First Steps Toward Linking Dialogues: Mediating Between Free-text Questions and Pre-recorded Video Answers

Sudeep Gandhe Andrew Gordon Anton Leuski David R. Traum
USC Institute for Creative Technologies, 13274 Fiji Way, Marina del Rey, CA 90292

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742

Abstract

Pre-recorded video segments can be very compelling for a variety of immersive training purposes, including providing answers to questions in after-action reviews. Answering questions fluently using pre-recorded video poses challenges, however. When humans interact, answers are constructed after questions are posed. When answers are pre-recorded, even if a correct answer exists in a library of video segments, the answer may be phrased in a way that is not coherent with the question. This paper reports on basic research experiments with short "linking dialogues" that mediate between the question and answer to reduce (or eliminate) the incoherence, resulting in more natural human-system interaction. A set of experiments were performed in which links were elicited to bridge between questions from users of an existing training application and selected answers from the system, and then comparisons made with unlinked answers. The results show that a linking dialogue can significantly increase the perceived relevance of the system's answers.

1 Introduction

Large video collections are becoming increasingly available, resulting in greater use of pre-recorded video in a broad array of interactive applications. Because video is a linear medium, interaction with a collection is normally achieved by selecting relatively brief segments to present in response to a request.

One simple approach to answering questions that is often used in museum kiosks is to preselect a set of questions from which the searcher can choose using a touch screen interface. Touching a question then calls up the associated video segment.

Information retrieval and question answering offer two alternative paradigms for crafting an answer selection process. The information retrieval paradigm relies on formulation of queries using whatever query language is supported by the system, (e.g., keyword matching), followed by interactive selection of interesting video segments from a list of possible matches. The CMU Informedia system is perhaps the best known example of information retrieval research for video collections (Hauptmann et al., 2003). The question answering paradigm, by contrast, seeks to interpret questions that are posed in natural language and respond with a segment of video that is directly responsive to the question (Voorhees, 2002).

Regardless of the answer selection method, we will naturally sometimes get questions for which we have no perfect pre-recorded answer. Two types of problems can be anticipated: (1) cases in which we simply have no answer available that is on topic, and (2) cases in which we have an answer that is topically suitable, but which is not phrased in a manner that makes it a good answer to the question. The first of these problems can be mitigated (although perhaps not entirely eliminated) by using conversational prompts to constrain the domain of discourse (e.g., "who has a question about ...?") and/or by working with a larger set of answers (oral history collections with more than a million answers now exist). The second problem would likely persist even with very large answer collections, however; there are simply too many ways in which people might pose their questions.

In this paper we look at the problem of coherence mismatch between a question and an answer. We propose that such a mismatch can be reduced (and often eliminated entirely) by constructing a "linking dialogue" between the question and the answer. Figure 1 shows an example linking dialogue from an Army training application prototype, where a question from a human user is linked to a pre-recorded response from a character in a fictional storyline. In this case the answer from Omar does not seem responsive to the question about Captain Young. The linking dialogue serves as a bridge from the question to the answer by explaining a reason for a negative answer, then setting up Omar's response as an amplification of that reason. The linking dialogue both introduced the responding speaker and established an aspect of the context that was not self contained within the answer. Identifying these key goals is the task of the link *construction* process. Once we have that information we can proceed to the link *generation* – producing a fluent utterance such as that shown in the example.

If we are to automate the link generation and construction processes, we must first understand the nature of those tasks. We have therefore conducted a series of elicitation experiments using actual questions and pre-recorded answers from users of an existing multimedia training application. In the next section, we describe that application. We then present our elicitation pro-

Question: Did CPT Young have time to ask Omar about Omar's ethnic group, clan, tribe, village, and family in this situation?
Link: Omar may not have wanted Captain Young to ask about these things...so what did he do to avoid them? Let's ask him.
Answer <Omar>: I did not want your commander questioning who I was or why I was here. So I presented him with a larger problem, one he would feel equipped to solve. A fight. In America, your armies don't fight with each other. Your commander assumed that is true in my country, too. It was an unfortunate assumption.

Figure 1: An example of a linking dialogue.

cess and an analysis of the human-created linking dialogues that resulted. The paper concludes with an assessment of the feasibility of automating the link construction process and some thoughts on future work.

2 Initial Testbed: TLAC-XL

A persistent goal of the US Army has been to successfully author computer-based training applications to support the development of US Army soldiers with strong leadership ability. The Army Excellence in Leadership project at the University of Southern California's Institute for Creative Technologies has been developing technologies that support a case-analysis model of instruction in software-based training. The first prototype developed in this project was the Think Like a Commander: Excellence in Leadership (TLAC-XL) prototype application, described by (Hill et al., 2003). Targeted at junior US Army officers, the TLAC-XL system presents users with short video depicting a fictional scenario where leadership challenges arise, and then engage users in an after-action analysis with a virtual mentor and virtual characters from the fictional storyline. In the first TLAC-XL prototype, the video depicts a security mission for a food distribution operation in Afghanistan, entitled *Power Hungry.* In the scenario, a new company commander (CPT Young) is seen making a number of questionable leadership decisions, which ultimately lead to the failure of the mission as a food riot erupts among the civilian population.

After watching the video, users of the TLAC-XL application engage in a natural-language dialogue with a virtual mentor and virtual characters from the fictional storyline. The virtual mentor (visualized as an animated talking head) directs this dialogue by questioning the user about the fictional scenario from a tactical military perspective, and then introduces storyline characters (e.g. CPT Young) that are interviewed by the user to gain a deeper understanding of the leadership issues that were embedded into the storyline. Users of the TLAC-XL application interview six storyline characters, where natural-language questions typed by users are answered with pre-recorded video clip responses. Selecting the appropriate pre-recorded video clip to play as a response is accomplished using a statistical approach to text classification based on a Naive Bayes machine-learning algorithm. (Hill et al., 2004) report an average accuracy of 52% in selecting the most appropriate available response to users questions (percent agreement, 10-fold cross-validation) for 16-class classification tasks with very modest amounts of training data (an average of 356 examples per training set).

The TLAC-XL prototype of the Army Excellence in Leadership project is a good initial testbed for piloting linking dialogues. The question analysis and answer selection components already exist, as does a corpus of questions culled from user logs. For our experiments, we operated the system in a mode where the user asked questions to all six of the storyline characters as a group. Our aim was to investigate how these dialogues could be improved through linking dialogues to better meet the training goals of this prototype application.

3 First Experiment: Human Off-line Creation of Linking Dialogues

Our ultimate goal is to automatically construct linking dialogues, but before that can be done we need a way of characterizing what constitutes a good linking dialogue. We have identified several concrete steps along the way to producing automated link dialogues to improve the coherence between spontaneous questions and pre-recorded video. First, the big question is whether links will actually improve the experience and whether it is possible to generate (good) links. We start with human created links, and evaluate their utility. The next step is to analyze these links and uncover what it is about their content and structure that improves the linking dialogue over the question answer pair. From that point we will be in a position to start to automate the processes. Finally, we must measure whether the computer generated links also have a benefit. As of this writing, we are still in the first phases of the process — we have enabled construction of human links, have shown that they can improve the experience, and have started to identify types of successful links.

Our first question is whether links, even human created links, can improve on the coherence of the interactions. We started with an off-line experiment in which five volunteers familiar with the domain built linking dialogues for the same set of questions. We selected 20 questions from the TLAC-XL user logs and presented each volunteer with the responses associated with the five most probable classifications, as shown in Figure 2. The volunteers then chose what seemed to them to be the best available response among the five and manually constructed a linking dialogue that could be interposed between the question and the answer (and optionally following the answer). Each volunteer created linking dialogues for the same set of 20 questions. One interesting result of this study was that we observed that different answers could be selected for the the same question without loss of coherence (given suitably crafted links).

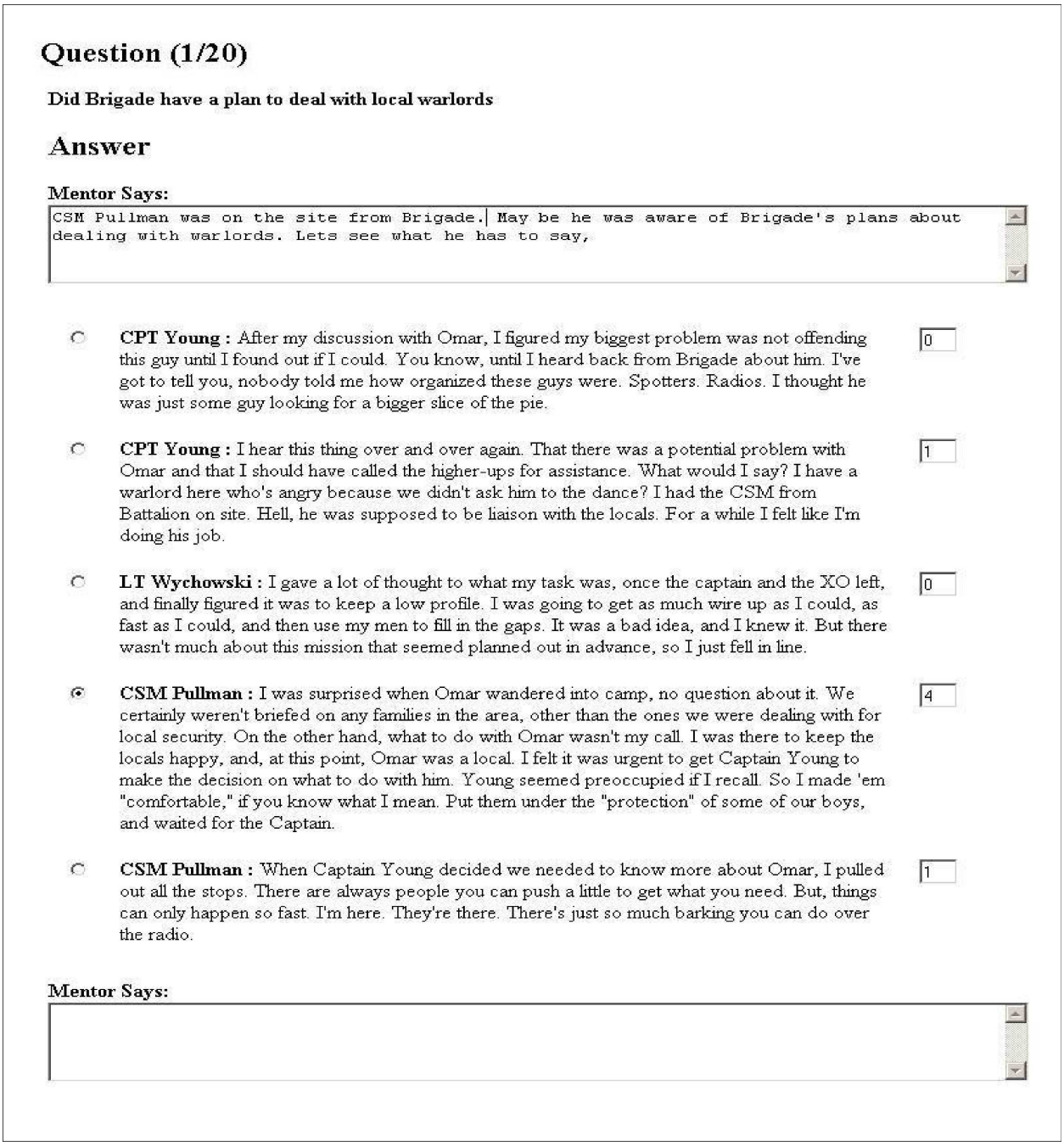

Figure 2: Mediation Interface for first experiment

4 Types of linking Dialogues

From analysis of the first experiment, we observed that Linking dialogues were created to perform several types of functions. In the types below, dialogues consist of three parts: (Q) - the original question asked by a TLAC-XL user, (A), an answer included in the set provided by the TLAC-XL system, and (L), a link produced by our link-writers to connect (Q) to (A). Not all of the responses (L+A) are perfect answers to the questions, but they tend to improve over just (A) (see also section 5.

4.1 Speaker-intro

These links introduce the speaker that will provide the answer, sometimes explaining why that speaker was chosen. For example,

Q: Could LT Perez, as the XO, reshaped the command climate with a quick one-on-one chat with the CPT?

L (mentor): SGT Jones had worked with Lt Perez for quite some time, lets hear from him what Perez could have done.

A (SGT Jones): I've worked with the XO and Lieutenant Wychowski for about two years, give or take. I can pretty much tell you what they're going to do in any given situation. The XO's a good organizer, but a bit weak on strategy. He needs a CO who sees the big picture. Wychowski tries hard, but he's not much for spontaneity. He's got to go through a situation once or twice, figure out what's going on, before you can cut him loose.

4.2 Coherence

These links bridge the gap between the question and an answer to that question in order to improve the coherence of the answer. Coherence links focus on explaining why the chosen answer is most appropriate for the question. For example, Omar's answer below does explain (from his point of view) why the CPT did not meet Mohammed, however some of the framing is missing, and filled in with the link.

Q: Why didn't CPT meet with Mohammed?

L (mentor): CPT Young did not know anything about Omar and Mohammed. Mohammed distracted CPT Young. And this is what Omar has to say about it.

A (Omar): I did not want your commander questioning who I was or why I was here. So I presented him with a larger problem, one he would feel equipped to solve. A fight. In America, your armies don't fight with each other. Your commander assumed that is true in my country, too. It was an unfortunate assumption.

4.3 Reformulation

When no good answer is found, a reformulation link offers an answer to a related question. In many cases, we will know the original question that prompted a prerecorded answer, so a reformulation link can be thought of as bridging between the asked question and the target question and then exploiting the natural coherence of the target question with its answer. In the following example, the question is changed from "has Pullman worked with the warlords" to "Did Omar knew Pullman".

Q: Has CSM Pullman ever worked with the warlords before?

L (mentor): Omar would know whether CSM Pullman had worked with warlords before. Let's ask him if he knew anything about CSM Pullman.

A (Omar): This CSM Pullman has been in many of our villages before. It was possible he might recognize me. Had he asked to drink tea with me I would have to agree or risk insulting him. But, so what? We would become friends, and I would ask him for protection from an angry Mohammed, my brother-in-law, and my greatest enemy. In the end, nothing would change.

4.4 Summarizing

A link can also appear after the answer, to focus on the important part or draw out an implicit connection. For example:

Q: Would more time have helped?

A (Pullman): This mission went to hell in a hand basket pretty quick. To tell you the truth, I was surprised the Captain didn't ask me to step in as his XO, given Lieutenant Perez was someplace north of nowhere. I didn't want to look like I was throwing my weight around, but maybe I should have jumped in.

L (mentor): If CPT Young had time to think about either calling back LT Perez or asking CSM Pullman to step in, the situation might have been better handled.

4.5 Guiding

When a system has conversational goals (as in our training scenario), the system could use linking dialogues to bring the user back on track if the conversation starts to drift from the intended focus. Even when no conversational goals are embedded in the system (as might be the case in an oral history collection, for example), knowledge of the topical coverage of the collection could be used as a basis for creating guiding links. Guiding links are generally focused on influencing subsequent questions. For example, the phrase "even CSM Pullman thought so" in the example below helps guide the user to ask the next question about CSM Pullman.

Q: Why did CPT Young spend so much time with the warlords and comparatively less time with his subordinates?

L (mentor): Local customs require him to spend time with warlords talking about their history & background. Even CSM Pullman thought so. Lets hear more from our CPT...

A (CPT Young): Use of your time is one of the things you have to learn as a commander. I look at the mission and I see what I can do and what my men can do. Warlords, they like to deal with the chief. That's their culture. I can't delegate that. But, standard company tasks, like crowd control, logistics, security, that's something I expect my men can handle. If not, I'm working with the wrong people.

These categories are not exclusive; for example, speaker-intro and coherence linking often occurs together. This is not yet an exhaustive set of functions, but it begins to reveal the range of conversational goals that might be explored. Another question that we need

to explore is whether different applications (e.g., interaction with oral histories in a museum setting) would yield different types linking dialogs.

Our current work is involved in discovering the most useful link functions and formalizing them sufficiently to be able to consistently identify them and generate them automatically.

5 Second Experiment: Quality of linked vs unlinked dialogues

In our second set of experiments, we sought to assess the effect of linking dialogues on perceived quality of the interaction. There are many ways in which we might measure "quality," including the (perceived) relevance of the reply, the coherence between the question and the linking dialogue, the coherence between the linking dialogue and the answer, and overall satisfaction. For the initial experiments reported in this paper we chose an overall satisfaction measure: do links, on the whole, improve the user's satisfaction with the response when compared to the same pre-generated answer without a linking dialogue? To answer this quantitatively, we selected a set of dialogue instances to be evaluated by multiple judges. 40 question-link-response triples were examined. For each triple, there was one instance with all three elements and one with only the question-answer pair, for a total of 80 instances to be evaluated. We present each human judge with 20 instances: 10 with links and 10 without. No subject saw both instances of the same triple. Order of presentation was balanced with respect to link vs non-link instances and overall presentation of an within instance the sequence.

1	Response is not related in any way the question
2	Response contains some discussion of people or objects mentioned in the question, but does not really address the question itself
3	Response partially addresses the question, but little or no coherence between the question and response
4	Response does mostly address the question, but with major problems in the coherence between question and response; seems like the response is really addressing a different question than the one asked.
5	Response does address the question, but the transition is somewhat awkward
6	Response answers the question in a perfectly fluent manner

Table 1: Dialogue Quality Scale

Each instance is rated from 1 (bad) to 6 (good), using the guide in table 1. The instructions include examples of instances with and without links for each score value. Evaluators also had an optional "comment" field for further observations on the question-response pair. Figure 3 shows preliminary quantitative results, with the distribution of judgments for instances with and without the links. Linking dialogues have significantly higher scores ($t = 2.032, DF = 118, p < 0.05$). Based on the evaluator comments, it seems that the perceived difference in quality is even greater than the numerical value, because of the conflation of response accuracy with response coherence — improved coherence for not very good answers sometimes did not lead to higher scores.

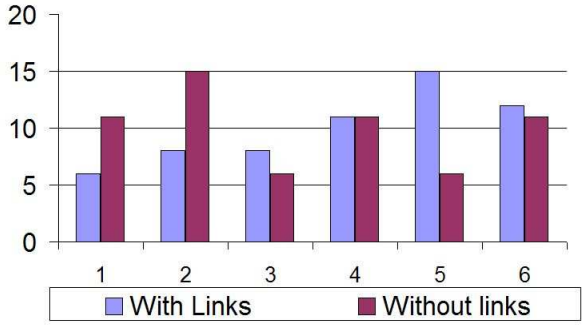

Figure 3: Quality Score Frequencies: the number of responses for each rating are shown for responses with links vs without links

6 Third Experiment: on-line linking dialogue creation

Finally, we also conducted a third set of experiments in a "Wizard of Oz" setting, with live TLAC-XL users receiving manually constructed linking dialogues accompanying the presentation of the pre-recorded video clip in response to their question. This was useful for several purposes. First, we needed to address the issue of whether users would find mediated answers to their own questions more satisfactory than just selected answers. Experiment 2 had evaluators judging dialogues created by others, whereas here the users got immediate feedback for their own questions in a manner parallel to the standard TLAC-XL usage. Secondly, this design allowed for subsequent questions to be influenced by previous replies, including the linking dialogues. The use of a human-in-the-loop could also result in better classification accuracy than would have been achieved in our present fully automatic system. The drawback of this design, however, is that it significantly increased the response time between when users finished typing their questions and when they saw the responses. Figure 4 shows the user's view – it is the same as the normal view for TLAC-XL user, except that the human-generated link from the mentor appears after the question and before the answer. To cover the time needed for composing and writing the links, a request for the user to wait for an answer is also produced. Figure 5 shows the view of the link-producer. Whenever a question is typed in by the user, the question and set of 5 most likely answers are provided to the link-producer. He then must choose the best answer (confidence scores are given on the right, as well as an answer-identifier), and (optionally) type in a link, either before or after

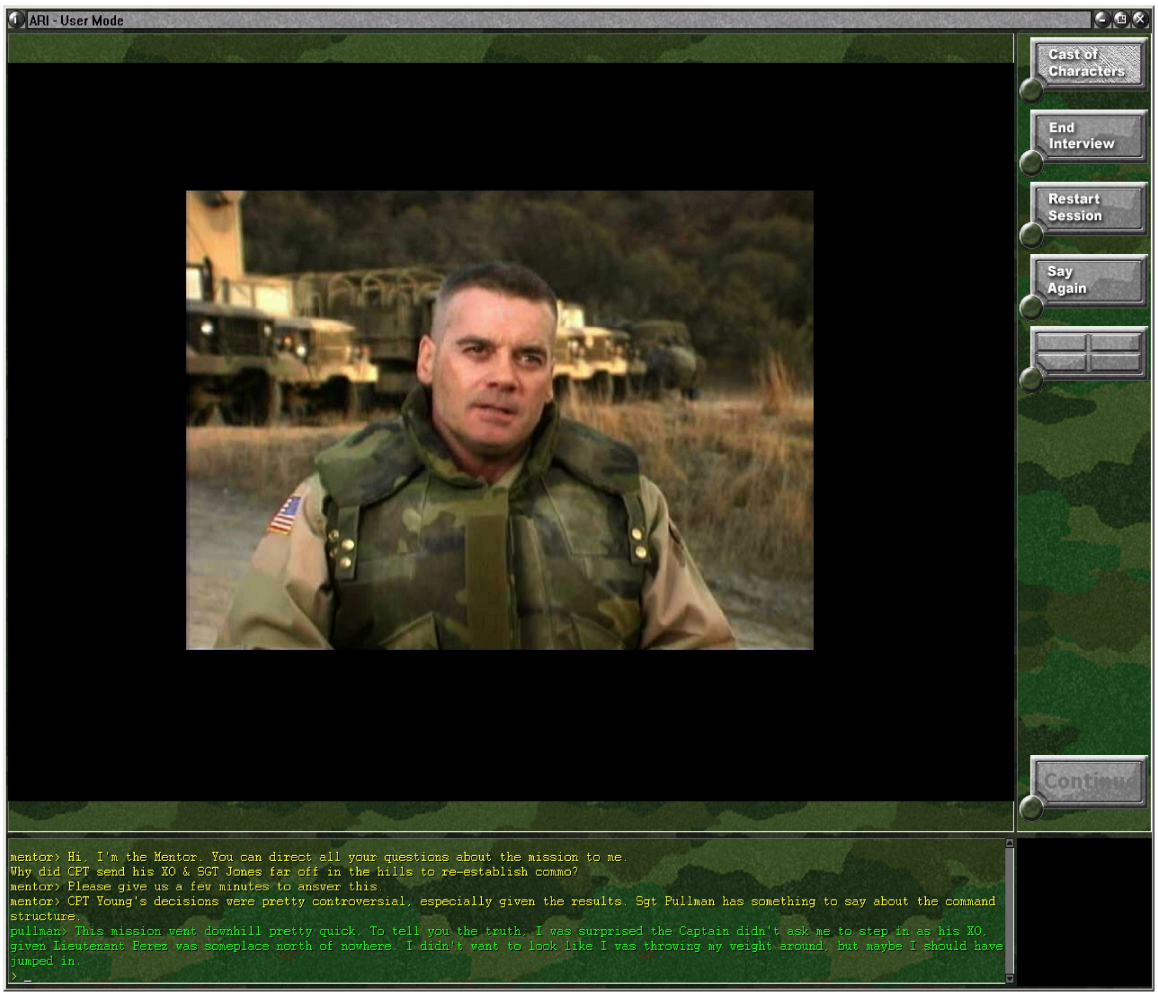

Figure 4: User View in Experiment 3

the answer. When the link-producer completes the answer, the link (if any) is shown to the user, and then the selected answer is played.

To date, we have collected four such dialogues (averaging 10 exchanges per dialogue). Although this is not a sufficient basis for quantitative evaluation, post-session interviews with the users of the system suggest that the link dialogues do have perceived benefits in spite of the longer delays.

7 Discussion and Future Work

There are several next steps for this work. First, more evaluation of the quality of specific types of links is needed. Our next step will be to look in more detail at which kinds of links improve which kinds of question/answer pairs. We plan to start by labeling our links according to an expanded set of the functions described in the previous section.

We also need to explore the limiting cases in more detail. When the coherence between the question and answer is high, adding a link may actually be harmful. So we need to investigate how to determine whether or not to add links. At the other extreme, it may happen that a question can't be answered by any available response. In such cases, an explicit topic change may be the most satisfactory outcome. For example,
Question: *What happened to all that food after the fight?*
Link: *I don't see how that's relevant. Let's instead listen to Captain Young how he thinks this outcome could have been avoided.*
Answer: < CPT Young > *"If I had the chance to do one thing differently, I would bailed out of that site. Postponed the whole operation until we found a location we could defend. My understanding was I didn't have that...discretion here. But, it was ultimately my responsibility to make the call."*

Once we have a better understanding of the nature of links and their applicable uses, we will be in a position to tackle the difficult problem of automatic link dialogue generation. While, in the extreme, the problem will be very challenging (combining some of the toughest problems in question answering, dialogue management and text generation), we expect to be able to make progress by starting with simple versions of existing technology

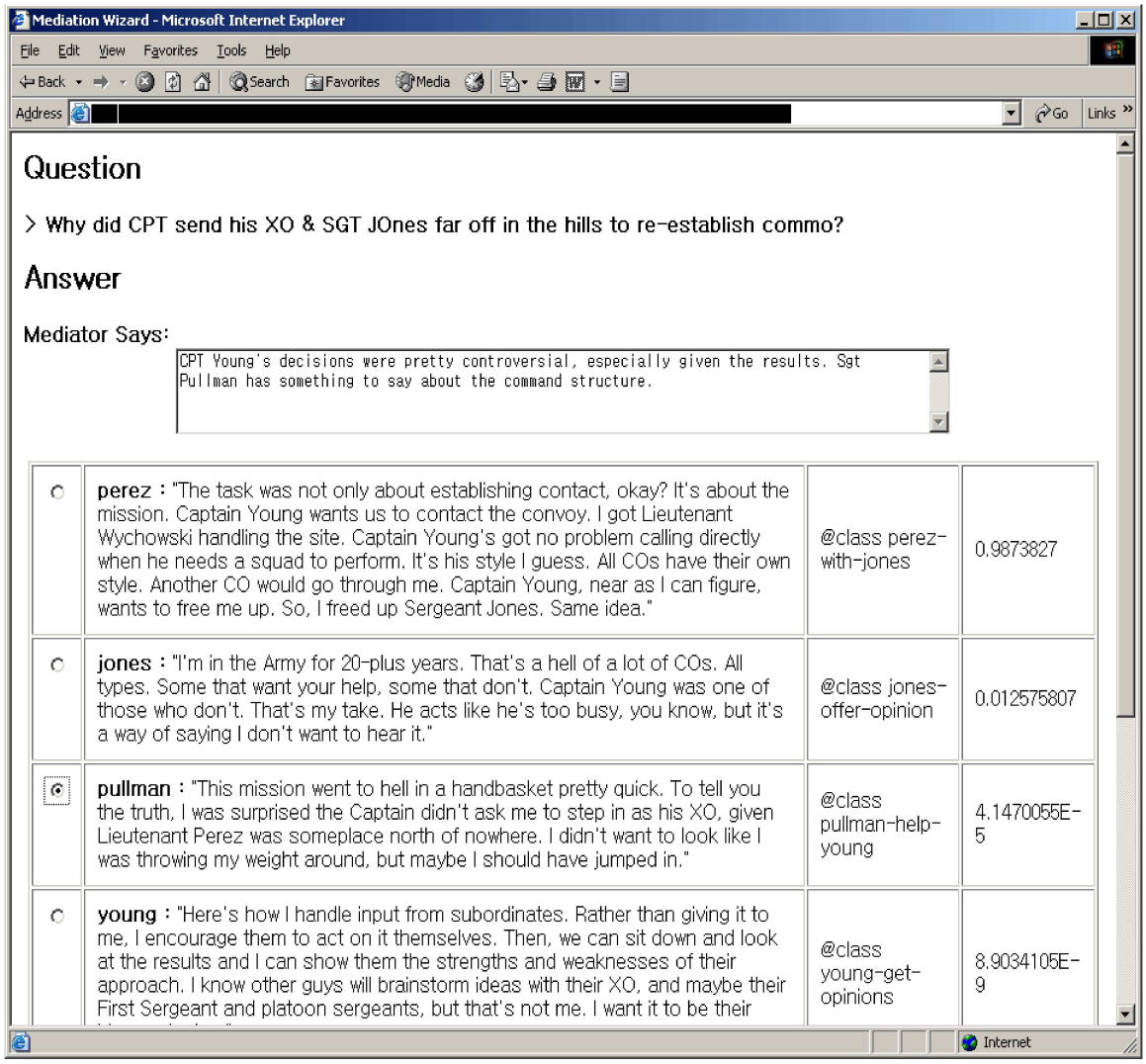

Figure 5: Wizard View in Experiment 3

components for these problems.

In the long run, we are interested in expanding this work in two other directions. The first is to work with other types of collections. We are presently working with a large (116,000-hour) collection of extensively annotated oral history interviews using an information retrieval paradigm, and recasting that problem as question answering would open the potential for new application scenarios. Finally, we want to extend our work to accommodate spoken questions. If we can ultimately combine large collections, accurate answer selection, fluent construction of linking dialogues, and correct recognition of spoken questions, our goal of creating effective new ways to interact with prerecorded media would be realized.

8 Acknowledgments

This paper was developed in part with funds of the Department of the Army under ARO contract number DAAD-19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army. NSF IIS award 0122466 (Malach) also supported this work. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Hauptmann, A. G., Jin, R., and Ng, T. D. (2003). Video retrieval using speech and image information. In *Electronic Imaging Conference: Storage and Retrieval for Multimedia Databases*. http://www.informedia.cs.cmu.edu/documents/ei03_haupt.pdf.

Hill, R. J., Douglas, J., Gordon, A., Pighin, F., and van Velsen, M. (2003). Guided conversations about leadership: Mentoring with movies and interactive characters. In *Proceedings of the Fifteenth Innova-*

tive Applications of Artificial Intelligence Conference (IAAI-03).

Hill, R. J., Gordon, A., and Kim, J. (2004). Learning the lessons of leadership experience: Tools for interactive case method analysis. In *Proceedings of the Twenty-fourth Army Science Conference*.

Voorhees, E. M. (2002). Overview of the TREC 2002 question answering track. In *The Eleventh Text Retrieval Conference (TREC-2002)*. http://trec.nist.gov.