Evaluating a Script-Based Approach for Simulating Patient-Doctor Interaction

R. Dickerson, K. Johnsen, A. Raij, B. Lok
Computer and Information Science and Engineering Dept.
University of Florida

J. Hernandez, A. Stevens
VA Hospitals
University of Florida

Keywords: Virtual Characters, Natural Language Processing, Human-Computer Interaction.

I. INTRODUCTION

Good communication between a patient and a doctor has long been accepted as essential for quality health care. “Good patient-clinician communication leads to better clinical outcomes and more satisfied patients. Poor communication leads to poor outcomes, dissatisfaction, and malpractice litigation [Coulehan and Block 2001].” *How* to educate medical students on these communication skills is the difficult issue for medical educators. To aid in teaching, practice, and formal evaluation of the patient-doctor interview, an ordered interview structure is commonly taught. The patient-doctor interview generally progresses through pre-defined *stages*, however, the conversation can diverge as new information is disclosed. Medical students are trained to follow this predictable path when taking a medical history (Table 1).

Greeting
Chief Complaint
History of Present Illness
Other Active Problems
Past Medical History
Family History
Social History
Review of Symptoms
Physical Exam
Impression and Plan

Table 1 - Medical Interview Stages

Until the 1970s, interviewing skills were not a part of most medical school curriculums. The prevailing attitude was “you’ll pick it up as you go along” [Coulehan and Block 2001]. Educators now believe that the “art of medicine” can be taught. The primary method is through textbooks and lectures [Bickley et. al. 2002]. As one might expect, these approaches have difficulty in presenting the subtleties, feedback, and practice required for advanced proficiency in interpersonal communication.

Another pedagogical approach involves employing standardized patients (SP’s). Medical students practice diagnosis with SP’s - actors that represent a condition. SP’s responses are based on a set script for a given condition. Over 94 medical schools in the US and Canada have integrated SP’s into their curriculum. Educators agree that good diagnostic skills can come only through repeated exposure to (initially) SP’s and (eventually) real patients.

Today, “actor training and availability, reproducibility, changing evaluation criteria, and implementation cost” have spurred research into using virtual SP’s as an alternative to hiring actors. Further, repeated practice in virtual environments leads to good decisions in the job [Hubal et. al. 2000]. However, computer systems are not capable of simulating interpersonal communication at a high level of fidelity. Can current simulations provide an adequate level of simulation to enable training, teaching, and evaluation of communication skills? Would the predictable structure of taking a medical history for basic conditions allow them to be simulated so that students can be educated on the interpersonal skills?

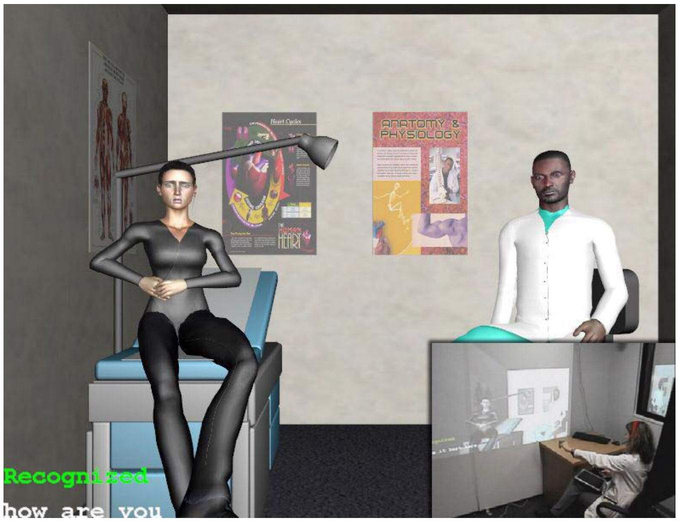

Figure 1 - Virtual patient and instructor application. (Inset) The user interacts naturally with the characters with speech and gestures.

We have applied an immersive virtual character system to this task. Medical students interact with life-sized virtual characters using natural speech and gestures (Figure 1). We believe this high level of immersion greatly facilitates the system effectiveness. However, the virtual characters animations, behaviors and responses are driven by a straight-forward question-and-answer model with scripted responses. Further, as the system was still in development, the virtual character's script matched only 60% of student queries. Surprisingly, a pilot study suggests that even with this limited, basic approach, effective communication skills education could still occur!

In this paper, we detail our approach to modeling the patient-doctor interview using scripts and multimodal interaction. Then, we explore the lessons learned from an initial group of medical students, their performance and end-user feedback on the applicability of the system. Finally, we try to derive insight into why this approach appears to be a fruitful path to follow, the possible extents, and future development directions.

II. PREVIOUS WORK

Simulating interpersonal situations, such as conversations and interviews, has recently sparked interest in researchers, companies, and end-users alike. Interactive Drama Inc.'s Conversim [Harless et. al. 2003] simulates an interview using video clips of expert or patients. Scenarios include HIV risk assessment, breast cancer information, brain injury education, and talks with amputees and family caregivers. The system displays the choices of questions that can be asked at each stage.

Systems that specifically simulate the patient-doctor interview have also been created. IDI's "Medical Spanish" gives doctors practice taking a medical history in Spanish. A virtual instructor helps the student with pronunciation and learning medical history questions in Spanish. Other patient-doctor scenarios include diagnosis depression and US Army National Guard medical training.

Research Triangle Institute's AVATALK system provides natural language processing, emotion and behavior modeling, and facial expression and lip shape modeling for a natural patient-practitioner dialogue [Hubal et. al. 2000]. The scenarios are pre-defined, but the interaction itself is unscripted. The conversation flow varies from interview to interview. Other RTI projects provides law enforcement personnel practice conversing with the mentally ill. RTI's products use advanced natural language processing (NLP) for processing speech input [Hubal et. al. 2003].

Conversational agents are a growing research field and employ many different approaches to simulating the flow of a conversation. The popular A.L.I.C.E. chatbot uses a markup language, AIML (Artificial Intelligence Markup Language) to store over 41,000 categories in its knowledge-base in stimulus-response architecture [Wallace 2002]. AIML organizes knowledge into "categories" as a basic unit of knowledge. The multiple input "patterns" are matched to response "templates" by *symbolic reduction*.

III. SYSTEM DESCRIPTION

System Overview

A 19 year-old Caucasian female looks at you and says, "my side hurts, please do something!" This is a common situation practiced with SP's. We simulated this condition using interactive virtual characters. The patient is DIANA (Digital Animated Avatar), and she is complaining of abdominal pain. Accompanying the student in the encounter is an instructor, VIC (Virtual Interactive Character). DIANA and VIC's scripts were written by medical faculty at the Shands Hospitals at the University of Florida.

The student interacts with DIANA and VIC using speech and gestures. To capture this information, the student wears a headset microphone and colored markers on the headset and hand. Two webcams track the color markers for proper perspective-based rendering and gesture recognition. A tablet PC is used to deliver the patient's vital signs on entry, and for note taking. The scene is rendered and projected at life-size. The setup is installed at the Harrell Professional Development and Testing Center at Shands Hospitals at the University of Florida, the current site for testing and training with real standardized patients. The students are given 10 minutes to arrive at a differential diagnosis for DIANA. The system (which is not the focus of this paper) is more fully discussed and evaluated in [Johnsen et. al. 2005].

The Dialog Manager handled the conversation flow between the student and virtual characters. It is composed of a Speech Manager and Gesture Manager.

Speech Recognition

The Speech Manager runs on the Scansoft's Dragon *NaturallySpeaking* engine. Two modes of speech recognition, dictation and command and control (C&C), work simultaneously in a tiered architecture (Table 2). First, C&C captures short expressions, like "uh huh", "hello", "yeah" and matches them directly to the script. For longer expressions, dictation mode is used. In dictation mode, instead of matching an audio input with expected responses directly, the audio is translated into a text string first, and then matching between the text string and expected responses is done. Since the match with dictation does not have to be exact, this allows us to avoid having to enumerate all possible queries like in C&C. We initially tried applying dictation to all audio input, however, since context-sensitivity is often used for short utterances in real conversations, direct matching to many short queries produced many errors.

A disadvantage in using dictation mode is that it requires voice training. Speaker independent systems, such as Sphinx, might be worth pursuing as the patient-doctor conversation typically has a limited vocabulary set. In addition to voice training, Dragon was also trained to the script file. This ensures that the speech engine has the full vocabulary and improves dictation predictions. Once the user has finished an utterance, the recognized text appears on the screen as feedback so that the student is confident that the speech recognition heard them correctly.

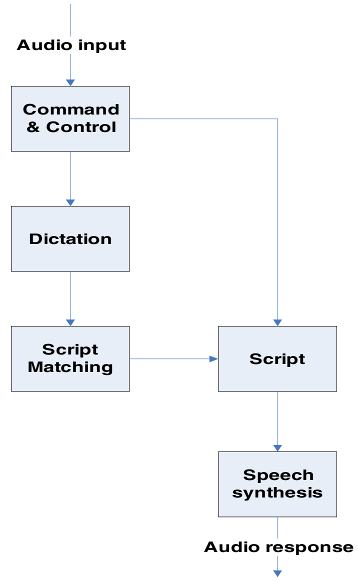

Table 2 - Flow of information in the Speech Manager

Script Matching

The script matching system is similar to *text-database searching [TDS]*, where filters and parsers are used to process natural language queries. "Parsers eliminate noise words (for example, the, of, or in), provide stemming (plurals and alternate endings, and produce a relevance-ranked list of documents based on term frequencies. The systems do not deal with negation, broader or narrower terms, and relationships [Shneiderman and Plaisant 2004]." Many popular web search engines use similar parsers.

It is difficult to anticipate the polysemous nature of natural language queries, but TDS works well for slight variations of similar queries. For example, the query for the chief complaint, can be phrased "how can I help you today?", "can you tell me about you problem?", or "what brought you into the clinic today?" These questions are grouped into a similar semantic block structure, inspired by the stimulus-response structure of AIML.

Like AIML, the script is constructed in XML. Although we do not use wildcard patterns, triggers do not need to be said verbatim for the program to match the input. A matching heuristic is used to determine the similarity of the input ("Could you tell me how old you are now, DIANA?") to an entry in the script ("How old are you?"). The highest matched script entry is used as the audio response. There is a minimum matching threshold, so that some inputs will never get mapped to a script entry. This threshold is difficult to determine, and was empirically determined through testing. The match score is calculated by morphing the input phrase to the matched entry. The costs for adding or subtracting each word are determined by the British National Corpus of word frequencies [Leech et. al. 2001]. For example, the article "the" is the highest frequency word in the English language, and thus the cost of adding or subtracting "the" to a query is very little. "Pregnancy", however, is a relatively rare word and its existence (or lack thereof) in a query is significant information to which script query it should match. This is loosely inspired by other NLP work in Latent Semantic Analysis where semantics are determined through lexical content and not through syntax.

The last response that the character spoke is recorded for two reasons: the student may ask the patient to repeat the last response ("can you please repeat that?") or use minimal facilitators, ("yes?", "uh huh?", "and?", "what else?") to encourage the patient to give up more information. The script contains additional information about addressing these commands.

Gestures

To capture the student's gesture, such as handshaking and pointing ("Does it hurt here?"), a separate color tracking system was used to follow markers on the student's head and hand. The tracking system continuously transmitted tracking information to the virtual character system. By following the path of the student's hand and head, gestures could be detected and a gesture trigger sent to the Dialog Manager. Gesture triggers are mapped in the script like the speech triggers. We have currently implemented handshaking and figure pointing gestures. Many gestures are designed to work in tandem with speech. When the Speech Manager handles "Does it hurt here?", it also queries the Gesture Manager for a contemporaneous gesture (ie. Pointed_to_lower_right_abdomen), before matching a response. Gestures can have targets since scene objects and certain parts of the patient's anatomy have IDs. Thus a response to a query could involve ***both*** an audio and gesture component.

Evaluation and Logging

VIC serves as an instructor and as an interface for guiding the scenario. VIC monitors whether the student properly greets the patient by exchanging names and a handshake. 11 essential questions that are crucial to the abdominal pain scenario are flagged in the script. At the end of the interview, VIC prompts the student for their differential diagnosis (the list of conditions that DIANA might have).

VIC then evaluates the student's performance by highlighting which (if any) of the critical questions were not asked. VIC comments about the questions that the student should have asked and critiques their differential diagnosis. This highlights one significant advantage of the system (as related to us by the students who tried out the system) is that it provides students with ***immediate*** feedback.

The speech recognition input, matched query, and responses are logged for reference and for later refining the script. The log files can be read by the student or human evaluators to gauge performance.

IV. SCRIPT-BASED APPROACH PERFORMANCE

Seven medical students participated in a pilot study to evaluate the effectiveness of a script-based approach to medical communication skills training. They all had previous experience with standardized patients, and most very familiar with the AAP scenario.

The participants' performance and reaction to the study are discussed in more detail in [Johnsen et. al. 2005]. Here we wanted to evaluate specifically how the script-based approach fared in the patient-doctor interview scenario.

The initial script was written by Jonathan Hernendez, 4th year medical student, and Dr. Amy Stevens. The script was iteratively developed through emails and occasional demos. There was significant difficulty in anticipating the effect of script changes on the experience from both the medical and computer science perspectives. The resulting script was primarily geared to handling queries that led to the 'correct diagnosis'. As all participants had significant experience in abdominal pain, it was assumed they would all perform well.

Script Performance

We reviewed the video tapes of all the interviews and categorized queries into different categories. The script-based approach correctly handled 60% of all queries. As all the students praised the system – some even claiming 100% recognition (!) – we were surprised that the formal analysis revealed a relatively modest script matching performance. Thus we carefully examined *what type of errors* occurred with the script-based approach (Table 3). The following include the types of queries/statements that presented trouble for the straightforward approach employed.

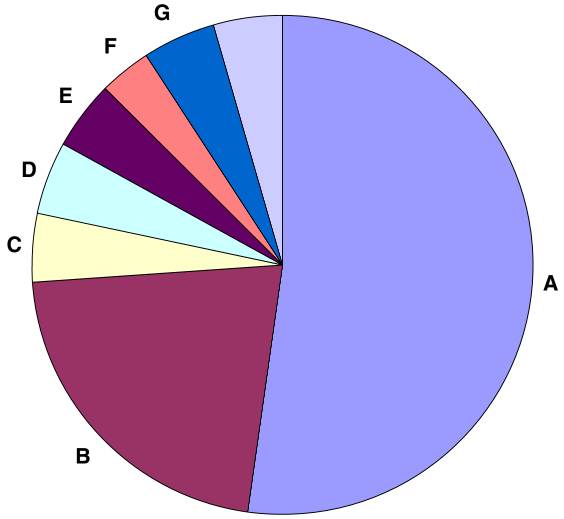

Table 3 - Script matching failure by type.

Failure Cases

We have identified several cases where DIANA responded either incorrectly, or not at all, to the student's query:

A) Entry does not exist (21% of all queries) The student asked a question that was not anticipated, and thus the semantic block did not exist. For instance, we were not expecting the student to pursue the possibility that DIANA may have a problem with her gallbladder, since the script was written in mind for appendicitis. Straight forward inclusion of more queries would alleviate this problem.

B) Variation of Query Phrasing (9% of all queries) It is difficult to anticipate every possible way of phrasing a query. Even tense and contractions cause the script matcher to fail. For example, "Have you felt" and "Are you feeling" are very similar semantically. Applying "stemming" to extract the root of the word through employing a thesaurus or database look-up would alleviate most failures of this type.

C) Joined Questions (2% of all queries) The student joined multiple questions together. ("Have you had any nausea or bowel problems?") However, responses to this query exist in two semantic blocks. Students are taught to avoid stringing together questions. We believe the dictation parser could detect to key words and separate the question, and either have VIC remind the user to keep the questions singular or combine DIANA's responses.

D) Declarative Statements (2% of all queries) The Speech Manager had assumed all speech would be in the form of a question. However, declarative statements are prevalent and important. Students begin the encounter with a greeting. "Hello, Diana, I am a 3rd year medical student and I am here to help you with your pain. I understand you are feeling abdominal pain, correct?" Long statements that precede the query will usually cause a high cost analysis for that input phrase. We are evaluating parsing of speech into statement and query (searching for interrogatives pronouns) components. This is still an imperfect solution, such as in the case of "You are hurting now?" It is unclear if an effective solution could be developed for this case.

E) Empathetic Statements (2% of all queries) Students respond empathetically with their patients, and it is interesting to see them do the same with virtual characters. Students comforted DIANA by saying statements like, "I understand how this can be scary for you". Concern for the patient's pain is an important clinical skill. This failure is similar to a declarative statement failure, and handling this might prove problematic.

F) Summarization (1% of all queries) Similar to empathetic statements, summarization is extremely useful for clarification. "Let me get this right, you have been feeling the burning pain, for the last three days? Then you decided to come into the clinic when …" These are not typical queries, and require logical reasoning to determine if the summary is correct or not. Again handling this case could be very difficult.

G) Incomplete sentences (2% of all queries) Sometimes when phrasing a question, one looses track, stops, and begins phrasing it again. "Uh, so what birth control… uh.. " The participants were surprised when DIANA began answering their question before they even finished asked it based on hearing keywords. We anticipate that accurately handling this case will be very difficult.

H) Pronoun use (2% of all queries) The use of pronouns require an antecedent. "How many days have you had that?" The "that" of the conversation happens to mean her nausea, but could mean many things. A solution includes storing the antecedent in variables for future linking based on feminine, masculine, or neuter instances. Yet context switch errors could still exist (as they do in real world conversations).

Participants believed they could perform their tasks adequately; and even when speech recognition could not understand a particular phrasing, they quickly learned how to rephrase their questions so that DIANA could respond.

Analyzing the types of errors show that a more complete script with stemming will help address 30% of the failure types can be easily addressed. Only about 8-9% of the failure cases, which surround breaking down queries semantically, require potentially much more complex approaches. Our goal is to handle 90% of all queries, and we believe at this level, the education goals of basic communication skills can be achieved.

We caution that the results are based on a relatively small number of participants. We have implemented a larger script based on reviewing audio logs of abdominal pain SP's interviews. We have identified over nine-hundred new queries to add to the system, and are running an additional group of medical students (*n* = 24) through the system in December 2004.

V. DISCUSSION

We were not expecting that a script-driven conversation engine – especially one working from an incomplete script - would be directly applicable to a task as complex as an interpersonal scenario. After debriefing the medical students and reviewing video of the conversations, we derived three primary reasons as to why the 'simple approach' seems effective and beyond our initial notions.

Taking a medical history, especially for relatively simple medical conditions, is a *constrained problem.* The conversation flow is predictable with a relatively limited set of dialogue options. The types of acceptable errors are also important. DIANA not responding to a question – and the question was not pertinent to deriving the correct diagnosis –did not hurt the learning objectives of the scenario. However, if DIANA responded incorrectly to a question, the student usually understood that the system had made a mistake and tried to rephrase the question to get to the desired information. Further, since the participants were intimately familiar with the environment, the task at hand, and goals of the scenario were clearly defined, there were few instances of students asking a question that would not lead to diagnosis ("did you catch the game last night?"). Moving forward, DIANA not responding at all (raising the matching threshold requirements), or responding that she is unsure of what the student asked, might be a better solution to unmatched or poorly-matched questions.

Medical student performance in basic standardized patient scenarios is *easy to measure and evaluate*. The communication skills taught to first and second year medical students are relatively basic. A student's passing of a scenario is based on the number of 'core questions' that the student asked the standardized patient (for the abdominal scenario, asking seven out of eleven core questions was needed to pass). These core questions are required to lead to a correct differential diagnosis. This 'question checklist' metric is simple to implement, and thus performance evaluation is similar to that of the actual standardized patients. This system would be less applicable to more advanced communication topics, such as friendly rapport, approachability, and maintaining eye-contact.

Finally, the *value add of the system overcame the technical limitations*. Feedback of student performance with standardized patients is limited. The logistical challenges in providing timely, personalized feedback on a per-student basis. These include having medical experts review hours of interview video, and the lack of an expert directly involved in the interview (recall that standardized patients are typically paid actors). Students reported that the lack of immediate feedback resulted in a reduced learning, reduced trust in result validity, and reduced overall effectiveness of standardized patients. The students were so hungry for feedback, that the technical and fidelity compromises were easily overlooked.

> "In terms of providing feedback, this interaction has been better than any person has given me in three years of medical school" (4th year medical student participant)

This again points to the impact of interactive systems on education. We believe selecting scenarios with similar properties would also result in effective use of a script-based approach.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the applicability of a straight-forward question-and-answer with scripted responses approach to simulating the patient-doctor interaction for basic scenarios. For first and second-years medical students who focus on learning the basics of diagnosis (asking the right questions and obtaining the proper diagnosis), a straight-forward system allows them to practice these critical skills. Important to the experience was to naturally interact with a believable entity (speak and gesture to a life-sized person).

We believe the surprisingly adequate performance of a script-based matching approach is due to three major properties: a *highly constrained scenario*, focus on *communication skills that are easy to measure and evaluate*, and *providing important feedback* not easily obtained in the real world situation.

The medical students who participated in the pilot study agreed that the script-based matching worked sufficiently well. The frequency and types of errors encountered suggests that augmenting and refining the script could significantly reduce system response errors. We have transcribed sixteen interviews of students practicing with standardized patients, and have identified over 900 queries. We expect by incorporating these queries to have created a rather complete scenario that can handle over 90% of potential queries and create an experience with limited breaks in presence.

Future work will focus on evaluating the student's perception of the virtual character as a teaching tool, as well as the effect of the system on the student. A series of planned studies will evaluate the educational equivalency between real and virtual patients, and if students who repeatedly use the system report a reduced anxiety with standardized patients.

We also are exploring how to evaluate more complex communication skills, such as determining if the student is talking 'down' to a patient by keeping noting the frequency of overly technical terms, the effect of tracking eye-gaze as a measure of attentiveness, and extending the dialogue format to handle questions initiated by the virtual patient.

The acute abdominal pain scenario was a good first-choice for the script-based engine. However, other medical interview skills such as grief counseling or empathy are less straightforward, and require a much more complex system. We hope to eventually address these in future versions of the system.

VII. REFERENCE LIST

BATES' Guide to Physical Examination and History Taking: Eighth Edition Eds. Lynn S. Bickley, Peter G. Szilagyi, John Stackhouse. Lippincott Williams & Wilkins, 2002.

Coulehan J, Block M, *The Medical Interview: Mastering Skills for Clinical Practice, F.* A. Davis Company, 2001.

Harless W, Zier M, Harless M, Duncan R, "Virtual Conversations: An Interface to Knowledge" in IEEE Computer Graphics and Applications Special Issue on Perceptual Multimodal Interfaces. 2003.

Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., & West, S.L. (2000). The Virtual Standardized Patient-Simulated Patient-Practitioner Dialogue for Patient Interview Training. In J.D. Westwood, H.M. Hoffman, G.T. Mogel, R.A. Robb, & D. Stredney (Eds.), Envisioning

Healing: Interactive Technology and the Patient-Practitioner Dialogue. IOS Press: Amsterdam.

Hubal R., Frank G, Guinn C, “Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual Humans for Training”, In *Proceedings of 8th International Conference on Intelligent User Interfaces,* 2003.

Hubal, R.C., Frank, G.A., & Guinn, C.I. AVATALK Virtual Humans for Training with Computer Generated Forces. Proceedings of the Ninth Conference on Computer Generated Forces. Institute for Simulation & Training: Orlando, FL, 2000.

Johnsen K, Dickerson R, Raij A, Lok B, Jackson, J., Shin, M., Hernandez, J., Stevens, A., Lind, D. “Experiences in Using Virtual Characters to Educate Medical Communication Skills” Submitted to IEEE Virtual Reality 2005.

Leech, Geoffrey, Rayson, Paul, and Wilson, Andy. *Word Frequencies in Written and Spoken English: based on the British National Corpus*, Longman, London, 2001.

Shneiderman B, Plaisant C “Designing the User Interface: Fourth Edition” Addison-Wesley, Boston, 2004.

Wallace, Richard S. “Chapter 00: The Anatomy of A.L.I.C.E.” A.L.I.C.E. Artificial Intelligence Foundation, Inc. (http://www.alicebot.org/anatomy.html) Retrieved October 15, 2004.