Challenges in evaluating spoken dialog systems that reason and learn

Gregory Aist[*,1,2], *Phil Michalak*[1], *George Ferguson*[1], *James Allen*[1,2]

[1]Computer Science Department, University of Rochester, USA
[2]Institute for Human and Machine Cognition, Pensacola, Florida, USA
*Contact address: gsa@gregoryaist.com

Abstract

Evaluation of a spoken dialog system's behavior – both understanding and generation – is a central aspect of dialog system research. At an abstract level, a dialog system can be viewed as a system for (a) listening to a user talk, (b) deciding what to say in response, and (c) saying it. We present developments from an aspect of dialog system behavior over time that challenge such a simplistic model: learning and adaptation on the part of the dialog system. Rather than having a fixed set of dialog policies that can then be evaluated as a complete set, an adaptive dialog system changes its behavior as it learns from experience with users. This shift in thinking about how dialog systems operate raise a number of challenges which we discuss in this paper.

1. The Standard Model of Dialog Systems and Modifications for Learning

Consider the standard model of dialog systems:
1. Listen to the user talk.
2. When the user is done speaking, decide what to say (or do) in response.
3. Say it.
4. When the system is done speaking, go to step 1.

The function described in step 2, in a conventional dialog system, is static: that is, it is the same throughout a single conversation, and it is the same from one conversation to another. In a dialog system that learns, we have something more like the following:
1. Listen to the user talk.
2. When the user is done speaking, decide what to say (or do) in response.
3. Say it.
4. When the system is done speaking, go to step 1, and watch the subsequent dialog to adjust the function in step 2.

Any dialog system that learns will (by definition) undergo some change or growth in its knowledge. In addition to direct measurement of such change, there are many different ways to look at how to evaluate the effectiveness of the dialog system's learning.

2. What does it mean for a dialog system to learn?

A solid, basic definition of learning is "change in state over time in response to experience." Any dialog system that learns will thus experience some change and/or growth in its knowledge. These changes may take several forms, including:
1. Tuning or adapting existing parameters to account for seen data or to better predict unseen data. For example, during training of an artificial neural network, the numerical weights on connections between nodes are adjusted. In the dialog system literature, learning such weights has been done using techniques such as reinforcement learning.
2. Instantiating a new set of parameters to account for the preferences of a particular user or set of users. In its most basic form this is provided by explicit creation of a user account; in more sophisticated versions this sort of learning in interactive systems has given rise to the discipline of user modeling, which has its own conferences, journals, techniques, and experts.
3. Learning a new set of conditions to trigger a previously existing behavior.
4. Learning a new behavior to be triggered at a previously known condition.
5. Learning a new behavior complete with novel triggering conditions; when such behaviors can recurse or otherwise trigger other behaviors, this type of learning beomes learning a new skill.

Most spoken dialogue systems have focused on performing relatively simple tasks for a wide range of occasional users. As such, the prime issues in constructing robust dialog systems concern improving the speech recognition rate and developing robust language interpretation and dialog techniques to compensate for speech recognition errors. We have been building dialog systems in settings where regular, motivated users interact with the system to help them perform complex problem solving tasks.

There are significant differences between such dialog systems and the more common casual-user systems. On one hand, problems with speech recognition are helped by the fact that users will naturally adapt their speech to improve recognition and understanding with continued use. On the other hand, the tasks that the user is performing are quite complex and require deep language understanding and significant reasoning about the intentions underlying the utterances.

3. Learning User Preferences in CALO

The CALO project is a large multi-site research and development effort aimed at building a personal assistant that learns. Aspects of CALO include applications such as calendar management (Gervasio et al. 2005, Berry et al. 2004) and fundamental research such as computer vision (Duong et al. 2005) and temporal planning (Venable and Yorke-Smith 2004). We are discussing here a small part of the overall endeavor: how to evaluate a spoken-language interface to (some parts of) the CALO system.

3.1. What goes in to a user model?

What kind of information should the system store and track about the user in order to enhance the user's experience: that is, what goes in to the user model? Possibilities include:
1. Static information such as birthdate. In some cases (such as due to grant reporting requirements for representativeness of the subject pool) software must be designed to collect demographic information such as gender, race, and ethnicity.
2. Dynamic information such as sequences of user-system interaction, or aggregate measures derived from such interactions (e.g. Pohl et al. 1995, Murphy et al. 1997, Beck et al. 1997).
3. Task- or learning-based information such as what subtasks have been completed, or what subskills have been learned (e.g. Anderson et al. 1990, Brusilovsky and Eklund 1998).
4. Information on user preferences and goals, either inferred from user-computer interactions (e.g. Seo and Zhang 2000) or explicitly mentioned.

3.2. Dimensions of data: Change, Type, and Evidence

Reconsidering these various kinds of user modeling information, we can describe them along several axes.
1. CHANGE: Static vs. dynamic information. Some information is clearly in the static category, such as place and date of birth. (Corrections to mistakes in input, and changes of political geography, we'll set aside for the purposes of this paper.) Other information is likely to remain static for periods of time that are relatively long with respect to a user-computer interaction, such as job title or military rank, but is still open to change on a larger time scale such as months or years.
2. TYPE: Person-based vs. task-based vs. knowledge-based vs. preference-based information. Some information is straightforwardly demographic (if not straightforwardly defined), such as gender, race, ethnicity, and socio-economic status (SES). Other information is clearly related to the task at hand, such as whether the user has completed a particular step in a process – for example, one step in a math problem might be to multiply out the expression (x+3)(y-2) to yield (xy+3y-2x-6). Other information is related to the cognitive state of the user, such as estimates of individual skills such as factoring a number into its prime factors – for example, how to use the FOIL (First-Outer-Inner-Last) method to multiply out expressions. Other information is related to the preferences or affective state of the user – for example, this user might prefer to use mnemonic acronyms such as FOIL to refer to mathematical techniques. (Another user might prefer an algorithmic natural langauge description such as "multiply each term in the first expression by each term in the second expression, and add the results.")
In terms of how information about the user is learned and verified we can add a third dimension:
3. EVIDENCE: Some information is typically acquired explicitly, such as the user's name. Other information is typically acquired implicitly, such as a particular user's link preferences while web browsing. And, of course, much information is amenable to evidence from both explicit and implicit sources.

3.3. Dimensions of data: Change, Type, and Evidence

For evaluating the spoken interface to CALO we plan to use metrics such as the following.

CHANGE: Static
TYPE: Person-based
EVIDENCE: Explicit
Q: Can CALO successfully acquire explicitly communicated information about a user's demographic information? (e.g. name and serial number.)

CHANGE: Dynamic
TYPE: Person-based
EVIDENCE: Implicit
Q: Can CALO successfully acquire implicit information about a person's demographic information? For example, if purchasing a $1000 item requires approval from a superior, and the user says to get approval from Pat Smith, then Pat Smith is likely to have a rank (or job title) that outranks the user's.

CHANGE: Dynamic
TYPE: Preference-based
EVIDENCE: Implicit
Q: Can CALO successfully infer user equipment preferences from requests made on specific purchases? For example, if the user chooses the highest amount of RAM available, CALO could guess that RAM is a high priority and store that preference for use in later procedures.

One interesting direction to pursue is to look at making the information about these learned preferences explicitly availab;e to the user, in a sort of visible user model. (Imagine being able to directly edit the information that CALO has acquired about who your colleagues are, or to turn a "knob" to adjust CALO's opinion about whether you prefer lots of RAM or a fast processor when you're buying a new computer.) And furthermore, due to the large complex nature of CALO these evaluation criteria will continue to evolve in consultation with other team members over time.

4. Learning New Skills in PLOW

Another system we are currently developing, PLOW, is a system that can learn to perform various everyday tasks on a desktop computer by watching the user perform the action as they explain what they are doing. The tasks we have looked at so far typically require requiring finding information on the web and then using this information to perform some office task (e.g., fill in a purchase order, book a ticket, send an email, etc.). For more details see Jung et al. (2006) and Chambers et al. (2006). For the purposes of this position paper, we comment on the challenges of evaluating such systems that combine spoken language understanding, reasoning and learning. We address some of the difficulties in evaluating the PLOW system, which learns task models for performing everyday tasks on a dekstop computer. These include determining how to evaluate dialogue that is tightly coupled with the results of reasoning and learning, and the selection of meaningful metrics for results of learning. We believe that these difficulties are not specific to the PLOW system, and will be issues for any dialogue system that reasons and learns.

4.1. PLOW Implementation and Example Dialog

The PLOW (Procedural Learning On the Web) system is implemented in the TRIPS dialogue system framework (Allen et al. 2001), which has been adapted to a number of different domains requiring mixed initiative dialogue capability. In order

to inform the continuing work on PLOW, we collected twelve dialogues in which subjects were asked to teach a human learner to perform five typical tasks involving a web browser. Subjects were told to assume that the person learning was familiar with web browser usage but not with the tasks being demonstrated. The subjects were given time to familiarize themselves with the tasks prior to demonstration in order to avoid unnecessary complications. In these trials, the demonstrator and the learner are situated in separate rooms and communicate verbally via headset. The learner can also see a copy of the demonstrator's screen. We chose to have an experimenter play the role of the learner in order to control feedback and question asking policy. The learner was permitted to acknowledge demonstrator utterances, to ask where on the screen information was located, and to indicate that a step was confusing. This constraint models interactions of the current PLOW system.

In addition to the dialogue generated during these trials, we collected feedback from the subjects regarding the degree to which their interaction felt "natural". In the remainder of this section we'll make reference to the data from this corpus where it provides insight. A typical interaction with the PLOW system that we'll use for illustration purposes is shown below.

Let me teach you to buy a book
You go to the purchase form
Enters a URL in the browser and hits enter
We fill in the author field
Types an author into the author field
And the title field
Types a title in the title field
Now let me show you how to find the other information
Go to Amazon
Opens a new tab, enters a URL in the browser
We select the books tab
Clicks the tab labeled BOOKS
Then select advanced search
Clicks the link labeled ADVANCED SEARCH
Put the title here
Types the title into the title search field
And put the author here
Types the author into the author search field
Then click the search button
Clicks the search button
Now we select the page for the book
Clicks the link with the title of the book
This is the price
Highlights the price
We're done here
Switches back to the book purchase form
We put the price here
Types the price into the PRICE field
Now submit the form
Clicks the submit button
OK, that's it

This type of dialogue is representative of expert/apprentice interactions; the PLOW system is a novice learning how to perform various tasks from a demonstration and accompanying verbal description. As such, it must perform deep reasoning on each utterance in order to properly situate the propositional content of the utterance in the developing task representation. This reasoning can manifest itself in two distinct ways: in the content of the system generated language, and in the learned task models. Evaluating the dialogue produced by the system in isolation is therefore only a partial evaluation of the system. We believe that a complete evaluation of the system must also include the output of the system, in this case the task models.The dialogue listed above is typical of interaction with the PLOW system. System feedback during task learning is largely back-channel acknowledgement except in cases where the system doesn't understand the user's utterance, in which case it indicates confusion. Survey results from subjects that participated in a small study in which they taught a human learner how to perform various web browser based tasks indicates that this type of interaction is not completely unnatural. Nine of twelve subjects felt that their interaction under these conditions was natural, while the remaining three indicated that the lack of questions was a little strange.

4.2. Considerations for Evaluation of PLOW

Future versions of the PLOW system will incorporate more sophisticated reasoning about the task model being constructed through dialogue with the user, which will enable focused questions to further refine the structure of the task. For instance, the steps of a demonstration are inherently ordered temporally even though the task itself may not place such constraints on their execution. The order in which the author and title fields are filled in the book buying dialogue is not important, even though the demonstration provides an explicit order. The PLOW system will attempt to detect such cases and to generalize the resulting task model. In new situations where the inferred constraints are considerably weaker than the demonstration order, it might be beneficial to verify with the user that the inference is correct. Such a change in initiative could drastically affect whether or not the resulting interaction is perceived to be natural. While human learners are expected to ask questions, the type and timing of the questions will likely influence the quality of the dialogue. Consequently, any system evolution that involves a shift in initiative should be evaluated for efficacy and aesthetics.

An additional consideration in evaluating systems that learn is how, if at all, to represent the changing state of the system. PLOW builds task models as it interacts with a user. It seems natural to convey these developing task models since the user desires an appropriate end result. But is spoken language the most natural and efficient modality for this communication? Complex dialogue systems are becoming increasingly multi-modal, and all aspects of system communication need to be evaluated. We've chosen to visually represent the state of the learned procedure because it is convenient to persistently depict the substructure of the task throughout the duration of the interaction. The question of how to represent such change is secondary to whether or not to represent it at all. In the case of the PLOW system it seems natural to do so, but even here it isn't clear from introspection alone that it is beneficial. This becomes increasingly apparent as more sophisticated reasoning is incorporated in the system. Will it be disconcerting for users to see a different structure in the developing task model than the surface structure that is being demonstrated? Do users even expect to have access to this type of internal mental state? They certainly don't have direct access to it when the demonstration is being made to another human. Evaluating any communication that arises from changing state is a difficult task in its own right. In large part, the content of the communication will determine its understandability, so evaluation of the dialogue is at the very least an implicit evaluation of the underlying state. We believe that a complete system evaluation should explicity evaluate the underlying state, and we discuss some potential evaluations of task structure (an aspect of PLOW's mental state) below.

Examining the restrictions on the language that PLOW understands provides further evidence that the quality of the learned procedures are an important aspect of system evaluation. PLOW currently expects key phrases to trigger some of its learning capacity. For instance, one must use a variant of the phrase “Let me show you how to - ” in order to indicate that the following sequence will be part of a separate subtask, and a variant of “We're done” to indicate completion of the current subtask. We hope to eliminate some of these restrictions by including more sophisticated intention recognition. From a dialogue evaluation perspective, this is a clear improvement, but a proper evaluation must examine the quality of the resulting task models. Here again, reasoning and learning have muddied the waters by transforming the task of evaluating system dialogue so that it includes an evaluation of the resulting task models. In situations where the instructor takes most of the initiative, this is perhaps the only effective measure of system performance.

4.3. Procedure Learning in PLOW

A simple evaluation of PLOW performance is to check whether it is capable of performing the task that has been demonstrated. This simple evaluation has already been performed in initial tests of the system's feasibility. On the book buying task from which the dialogue example above is taken, after a single demonstration the system is able to correctly retrieve the price of more than 95% of one hundred and sixty-two randomly selected book titles. (The books were randomly selected from Amazon.com's inventory, and the task models were applied to Amazon.com and BarnesAndNoble.com. See Jung et al. 2006 for details.)

This simple metric, however, fails to capture key aspects of the task learning problem. One important feature of this learning problem is that there is no single “correct” answer. Some solutions may be preferrable to others, but criteria for distinguishing ``good'' solutions from ``bad'' ones are difficult to specify quantitatively. We consider in this section some of these criteria and ways that they might be quantified.

One subjective evaluation of PLOW performance is to measure how understandable or intuitive the learned procedures are to humans. This is, of course, an important evaluation because it speaks directly to the ability of the system to effectively converse with users about the results of learning. This is a difficult criterion to quantify, so we hypothesize that “intuitive” solutions will be ones that are reusable and generalizable, criteria which are more easily quantified. In addition to approximating procedure understandibility, procedure reuse and generalizability measure more directly another desirable system feature.

One goal of the PLOW system is to learn incrementally over time in addition to learning at individual points of demonstration. The ability to reuse previously learned procedures while learning new procedures demonstrates this facility, as does the more complex ability to abstract from a particular demonstration to a broader class of situations.

We will consider the following metrics for reusability:

1. Count the average number of times that any task model is reused during the learning of a particular procedure while varying the number of distinct prior demonstrations. Ideally this quantity will increase with the number of prior task demonstrations if the learned procedures are reusable.
2. Count the number of times that a previously unseen subtask is created during the learning of a particular procedure while varying the number of distinct prior demonstrations. Ideally this quantity will decrease with the number of prior task demonstrations, indicating that previously learned tasks can be composed to perform new tasks.
3. Count the number of times that the subtasks learned during a demonstration from one instructor on a particular task are reusable during a demonstration from another instructor on the same task. This measure of reusability is stricter than the previous two, because the task is fixed while instructors vary. This quantity indirectly measures the quality of the reasoning algorithms that generate subtask structure, increasing when the algorithms perform well enough to ignore spurious differences in the surface structure of demonstrations.

We will consider the following metrics for generalization:

1. Count the average number of times that a procedure can be applied without modification to different classes of parameters than those for which it was demonstrated. This measure is somewhat dependent on the nature of the procedure and will often measure the raw capacity for generalization of the method encoded by the procedure rather than capability of the learning algorithms for generating generic procedures. For example, a procedure for buying books that uses a book store website will inherently be limited in the types of objects to which it can apply. For this reason, we consider the following metric as well.
2. Count the average number of times that a procedure can be specialized for parameters of different ontology types than the ones for which it was demonstrated by simple modification. As a first approximation, we limit the modification to substitution of subtask models, e.g. substituting a procedure that finds the price of a piece of office equipment for a subtask that finds the price of a book in the in the book purchasing dialogue above.

5. Conclusion

In this position paper we have covered several areas related to the evaluation of dialog systems that learn. First, we have described the “standard model” of dialog system behavior and how to modify that model to account for learning and adaptation on the part of the dialog system. Second, we have given a taxonomy of various kinds of learning in which a dialog system can engage. Third, we have described our approach to evaluation in the spoken dialog interface to CALO, where a key part of the system's performance is its adaptation to users over time. Finally, we briefly summarized some of the challenges inherent in evaluating PLOW, a dialogue system that reasons and learns. The language output of the system becomes more variable with the introduction of more complex reasoning capabilities, as is the case when the system gains the capability to ask focused questions to aid its learning process. Evaluating the questions that the system asks and the effects of relaxed input speech requirements moves to the realm of evaluting the system's reasoning. Such evaluations are non-trivial, and require indirect evaluation of the system output, in this case the user model or learned procedures. We believe that our experiences are not unique to the CALO system or to the PLOW system. As dialogue systems incorporate more complex reasoning, they will naturally become more difficult to evaluate. Specifically, we expect that these systems will face the dilemna of representing dynamic internal state, and the problems associated with evaluating dialogue that is intrinsically tied to the reasoning performed by the system.

6. References

[1] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A. 2001. Towards conversational human-computer interaction. *AI Magazine* **22**(4): 27-38.

[2] J.R. Anderson, C.F. Boyle, A.T. Corbett, M. Lewis. 1990. Cognitive modeling and intelligent tutoring. Artificial Intelligence 42:7-49.

[3] Beck, J., Stern, M. and Woolf, B. P., (1997), Using The Student Model To Control Problem Difficulty in *User Modelling; Proceedings of the 6th International Conference* eds. A. Jameson, C. Paris and C. Tasso, Springer Wien.

[4] Berry, P.M. and Gervasio, M. and Uribe, T.E. and Myers, K. and Nitz, K. . *A Personalized Calendar Assistant*, in AAAI Spring Symposium Series, Stanford University, March 2004.

[5] Brusilovsky, P., and Eklund, J. 1998. A study of user model based link annotation in educational hypermedia. Jnl of Unversal Computer Science 4(4).

[6] Chambers, N. and Allen, J. and Galescu, L. and Jung, H. and Taysom, W., ``Using Semantics to Identify Web Objects", Proceedings of the National Conference on Artificial Intelligence: Special Track on AI and the Web, 2006.

[7] Duong, T. and Bui, H. and Phung, D. and Vekatesh, S. *Activity recognition and abnormality detectiong with the switching hidden semi-Markov model*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[8] Gervasio, M. and Moffitt, M. and Pollack, M. and Taylor, M. and Uribe, T. *Active Preference Learning for Personalized Calendar Scheduling Assistance*, in Proceedings of the 2005 International Conference on Intelligent User Interfaces, San Diego, CA, Jan 2005.

[9] Jung, H. and Allen, J. and Chambers, N. and Galescu, L. and Swift, M. and Taysom, W., ``One-Shot Procedure Learning from Instruction and Observation", Proceedings of the International FLAIRS Conference: Special Track on Natural Language and Knowledge Representation, 2006.

[10] Murphy, M. and McTear, M., (1997), Learner Modelling for Intelligent CALL in *User Modelling; Proceedings of the 6th International Conference* eds. A. Jameson, C. Paris and C. Tasso, Springer Wien.

[11] Pohl, W., Kobsa, A. and Knutter, O., (1995), User Model Aquisition Heuristics Based on Dialogue Acts in *International Workshop on the Design of Cooperative Systems,* 471-486, France.

[12] Seo, Y. and Zhang, B. 2000. "Learning user's preferences by analyzing web-browsing behaviors," Proc. of Int'l Conf. on Autonomous Agents 2000 (AA '2000), pp. 381 - 387.

[13] Venable, K. B. and Yorke-Smith, N. *Disjunctive Temporal Planning with Uncertainty*, in Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, UK, pp. 1385–1386, Aug 2005.