Knowledge Consistent User Simulations for Dialog Systems

Hua Ai [1], *Diane J. Litman*[1,2]

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
[2]Department of Computer Science and Learning and Research Development Center, University of Pittsburgh, PA, USA
hua@cs.pitt.edu, litman@cs.pitt.edu

Abstract

We propose a novel model to simulate user knowledge consistency in tutoring dialogs, where no clear user goal can be defined. We also propose a new evaluation measure of knowledge consistency based on learning curves. We compare our new simulation model to real users as well as to a previously used simulation model. We show that the new model performs similarly to the real students and to the previous model when evaluated on high-level dialog features. The new model outperforms the previous model when measured on knowledge consistency.

Index Terms: spoken dialog, user simulation, evaluation measures, knowledge consistency

1. Introduction

User simulation is increasingly being used in the development of dialog systems. In contrast to experiments with human subjects, which are usually expensive and time consuming, user simulation generates a large corpus of user behaviors in a low-cost and time-efficient manner. Many studies (e.g., [1][2][3]) have verified that simulation models of real user behaviors can be successfully trained from small real corpora. Nevertheless, how well these models can simulate realistic human behaviors and how realistic the models need to be for various tasks are still open questions. On the one hand, it is necessary to build simulation models that can explore some unseen user behaviors to generate a rich training corpus for dialog strategy learning [4]. On the other hand, it is also important to build simulation models that can generate realistic user behaviors for dialog strategy evaluation [5] and user tracing [6].

Many studies in the user simulation literature emphasize that *consistency* is one important feature of realistic user behaviors [2][3][7]. "User goal" is often defined in order to ensure that the simulation model generates realistic behaviors in a consistent and goal-directed manner. [2] introduces fixed goal structures to hard-code all the possible paths of users' actions into a network. [3] explicitly models the dependencies between a user's actions and his/her goal by conditioning the user's action on a representation of the user goal. [7] proposes an agenda-based simulation model in which the user's goals are kept in a dynamic stack to drive the next user action. A commonality of these systems is that they are all information-providing systems which help the user to complete certain tasks set up by the user. It is natural for these systems to assume a user goal because the goal can be viewed as an abstraction of the task. For example, assuming a user calls the system to complete a task of booking a flight ticket, the user goal can then include providing basic information regarding departure/destination cities and constraints of dates. Besides the user goal, the simulation model also keeps tracks of the dialog. [2][3][7] use different features to represent dialog history, including previous system/user actions, the satisfaction of the user goal, and so on. Then, the simulation model decides the next user action based on the user goal and the current status of the dialog.

While this goal-directed simulation approach performs well with information-providing dialogs, it is less applicable for other dialog genres where it is harder to define clear user goals. For example, when using a physics spoken dialog tutoring system [8], students usually do not have a clear goal of what physics concepts they want to learn. They may have a general goal to learn some physics, but this kind of goal is too general to be helpful in building the simulation model since it cannot help with deciding the next user action in the dialog. Unlike information-providing systems, which collaborate with users to accomplish pre-existing user goals, computer tutors set up the learning goals for students in a tutoring dialog. Similarly, in a training system which trains users to use speech commands to manipulate an in-car driving assistance system [9], users do not have detailed goals to learn specific speech commands before the dialog. Instead, the system decides the steps and the pace of training.

In this paper, we propose a new simulation model that generates consistent dialog behaviors by capturing user knowledge consistency when no clear user goal can be defined. This is inspired by the findings in learning research [10] that knowledge acquisition is consistent and can be visualized by learning curves. We suggest that a good simulation model should be able to model a student's knowledge consistency, and propose a new measure to evaluate simulation models based on the goodness-of-fit of simulated and observed learning curves. In contrast to most prior work which primarily evaluates user simulations with respect to dialogue characteristics, our new measure evaluates simulations with respect to a mathematical model of user cognitive processing.

2. Knowledge Consistency

In a tutoring dialog, the computer tutor trains the student to master any fragment of the persistent, domain-specific information that should be used to accomplish tasks. These task domain concepts, principles, and facts are called *knowledge components* [11]. During the dialog, the tutor helps the student to construct and apply knowledge components. Eventually, a practice effect should be observed in which the students remove the flaws in their understanding of knowledge components with more practice. Research on learning [12] suggests that the learning process proceeds smoothly without sudden gain or loss of knowledge components. In other words, once the student acquires

ITSPOKE1:	Do you recall what Newton's third law says? [**3rdLaw**]
Student1:	No. [**ic**]
ITSPOKE2:	Newton's third law says ... If you hit the wall harder, is the force of your fist acting on the wall greater or less? [**3rdLaw**]
Student2:	Greater. [**c**]
	Dialog goes on

Table 1: Sample coded dialog excerpt.

certain knowledge components, his/her performance on similar problems that require that knowledge component will become stable. Based on this theory, we propose to model student knowledge consistency in our new simulation method by constraining student performance on similar problems that require the same knowledge component. Our approach is further explained in Section 4.

Learning researchers [11] find a power relationship between the error rate of performance and the amount of practice. This relationship can be depicted by a learning curve, which represents that the error rate decreases according to a power function as the amount of practice increases. They also observe that the power relationship might not be readily apparent in a whole learning event, but holds if the learning event is decomposed into sub-components. In this study, we follow the standard way adopted by learning researchers [12] to plot the learning curve. First the learning curves for each knowledge component are plotted; then, an overall learning curve is obtained by averaging these curves. We use the learning curve to visualize the learning of knowledge components observed in the real corpus or modeled in the simulated corpus.

3. Experimental System, Data, and Knowledge Component Analysis

ITSPOKE [8] is a spoken dialog tutor that helps students to understand qualitative physics problems. During the interaction, the system first asks a question and analyzes the student's answer. Then, the system initiates a spoken tutoring dialog to correct misconceptions and to elicit further explanations. The tutoring dialog strategy is hand-crafted in a finite state paradigm. A sample dialog is given in Table 1. In a prior study [8], we collected 100 dialogs between 20 students and the system. The student takes a pretest before interacting with the system and a posttest afterwards. The normalized learning gain is computed using the following formula: NLG=(posttest score-pretest score)/(1-pretest score). 210 different tutor questions are asked in this dialog corpus. Correctness (correct ([**c**]), incorrect ([**ic**])) of student answers is automatically judged by the system and kept in the system's logs.

We explain how we represent knowledge components in our data in Section 3.1, and then we show how we plot the learning curve using these knowledge components in Section 3.2.

3.1. Knowledge Component Representation

Most tutoring systems depend on human experts to determine the knowledge components that most accurately represent students' cognition [11], since the choice of grain size is usually determined by the instructional objectives of the designers. Following this standard approach, a domain expert was assigned to define knowledge components for our data, by manually clustering tutor questions that discussed the same physics concepts together. For example, the domain expert read the two system questions in the dialog shown in Table 1, and judged that both of them talk about Newton's third law. Thus, both these tutor questions were tagged as **3rdLaw** and were added into a mapping table, as illustrated in Table 2. In this table, the first column shows the names of two example knowledge components (KC). **3rdLaw** stands for Newton's third law, and **accel** stands for acceleration. The second column shows examples of some of the tutoring questions associated with these knowledge components. 20 knowledge components were created from the 210 tutor questions in our tutoring dialogs. We call these clusters of questions associated with knowledge components the *manual clusters*.

KC	Tutoring Questions
3rdLaw	Do you recall what Newton's third law says?
	If you hit the wall harder, is the force of your fist acting on the wall greater or less?
accel	What is the definition of acceleration?
	Acceleration is the rate of change of what quantity?

Table 2: Examples of knowledge components.

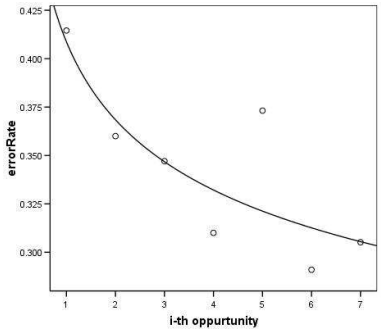

Figure 1: *Overall learning curve for high-learners.*

3.2. Knowledge Component Learning Curve

In a learning curve figure, the x-axis stands for the i^{th} opportunity the student has to practice a certain knowledge component; the y-axis stands for the error rate, which is the percentage of students who failed to use the knowledge correctly. For example, in the dialog shown in Table 1, this student failed to use **3rdLaw** at the first opportunity, but was successful at the second opportunity. Assume there is another student who uses **3rdLaw** correctly at both of the opportunities. Given these two students, on the learning curve for **3rdLaw**, the error rate would then be 50% for the first opportunity and 0% for the second opportunity. We first compute separate learning curves for each of the 20 knowledge components. Then, we get an overall learning curve (as in Figure 1) by computing the average error rates of all the knowledge components for each practice opportunity.

When using all data from the 20 students, we did not observe a decreasing power curve. However, when we split the 20 students into 10 high learners and 10 low learners according to the median of the normalized learning gain, we observe a decreasing power learning curve from the high learner data which is shown in Figure 1. The equation of this curve is $ErrorRate = 0.409 * i^{th}Opportunity^{(-1.50)}$. The adjusted

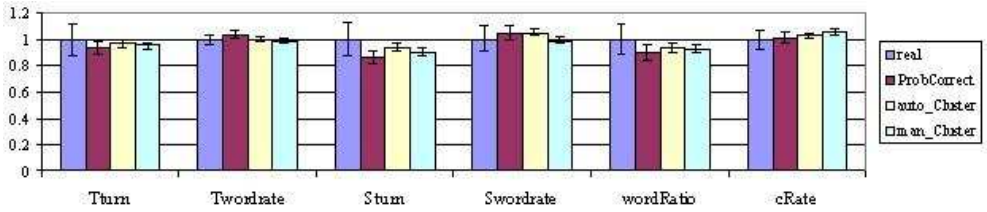

Figure 2: *Evaluation of real and simulated dialogs measured by high-level dialog features.*

Model	probCorrect	auto_Cluster	man_Cluster
R^2	0.252	0.352	0.564
adjusted R^2	0.102	0.223	0.477

Table 3: The goodness-of-fit of simulated and observed learning curves measured by R^2

R^2 value is 0.631, which represents that 63% of the variance in the data is explained by the curve. We do not see a learning curve when using the low learner data. This is not surprising since previous research also reports that learning occurs differently among high/low learners [11]. Therefore, we confirm that in our ITSPOKE data from real users, the performance of high-learners can be represented by a learning curve in terms of our domain expert's representation of knowledge components. Since the focus of this study is to model student behaviors while observable learning is taking place, in Section 4 we train the simulation models on the high-learner data only and compare the simulation models only with the high-learners.

4. Experiments

In Section 4.1, we introduce a new simulation model that captures student knowledge consistency. We also describe a previously used probabilistic model that we compare the new model with. In Section 4.2, we propose a new measure to evaluate the simulation models in terms of user knowledge consistency. In addition, we review previously proposed measures from the literature that we use in the comparison.

4.1. Simulation Models

We train the simulation models from the real corpus we collected. The recognized student utterances in the real dialog corpus are used as the candidate answers for the simulated students.

The Cluster Model. This model generates a student answer based on both the content of the tutor's question and the student's previous answer to a similar question. The answer selection probability can be represented as: $P(A|KC, c)$. **A** stands for the simulated student's action of picking a correct/incorrect answer from the candidate answer set; **KC** stands for the knowledge component of the tutor's question; and **c** stands for the correctness of the student's answer to the last previous question that requires the same knowledge component. When there is no previous student answer, the answer selection probability is computed as $P(A|KC)$. In general, this model assumes that a student will have a higher chance to give a correct answer to the question of a cluster in which he mostly answers correctly before, and a lower chance to do so otherwise.

Besides the manual clusters we described in Section 3.1, we also automatically cluster the tutor questions into 20 clusters based on the lexical items of the questions in order to investigate how well machine clustering can replace the manual clustering. We use the RBR clustering algorithm provided by CLUTO [14]. We use *automatic clusters* to refer to these clusters created by machine, and **auto_Cluster** to refer to the Cluster simulation model based on the automatic knowledge component clusters. We use **man_Cluster** to refer to the Cluster model based on the manual clusters.

The ProbCorrect Model. This previously used model [15] is designed to give a correct/incorrect answer with the same probability as the average of the real students. This model is also similar to the bigram model described in [16]. The answer selection probability can be represented as: $P(A|Q)$. In this equation, **Q** stands for the tutor question. For each tutor question, we automatically compute the average correctness rate of real student answers from the system logs. Then, following this distribution, a correct/incorrect answer is randomly chosen from the correct/incorrect answer sets for this question. A back-off mechanism was used to smooth the probability.

4.2. Evaluation Measures

Knowledge consistency measures. Since we believe student knowledge consistency is an important feature to characterize realistic behaviors and learning curves can visualize this feature, we plot the learning curves for the simulated corpora to compare with the curve observed in the real corpus. If the simulation can model the student learning exactly as what we observed in real data given the knowledge components we defined, the simulated learning curve should mirror the observed learning curve. We use the R^2 to measure the goodness-of-fit between each simulated curve and the real curve. We also report the adjusted R^2 since it is believed to be more accurate for allowing the degrees of freedom to be associated.

Previously proposed Evaluation Measures. [17] propose a comprehensive set of measures to evaluate simulation models. We also compare the models proposed in Section 4.1 using the subset of the measures that are applicable to our data as in our previous study [15]. We use high-level dialog feature measures including the number of student turns (Sturn), the number of tutor turns (Tturn), the number of words per student turn (Swordrate), the number of words per tutor turn (Twordrate), the ratio of system/user words per dialog (WordRatio), and the percentage of correct answers (cRate).

5. Results

We let both of the described simulation models interact with the ITSPOKE system, generating 500 dialogs for each model. This provides us with simulated corpora of comparable size to previous studies [17] that compare simulated and real corpora.

We first evaluate with respect to the previously used measures. In Figure 2, the x-axis shows the evaluation measures; the y-axis shows the mean for each corpus normalized to the

mean of the real corpus. The error bars show standard deviations of the mean values. We can tell how different two corpora are from the overlapping between the error bars. The less overlapping the error bars, the greater the difference between the two corpora. We can see that all the models do not significantly differ from the real students on all the measures, which suggests that they can all simulate realistic high-level dialog behaviors.

Then, we evaluate the simulation models with respect to our new knowledge consistency measure. Table 3 shows the R^2 and adjusted R^2 value of the simulated curves. The first row lists the name of the models. A higher R^2 or adjusted R^2 implies that the simulation models student knowledge consistency better. We can see that both the **man_Cluster** model and the **auto_Cluster** model outperform the **probCorrect** model. When using automatic clusters, the adjusted R^2 performance of the Cluster model decreases relatively 53.2%.

In summary, although the three models perform equally well when measured by high-level dialog features, they show quite different abilities in modeling knowledge consistency. This implies that the knowledge consistency measure is different from the previously used dialog metrics and can be used to distinguish different simulation models.

6. Conclusions and Future Work

Modeling realistic user behavior is very important for applying user simulation to dialog system development. Instead of depending on fixed user goals as in information-providing dialogs, we propose a knowledge consistency model to simulate consistent user behaviors when no clear user goal can be defined. We investigate two variations of this model: one using manually tagged knowledge components, and the other using automatically detected knowledge components. We first evaluate the new models on the dialog behavior level, using previously used measures. Our results demonstrate that the new models as well as a previously proposed probabilistic model do not significantly differ from real users when tested on these measures. In addition, we propose a new knowledge consistency measure to evaluate the simulated user behaviors on the cognitive level. Our comparisons show that the new models outperform the probabilistic model when using either the manual or the automatic clusters, although using the manual clusters gives better performance.

This study is an attempt to simulate consistent user behaviors based on user knowledge consistency rather than fixed user goals. User knowledge can be much broader than the understanding of knowledge components in tutoring dialogs. In user training dialogs, a knowledge consistency model could be used to simulate user's learning and forgetting of speech commands, which are also believed to follow well-known mathematical models [9].

In the future, we intend to explore other ways of constructing a knowledge consistency model, for example, constructing the model based on the learning curve observed in the real user corpus, or taking into account the user's previous performance in a longer time period. We would also like to examine the utility of modeling user cognitive behavior consistency in the information-providing dialogs.

7. Acknowledgements

This work is supported by NSF 0325054. The authors wish to thank Amruta Purandare, Mihai Rotaru, Joel Tetreault, and the three conference reviewers for their insightful suggestions, Scott Silliman for his support on building the simulation system, and Kyle Cunningham, Kenneth Koedinger, and Kurt Vanlehn for their valuable help on plotting the learning curve.

8. References

[1] Georgila, K., Henderson, J., and Lemon, O., "Learning user simulations for information state update dialogue systems", In Proc. of EuroSpeech, 2005.

[2] Scheffler, K., "Automatic Design of Spoken Dialog Systems", Ph.D. diss., Cambridge University, 2002.

[3] Pietquin, O., "A Framework for Unsupervised Learning of Dialog Strategies", Ph.D. diss., Faculte Polytechnique de Mons., 2004.

[4] Ai, H., Tetreault, J. R., and Litman, D. J., "Comparing User Simulation Models for Dialog Strategy Learning", In Proc. of NAACL-HLT, 2007.

[5] López-Cózar, R., De la Torre, A., Segura, J., and Rubio, A., "Assessment of dialog systems by means of a new simulation technique", Speech Communication (40): 387-407, 2003.

[6] Corbett, A. and Anderson, J. R., "Knowledge tracing: Modeling the acquisition of procedural knowledge", User Modeling and User-Adapted Interaction, 1995.

[7] Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young. S., "Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System", NAACL-HLT, 2007.

[8] Litman, D.J., Rosé, C.P., Forbes-Riely, K., Vanlehn, K., and Bhembe, D., "Spoken Versus Typed Human and Computer Dialog Tutoring", International Journal of Artificial Intelligence in Education, 16:145-170, 2006.

[9] Hof, A., Hagen, E., and Huber, A., "Adaptive Help for Speech Dialogue Systems Based on Learning and Forgetting of Speech Commands", In Proc. 7th SIGdial Workshop, 2006.

[10] VanLehn, K., "Cognitive Skill Acquisition", Annu. Rev. Psychol. 47:513-39, 1996.

[11] VanLehn, K., "The Behavior of Tutoring Systems", International Journal of Artificial Intelligence in Education, 2006.

[12] Cen, H., Koedinger, K., and Junker, B., "Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement", In Proc. the 8th International Conference on Intelligent Tutoring Systems, 2006.

[14] Karypis, G., "CLUTO - a clustering toolkit". Technical Report 02-017, University of Minnesota, 2002.

[15] Ai, H., and Litman, D. J., "Comparing Real-Real, Simulated-Simulated, and Simulated-Real Spoken Dialogue Corpora", In Proc .of AAAI workshop, 2006.

[16] Levin, E., Pieraccini, R., and Eckert, W. 2000, "A Stochastic Model of Human-Machine Interaction For learning Dialogue Strategies", IEEE Trans. On Speech and Audio Processing, 8(1):11-23, 2000.

[17] Schatzmann, J., Georgila, K., and Young, S., "Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems", In Proc .of 6th SIGdial, 2005.