Establishing Trust in a Deployed Spoken Language System

Jody J. Daniels, Daniel Miksch, and Helen Wright Hastie

Lockheed Martin Advanced Technology Laboratories
3 Executive Campue, 6th Floor
Cherry Hill, NJ 08002
{jdaniels, dmiksch, hhastie}@atl.lmco.com

Abstract

Lockheed Martin Advanced Technology Laboratories has been designing, developing, testing, and evaluating spoken language understanding systems in several unique operational environments over the past six years. Through these experiences we have encountered numerous challenges in making each system become a useful and trustworthy tool that is an integral part of a user's operations. In this paper, we discuss the challenges of building trust and report how we overcame them with respect to a number of domains.

Introduction

Lockheed Martin Advanced Technology Laboratories (ATL) has been designing, developing, testing, and evaluating spoken language understanding systems (SLS) in several unique operational military environments over the past six years. This model of human interaction is referred to as Listen, Communicate, Show (LCS). In an LCS system, the computer listens for information requests, it communicates with the user and networked information resources to compute user-centered solutions, and it shows tailored visualizations to individual users. While developing these systems, we were challenged to make each system become a useful and trustworthy tool that is an integral part of a user's operations. For example, Figure 1 shows the deployment of a dialogue system for placing Marine supply in a tactical HMMWV.

One of the challenges of creating spoken language systems is giving appropriate system responses. There is a tension between brevity and providing sufficient context. Military users typically want system utterances to be brief and concise for a number of reasons. These users generally want to accomplish their tasks as rapidly as possible and lengthy responses can become frustrating. They also want the system to emanate the shortest radio signal possible, so the transmissions are neither intercepted nor used to locate the source of the signal. Similarly, users do not want to have to repeat redundant information, which would add to the electronic signature of the soldier. On the converse side, users must have confidence that the system understood what the user was attempting to convey. This implies that the system must give enough context in its response to build the user's trust that it did correctly understand the user.

Figure 1. LCS Marine mounted in a HMMWV.

Architecture

The LCS Spoken Language systems use the Galaxy architecture (Seneff et al., 1999). This Galaxy architecture consists of a central hub and servers. Each of the servers performs a specific function, such as converting audio speech into a text translation or combining the user's past statements with recent statements. The individual servers exchange information by sending messages through the hub. These messages contain information to be sent to other servers as well as information used to determine what server or servers should be contacted next.

Various Galaxy servers work together to develop a semantic understanding of the user's statements and questions. The sound spoken into the microphone, telephone, or radio is collected by an audio server and sent to the recognizer. The recognizer translates this wave file into text and sends it to a natural language parser. The parser converts the text into a semantic frame, a representation of the statement's meaning. This meaning representation is sent on to the dialogue manager that monitors the current context of a conversation and, based on this context, can prompt the user for any necessary clarification and present intelligent responses to the user. Since the dialogue manager is aware of what information has been discussed thus far, it is able to determine what information is still needed. The dialogue manager creates a semantic frame. This frame is sent through the language generation server to generate a text response that is then spoken to the user through a speech synthesis server.

The dialogue manager can be used to establish the user's trust by making explicit or implicit confirmations. Different confirmation strategies can be deployed for different domains and for users of varying expertise, such as, confirming after every turn or leaving the confirmation until the end of the dialogue. The user can make corrections if the system indicates a mistake has been made.

To solve the problem of retrieving or placing data from/in remote and local sources, we gave the systems below the use of mobile software agents (Daniels, 2000). If user-requested information is not immediately available, an agent can monitor the data sources until it is possible to respond. Users may request a notification or alert when a particular activity occurs. Because of the potentially significant time lag, it is important to manage dialogue activity so the user is only interrupted when the need for information is more important than the user's current task. This active management of interruptions aids task management and lightens cognitive load (Daniels, 2000).

Domain

LCS Marine

One of the first LCS systems to be tested in the field was our Marine Logistics spoken dialogue system. This application connected the Marine in the field to the Small Unit Logistics (SUL) database, which maintains current information about supply requisitions. Warfighters wanted to place requests as well as check on the status of existing requests without the need of additional hardware or third party communications. It was also highly desirable to use existing communications procedures to minimize the training time to use the system. The system needed to be robust, speaker-independent and mixed initiative, enabling the warfighters to develop a sense of trust in the technology.

Marines using the system were able to perform several tasks. They could create new requests for supplies with the system prompting them for information needed to fill in a request form. They could also modify and delete previously placed requests and had two ways to check the status of requests. They could directly ask about the current status or they could delegate an agent to monitor the status of a particular request. It was easy to build a constraint to have the agent return after a specified time period if no activity occurs on the request, which was valuable information for the Marine. These delegated agents travel across a low-bandwidth SINCGARS radio network from the Marine to the database, accessing that database to place, alter, and monitor supply requisitions.

The challenges in deploying this system to the field were twofold - building user trust in the system so it would become part of normal operations and dealing with the unique environmental factors that contribute to user frustration in an SLS. Marines have the conflicting goals of wanting to restrict their time on the radio net as much as possible and wanting to ensure that what they request is what they receive. The interaction with the system had to be realistic, robust, and intelligent, or the Marines would quickly become frustrated and discard the system. This was shown during Marine exercises where they simulated real battle situations and the LCS Marine system was used as an integral part of the operation. The system was defined and refined to meet both needs.

ATL engineer's continuously worked with Marine Corps logistics subject matter experts to create realistic dialogs that enabled Marines to rapidly and intelligently make supply requests. This involved building several prototypes and participating in multiple sessions with numerous Marines who would evaluate user inputs and dialogue responses. Through this iterative process we were able to incorporate the input from the Marines to create a system that presented a familiar interface that the warfighter was comfortable interacting with. We also spent time ensuring the Marines that LCS Marine could handle both proper and non-standard radio protocols. Broad coverage of potential expressions, especially those used under stress, like curse words, led to greater user ability to successfully interact through the system. By involving the Marines in the development and evaluation of the system dialogues, the LCS Marine system evolved quickly into a realistic system that the Marines were comfortable using and quick to rely on, even in extreme conditions.

In addition to the iterative process discussed above, we measured the effectiveness of the LCS Marine system under operational conditions with real users placing real requests, accessing live databases, and using existing communications links. We participated in four Integrated Feasibility Demonstrations (IFDs) over a twelve-month period. The IFDs ranged from scripted dialogue, replicated databases, and testing in the lab with prior military personnel, to active duty Marines using the system operationally over a series of days as their sole means of interaction with the logistics system for rapid requests. At the conclusion of each IFD we evaluated the effectiveness of the LCS Marine using a three-tier metric system (Stibler et al., 2001). The first level of the approach measured overall user satisfaction derived from a collection of user reactions on a Likert-scaled questionnaire. The questions were associated with eight user satisfaction metrics: ease of use, system response, system understanding, user expertise, task ease, response time, expected behavior and future use. The middle tier metrics captured the ability of users to successfully complete their domain tasks in a timely manner. The lowest level tier measured the effectiveness of individual system components along specific dimensions, including component error rates (Stibler et al., 2001). The results of these evaluations gave greater insight into the effectiveness of the LCS Marine system, while demonstrating where users were either satisfied or frustrated with their LCS Marine System interactions.

The second set of challenges, unique environmental factors, included access while on the move, battlefield noise, and coping with adverse conditions such as sand storms, while not creating the burden of additional hardware or frustration to the end warfighter. Accessing LCS Marine while on the move meant using a SINCGARS radio as the input device. Attempts to use the system by directly collecting speech from a SINCGARS radio were dropped due to the technological challenges presented by the radio's distortion. Instead, we installed the majority of the system on laptops and put these into the HMMWV. The warfighter would then use the LCS Marine system to task mobile agents to access the SUL database. The mobile agents were sent over the SINCGARS data link back to the data sources. This meant securing two laptops and SINCGARS data modem in a HMMWV and powering them off of the vehicle's battery as illustrated in Figure 1. Since the equipment was secured in the HMMWV, there was no additional hardware for the warfighter using the system. The mobile agents were able to easily traverse a retransmission link and reach the remote data source.

Dealing with hugely varying background noises was less of a problem than originally predicted. Most of the time when loud events occurred, users would stop talking. Their hearing was impaired so they would wait for the noise to abate and then continue the dialogue. We did encounter several users who, because of the Lombard effect, insisted upon always yelling at the system. Our recognizer was able to handle these situations. There were a few times when the system was not able to understand the user because of background noise, such as a jet flying overhead, however we did not measure decibel levels to determine a noise threshold .

Shipboard Information

An LCS system has been developed to monitor shipboard system information aboard the Sea Shadow (IX 529), a Naval test platform for stealth, automation, and control technologies. From anywhere on the ship, personnel use the on-board intercom to contact this system, Shipboard Ubiquitous Speech Interface Environment (SUSIE), to ask about the status of equipment that is located throughout the ship. Crewmembers do not have to be anywhere near the equipment being monitored to receive data. Figure 2 illustrates a sailor using SUSIE through the ship's intercom. Personnel can ask about pressures, temperatures, and voltages of various pieces of equipment or can delegate monitoring those measurements (sensor readings) to the system. A user can request notification of an abnormal reading by a sensor. This causes the LCS system to delegate a persistent agent to monitor the sensor and report the requested data. Should an abnormal reading occur, the user is contacted by the system using the intercom.

Figure 2. Sailor interacting with SUSIE through the ship's intercom.

This domain presented several challenges and opportunities. Through numerous discussions with users and presentation of possible dialogues, we learned that the users benefit from a system's ability to remember between sessions the most recent activity of each user. This would permit a user to simply log in and request: "What about now?" SUSIE would determine this user's most recent monitoring activity and would seek out and report the current status. While this seems quite simple, there is significant behind-the-scenes work to store context and make the interaction appear seamless. In addition to being useful, the end user would also have a greater amount of confidence in the system and its capabilities for intelligent understanding.

Another significant challenge was gaining the support and trust of the crew to allow SUSIE to use the ship's intercom. The Sea Shadow is equipped with one intercom. Since crewmembers are spread all over the ship, the intercom is the only available link for crew member communications. The operational protocol of the ship demands that only one conversation occur on the intercom at a time. A conversation should only be interrupted in the case of a serious emergency. This created two problems for SUSIE. Firstly detecting when a crew member was requesting information from SUSIE rather then another crew member and secondly, determining when to alert a user of an abnormal sensor reading without interrupting a current conversation.

All conversations that occur onboard the ship begin by the crewmember "logging" on. For example, if crew-member A wanted to talk to crewmember B, A would say "B this is A," beginning the conversation. We adopted this protocol to "log–on" to SUSIE. If crew member A would like to ask SUSIE about the status of a ship board system, the crew member would say, "SUSIE, this is A." The Dialogue Manager becomes aware that this is a conversation between A and the system. At the conclusion of the conversation, the crewmember would logout by using a phrase such as "Out" or "Goodbye."

Determining when it is appropriate to alert a user of an abnormal sensor reading is a more difficult problem, and is being addressed in our current work. Our initial interruption strategy was to wait for periods of silence of at least thirty seconds in length before notifying a user. This solution was initially acceptable by the crew, as conversations between crewmembers are normally kept brief without long delays. This does not adequately solve the problem, however, because if multiple users continue to carry on conversations back-to-back, there may not be a thirty second silence for a couple of minutes. In this situation, the system would have been unable to alert the user of the abnormal sensor reading, perhaps resulting in a critical system failure on board the ship. This system behavior would decrease confidence levels and trust in the SLS. To ensure that this does not happen, we developed a Galaxy server to handle interruption management. The Interruption Manager determines the best time to interrupt the user based on the level of criticality of the abnormal sensor reading. This requires the Interruption Manager to have an understanding of the user's current task and the priority. Based on this understanding, the SLS can determine the appropriate interruption scheme.

Collecting speech data through the intercom system to pass to SUSIE required linking two digital signal processors (DSP) and adjusting them to the hardware for the SLS. Once connected, the next significant challenge was varying noise levels. Background noise varied from one room to the next and even within a single space. We were not able to use a push-to-talk or hold-to-talk paradigm because of the inconvenience to the crew members; they leave the intercom open for the duration of the conversation. Fortunately, the recognizer, based on MIT's SUMMIT (Glass and Hazen,1998) is able to handle a great deal of a noise and still hypothesize accurately. To improve the system accuracy, we will incorporate automatic retraining of the recognizer on noise each time that a new session begins.

Although no formal evaluation has taken place to date, through in depth interviews with crew members, it is clear SUSIE has been well received by the crew. The crew members evaluate SUSIE while underway and provide feedback to ATL engineers. Based on this feedback, SUSIE is modified. Crew members are satisfied that they can ask SUSIE about sensor readings and trust that the answer provided is accurate. The majority of the feedback comments we receive from the crew are requests for expanded functionality and additional sensors to be added to the vocabulary.

Future Work

Areas of research needed for more dynamic and robust systems include better, more robust or partial parsing mechanisms. This allows the system to ask the user to clarify the portion of an utterance which was either misunderstood or not included in the lexicon, rather than discarding the entire utterance. Systems must be able to cope with multi-sentence inputs, such as humans do. Most current systems only permit one utterance at a time to be given by a user. With longer utterances, the system should provide backchannels to confirm that the system is still listening.

Intelligent interruption management and capabilities must continue to evolve as well to increase user trust. Measuring user workload and stress, taking advantage of multiple, perhaps less intrusive modalities, and better system understanding of user tasks will help gain user satisfaction and trust with the system. Ease of domain expansion continues to be important as systems evolve. Requiring that system modifications be done solely by developers will inhibit the expansion, adaptation, and utility of an SLS over time.

Conclusions

We have discussed the pragmatics involved with taking an SLS system out of the laboratory or away from telephony and placing it in a volatile environment. To be accepted and trusted by the end users, these systems have to be robust, able to cope with varying input styles and modes, and able to modify their output to the appropriate situation. In addition, the systems must be an integral part of the technology currently in use and be able to withstand adverse conditions. Satisfying all of these constraints involves active participation in the development process with the end-users as well as creative solutions and technological advances.

Acknowledgements

These systems were developed and tested by a number of engineers, including Steve Knott, Mike Orr, Jose Rivera, Kath Stibler, and Mike Thomas.

References

Jim Glass and T.J. Hazen. 1998. Telephone-based conversational speech recognition in the jupiter domain. In *Proceedings of ICSLP 98*, Sydney, Australia, November.

Stephanie Seneff, Ray Lau, and Joe Polifroni. 1999. Organization, communication, and control in the galaxy-II conversational system. In *Proceedings for Eurospeech '98*, Budapest, Hungary.

Jody Daniels. 2000. Integrating a spoken language system with agents for operational information access. In *Proceedings of Innovative Applications of Artificial Intelligence (IAAI-2000)*, Austin, TX, August.

Kathleen Stibler and James Denny, 2001. A Three-Tiered Evaluation Approach for Interactive Spoken Dialogue Systems. In *Proceedings of the Human Language Technology Conference (HLT-2001),* San Diego, CA, March.