PROBABILISTIC SIMULATION OF HUMAN-MACHINE DIALOGUES

Konrad Scheffler, Steve Young

Department of Engineering, Cambridge University, Cambridge, UK
email: {khs22,sjy}@cam.ac.uk

ABSTRACT

The field of spoken dialogue systems has developed rapidly in recent years. However, optimisation, evaluation and rapid development of systems remain problematic. This paper describes a method of producing a probabilistic simulation of mixed initiative dialogue with recognition and understanding errors. Both user behaviour and system errors are modelled using a data-driven approach, and the quality of the simulations are evaluated by comparing them to real human-machine dialogues.

The simulation system can be used to perform rapid evaluations of prototype systems, thus aiding the development process. It is also envisaged that it will be used as a tool for automation of dialogue design.

1. INTRODUCTION

Recent advances in the field of spoken dialogue systems include automatic evaluation by means of dialogue simulation [1, 2, 3]. Building on this work, we have developed a dialogue simulation system that incorporates a model of goal directed user behaviour in mixed initiative dialogues, as well as system recognition/understanding errors. This makes it possible to obtain realistic simulations of dialogues with complex structure (requiring user behaviour that is consistent from one utterance to the next). Since the principles on which the system is based are domain independent, it can be applied to any co-operative, task-oriented dialogue.

By simulating dialogues and measuring performance on the simulated rather than real dialogues, the expense and effort of running tests with real users can be avoided. Repeating the simulations many times during the development process can help developers to optimise various system aspects and generally speed up development. It is also hoped that the system will prove useful as a tool for automatic design of dialogue systems [4, 5, 6].

For evaluating the proposed approach, an existing dialogue system was used. This system provides a telephonic, speech based interface for a banking application, supporting transfer transactions between accounts, enquiries of the caller's account balances and stock quote enquiries. The system is implemented using a finite state dialogue structure. The allowable syntax for speech input is defined for each dialogue state, by means of finite state word networks designed to support natural language, mixed initiative speech input. In this paper, this particular system will be referred to as the *banking application.*

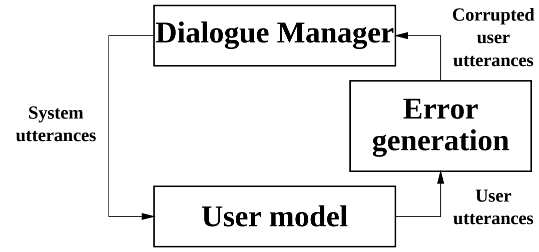

Figure 1: Process of simulating a human-machine dialogue.

2. SIMULATION METHODOLOGY

In this research, we consider systems with dialogue managers that can be described as finite state automata, where the states correspond to the dialogue circumstances. Such systems implement a dialogue strategy that maps states into actions. We are interested in applications that are complex enough to require both loops and confirmation subdialogues in the dialogue structure.

The simulation is performed by combining the dialogue manager with simulation components as in Figure 1.

2.1. Interaction level

The interaction between the system and the user simulation takes place solely at the intention level, rather than the word or the speech signal level [1]. An intention represents the actual information (concepts) that a dialogue participant wants to convey and can be defined as the minimum unit of information that can be expressed independently within a given application.

An example of an intention would be "transfer". This could correspond to any of a large number of possible word strings, such as (in the context of an open "How may I help you?" prompt):

1. Transfer.
2. I'd like to transfer some money.
3. 200 pounds to my checking account, please.

In (3), the "transfer" intention is inferred in spite of the absence of "transfer" or any synonym in the utterance, and occurs along with the intentions "cash amount (200)" and "to the checking account".

The following arguments motivate the approach of working solely on the intention level:

- Constructing a reasonable model of user behaviour at the word level may be infeasible, because of the large range of possible user outputs for even a simple utterance. Data sparsity would make the task of parameter estimation impractical if not impossible.
- Modelling user behaviour at the word (or lower) levels is unnecessary for the purpose of testing the high level design of a dialogue system. Furthermore, by incorporating performance statistics from the lower levels (such as natural language understanding), the resulting system can simulate the performance of the entire dialogue system.
- The developed system is equally applicable to non-speech modalities. The only requirement is that errors made by the input system can be modelled.
- By eliminating the linguistic component, the simulation system becomes largely domain and language independent.

2.2. User model

Eckert, Levin and Pieraccini [1, 2] propose a bigram user model that specifies the probability of each possible user utterance conditioned only on the preceding system utterance. However, such a model takes neither the history of the dialogue (other than the most recent utterance) nor the aims of the user into account, which is insufficient for dialogues where different user utterances are interdependent, or where confirmation subdialogues are required. In such cases a purely probabilistic model results in inconsistencies between user utterances, aimlessness on the part of the user and unnecessary repetition (or respecification with different values) of transaction details that had already been specified.

For this reason, we impose an overall structure on the user's behaviour by constraining the user utterances to be consistent with a predefined *user goal*. In addition, aspects such as information provided by the system for direct or indirect confirmation, feedback on the actions taken by the system, and previous user utterances are taken into account. A probabilistic model of user behaviour is used when the user has more than one option that are consistent with the current goal.

2.3. Error modelling

Human-machine dialogues are contaminated by speech recognition and understanding errors, as well as formulation errors on the part of the human user. For the purpose of modelling, we group the different error types together and consider only the total distortion that takes place between the original intention[1] in the mind of the user and the eventual interpretation arrived at by the system. A probabilistic model for these errors can be created by comparing reference transcriptions with recognition output at the intention level.

Transaction type:	Transfer
Amount:	500
From account:	Savings
To account:	Checking
Balance account:	NA
Stock name list:	NA

Figure 2: Example of a user goal

2.4. Utterance construction

A dialogue is seen as consisting of one or more *transactions* between the user and the system. The different transactions that are supported by the application are defined by means of the transaction type and variables for type-specific details. By filling in the relevant variables, a transaction task can be specified. An example task in the banking application might be to transfer 500 pounds from the savings account to the checking account. These details are used to initialise a goal structure as illustrated in Figure 2, which is then used as a constraint during the simulation.

A transaction is carried out in a sequence of dialogue turns (*utterances*), where each user utterance can be viewed as a sequence of *intentions*. Although intentions are autonomous units, they can only occur in specific structures, which are predefined by the recognition and understanding components of the dialogue system.

During the construction of an utterance, the user may be faced with a series of choices. The estimation of parameters for a probabilistic user model will amount to estimating the probabilities with which, when faced with a given choice, a user chooses the different available options.

Once a "correct" intention has been added to an utterance, the error model is used to generate valid substitutions of these intentions. Insertion and deletion errors are treated as a special case of substitutions. By performing error generation concurrently with the generation of the utterance, construction of syntactically invalid utterances can be avoided.

Errors are generated on individual intentions, taking dialogue context (state) into account, but neglecting the effect of neighbouring intentions. This is done because the amount of data required would make estimation of a fully context specific model impractical.

In order to estimate substitution probabilities it is necessary to align the reference intention sequence with the recognised sequence so that corresponding intentions in the two sequences are matched. This is done by considering only substitutions within restricted groups of intentions. These groups, called *intention groups*, are chosen to correspond with the intentions allowed at specific points in the syntax.

3. EXAMPLE

To demonstrate how the concepts introduced above are used in an actual application, the utterance construction for the banking application is illustrated in Figure 3. A detailed description of the algorithm can be found in [7].

[1] We assume the existence of unambiguous user intentions which are expressible in the semantic language of the system. Intentions that are beyond the system capabilities are classified as "mumbles".

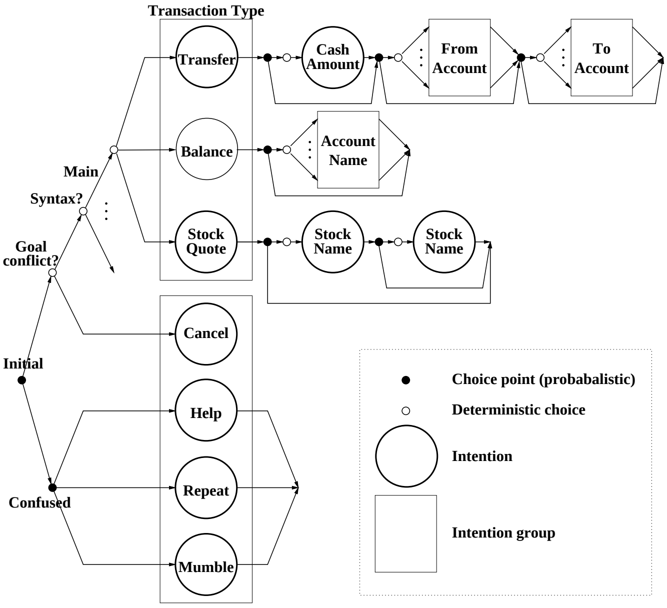

Figure 3: Partial structure for utterance construction in the banking application.

The process starts at a *choice point* (labelled "Initial"), where a choice is made according to the probabilities (stored in the user model) of the available options. The choice in this case determines whether the utterance will be directed towards achieving the current goal, or indicate confusion on the part of the user. The former option leads to a deterministic choice based on whether there is a conflict between the user goal and the inferred system goal, which would cause the "cancel" intention to be added to the utterance. If there is no conflict, the next decision depends on the syntax allowed by the speech recogniser, with each syntax determining the structure of what follows. Shown in the figure are the possible choice sequences for the "Main" syntax, which is used when the user has not yet specified a type for the current transaction. From this point onwards the deterministic choices specify the contents of a particular goal field, while the probabilistic choice points correspond to mixed initiative in the system, where the user can choose which of the possible details to specify and which to withhold. For example, after specifying that the desired transaction type is a transfer, the user may or may not specify any details connected to that transaction in the same utterance. In the case of the intentions "Cash Amount" and "Stock Name", the deterministic choice determines the specific details.

If, instead, the initial choice leads to the "confused" choice point, the user model does not attempt a direct move towards the goal. This corresponds to a situation where, for example, the user has misheard something, is uncertain about what to do next, or says something that is beyond the system capabilities. The choice between these options is again made probabilistically.

The figure also shows how intentions are grouped together in intention groups for substitution purposes. Note that the "cancel" intention belongs to the same intention group as the "confused" intentions. This encodes the fact that these intentions can always be substituted for each other during error generation. The options for each choice from a particular goal field also belong to the same intention group.

4. PARAMETER ESTIMATION

Probabilities for the choice points are estimated by first producing counts of the number of times the choices in question are taken in a set of training data. Once the counts have been obtained, maximum likelihood estimates of the choice point probabilities are calculated:

$$P(O_{ij}|CP_i) = \frac{cnt(O_{ij})}{cnt(CP_i)}, \qquad (1)$$

where CP_i is the ith choice point, O_{ij} is the jth option of the ith choice point, and $cnt(\cdot)$ is the counted number of occurrences.

Similarly, substitution probabilities are estimated by counting how often any given intention in the reference transcription is replaced by the possible substitutes in the recognised version.

Some of the parameters estimated in this way were subject to data sparsity. This problem is treated by generalising those choice points for which the number of instances in the data was below an arbitrary threshold. Instead of using state dependent probabilities for these choice points, data from different dialogue states were pooled together to enable estimation of state independent probabilities.

5. RESULTS

The purpose of the experiments reported here was to ascertain to what extent the simulated dialogues give an accurate picture of the dialogues produced by real users applying the system in the task domain. Completion times (in number of dialogue turns) for various goal types were measured on a corpus of real dialogues, and compared to those obtained by simulation.

5.1. Data corpus

User identification	490
Transfer transactions	243
Balance enquiries	282
Stock quote enquiries	575
Dialogue exit	133

Table 1: Distribution of goal types in the corpus.

The dialogue system on which these experiments are based, was tested by a group containing both system developers and members of the public. The resulting corpus contains 490 dialogues in which at least one user goal was completed, and a total of 4201 parseable utterances. The numbers of different user goals completed are shown in Table 1. The user identification (ID) goal is achieved whenever the user enters a valid pin number and is accepted by the system. The dialogue exit goal is achieved when the dialogue terminates normally, after the system outputs its final greeting prompt.

A subset of the corpus was isolated for testing purposes, with the remainder being used as a training set to estimate the system parameters. The test set contained 140 dialogues in which at least one goal was accomplished, and a total of 1037 parseable utterances. The training set contained 350 dialogues accomplishing at least one goal, and a total of 3164 parseable utterances.

5.2. Parameter estimation

The estimated parameters included a total of 30 probabilities for 9 mixed initiative choice points, some of them using more than one context, and a total of 418 probabilities for 85 context sensitive substitution choice points.

Some of the parameters were found to be very sensitive to context. For example, during specification of a transfer transaction, the "to" account was specified 17% of the time after prompts for the "from" account, but only 5% of the time after prompts for the cash amount. Variations across different types of transaction seemed to be smaller, which indicates that it may be possible to obtain reasonable simulations even when the parameters have been estimated on a slightly different task.

5.3. Dialogue simulation

Three user scenarios were created for the simulation experiments, so that the number of turns taken for all the subgoals in the system could be measured. The scenarios were:

1. **Transfer:** Transfer 500 units from the savings account to the checking account.
2. **Balance:** Find out the balance on the savings account.
3. **Stock quote:** Find out the stock prices for 3 specified stocks (thus three goals are achieved in this scenario).

Goal	Real data	Simulated data
User identification	2.01	1.51
Transfer transactions	7.97	6.20
Balance enquiries	3.38	2.20
Stock quote enquiries	3.14	2.09
Dialogue exit	2.17	1.29

Table 2: Goal achievement times (average number of turns) for real and simulated dialogues.

A set of 1000 simulated dialogues was created for each scenario. The resulting average number of turns for the different goals are given in table 2, along with those obtained for the real data. While the numbers are not identical, the ordering of the different goal types correspond (except for dialogue exit, which is quicker than user identification in the simulated dialogues). It is clear from both data sets that transfer transactions take much longer than simple enquiries.

6. CONCLUSIONS

By introducing goal constrained choices and error modelling at the intention level, realistic simulations were produced for dialogues with nontrivial complexity. The simulated dialogues showed reasonable agreement with real dialogues in terms of average length for individual dialogue tasks.

Goal achievement in the simulations is consistently faster than for real users, by up to one turn per goal. The main reason for this seems to be that real users are considerably less goal directed and more confused about what they want to do than simulated users, especially since many of the callers were taking part in the trial of a new and unfamiliar system. It is not clear to what extent this phenomenon would be observed if the data was gathered during actual use of the system.

The simulations revealed that an unacceptably high failure rate of 4.8% was experienced for transfer transactions. This was due to a high recognition error rate on cash amounts, and the fact that the system strategy causes the wrong amount to be transferred when an amount is misrecognised twice in succession. The latter point is a flaw in the dialogue strategy that escaped the developers' attention until it was highlighted by the simulation. Thus it is clear that the simulation system can be a useful tool for the development process.

The immediate applications of the system include optimisation, evaluation and testing of dialogue systems during the development phase. Apart from these applications, an accurate user simulation makes it feasible to use reinforcement learning techniques to optimise dialogue strategies without requiring large amounts of corpus data [4, 5, 6], which will be a key focus of future work.

7. REFERENCES

[1] W. Eckert, E. Levin, and R. Pieraccini, "User modelling for spoken dialogue system evaluation," *Proc. IEEE ASR Workshop*, 1997.

[2] W. Eckert, E. Levin, and R. Pieraccini, "Automatic evaluation of spoken dialogue systems," Tech. Rep. TR98.9.1, AT&T Labs Research, 1998.

[3] M. Araki and S. Doshita, "Automatic evaluation environment for spoken dialogue systems," in *Dialogue Processing in Spoken Language Systems* (E. Mayer, M. Mast, and S. LuperFoy, eds.), pp. 183–194, Springer, 1997.

[4] E. Levin and R. Pieraccini, "A stochastic model of computer-human interaction for learning dialogue strategies," *Proc. Eurospeech*, pp. 1883–1886, 1997.

[5] E. Levin, R. Pieraccini, and W. Eckert, "Using markov decision process for learning dialogue strategies," *Proc. ICASSP*, 1998.

[6] S. Young, "Probabilistic methods in spoken dialogue systems," *Proceedings of the Royal Society, London*, Sept. 1999.

[7] K. Scheffler and S. Young, "Simulation of human-machine dialogues," Tech. Rep. CUED/F-INFENG/TR 355, Cambridge University Engineering Dept., 1999.