Evaluation of an Information State-Based Dialogue Manager

Antonio Roque, Hua Ai[†], and David Traum
USC Institute for Creative Technologies
13274 Fiji Way
Marina Del Rey, CA 90292
{roque,traum}@ict.usc.edu

[†]Intelligent Systems Program
University of Pittsburgh
210 S. Bouquet, Pittsburgh PA 15260
hua@cs.pitt.edu

Abstract

We describe an evaluation of an information state-based dialogue manager by measuring its accuracy in information state component updating.

1 Introduction

Evaluation of dialogue managers is essential for the development of dialogue systems. However, it can be difficult to separate the performance of a dialogue manager from the performance of the system as a whole. Here we describe an approach towards evaluating the performance of an Information State-based dialogue manager separately from the other components of the dialogue system and the system as a whole.

Our testbed system, Radiobot-CFF (Roque et al., 2006), is a military virtual reality environment designed to train soldiers in artillery strike requests. The trainees hold a radio dialogue with Radiobot-CFF during which an enemy target is located and attacked. Radiobot-CFF includes a speech recognition component, a dialogue move interpreter, and an information state-based dialogue manager (Roque and Traum, 2006). We ran an evaluation of the system from which we calculated task completion rates and time-to-task measures for the system as a whole, as well as error rates for the speech recognition and interpreter components (Robinson et al., 2006). However, we lacked an analysis of the dialogue manager component's performance.

2 Evaluation

Radiobot-CFF uses an information state-based (Traum and Larsson, 2003) dialogue manager, and therefore works by firing update rules which are dependent on and which change information state components. For example, Radiobot-CFF uses information state components to track whether it has received a target's location and what that target location is, as well as whether it has enough information to send a fire. To evaluate the performance of our dialogue manager, we studied how well it updated its information state components.

2.1 Approach

Our approach is to use human coders to decide how the information state components should be updated, given a sequence of utterances, and to compare that to how the system actually does update its information state components.

We develop a coding manual of guidelines for updating the information state components based on the kind of input received. We then use a sequence of trainee utterances (produced by hand-transcribing audio logs and hand-correcting system dialogue move interpretations of those utterances) to produce a sequence of hand-coded information state components. That sequence is our gold standard, and represents the output of the dialogue manager if the speech recognition, interpreter, and dialogue manager components are all performing to the level of a human.

We compare our system's performance to this gold standard corpus in two conditions. First, we run the dialogue manager on perfect input by feeding it the hand-corrected interpreter output, recording the information state components after every utterance, and comparing that to our gold standard. This allows us to evaluate the dialogue manager separately from the rest of the system, so that errors in the speech recognition and interpreter components do not affect its performance. Secondly, we compare the gold standard to the system's information state components when updated by the system on actual speech recognition and interpreter input. This allows us to evaluate the dialogue manager's performance given noisy input.

IS Component	*Accuracy, corrected input*	*Accuracy, noisy input*
has warning order	0.76	0.67
has target location	0.98	0.90
has grid location ‡	0.99	0.96
has polar direction	0.83	0.80
has polar distance	0.99	0.91
has target descript.	0.93	0.76
has enough to fire	0.99	0.52
method of control	0.71	0.71
method of fire †	0.38	0.44
grid value ‡	0.98	0.96
direction value	0.83	0.79
distance value	0.99	0.91
adjust fire	0.88	0.65
repeat FFE *	0.89	0.97
LR adjustment	0.99	0.92
AD adjustment	1.00	0.97
end of mission	0.93	0.91
disposition	0.93	0.78
number of casualties	0.95	0.83
mission is polar	0.99	0.85
last method of fire †	0.90	0.61
missions active	0.81	0.67

† Kappa was less than 0.8 and greater than 0.67
‡ Kappa was less than 0.67
* Kappa could not be calculated, as its value never changed in the data over which kappa was measured.

Table 1: Accuracy per IS Component

2.2 Results

We worked with a corpus of 17 sessions consisting of 407 utterances, representing a total of 8954 information state components to be updated. A pair of human coders coded several sessions by consensus to develop a set of guidelines, then individually coded the rest of the corpus. Several sessions were held out for concurrent coding by both coders, from which a kappa score was calculated per information state component. Components had kappa values above 0.8 except as noted in Table 1.

We then fed the corrected utterance interpretations into the dialogue manager to get sequences of IS component updates for corrected interpretations, and processed log files from the full system evaluation to get sequences of IS component updates for noisy interpretations. Accuracy results (measured by number of times the dialogue manager agreed with the human coder) for both are shown in Table 1.

3 Future Work

Because the input used in the corrected input condition is not reacting to the dialogue manager's responses, the dialogue may take an unnatural direction; for example, in which the dialogue manager is repeatedly prompting or correcting the trainee, but the trainee is proceeding as if there is no problem.

Also, a component's value may be more important at certain parts of a dialogue than at others. For example, as shown in Table 1, the "method of fire" component's accuracy is low, but the dialogue manager and humans disagree on its value most often at a phase of the dialogue in which the "method of fire" value is never used in decisions or output.

We hope to quantify and address these problems in future work.

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

Susan Robinson, Antonio Roque, Ashish Vaswani, Charles Hernandez, Bill Millspaugh, and David Traum, "Evaluation of a Spoken Dialogue System for Virtual Reality Call For Fire Training," Submitted, 2006.

Antonio Roque, Anton Leuski, Vivek Rangarajan, Susan Robinson, Ashish Vaswani, Shri Narayanan, David Traum, "Radiobot-CFF: A Spoken Dialogue System for Military Training," 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), Pittsburgh, PA, September 17-21, 2006.

Antonio Roque and David Traum, "An Information State-Based Dialogue Manager for Call for Fire Dialogues," 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia, July 15-16, 2006.

David Traum and Staffan Larsson, 2003. The Information State Approach to Dialogue Management. In R. Smith & J. van Kuppevelt (eds.) Current and New Directions in Discourse and Dialogue. Dordrecht: Kluwer, 325-353.