduced by employing an *automatic evaluation process* which leads to less transcription errors and

higher processing speed, and (2) comparative evaluation is possible because of a *simulated user* interacting with the system in a consistent way following stochastic constraints.

We showed that a sound model of user behavior can be deduced using stochastic properties that can be determined from a corpus. Factorization of conditional probabilities into the layered levels of acoustic, syntactic, and semantic knowledge representations gives us a decomposition into tractable sub-problems. We started our experiments on the level of semantic information, i.e. the dialogue system and the user simulation both communicate descriptions of their intentions. Experiments to extend this to the next level of text strings are in progress. We used the ATIS task as a test bed for our experiments and connected this evaluation environment to our AMICA system. Our experiments were successful in the sense that we found (and fixed) several shortcomings of the dialogue system and of the natural language understanding module. Even with very rough approximations of user models we were able to show very reasonable behavior of the user simulation. Additionally, we employed Reinforcement Learning to obtain a dialogue strategy that is optimal with respect to the characteristics of the simulation population.

Future work will tackle the problem of obtaining good simulation models. A very interesting approach is to extract the model while interacting with a real user. This user model would be exploited in idle times to provide the grounds for "what-if" simulations which would improve the system's performance for previously unseen dialogues. One would need to leap from hand crafted dialogue strategies to automatic, learning systems. Finally, there is hope that speech dialogue systems can become intelligent.

References

Danieli, M. and Gerbino, E.: 1995, 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*, pp 34–39

Eckert, W., Levin, E., and Pieraccini, R.: 1997, User Modeling for Spoken Dialogue System Evaluation, in Furui et al. 1997, pp 80–88

Fraser, N. M. and Dalsgaard, P.: 1996, Spoken dialogue systems: A European perspective, in ISSD 96 1996, pp 25–36

Furui, S., Juang, B.-H., and Chou, W. (eds.): 1997, *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, Santa Barbara

Gibbon, D., Moore, R., and Winski, R. (eds.): 1998, *Handbook of Standards and Resources for Spoken Language Systems*, Vol. 3, Spoken Language System Assessment, Walter de Gruyter, Berlin

Gorin, A. L., Parker, B. A., Sachs, R. M., and Wilpon, J. G.: 1996, How may I help you?, in IVTTA 96 1996, pp 57–60

Hirschman, L. and Thompson, H.: 1996, Overview of evaluation in speech and natural language processing, in R. Cole (ed.), *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, Cambridge

ISSD 96: 1996, *Proceedings International Symposium on Spoken Dialogues*, Philadelphia, Acoustical Society of Japan

IVTTA 96: 1996, *Proceedings of the IEEE Third Workshop on Interactive Voice Technology for Telecommunications Applications*, Basking Ridge, IEEE Communication Society

Kellner, A., Rueber, B., and Seide, F.: 1996, A voice-controlled automatic telephone switchboard and directory information system, in IVTTA 96 1996, pp 117–120

Lamel, L. F., Gauvain, J. L., Bennacef, S. K., Devillers, L., Foukia, S., Gangolf, J. J., and Rosset, S.: 1996, Field trials of a telephone service for rail travel information, in IVTTA 96 1996, pp 111–116

Levin, E., Pieraccini, R., and Eckert, W.: 1997, A Stochastic Model of Computer-Human Interaction for Learning Dialogue Strategies, in Furui et al. 1997, pp 72–79

Pieraccini, R., Levin, E., and Eckert, W.: 1997, AMICA: The AT&T Mixed Initiative Conversational Architecture, in *Proc. European Conf. on Speech Communication and Technology*, pp

1875–1878, Rhodes, Greece

Sadek, M. D., Ferrieux, A., Cozannet, A., Bretier, P., Panaget, F., and Simonin, J.: 1996, Effective human-computer cooperative spoken dialogue: The AGS demonstrator, in ISSD 96 1996, pp 169–172

Simpson, A. and Fraser, N.: 1993, Black Box and Glass Box Evaluation of the SUNDIAL System, in *Proc. European Conf. on Speech Communication and Technology*, pp 1423–1426, Berlin, Germany

Sutton, R. and Barto, A.: 1998, *Reinforcement Learning, An Introduction*, MIT Press

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A.: 1997, PARADISE: A Framework for Evaluating Spoken Dialogue Agents, in *Proc. Conf. of the Association for Computational Linguistics*, pp 271–280, Madrid, Spain