NATURAL LANGUAGE PROCESSING IN VIRTUAL REALITY TRAINING ENVIRONMENTS

Curry I. Guinn and R. Jorge Montoya
Research Triangle Institute
Research Triangle Park, NC 27709

ABSTRACT

Technological advances in areas such as transportation, communications, and science are rapidly changing our world--the rate of change will only increase in the 21st century. Innovations in training will be needed to meet these new requirements. Not only must soldiers and workers become proficient in using these new technologies, but shrinking manpower requires more cross-training, self-paced training, and distance learning. Two key technologies that can help reduce the burden on instructors and increase the efficiency and independence of trainees are virtual reality simulators and natural language processing. This paper focuses on the design of a virtual reality trainer that uses a spoken natural language interface with the trainee.

RTI has developed the Advanced Maintenance Assistant and Trainer (AMAT) with ACT II funding for the Army Combat Service Support (CSS) Battlelab. AMAT integrates spoken language processing, virtual reality, multimedia and instructional technologies to train and assist the turret mechanic in diagnosing and maintenance on the M1A1 Abrams Tank in a hands-busy, eyes-busy environment. AMAT is a technology concept demonstration and an extension to RTI's Virtual Maintenance Trainer (VMAT) which was developed for training National Guard organizational mechanics. VMAT is currently deployed in a number of National Guard training facilities. The AMAT project demonstrates the integration of spoken human-machine dialogue with visual virtual reality in implementing intelligent assistant and training systems. To accomplish this goal, RTI researchers have implemented the following features:

- Speech recognition on a Pentium-based PC,
- Error correcting parsers that can correctly handle utterances that are outside of the grammar,
- Dynamic natural language grammars that change as the situation context changes,
- Spoken message interpretation that can resolve pronoun usage and incomplete sentences,
- Spoken message reliability processing that allows AMAT to compute the likelihood that it properly understood the trainee (This score can be used to ask for repeats or confirmations.),
- Goal-driven dialogue behavior so that the computer is directing the conversation to satisfy either the user-defined or computer-defined objectives,
- Voice-activated movement in the virtual environment, and
- Voice synthesis on a Pentium-based PC.

AUTHORS BIOGRAPHIES

Curry Guinn, Ph.D., is a Research Engineer and lead developer of the Advanced Maintenance Assistant and Trainer at RTI. Using pioneering spoken human-computer dialogue algorithms developed as part of his Ph.D. research at Duke University, he has integrated advanced spoken dialogue capabilities into computer-based training and operational environments. These systems allow a trainee or mechanic to talk to a computerized assistant in a VR environment during the diagnosis and repair of equipment. These systems have been implemented on a variety of platforms; off-the-shelf PCs, wearable computers, hand-held computers, and high-end workstations. Specific project domains include maintenance of the M1A1 Abrams tank, orientation training for ship engineering officers, flight training, and point-of-sale tasks.

R. Jorge Montoya is a Senior Research Engineer and Head of the Virtual Reality Group at RTI. He has a Master of Science degree from North Carolina State University. Mr. Montoya led the effort to specify, acquire, and integrate the hardware and software infrastructure that forms the basis of RTI's Virtual Reality Laboratory. Mr. Montoya was the project manager for AMAT.

NATURAL LANGUAGE PROCESSING IN VIRTUAL REALITY TRAINING ENVIRONMENTS

Curry I. Guinn and R. Jorge Montoya
Research Triangle Institute
Research Triangle Park, NC 27709

1.0 INTRODUCTION

Computer-based trainers (CBTs) are gaining a great deal of acceptance in the training community as their effectiveness is demonstrated. With the advent of multimedia and virtual reality (VR), many CBT systems are now being either re-engineered or are being designed with these features built in.

An exciting recent development, Virtual Reality (VR) represents a culmination of technological advances in real-time computer graphics hardware and software that support the generation of high-quality, photo-realistic, images in real time. Augmented by immersive capabilities provided by helmet mounted displays (HMDs) and other auxiliary dimensions such as touch (data gloves), voice recognition, voice synthesis, 3D sound, tracking (both head and hand), and others, this technology holds great promise as a cost effective training and teaching tool. In its simplest form, VR is the presentation of and interaction with a synthetic, computer generated 3D world, so realistic that the user feels as if he/she were experiencing the real thing. VR supports a new way for humans to interact with computers that is multi-sensorial that approaches the way in which humans interact with real environments. These interactions include visual, haptic, sound, speech, and olfactory.

Natural Language Processing (NLP) is a technology that supports a spoken human-computer dialogue. Using speech recognition as the input modality, NLP parses the inputs and extracts application-specific, grammatically correct content that then it matches with a specific domain knowledge base. The result of the match produces a reply from the system that is synthesized through a speech synthesizer.

It is clear that for applications that are hand-busy and eye busy, a CBT trainer that incorporates VR and NLP technologies should be very useful. Further, the knowledge base can be implemented with a static and a dynamic component so that the system can be used as an assistant or as a trainer.

This paper describes the implementation of the Advanced Maintenance Assistant and Trainer (AMAT), a VR-based CBT which includes natural language processing as the primary interface between the trainee and the system. The paper identifies the characteristics of VR simulators and natural language processing in training environments. It describes the specifics of the AMAT project, including the training requirements, a description of virtual reality technology applicable to the project and a description of the natural language processing technology as applied to the project. The paper concludes with a discussion of future directions for NLP research and potential applications of NLP in training.

1.1 Virtual Reality Simulators in Training Environments

Virtual Reality Multimedia Training can dramatically reduce the cost of delivering training by decreasing learning time for students and instructors, the need for expensive and dedicated training equipment (physical mock-ups, labs, or extra equipment for training purposes) , and travel expenses. Students can work in simulated worlds that may be too expensive or too dangerous to practice in reality.

There are also great learning benefits that translate into productivity gains. In an interactive virtual world, students must act rather than just observe and answer questions. A number of studies indicate that

VR-based learning systems when compared to traditional instructor-led classroom and laboratory methods, have been shown to:

- increase retention
- reduce learning time
- increase access to training
- promote conceptual and procedural learning
- reduce errors in performing skills, particularly for complex tasks

Virtual Reality is a Human Computer Interface (HCI) technology the use of which has been made practical by recent advances in real-time computer graphics hardware and software. VR trainers support interactions with synthetic environments developed specifically to replicate training environments. As training needs increase and the availability of both instructors and equipment to train with decreases, VR-based trainer simulators are gaining popularity and acceptance. This class of CBTs are flexible (easy to reconfigure) and support on-demand training. Moreover, when implemented on laptop computers, they provide a portable training tool. VR trainers also eliminate the need to use real equipment in training.

In its most basic form, a VR trainer provides a virtual environment with which the trainee can interact. How faithful must this environment be depends on the training objectives. A typical VR trainer implementation starts with a requirements definition phase in which system fidelity is established as a function of training objectives, system performance, and system cost.

VR trainers are extensions to the traditional CBT training systems. CBTs have evolved from primarily textual and static systems to ones which are based on multimedia technology and which are dynamic. Virtual reality adds to these systems true immersive 3D with human-like interactive features.

NASA's virtual space shuttle was used to train astronauts and flight controllers to repair the Hubble space telescope. A post-training survey indicated that trainees graded overall effectiveness at "slightly over effective" while training activity time was reduced from hours to minutes (Kenney and Saito, 1994). A study by the Canadian Defense & Civil Institute for Environmental Medicine compared training ship operators using VR versus using active vessels. Results from sea maneuvers showed that the officers who trained with the simulator generally performed better than those who trained on active ships (Magee, 1995). Motorola conducted a study of robotic manufacturing plant operations by comparing three groups: (1) trainees in an immersive VR environment, (2) trainees who watched the environment be simulated, and (3) a control group who worked directly in the manufacturing laboratory. VR test scores were equal to the control group test scores and the fewest number of errors were made by the students using VR particularly during complex tasks (Adams, 1995).

Over the last three years, RTI has implemented a series of VR-based trainers for military and commercial applications. Examples of these include the Virtual Maintenance Assistant Trainer (VMAT for the U.S. National Guard), the Virtual Medical Emergency Trainer (VMET for the U.S. Army), the Explosive Ordnance Disposal Trainer (EOD for the U.S. Navy), the M1A2 D/T Virtual Trainer (for the U.S. Army), Radiation Safety Trainer (for Carolina Power and Light), and training modules to support the development of team problem solving skills (for the Ford Motor Company).

The work described in this paper is an extension to the VMAT trainer.

1.2 Training Requirements

The primary goal of the VMAT system is the training of the M1A1 organizational tank mechanic (45T) to stay proficient in the diagnosing, isolation, and repair of faults in the M1A1 tank and its derivatives to the line replaceable unit (LRU) level. Traditionally, the National Guard (NG) tank mechanic goes through a two week rotation at NG Regional Training Schools (RTSs) approximately every 12 to 18 months. Because in the interim, the NG mechanic has had very little, if any, interaction with a real tank, it takes a large portion of the eighty hours of training at the RTS to get re-familiarized with the diagnostic and test procedures associated with the M1A1.

VMAT (and AMAT by extension) was specified to provide a virtual tank in which to diagnose and isolate all faults associated with the work of the 45T mechanics. The faults were organized in a course consisting of sixteen lessons implemented in a CBT structure. The course was implemented with self-paced, sequential, secure, and dynamic features.

The trainer was specified to support the access of individual trainees through a personalized process and to protect the performance records of the individual. Time spent in a given lesson was unlimited but a proficiency test had to be passed (passing grade can be varied by the instructor) in order to advance to the next lesson. In addition, the courseware was designed to allow for modifications to the training procedures resulting from modifications and changes to the equipment.

1.3 Natural Language Processing in Training Environments

There are several factors that make adding spoken natural language processing (NLP) to training environments beneficial.

- *Ease of interactively*. Natural language is a natural mode of communication for humans. It lessens the demand on the student to learn how to manipulate the computer environment. In a series of human--computer experiments, human users in a database query environment overwhelmingly chose voice input over keyboard or scroller input in preference [Rudnicky, 1993]. This preference surfaced despite the fact that the total time to complete the task took slightly longer when voice input was used.

- *Expressiveness*. Natural language is extremely expressive. A well-designed NLP interface allows the user to express things that might be extraordinarily difficult using a menu- or button-based interface.

- *Multi-modal*. Spoken NLP frees other channels of communication. In an intensely graphical environment like a virtual world, the computer screen is already loaded with information. Presenting more information with pop-up boxes and graphics can quickly clutter the screen and overwhelm the sensory field of the user. Spoken dialogue allows another channel of communication with the computer.

- *Hands-free, eyes-free*. Spoken language may be the only available mode of communication. In an immersive virtual environment, for instance, keyboard entry or mouse entry may be difficult. In operational environments, it may be desirable to have the hands and eyes free.

- *Novelty*. Clearly one of the tasks of instructional designers is to engage the student. The current novelty of being able to talk with the computer helps to engage the student's interest. As the technology progresses, this factor will lessen in importance (but probably not within the coming decade).

1.4 Natural Language Processing Technology

The AMAT system uses a modular technology that allows for flexibility in future designs. The overall architecture is represented in Figure 1. In the following sections we will discuss each component.

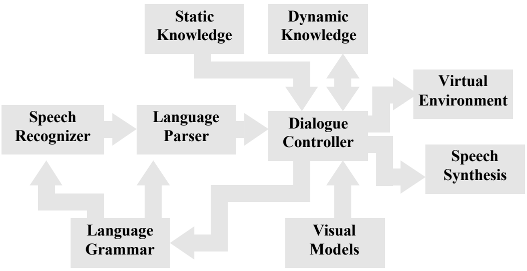

Figure 1. Advanced Maintenance Assistant and Trainer Architecture

1.4.1 Speech Recognition

The front-end to the dialogue system is a **speech recognizer**. The purpose of this module is to take the spoken audio signal and convert it to text. There are several dimensions of a speech recognition that must be analyzed for a given application to determine which recognition system is best suited for the domain.

- Speaker dependence vs. Speaker independence. A speaker-dependent recognizer is trained to recognizer a particular voice. Each user of the system must speak words, phrases and sentences in a training session that may range from a half hour to three or more hours depending on the recognizer and vocabulary size. A speaker-independent system requires no training for individual users. Obviously speaker independence is a desirable feature; the tradeoff is a reduction in the size of vocabulary that can be used. Current state-of-the-art techniques indicate a rough order of magnitude difference between vocabulary sizes of speaker independent versus speaker-dependent systems.

- Discrete speech vs. Continuous speech A discrete speech recognizer requires users to place pauses between spoken words. A continuous speech recognizer allows users to speak in their natural cadence. Again the tradeoff is in vocabulary size. Continuous speech recognizers tend to have vocabulary sizes an order of magnitude smaller than discrete speech recognizers. However, recent advances in the area have reduced this gap in speaker-dependent systems [Ryan and Mokhoff, 1997].

- PC platform vs. UNIX workstation. Recent advances in PC architectures and lower memory cost have helped equalize the difference in speech recognition on these two platforms. Nonetheless, several high-performance speech recognition engines are available only on UNIX workstation platforms that are not available on PC platforms.

For the AMAT domain, we selected a speaker independent PC-based recognizer, IBM's VoiceType Application Factory. This system is speaker independent, continuous speech, and works on a PC platform through C API function calls. Vocabulary size is limited. For the grammars that we used, we found an *active* vocabulary of approximately 200 words was the limit before unacceptable performance degradation. The relatively small vocabulary size may not prove a hindrance in many domains. Fink demonstrated that there was little functional difference between user vocabularies of 100 words versus 500 words in simple assistant repair domains [Moody, 1988]. Further, the *effective* vocabulary size of the system can be made larger than 200 words by changing the *active* vocabulary based on dialogue context (see discussion below on Dynamic Grammars). We opted for a speaker independent recognizer as one of the potential uses of the system is a refresher trainer at National Guard home stations. Thus, training time is limited. Further, the PC platform was chosen for cost; there is no special hardware for this system.

1.4.2 Parsing

AMAT uses a **Minimum Distance Translator (MDT) parser** [Hipp, 1992]. This parsing technique tries to match the spoken words to the closest grammatical sentence as defined by the currently active language grammar(s). Thus, a user could speak an utterance that is out-of-grammar and be understood. AMAT's Language Parser may be able to correctly interpret the utterance if it is close to something in the grammar. For instance, the user's utterance "tank working" might be correctly matched with "the tank is working."

The main advantage of using a minimum distance translator parser is its ability to handle out-of-grammar utterances. Not only do speakers frequently leave out articles and "unimportant" words (particularly when talking to computers ([Chapanis, 1981, Eastman and McLean, 1981]), but speech recognizers often insert or delete short words like "the", "a" and "and". Most parsing strategies require accounting for every word. Thus, a misrecognized utterance such as “the switch is the up” would be unintelligible by most parsers.

1.4.3 *Dynamic Grammars*

A grammar specifies the language to be accepted by the parser. AMAT’s Language Grammar is a model of acceptable spoken statements. In the following example grammar,

```
S -> WHEREIS COMPONENT’ : ask(location(COMPONENT’ )).
WHEREIS -> help me LOCATE .
WHEREIS -> where is .
LOCATE -> locate.
LOCATE -> find.
COMPONENT -> the laser range finder : lrf.
COMPONENT -> the range finder : lrf.
COMPONENT -> the gunners control handle : gch.
```

the sentence “help me find the gunners control handle” will return the semantic statement “ask(location(gch)”. Note that the above grammar is a **semantic grammar**. A semantic grammar uses semantic categories to categorize syntactic components. This creates a more efficient parsing strategy and greatly assists in handling ambiguity. The disadvantage of semantic grammars is that they tend to be domain specific. The representation language is quite free; literally any sentence can be encoded in the grammar. A statement in the grammar has the form

S -> <token$_1$> <token$_2$> ... <token$_m$> : M

where S stands for some non-terminal symbol, <token$_i$> stands for either a non-terminal or a terminal symbol and M stands for the semantics of the statement that will be returned from the parser. M may contain variables instantiated from parsers of the non-terminal tokens.

The Dialogue Controller dynamically selects which grammars should be active based on the current context. This increases the reliability of the speech recognition and also speeds up the parsing process. The actual process by which AMAT selects which grammars are active is proprietary. However, the essence of this selection is that the system examines its goal stack and its model of the user’s goal and chooses grammars corresponding to those goals. Thus, during knowledge base development links must be made between the domain knowledge and subsets of the total grammar. Automated tools to assist in this process are one of the important future goals.

1.4.4 Language Interpretation

A difficulty in processing natural language is the amazing amount of ambiguity inherent in human language. Humans are quite adept at understanding utterances spoken in context even when the actual utterances may be quite cryptic standing alone. For example, taken alone “It is” has very little meaning. However, if this statement follows a command “Put the switch up”, a very reasonable interpretation might be “The switch is up”. Note that “the switch” is also ambiguous in an environment where there are multiple switches. One solution to the problem of pronouns and anaphora is to disallow them in the dialogue. However, this restriction on spoken interaction makes the speech so unnatural that the benefit of having a spoken interface may be lost.

The AMAT system handles anaphora via two mechanisms: context switching and expectation. As mentioned in the Dynamic Grammars section, AMAT changes the active grammars based on the current goals to be solved. One ramification of this is the subject in the statement “The switch is up” may be resolved quite naturally if there is only one switch in the current set of active contexts. However, context-changing grammars would not necessarily resolve the reference “the switch” if there are multiple switches that can be talked about in the current situation. Therefore, AMAT also uses utterance *expectations* that are based on techniques developed by Smith and Hipp [Smith and Hipp, 1995]. Recent utterances in the dialogue define certain expectations of what will be said next. These expectations are sorted by likelihood and can be fed into the parser to resolve ambiguities. For instance, the utterance “Set the laser range finder safety switch to safe” defines a set of expected replies with semantics such as

“fact(switch(lrf,safe))”,
“fact(switch(lrf,armed))”,
“ask(location(switch,lrf))”,

and so on. An utterance “The switch is safe” would be parsed as “fact(switch(*,safe))” where ‘*’ is a wildcard. This would then be matched with the expectation “fact(switch(lrf,safe))”.

1.4.5 Reliability Scores

One of the difficulties in spoken dialogue systems is the recovery from errors. When the system misunderstands something the user has said, the fact that the system misunderstood may go unnoticed for several dialogue turns. Therefore, a great amount of work may have to be undone if that is even possible given the situation. Therefore, it is tempting to have the system parrot back what it believes the user said in order to confirm understanding. In practice, this procedure would be extremely annoying and may make the user of spoken interaction cumbersome enough that another mode of communication may be more appropriate. An intermediate position is to have the system compute the likelihood of having correctly understood a particular utterance. In determining the likelihood of understanding an utterance, several factors may be taken into account:

- Speech recognizer score. Many speech recognizers return a score of the “goodness of fit” between the audio signal and the acoustic model it has of spoken language.
- Parser score. A minimum distance parser returns a score based on the number of insertions and deletions needed to make the string of words parseable by the language grammar.
- Expectation score. Based on context some utterances might be high, medium or low probability. For example, if the computer and the mechanic are buried deep in a particular troubleshooting procedure after it has been determined that the laser range finder is broken, the utterance “The tank is working” is low probability. On the other hand, if the computer asks “What is the position of the switch?”, the replies of “The switch is up” or “The switch is down” are fairly high probabilities.
- User competency. Users may be rated on how competent they are in completing certain goals. For instance, if it has been determined that there is a problem with the laser range finder and we are dealing with a novice mechanic, the statement “I have diagnosed the fault” will have a low probability as compared to an expert mechanic.
- Recognition rate. Simply put, if there have been many previous recognition errors, there are likely to be more.

We combine these factors by normalizing each score to be between 0 and 1 and then multiply them together. If the resulting score falls below a certain threshold, the system asks the user to repeat. If the score falls into the range, the system paraphrases what it believes the user and asks for a confirmation (i.e., “Did you mean to say you have fixed the tank?”). If the score falls above a certain threshold then the system accepts its interpretation and continues.

We add an additional factor to allow to compensate for critical goals; specifically, we change the thresholds used for requests for restatement, confirmation, and acceptance depending on the “weight” of a goal. Certain goals in a task may be critical. Successful understanding of the results of the goals may require more exacting verification. These weightings are assigned during knowledge base creation.

1.4.6 Goal-driven Dialogue Behavior

The **dialogue controller** is the heart of the AMAT system. This dialogue controller is goal-driven and is based on the model of dialogue developed by Guinn [Guinn, 1994]. Goal-driven dialogue behavior provides a link between expert system technology, automated planning, and natural language processing. The system’s *raison d être* is to solve goals. To solve goals, it may decompose a goal into simpler subgoals. If it cannot solve a goal itself, it may ask its collaborator (the user) for assistance. The system also understands that its collaborator (the user) is solving goals and it maintains a model of what it believes is the user's intentions. Depending on its mode of operation, whether AMAT is acting as a teacher or as an assistant, and the context, it decides whether to pursue its own goals or the goals of the user. When the user asks a question, this modifies the system's model of the user's goals. When the user provides information, this modifies the system's dynamic knowledge.

Static and Dynamic Knowledge. Certain modules of knowledge within the system are static; they do not change during the dialogue. For instance, the troubleshooting tree that might accompany a particular fault is static as is the location of the major components of the M1A1. On the other hand, situation specific knowledge is dynamic. The position of switches, what procedures have been carried out, the model of the user's knowledge all change during the course of a dialogue.

1.4.7 Interaction with the Virtual Environment

There are two dimensions to the user and system's interaction with the virtual environment: three-dimensional movement and object manipulation. Because of the sometimes awkwardness of navigating with a multi-button mouse in a 3D environment, AMAT also allows the user to *navigate* using voice commands. Examples of movement commands are "Turn right 45 degrees", "Look down", "Zoom forward quickly", "Go to the commander's chair", and "Show me the gunner's control handle." Similar processing could have allowed the user to use voice to *manipulate* objects. For example, "Put the laser range finder to the safe position" would cause the virtual switch to flip. However, after talking with educational specialists, it was decided that such an interaction would detach the trainee too much from the virtual environment[cite]. One of the big advantages of virtual reality training is the immersive effect; the trainee gets a sense of what it is actually like in the environment. Having the trainee actually use the mouse to flip a switch, open a door, or connect a wire all help to facilitate the immersive sensation.

The system treats all user actions in the virtual environment as potential *dialogue* inputs. Thus, the system maintains a knowledge of the visual environment. When a user opens a door or flips a switch, that information is passed on to the dialogue controller. In the training mode, where the system is having the user work through a lesson, inappropriate actions in the virtual environment are commented on and undone *by the computer*. This method allows the trainee to make an error but be corrected and continue down the lesson path. In the more free-form assistant mode, the system has the option of commenting on inappropriate actions or remaining quiet. Based on subject interactions in assistant mode, we found it preferable to give the trainee the opportunity to find his or her own errors. If they are not caught during the troubleshooting, the trainee will not successfully complete the task and the mistakes can be brought out in an after-action review.

1.4.8 Voice Synthesis

AMAT can communicate with the user by modifying the Virtual Environment (text boxes, arrows, movement, flashing components) and also by speaking. There are two methods by which a computer can talk using spoken language: recorded speech and synthesized speech. **Recorded speech** has terrific quality. However, it has the disadvantage of having to have every possible system utterance recorded. This recording can take significant disk space as well as making a less flexible development environment. A way of overcoming this is to have certain words and phrases recorded and then piecing together those recorded words/phrases into sentences. Currently, it is very hard to make this speech sound very natural because of the gaps between words. **Synthesized speech** involves taking written text and converting it to a corresponding audio sound. Text-to-speech systems are highly flexible as they generally do not need to be retrained as you add more words to a system's vocabulary. The disadvantage of these systems is that synthesized speech still sounds like a robot talking. While highly intelligible, no one would mistake a synthesized voice for a human voice. AMAT uses a Digital Equipment DECtalk text-to-speech synthesizer.

2.0 FUTURE DIRECTIONS FOR RESEARCH AND APPLICATIONS

Technologies in the area of speech recognition and natural language processing are rapidly changing and will continue to develop and a rapid pace over the next decade. There are four key areas of research: large vocabulary, speaker independent, continuous speech recognizers, large vocabulary parsing and understanding systems, and intelligent user modeling.

2.1 Large Vocabulary Recognizers

The state-of-the-art in speech recognition is changing rapidly. Specialized continuous speech recognizers in the 20,000 to 40,000 word range have been developed for specific domains (law, medicine). In the near term, we can foresee large vocabulary, general purpose speech recognition on PC platforms. These advancements will of course have a beneficial effect on spoken dialogue systems as one of the major hurdles is recognition rate. However, having a large vocabulary speech recognizer does not mean you will have a large vocabulary **dialogue** system. Dialogue requires understanding as well. The system must map the words to some semantics that it is capable of processing. The semantic grammars used in AMAT do not scale particularly well to large vocabulary systems — they are much more appropriate for specific domains.

2.2 Large Vocabulary Parsing and Understanding Systems

As vocabulary size increases, parsing and understanding becomes more difficult. The ambiguities in the language become much more complex to analyze. In the parsing domain, there have been a number of very successful large vocabulary parsing mechanisms [Bunt and Tomita, 1996]. However, the process of mapping these parsers to meaningful semantics is an ongoing research challenge. How do we represent knowledge in general? In the near term, it is unlikely that we will build a general purpose knowledge understanding system. Thus, our efforts will be focused on large vocabulary understanding in finite domains. It is quite possible that semantic grammars may be sufficient for this task particularly if the developer is provided with tools and libraries of grammars with which to work.

2.3 Intelligent User Modeling

The AMAT system only makes rudimentary models of the user's intentions and abilities. An initial assessment of the user's capabilities is based on a **stereotype** based on the user's experience in the domain (i.e., the mechanic's grade level). This assessment is modified based on the questions the user poses to the computer as well as how readily the user is able to accomplish tasks. We consider a flexible, user adaptable interface to be one of the most critical components in building a fieldable system. Perhaps the most critical ability the system needs is the capability of dynamically changing **dialogue initiative**. The dialogue initiative reflects which participant is in control of the dialogue. In task-oriented domains it is important for the most knowledgeable participant to be in control. For instance, it would be very frustrating for an expert mechanic to be reigned in by an inflexible expert system. On the other hand, an inexperienced mechanic may make undesirable errors if allowed to make decisions without the expert system's assistance. This switching back and forth of initiative will change during the dialogue based on the goals being tackled. Promising headway in this area has been made by comparing the system's Bayesian analysis of the troubleshooting with a probabilistic model of the user's capabilities [Guinn 1996].

3.0 CONCLUSIONS

Several novel technologies have been developed to make spoken natural language systems a reality:

- Dynamic grammars to allow small vocabulary recognizers to effectively have larger vocabularies.
- Error correcting parsing that allows the user to stray from the defined grammars and still be understood.
- Utterance reliability scoring to initiate clarifications.
- Fusion of virtual reality, expert systems, and natural language processing.

These technologies allow for spoken human-computer systems to emerge from laboratory settings and become useful for real-world applications.

4.0 REFERENCES

Bunt, H. and Tomita, M., *Recent advances in parsing technology*, 1996.

Chapanis, A., Interactive Human Communication: Some Lessons Learned from Laboratory Experiments, *Man-Computer Interaction: Human Factors Aspects of Computers and People*, 1981.

Eastman, C.M. and McLean, D.S., On the Need for Parsing Ill-Formed Input. American Journal of Computational Linguistics, 7:4(257), 1981.

Guinn, C. Mechanisms for Mixed-Initiative Human-Computer Collaborative Discourse, in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 1996.

Guinn, C,. *Meta-Dialogue Behaviors: Improving the Efficiency of Human-Machine Dialogue - A Computational Model of Variable Initiative and Negotiation in Collaborative Problem-Solving*, Ph.D. thesis, Duke University, available as Technical Report, CS-1995-02, 1994.

Hipp, D.R., *Design and development of spoken natural-language dialog parsing systems*. Ph.D. thesis, Duke University, available as Technical Report, CS-1993-15, 1992.

Lehman, J.F. and Carbonell, J.G., Learning the User's Language: A Step Towards Automated Creation of User Models, in *User Models in Dialog Systems, 1989.*

Moody, T.,. *The Effects of Restricted Vocabulary Size on Voice Interactive Discourse Structure*, Ph.D. thesis, North Carolina State University, 1988.

Psotka, J. and Davison, S., Cognitive Factors Associated with Immersion in Virtual Environments. 1993 Available on at http://205.130.63.7/cognition_of_immersion.text.

Psotka, J., Davison, S. A., and Lewis, S.A. Exploring immersion in virtual space. *Virtual Reality Systems*, 1(2), pp. 70-9, 1993.

Rudnicky, A. I., Mode Preference in Simple Data-Retrieval Task, *Proceedings of the ARPA Human Language Technology Workshop*, 1993.

Ryan, M. and Mokhoff, N., “Continuous-dictation system aims at the PC”, *Electronic Engineering Times*, April 14, 1997.

Adams, N. (1995, June). Lessons from the Virtual World. *Training*. 1995.

Kenney, P.J., & Saito, T. Results of a Survey on the Use of Virtual Environment Technology in Training NASA Flight Controllers for the Hubble Space Telescope Servicing Mission. VETL Technical Report. 1994.

Magee, L.E. Virtual Reality and Distributed Interactive Simulation for Training Ship Formation Maneuvers. Proceedings of the 36th NATO Defense Research Group (DRG) Seminar. 1995.