How to Talk to a Hologram

Anton Leuski
Institute for Creative Technologies USC
Marina del Rey, CA, 90292
leuski@ict.usc.edu

Jarrell Pair
Institute for Creative Technologies USC
Marina del Rey, CA, 90292
pair@ict.usc.edu

David Traum
Institute for Creative Technologies USC
Marina del Rey, CA, 90292
traum@ict.usc.edu

Peter J. McNerney
Institute for Creative Technologies USC
Marina del Rey, CA, 90292
mcnerney@ict.usc.edu

Panayiotis Georgiou
University of Southern California
Los Angeles, CA, 90089
georgiou@sipi.usc.edu

Ronakkumar Patel
Institute for Creative Technologies USC
Marina del Rey, CA, 90292
ronakkup@usc.edu

ABSTRACT

There is a growing need for creating life-like virtual human simulations that can conduct a natural spoken dialog with a human student on a predefined subject. We present an overview of a spoken-dialog system that supports a person interacting with a full-size hologram-like virtual human character in an exhibition kiosk settings. We also give a brief summary of the natural language classification component of the system and describe the experiments we conducted with the system.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection Process; H.4.m [**Information Systems Applications**]: Miscellaneous; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Natural language, Voice I/O*

General Terms

Design, Experimentation, Human Factors

Keywords

Spoken dialog, text classification

1. INTRODUCTION

In the recent Hollywood movie "iRobot" set in 2035 the main character played by Will Smith is running an investigation into a death of an old friend. The detective finds a small device that projects a holographic image of the deceased, delivers a recorded message and responds to the detective's questions by playing back prerecorded answers. The device responses are limited to the events preceding the accident. We are building virtual characters with similar capabilities in our lab.

These virtual characters are focused on providing educational and training aids of different specializations. For example, such a character can be designed to deliver a spoken message to a student and answer questions regarding the subject of the message.

In this paper we describe one of these characters called "Sgt. Blackwell" that serves as interface for an exhibition kiosk. It supports natural language dialog allowing the user to explore the system design and get an overview of the exhibition. There are two main contributions of this project. The first is the system integration aspect: the kiosk combines high-quality graphics, elements of physical staging, state-of-the-art automatic speech recognition system (ASR), and a text classification approach to provide the user with a complete experience of interaction with a virtual human-like character. The second aspect is the statistical natural language classifier at the core of the system that analyzes the user's speech and selects the appropriate responses.

We describe the system setup, the technology used to analyze the human speech and select the appropriate character response. We also describe a set of experiments we use to analyze the system's performance.

2. SGT. BLACKWELL

Figure 1 shows a photograph of the system setup. We see a soldier in a full combat gear standing in a doorway of a building. Everything we see here including the building walls, the stones in the front, the table inside of the room, the map, the radio, and the rest of the military gear are physical props. The soldier is a life-size pseudo-hologram. It is a high-resolution real-time graphics projection onto specialized transparent optical film that covers the doorway.

The graphics of Sgt. Blackwell is a high-quality computer rendering of a 3D model consisting of more than 60,000 polygons. The life-presence effect is further facilitated by a range of idle movement animations. The character periodically steps from one foot to another, flexes his hands, and moves his head.

A user talks to Sgt. Blackwell using a head-mounted close-capture usb microphone. The user's speech is converted into text using an automatic speech recognition system. We used the Sonic speech recognition engine from the University of Colorado [6] with our own acoustic and language models [7].

The character can deliver 83 spoken lines ranging from

IUI'06, January 29–February 1, 2006, Sydney, Australia.
ACM 1-59593-287-9/06/0001.

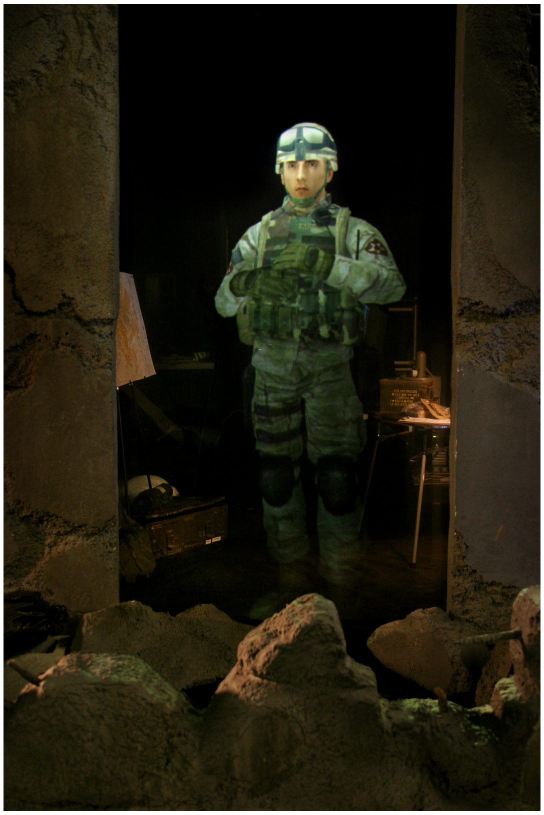

Figure 1: A photograph of the Sgt. Blackwell system setup.

one word to a couple paragraphs long monologues. The spoken lines were recorded by a voice actor and automatically transcribed to generate the character's lip movement animations. The delivery of the lines is accompanied by a variety of life-like gestures and body movements. These animations were recorded at a motion-capture studio and edited in our graphics group.

There are three kinds of lines Sgt. Blackwell can deliver: content, off-topic, and prompts. The 57 content-focused lines cover the identity of the character, its origin, its language and animation technology, its design goals, our university, the exhibition setup, and some miscellaneous topics, such as "what time is it?" and "where can I get my coffee?"

When Sgt. Blackwell detects a question that cannot be answered with one of the content-focused lines, it selects one out of 13 off-topic responses, – e.g., "I am not authorized to comment on that," – indicating that the user has ventured out of the allowed conversation domain.

In the event of the user persisting on asking the questions for which the character has no informative response, the system tries to nudge the user back into the conversation domain by suggesting a question for the user to ask: "You should ask me instead about my technology." There are 7 different prompts in the system.

One topic can be covered by multiple answers, so asking the same question again often results in a different response introducing variety into the conversation. The user can specifically request the alternative answer by asking something along the lines of "do you have anything to add?" or "anything else?" This is the first of two types command-like expressions Sgt. Blackwell understands. The second type is a direct request to repeat the previous response, e.g., "come again?" or "what was that?"

If the user persists on asking the same question over and over, the character might be forced to repeat its answer. It indicates that by preceding the answer with one of the four "pre-repeat" lines indicating that incoming response has been heard recently, e.g., "Let me say this again..."

3. TEXT CLASSIFICATION

A crucial part of the Sgt. Blackwell system is the language understanding module. It analyses the text strings coming out from the speech recognition system and selects the appropriate system responses. The main problem with language understanding is the uncertainty. There are two sources of uncertainty in a spoken dialog system: the first is the natural language ambiguity, making it difficult to compactly characterize the mapping from the text surface form to the meaning; and the second is the error-prone text output from the ASR module. When creating a language understanding system one possible approach is to design a set of rules that for an input text string select a response [10]. Because of the uncertainty this approach can quickly become intractable for anything more than the most trivial tasks. An alternative is to create an automatic system that uses a set of training question-answer pairs to learn the appropriate question-answer matching algorithm [1].

On the surface our answer selection problem is similar to the question answering scenario that has been studied in the context of Q&A track at TREC [9]. Our setup differs in that we have a fixed number of responses to choose from, while a typical Q&A system has to find the response from a large collection of textual material. Additionally such a system assumes that the questions are based on facts and the response has to be relevant to the question. The main focus in our settings is on the answer's *appropriateness* as opposed to its *relevance*. For example, an evasive or a misleading answer would be appropriate but not relevant.

Our question-answer matching task is similar to text classification tasks that have been studied in Information Retrieval for several decades [5]. Indeed, we have a fixed set of answers or classes, we have a training set of questions that correspond to individual answers and we are building a system that for a previously unseen question selects the appropriate class or the answer. The distinct properties of our study are the small size of the text we need to classify, – the questions are a few words long and the answers almost never exceed 2-3 sentences, – and the large number of classes (Sgt. Blackwell has 60 responses to choose from[1]). We should also note that the answers have a more informative structure than traditional class labels – the answer text has a particular meaning and simply replacing each answer with a class label would discard any information contained in the text representation.

For this project we have created three text classification approaches. The first approach is a state-of-the-art text classification system based on Support Vector Machines (SVM) [8, 4]. We represent each question as a feature vector of $tf \cdot idf$ weighted word unigrams, bigrams, and trigrams. The sys-

[1] 57 content-focused responses, 2 command-related responses, and one off-topic class

tem is trained to classify question vectors into classes labeled with individual answers. The other two systems are based on the statistical language modeling techniques used recently in mono-lingual (LM) and cross-lingual information retrieval (CLM) [2, 3]. The former approach uses the Relevance Model technique to compute the language model of the training questions assigned to each answer and compares it to the language model of the test question. The latter uses the training data to "translate" the test question – calculates the language model of the most likely answer for the test question – and then compares this model to language models of the individual answers.

We have compared all three approaches and observed a statistically significant advantage in performance of the language modeling techniques over the SVM approach. Both language modeling approaches create better text representations than the feature vectors used in the SVM system. In addition, the CLM system takes advantage of the answer text information.

4. EXPERIMENTS

To train our system we have developed a data set of matching question-answer pairs. First, a scriptwriter defined a set of questions Sgt. Blackwell should be able to answer and prepared answers for those questions. Then we expanded the set of questions by a) explicitly paraphrasing the questions and b) collecting questions from users by simulating the final system in a Wizard of Oz study (WOZ). The audio recorded during the WOZ studies were transcribed and the questions added to the data set. That way we collected 1261 questions total. We then annotated each question with the appropriate answers.

We conducted two sets of experiments. First we run an off-line evaluation of the classification algorithms comparing the language model classification techniques with the SVM method. These experiments followed the 10-fold cross-validation schema. We compared the systems using the classification accuracy or the proportion of times the system returned the appropriate answer. The SVM, LM, and CLM showed 53.1%, 57.8%, and 62.0% accuracy correspondingly. It is 16.67% improvement for the CLM approach over the SVM system. The numbers are statistical significant using t-test with the cutoff set to 5% ($p < 0.05$).

We have incorporated the best performing classifier (CLM) into the Sgt. Blackwell system and for the second set of experiments we conducted a user study where we asked each participant to talk to Sgt. Blackwell. Each person had to ask at least 20 questions. We specifically defined 10 of those questions by selecting from the training data set to test the effect of the speech recognition rate on the classification. For the rest of the study the users were free to ask any questions. 18 people participated in the study and we collected 378 question-answer pairs. We have recorded and transcribed the sessions. The preliminary analysis shows an average 77.8% classifier accuracy on the ASR output, 83.3% accuracy on the hand-transcribed data and 36.1% average word error rate (WER) for the ASR. Plotting the answer selection accuracy as a function of the ASR WER shows that the classification accuracy on the ASR output starts to degrade significantly relative to the accuracy on the hand-transcribed data only after approximately the 70% WER mark. It indicates that the classifier is very robust to the speech recognition errors.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented an overview of a spoken-dialog system that supports a person interacting with a full-size hologram-like virtual human character in an exhibition kiosk settings. We also gave a brief summary of the natural language classification component of the system and described the evaluation experiments we conducted with the system.

Preliminary failure analysis indicates a few directions for improving the system's quality. First, we could continue our efforts on collecting more training data and extending the question sets.

Second, we could have the system to generate a confidence score for its classification decisions. Then the answers with a low confidence score can be replaced with an answer that prompts the user to rephrase her question. The system would then use the original and the rephrased version to repeat the answer selection process.

Finally, we observed that a notable percent of misclassifications results from the user asking a question that has a strong context dependency on the previous answer or question. We are presently looking into incorporating this context information into the answer selection process.

6. REFERENCES

[1] J. Chu-Carroll and B. Carpenter. Vector-based natural language call routing. *Journal of Computational Linguistics*, 25(30):361–388, 1999.

[2] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR*, pages 175–182, Tampere, Finland, 2002.

[3] A. Leuski. Using statistical language models for a dialog response selection. Forthcoming.

[4] A. Leuski. Email is a stage: discovering people roles from email archives. In *Proceedings of the 27th annual international ACM SIGIR*, pages 502–503, Sheffield, United Kingdom, 2004. ACM Press.

[5] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th International ACM SIGIR*, pages 298–306, Zurich, Switzerland, 1996.

[6] B. Pellom. Sonic: The university of colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, CO, 2001.

[7] A. Sethy, P. Georgiou, and S. Narayanan. Building topic specific language models from webdata using competitive models. In *Proceedings of EUROSPEECH*, Lisbon, Portugal, 2005.

[8] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the 21st international conference on Machine learning*, Banff, Alberta, Canada, 2004.

[9] E. M. Voorhees. Overview of the trec 2003 question answering track. In *Proceedings of The 12th Text Retrieval Conference*, pages 54–69, 2003.

[10] J. Weizenbaum. Eliza–a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.