Improving a Virtual Human Using a Model of Degrees of Grounding

Antonio Roque and **David Traum**
USC Institute for Creative Technologies
Marina del Rey, CA, USA
{roque, traum}@ict.usc.edu

Abstract

We describe the Degrees of Grounding model, which tracks the extent to which material has reached mutual belief in a dialogue, and conduct experiments in which the model is used to manage grounding behavior in spoken dialogues with a virtual human. We show that the model produces improvements in virtual human performance as measured by post-session questionnaires.

1 Introduction

Human speakers conversing with each other risk experiencing various kinds of errors of understanding and have a number of strategies to recover from error. Studying these errors and strategies can be helpful when designing dialogue managers for spoken dialogue systems [Swerts *et al.*, 2000; Paek, 2003; Skantze, 2005; Litman *et al.*, 2006]. Similarly, dialogue system researchers have studied domain-independent error detection and recovery techniques [Bohus and Rudnicky, 2005a; 2005b] to determine when to provide confirmations or rejections, and to determine how to handle cases of complete non-understanding.

One approach to studying how humans manage errors of understanding is to view conversation as a joint activity in which **grounding**, the process of adding material to the common ground between speakers, plays a central role [Clark and Schaefer, 1989]. From this perspective, introduced by Clark and his collaborators, conversations are highly coordinated efforts in which participants work together to ensure that knowledge is properly understood by all participants. There is a wide variety of grounding behavior that is determined by the communication medium, among other things [Clark and Brennan, 1991].

This approach is further developed by Traum, who presents a model of grounding which adapted Clark and Schaefer's contributions model to make it more suitable for an online dialogue system [Traum, 1994]. Traum's approach uses **common ground units** (CGUs) [Nakatani and Traum, 1999] to represent the content being grounded, and grounding acts to describe the utterances that ground the common ground units. Other computational approaches to grounding use decision theory [Paek and Horvitz, 2000] or focus on modeling belief [Bunt *et al.*, 2007].

Grounding models generally consider material to be in one of three states: ungrounded, in the process of becoming sufficiently grounded, or sufficiently grounded. An exception is our **Degrees of Grounding** model [Roque and Traum, 2008], which provides a more detailed description of the extent to which material has become a part of the common ground during a dialogue. In this paper we describe experiments in applying that model to handle explicit grounding behavior in a virtual human. We begin by describing the model and the testbed domain, then describe the experimental methodology and the results, which showed improvements in several measures including appropriateness of response.

2 The Degrees of Grounding Model

The Degrees of Grounding model is made up of a set of types of **evidence of understanding** describing the cues that speakers can give to indicate that mutual belief is being achieved, a set of **degrees of groundedness** describing the extent to which material has achieved mutual belief, a set of algorithms for dialogue management (to identify evidence of understanding and degrees of groundedness in an utterance and to determine evidence to provide in reply,) and a **grounding criterion** for each type of information, describing the extent to which the material should be discussed before it is grounded enough for the current purposes.

Table 1 describes the set of types of evidence of understanding used in the Degrees of Grounding model. Several types of evidence are related to the order in which material is introduced and the speaker who introduced them. In a Submit, material is introduced for the first time; in a Repeat Back, material is presented again by the other speaker; and in a Resubmit, material is presented again by the speaker who originally introduced it, along with an indication of misunderstanding. Other kinds of evidence are identified by semantic interpretation, such as Acknowledgments, Request Repairs, and Uses. Move Ons are identified by a task model, and Lack of Response is identified by a length of time.

Table 2 shows the set of degrees of groundedness. These are produced by combinations of evidence of understanding, and represent the extent to which the material has reached mutual belief. The current degree of a piece of material is generally identified by the previous degree and the latest type of evidence given.

Evidence	Description
Submit	Material is introduced for the first time
Repeat Back	Repetition of material that was submitted by the other speaker
Resubmit	Material previously submitted is submitted again by the same speaker.
Acknowledge	A general statement of understanding without verbatim repetition
Request Repair	A participant states a lack of understanding and asks for a resubmission
Use	A submission is made in a way that indicates previously submitted material has been understood
Move On	A participant proceeds to the next step of a task that requires completion of the previous step
Lack of Response	Neither of the participants speak for a given amount of time

Table 1: Evidence of Understanding

Degree	Description
Unknown	Material has not yet been introduced
Misunderstood	A participant has requested repair of the material
Unacknowledged	A participant has submitted material but not heard an expected response
Accessible	Material has been submitted or resubmitted
Agreed-Signal	A participant has given an expected acknowledgment
Agreed-Signal+	A participant has given an expected acknowledgment, followed by futher evidence of understanding
Agreed-Content	A participant has given an expected repeat back
Agreed-Content+	A participant has given an expected repeat back, followed by further evidence
Assumed	Material has been grounded by other means

Table 2: Degrees of Groundedness

3 Testbed Domain

3.1 Virtual Humans for Tactical Questioning Training

In the testbed domain of Tactical Questioning, military personnel engage in dialogues to obtain information of military value [Army, 2006]. We are specifically interested in this domain when applied to civilians, when the process becomes more conversational and additional goals involve developing social bonds with members of the population and gathering general information about the area of operations. In the example of Iraqi Arabic culture, appealing to the subject's sense of honor and being aware of issues of interest to an influential person - such as a need for secrecy, or cooperation on Civil Affairs projects - can lead to success in acquiring information during a Tactical Questioning session [Paul, 2006].

We have investigated the use of virtual humans (spoken dialogue systems embodied in a virtual environment) for training individuals in conducting Tactical Questioning dialogues. This work is in the tradition of research in building virtual humans for training or tutoring purposes [Gratch and Marsella, 2005; Traum *et al.*, 2005]. The testbed for this research involves a virtual human named Hassan, which includes a model of emotions and social interactions [Traum *et al.*, 2007; Roque and Traum, 2007]. The next sections describe the training scenario, the system architecture, and how grounding works in the system.

3.2 Testbed Scenario

The scenario takes place in contemporary Iraq, where the trainee talks to Hassan, a local businessman. If the trainee convinces Hassan to help him, the trainee will confirm suspicions about an illegal tax being levied on a new marketplace. A successful trainee may discover that the tax has been placed by Hassan's employer, and even learn where to find that employer. But if Hassan becomes adversarial, he may lie or become insulting.

Figure 1 shows an excerpt from a typical dialogue with Hassan. Along with social and emotional considerations, Hassan has a set of goals that must be satisfied, such as assurances of protection, or in some cases, an offer of money. In utterance 1 the Trainee asks a question about the identity of the person collecting the tax. In utterance 2, Hassan indicatates that he might be interested in answering the question, if his needs were fulfilled. In utterance 3 the Trainee indicates that Hassan could have protection and money. In utterance 4 Hassan indicates that he is interested in the offer of money, and answers the question about taxation. Note that utterance 4 is made up of two sentences: in the first, Hassan repeats back the topic of the offer that he is interested in, and in the second he provides the information. The first sentence is an explicit grounding act, which provides to the Trainee an indication of the particular offer that Hassan is interested in.

1	Trainee	Well then why don't you just tell me who is collecting the tax?
2	Hassan	I might tell you what you want if there was something in it for me.
3	Trainee	We can protect you or offer you money.
4	Hassan	So, you offer me money. Indeed, you might say that I collect the taxes.

Figure 1: Tactical Questioning Dialogue

As a training application, Hassan logs utterances, language features, and emotional states at every turn, with the aim of producing a summary for an after-action review, at which time a human trainer and trainee may discuss the session. The notion is that although Hassan is an automated system and handles dialogues without human intervention, a trainer should be allowed to supervise the session and have the ability to intervene mid-session or review the session after the fact. For this reason, Hassan may react realistically to a trainee's bribes or threats of force, even though such actions are against policy for Tactical Questioning of noncombatants

[Army, 2006]: these behaviors would be reviewed by a human trainer during or after the training session.

3.3 System Architecture

Hassan's natural language components are shown in Figure 2. Voice input is translated into text by an Automated Speech Recognition (ASR) Component. An Anaphora Resolution component resolves pronouns and other anaphors based on dialogue context, and a Natural Language Understanding (NLU) component performs statistical dialogue act classification. The output of the NLU component, which is a semantic frame, is next used by two components: the Grounding component and the Compliance component.

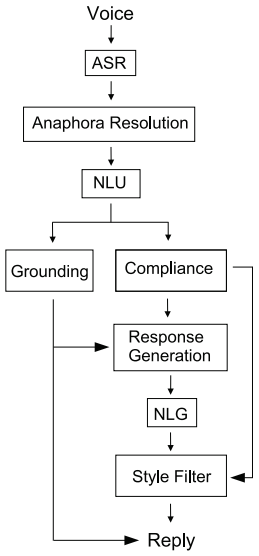

Figure 2: System Architecture

The Grounding component is discussed below. The Compliance component updates Hassan's emotional state, and is used in conjunction with the Response Generation to determine the semantics of the reply (apart from explicit grounding behavior) that Hassan will make. This response is effected by an Natural Language Generation (NLG) component to produce a text reply, which may be made less or more polite by a Style Filter, which is influenced by the model of emotions. This final reply is then spoken by the virtual human.

3.4 Grounding and Dialogue Management

As described above, the Grounding and Response Generation components work with an interpretation produced by the NLU component, an example of which is shown in Figure 3. It is a speech act expressed in XML; the example shown is a question asking why the market is not being used. For purposes of grounding, the topic of the relevant common ground unit is *market*.

The Grounding component handles the optional production of certain kinds of evidence of understanding: Acknowledgments, Repeat Backs, and Repair Requests, while other

```
<speech_act speaker="player">
  <whq>
    <object name="market">
      <attribute name="reason-not-used" />
    </object>
  </whq>
</speech_act>
```

Figure 3: Example Interpretation

task-related evidence of understanding (such as Move Ons and Uses) are produced by the Response Generation component. The Response Generation component always produces a response, but not until the Grounding component notifies it that it has produced its own reply or has decided not to produce a reply. So for this architecture, Hassan's utterances are made up of an optional grounding statement produced by the Grounding component followed by a reply produced by the Response Generation component.

The Grounding component tracks the degree of groundedness of material under discussion, and uses each material's current degree of groundedness and its grounding criterion to decide what kind of evidence of understanding to produce, if any. Offers or sensitive topics such as Hassan's employer have higher grounding criteria than less-threatening topics such as general social talk.

The algorithm used by the Grounding component is shown in Figure 4. In lines 2-3 the evidence of understanding provided by the utterance and the relevant material's subsequent degree of groundedness is computed by two sets of rules, and the material's associated CGU is updated. In line 4 a set of rules determines what kind of grounding behavior the virtual human will provide, if any, based on its degree of groundedness and whether it has met its grounding criteria yet, or whether a repair request is needed. In lines 5-7 the response is made and for each piece of evidence of understanding that was given in the response, a set of rules computes and updates the relevant CGU's new degree of groundedness.

```
1    given a speech act,
2      compute the evidence it provides
3      compute the CGU's new degree
4      determine evidence to provide

5    provide the evidence

6    for each evidence that was given
7      compute the CGU's new degree
```

Figure 4: Grounding Component Algorithm

4 Example Dialogue Excerpt

Figure 5 provides an example exchange, along with a depiction of how it would be interpreted by the Degrees of Grounding model. After the human Trainee asks the question in line 1, the NLU component identifies that the object of the question, and therefore the topic of the utterance, is Imam. The Grounding component examines its CGU record, which

tracks the material under discussion, and determines that this topic is being introduced for the first time. A rule therefore determines that the type of evidence of understanding provided by the utterance in line 1 is Submit. A second rule determines that the degree of groundedness of the Imam CGU is Accessible, because the latest evidence of understanding regarding it was Submit.

1	Trainee	Do you know the Imam? *Topic: Imam* *Evidence: Submit* *Degree: Accessible*
2a	Hassan	You take an interest in the Imam. *Topic: Imam* *Evidence: Repeat Back* *Degree: Agreed-Content*
2b		I would prefer to remain anonymous in these matters. *Topic: Anonymity* *Evidence: Submit* *Degree: Accessible*

Figure 5: Example Dialogue

At this point the Grounding component determines what reply to make. Because the topic of the Imam is a sensitive one, its grounding criterion is higher than the Accessible degree of groundedness, so its grounding criterion is not yet met. A rule determines that the appropriate type of evidence of understanding to provide is a Repeat Back. The Grounding component produces this reply in line 2a, and a rule determines that because of this the degree of groundedness of the Imam CGU has reached Agreed-Content, which is sufficient for the current purposes. The Grounding component then tells the Response Generation component to produce the rest of Hassan's response, which is a request for anonymity shown in line 2b.

Note that the response in line 2b, produced by the Response Generation component, contains no mention of the Imam. Without the Grounding component's explicit grounding utterance in line 2a, the Trainee would have no indication that Hassan understood that the question was about the Imam.

5 Experiments

5.1 Methodology

The purpose of these experiments was to determine whether the Degrees of Grounding model, when implemented as a grounding component to track grounding and handle explicit grounding responses, would provide increased performance in a virtual human. We measured performance as improvement in post-session ratings by human users in a number of measures to track their perception of the virtual human's grounding behavior and appropriateness of response. We performed two experiments to investigate two hypotheses. In the first experiment we tested the hypothesis that the grounding component using the Degrees of Grounding model would improve, rather than reduce or not affect, the performance of the virtual human. In the second experiment we tested the hypothesis that any improvements would be due specifically to the Degrees of Grounding model rather than to increased grounding of any kind.

In each experiment we had 10 subjects hold dialogues with Hassan, interacting with him in two conditions: a condition in which Hassan's grounding component implemented the Degrees of Grounding model, and a control condition which varied by experiment. The sequence of initial conditions alternated, to manage ordering effects. A different set of subjects was used for each experiment, and their dialogues were free rather than given. After each dialogue, subjects filled out a questionnaire that used 7-point Likert scales to measure the human's perceptions of Hassan's grounding behavior and appropriateness of dialogue responses.

The questions used for the first experiment are shown in Table 3 and the questions for the second experiment are shown in Table 4. The questions were slightly modified in the second experiment based on feedback from the first experiment. The questions were developed to target a number of issues in the use of grounding in dialogue, but are stated in a colloquial way to overcome the difficulty of handling the human subjects' non-technical understanding of words such as "belief" and "grounding." The first two questions are aimed at perceptions of mutual belief, which are integral to grounding. Question 1 addresses the human's perceptions of the virtual human's beliefs and goals, and Question 2 addresses the human subject's perceptions of the virtual human's beliefs about the human's beliefs. The next two questions measure the extent to which the human detected a difference in grounding behavior between the two conditions. Question 3 addresses whether the human subject can perceive the virtual human's attempts at understanding the human's beliefs, and Question 4 addresses whether the human can perceive the virtual human's attempts at having the virtual human's beliefs understood. (In the first experiment, Question 4 focused on whether the grounding behavior related to attempts to resolve problems were human-like.) The final two questions measure the extent to which the human detected improvements in the virtual human as a dialogue system. Question 5 is about quality of responses, and Question 6 is about human-like behavior, which is a desireable quality in a virtual human.

5.2 Improvement Produced By Grounding Component

To study the first hypothesis we compared a version of Hassan which had a grounding component implementing the Degrees of Grounding model to a control version of Hassan which had no grounding component, so all replies were made by the Response Generation component.

Table 3 shows that on average, subjects identified improvements in the grounding condition over the control condition in all measures. In particular, Question 2, how well Hassan seemed to understand what the human was talking about, and Question 3, how much effort Hassan seemed to put into trying to understand the human, both had $p<0.05$. Question 5, regarding appropriateness of response, had the strongest effect with $p<0.01$, and was identified as the focus for confirmation in the second experiment.

Q.	Text	Mean: Grounding	Mean: Control	p
1	Did you have a sense of what Hassan wanted?	3.7	2.5	0.084
2	How well did Hassan seem to understand what you were talking about?	3.6	2.8	**0.035**
3	How much effort did Hassan seem to put into trying to understand you?	4.3	3.5	**0.026**
4	When Hassan had problems understanding you, how human-like were his attempts to resolve the problems?	3.7	3.1	0.11
5	How appropriately did Hassan respond to what you were saying?	4.0	2.9	**0.0087**
6	Taken as a whole, how human-like was Hassan as a conversation partner?	3.2	2.7	0.19

Table 3: Questions and Results for Experiment 1

Q.	Text	Mean: Grounding	Mean: Control	p
1	How strong was your sense of what Hassan was thinking?	4.5	3.7	0.06
2	How strong was your sense of what Hassan thought you were saying?	4.4	3.2	**0.011**
3	In terms of the things he was saying, how good of an effort did Hassan make at trying to understand you?	3.5	3.6	0.44
4	In terms of the things he was saying, how good of an effort did Hassan make at trying to be understood?	4.1	3.8	0.17
5	In general, how appropriately did Hassan respond to what you were saying?	4.0	3.1	**0.015**
6	Taken as a whole, how human-like was Hassan as a conversation partner?	3.9	3.5	0.084

Table 4: Questions and Results for Experiment 2

5.3 Improvement Compared to Non-Methodical Grounding

To study the second hypothesis we compared a version of Hassan which had a grounding component implementing the Degrees of Grounding model, to a version of Hassan which had a baseline grounding component that produced grounding utterances as frequently as the Degrees of Grounding component, but in a non-methodical way. To produce this baseline, we studied the number of explicit grounding statements made in the first experiment and developed a control grounding component that would produce approximately the same distribution of grounding statements as the Degrees of Grounding component. This control component would produce grounding statements randomly but appropriately: it would not try to repeat back material that it did not understand, for example.

On average, subjects rated the Degrees of Grounding condition higher than the control condition in all measures with the exception of Question 3. In this experiment, Question 2 was shown to have a $p<0.05$, and Question 5, regarding the appropriateness of response, was confirmed as a statistically significant variable with a $p<0.05$.

6 Discussion

The advantage provided by the Degrees of Grounding grounding component is exemplified in Figure 6, which was taken from these experiments. In line 1 the Trainee asks who Hassan is working for. However, errors in the ASR and NLU components work together to misinterpret the utterance as an offer to protect Hassan's family.

Without the grounding component, Hassan would reply only with 2b: an acceptance of the offer. However, since the Trainee is unaware that Hassan misinterpreted the question,

1	Trainee	Who are you working for? *Topic: Protection for Family*
2a 2b	Hassan	So, you offer to protect my family. I do trust you, and hope that the U S can fulfill these expectations.
3	Trainee	Yes we will protect your family.

Figure 6: Example Dialogue

the Trainee would not be aware that Hassan expects his family to be protected, and the expectations mentioned by Hassan would make no sense. If this were just social talk such a misunderstanding might not matter, but here the topic is one of great importance. The Degrees of Grounding component helps this problem by identifying that because the topic of Protection for Family has a high grounding criterion, it should repeat back Hassan's understanding of the topic of the Trainee's utterance, as it does in line 2a. Because of this, the Trainee realizes that Hassan believes he has been offered protection, and in line 3 the dialogue is able to proceed productively despite the misunderstanding.

7 Conclusions and Future Work

A set of experiments has shown that the Degrees of Grounding model provides advantages over baseline models in several measures in terms of human perceptions of grounding behavior and appropriateness of response. This is also a demonstration that the Degrees of Grounding model, which was previously developed in the context of a corpus of radio-based dialogues [Roque and Traum, 2008], can also be applied to

dialogues with virtual humans.

Future work includes comparing the Degrees of Grounding model to other grounding models such as those of [Traum, 1994] and [Paek and Horvitz, 2000], which will require addressing several procedural issues such as identifying a fair testbed and determining appropriate evaluation metrics. The Degrees of Grounding model can continue to be applied to new domains to explore its use, in particular when other modalities such as those related to vision are added. Finally, another research direction involves determining how non-cooperativeness and personality should affect the grounding behavior of a virtual human: for example, if Hassan is unwilling to help the Trainee he might not wish to produce evidence of understanding to help bring material to the grounding criterion; alternately, if Hassan has a neurotic personality, he might produce explicit grounding behavior where others might not.

Acknowledgments

This work has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

[Army, 2006] Department of the Army. Police intelligence operations. Technical Report FM 3-19.50, Department of the Army, 2006. Appendix D: Tactical Questioning.

[Bohus and Rudnicky, 2005a] Dan Bohus and Alexander Rudnicky. Error handling in the RavenClaw dialog management architecture. In *Proceedings of HLT-EMNLP-2005*, 2005.

[Bohus and Rudnicky, 2005b] Dan Bohus and Alexander Rudnicky. Sorry, I didn't catch that! - an investigation of non-understanding errors and recovery strategies. In *Proceedings of SIGdial-2005*, 2005. Lisbon, Portugal.

[Bunt *et al.*, 2007] Harry Bunt, Roser Morante, and Simon Keizer. An empirically based computational model of grounding in dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[Clark and Brennan, 1991] Herbert H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. APA Books, 1991.

[Clark and Schaefer, 1989] Herbert H Clark and Edward F Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[Gratch and Marsella, 2005] Jonathan Gratch and Stacy Marsella. Some lessons for emotion psychology for the design of lifelike characters. *Journal of Applied Artificial Intelligence*, 19(3-4):215–233, 2005. Special issue on Educational Agents - Beyond Virtual Tutors.

[Litman *et al.*, 2006] Diane Litman, Julia Hirschberg, and Marc Swerts. Characterizing and predicting corrections in spoken dialogue systems. *Computational linguistics*, pages 417–438, 2006.

[Nakatani and Traum, 1999] Christine Nakatani and David Traum. Coding discourse structure in dialogue (version 1.0). Technical Report Technical Report UMIACS-TR-99-03, University of Maryland, 1999.

[Paek and Horvitz, 2000] Tim Paek and Eric Horvitz. Conversation as action under uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 455–464, 2000.

[Paek, 2003] Tim Paek. Toward a taxonomy of communication errors. In *Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 53–58, August 28-31 2003. Chateau d'Oex, Vaud, Switzerland.

[Paul, 2006] Matthew C. Paul. Tactical questioning: human intelligence key to counterinsurgency campaigns. *Infantry Magazine*, Jan-Feb 2006.

[Roque and Traum, 2007] Antonio Roque and David Traum. A model of compliance and emotion for potentially adversarial dialogue agents. In *The 8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[Roque and Traum, 2008] Antonio Roque and David Traum. Degrees of groundedness based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008.

[Skantze, 2005] Gabriel Skantze. Galatea: a discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Proceedings of SigDial*, pages 178–189), 2005. Lisbon, Portugal.

[Swerts *et al.*, 2000] Marc Swerts, Diane Litman, and Julia Hirschberg. Corrections in spoken dialogue systems. In *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP-2000)*, October 2000.

[Traum *et al.*, 2005] David Traum, William Swartout, Stacy Marsella, and Jonathan Gratch. Fight, flight, or negotiate: Believable strategies for conversing under crisis. In *5th International Conference on Interactive Virtual Agents*, 2005. Kos, Greece.

[Traum *et al.*, 2007] David Traum, Antonio Roque, Anton Leuski, Panayiotis Georgiou, Jillian Gerten, Bilyana Martinovski, Shrikanth Narayanan, Susan Robinson, and Ashish Vaswani. Hassan: A virtual human for tactical questioning. In *The 8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[Traum, 1994] David R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.