Non-Cooperative and Deceptive Virtual Agents

David Traum, University of Southern California

Virtual agents that engage in dialogue with people can be used for a variety of purposes, including as service and information providers, tutors, confederates in psychology experiments, and role players in social training exercises. It seems reasonable that agents acting as service and information providers, and arguably as tutors, would be be truthful and cooperative. For other applications, however, such as role-playing opponents, competitors, or more neutral characters in a training exercise, total honesty and cooperativeness would defeat the purpose of the exercise and fail to train people in coping with deception. The Institute for Creative Technologies at the University of Southern California has created several role-playing characters, using different models of dialogue and uncooperative and deceptive behavior. This article briefly describes these models, as used in two different genres of dialogue agent: interviewing and negotiation. The models are presented in order from least to most sophisticated reasoning about deception.

Most accounts of pragmatic reasoning in dialogue use versions of Grice's cooperative principles and maxims[1] to derive utterance meanings (which might be indirect in their expression). However, these maxims, such as "be truthful," don't cover situations in which conversationalists are deceptive or otherwise uncooperative, even though much human dialogue contains aspects of uncooperative behavior. Gricean accounts alone don't adequately cover cases in which conversational participants aren't cooperative—for example, why do they ever answer at all? The notion of *discourse obligations*[2] differentiates the obligation to respond from the mechanism of response generation, which could be either cooperative, neutral, or deceptive.

Creating Deceptive Characters

The simplest way to create deceptive characters is for the scenario author to program the deceptive answers directly into the dialogue policy. In this way, the deception comes only from the designer, not from the character itself, which can't distinguish deceptive from sincere utterances. We've used this technique for some simple interviewee characters (including C3IT[3]), where the trainee's goal was to uncover which of several suspects was guilty, and deceptive answers aided in the diagnosis.

To engage in more flexible behavior, characters must know the difference between truthful and deceptive or evasive information and be able to decide on an honest, evasive, or deceptive dialogue-management and response-generation policies. A more advanced question-answering architecture includes three levels of compliance (compliant, reticent, and adversarial), and the designer creates a different response set for each of these when creating a character.[4] When characters are *Compliant*, they provide information when asked, but fall short of Gricean cooperativity since they don't provide helpful information that was implicated rather than explicitly solicited. When characters are *Reticent,* they provide neutral information, but will evade any questions about important or sensitive information. When characters are *Adversarial*, they provide deceptive or untruthful answers. The characters maintain a set of social and emotional variables throughout the dialogue to determine their compliance level. These include *respect* (trainee for character and character for trainee), *social bonding*, and *fear*. When the variables change significantly, the characters change their compliance levels, letting the trainee experiment with different interview strategies, including empathy, threatening, or bargaining. Although these characters can reason about when to employ deception, the models operate globally, depending on the character's mood rather than on the specifics of the information itself. When in an antagonistic mood, the character will lie about everything it can. The scenario author must still decide explicitly which content can have deceptive answers.

Tactical Questioning and Negotiation

The **SASO-ST** system provides a virtual reality environment for a trainee to practice negotiation (for example, convincing a doctor to move his clinic).[5] It includes characters who have goals that might either align or conflict with those of a human dialogue participant. When engaged in negotiation, the characters will dynamically adopt a negotiation strategy, based on a similar, but more advanced, calculation of social and emotional variables, including trust (with subcomponents of solidarity, credibility, and familiarity), expected utility, and control. Several of these strategies are uncooperative, in that they try to achieve different outcomes than the trainee is trying to acheive. In the *avoidance* strategy, the characters will refuse to answer questions germane to the negotiation topic. However, they will answer questions that are considered irrelevant (meaning questions about issues that are not related causally to the topic through a plan structure). Characters also use trust to decide when to believe or doubt another character. Moreover, characters use the task model structure to infer implications of what has been said, both for Gricean cooperative purposes and to recognize hidden motivations that a speaker might like to keep private. Figure 1a shows negotiations between the character "Dr. Perez" and a human playing a captain who wants to move the doctor's clinic.

Figure 1. Simulations for tactical questioning and negotiationWe can see human and character interactions with the (a) SASO-ST, (b) Tactical Questioning Amani, and (c) SASO4 systems.

The third generation Tactical questioning architecture,[6,7] allows authors much more fine-grained control in crafting sophisticated policies for uncooperative and deceptive behavior than the models described above. Using the domain-editor software, an author constructs a domain-specific ontology of information (including false situations) and can create specific policies for conditions under which the character should be compliant, evasive, or deceptive about any information in the ontology. These conditions can include aspects of the emotional variables, as before, but also arbitrary aspects of the information state, including whether specific topics had previously been discussed, or whether specific incentives have been offered. Strategies for several kinds of responses have been constructed in the form of finite-state subdialogue networks, which the character can use to meet obligations while answering (truthfully or deceptively), eliciting an offer, or refusing to answer. This architecture has been used by students and other scenario authors to construct more than a dozen different characters used for purposes like training tactical questioning (for example, the Amani character in Figure 1b), negotiation, or for use as a virtual confederate in psychology experiments.[7] All of these characters engage in uncooperative behavior at times, and most include some deceptive aspects. Two characters in particular are used for training deception detection,[8] where characters' verbal and nonverbal behavior can differ along various dimensions. Students can learn to follow reliable cues and discount unreliable ones.

Finally, in more recent work on the SASO architecture, we have created an ability for characters to reason more thoroughly about secrecy. In the Tactical questioning architecture described above, authors must indicate each sensitive piece of knowledge separately. However, knowledge is often related, such that talking about one topic might reveal another closely related one. For the SASO4 scenario (see Figure 1c),[9] each virtual character has a shameful secret and wants to avoid revealing it to the other virtual character. To achieve a successful result, the human participants must separate the virtual characters and only then will the characters bring up the sensitive issues, so that the humans can address them. When the participants are all together, the virtual characters must appear to be seriously negotiating while keeping secret the shameful reason that they're against the proposal. To achieve this with a minimum of authoring overhead, we have designed and implemented a secrecy inference scheme in which the author designates only the secret concept and who it must be kept secret from. The inference scheme then automatically marks secret related actions and states that would reveal the secret. Secret items won't be discussed—as proposals, arguments in the negotiation, answers to questions, or justifications of other claims—if the character that the secret is to be kept from is in contact. The preliminary set of inference rules is as follows:

- The sole precondition for a secret action is secret.
- A task with a secret precondition is secret.
- A state that can be achieved only as the effect of a secret task is secret.
- The only task that can establish or remove a secret effect is secret.

These rules are currently at the level of heuristics rather than sound and complete guarantees of secrecy. In particular, the first rule might be overly conservative, because other reasons might exist for discussing the precondition. Moreover, revealing the secret assumes that other participants have similar task knowledge and inference ability. We have performed a preliminary evaluation within the SASO4 scenario.

We asked eight participants to read a brief description of the scenario and the main item to be kept secret, and then rate all elements of the domain as to whether it would reveal the secret (76 total concepts, 15 derived secrets). In all but two cases, the majority view agrees with the inferences that the system makes using the set of rules described above. In the two exceptions, there was a 50% split among the participants as to whether the concepts were secret (the inference rules marks them secret).

We have briefly presented several different architectures for creating characters that can engage in noncooperative and/or deceptive dialogue behavior. These vary in the authoring burden, the characters' ability to dynamically decide on and change their behavior, and their ability to perform inference about how sensitive material is related. Generally, a trade-off exists between simplicity of authoring, the amount of authoring needed, and the inference ability that the character can perform.

References

1. J.P. Grice, "Logic and Conversation," *Syntax and Semantics*, vol. 3, Speech Acts, P. Cole and J.L. Morgan, eds., Academic Press, 1975, pp. 41–58.
2. D.R. Traum and J.F. Allen, "Discourse Obligations in Dialogue Processing," *Proc. 32nd Ann. Meeting Assoc. Computational Linguistics*, 1994, pp. 1–8.
3. P. Kenny et al., "Building Interactive Virtual Humans for Training Environments," *Proc. Interservice/Industry Training, Simulation and Education Conference (I/ITSEC),* 2007, paper no. 7105.
4. A. Roque and D. Traum, "A Model of Compliance and Emotion for Potentially Adversarial Dialogue Agents," *Proc. 8th SIGDIAL Workshop on Discourse and Dialogue*, 2007, pp. 35–38.
5. D. Traum et al., "Virtual Humans for Non-team Interaction Training," *Proc. AAMAS Workshop on Creating Bonds with Embodied Conversational Agents*, 2005, pp. 70–75.
6. S. Gandhe et al., "From Domain Specification to Virtual Humans: An Integrated Approach to Authoring Tactical Questioning Characters," *Proc. 9th Ann. Conf. Int'l Speech Communication Assoc.* (InterSpeech 08), 2008, pp. 2486–2489.
7. S. Gandhe et al., "Evaluation of an Integrated Authoring Tool for Building Advanced Question-Answering Characters," *Proc. 12th Ann. Conf. Int'l Speech Communication Assoc.* (InterSpeech 11), 2011, pp. 1296–1299.
8. H.C. Lane et al., "Virtual Humans with Secrets: Learning to Detect Verbal Cues to Deception," *Intelligent Tutoring Systems*, LNCS 6095, V. Aleven, J. Kay, and J. Mostow, eds., Springer, 2010, pp. 144–154.
9. B. Plüss, D. DeVault, and D. Traum, "Toward Rapid Development of Multiparty Virtual Human Negotiation Scenarios," *Proc. 15th Workshop Semantics and Pragmatics of Dialogue* (SemDial 11), 2011, pp. 63–72.

David Traum *is a principal scientist at the University of Southern California's Institute for Creative Technologies.. Contact him at traum@ict.usc.edu.*