WIRE: A Wearable Spoken Language Understanding System for the Military

Helen Hastie **Patrick Craven** **Michael Orr**

Lockheed Martin Advanced Technology Laboratories
3 Executive Campus
Cherry Hill, NJ 08002
{hhastie, pcraven, morr}@atl.lmco.com

Abstract

In this paper, we present the WIRE system for human intelligence reporting and discuss challenges of deploying spoken language understanding systems for the military, particularly for dismounted warfighters. Using the PARADISE evaluation paradigm, we show that performance models derived using standard metrics can account for 68% of the variance of User Satisfaction. We discuss the implication of these results and how the evaluation paradigm may be modified for the military domain.

1 Introduction

Operation Iraqi Freedom has demonstrated the need for improved communication, intelligence, and information capturing by groups of dismounted warfighters (soldiers and Marines) at the company level and below. Current methods of collecting intelligence are cumbersome, inefficient and can endanger the safety of the collector. For example, a dismounted warfighter who is collecting intelligence may stop to take down notes, including his location and time of report or alternatively try to retain the information in memory. This information then has to be typed into a report on return to base. The authors have developed a unique, hands-free solution by capturing intelligence through spoken language understanding technology called WIRE or Wearable Intelligent Reporting Environment. Through WIRE, users simply speak what they see, WIRE understands the speech and automatically populates a report. The report format we have adopted is a SALUTE report which stands for the information fields: Size, Activity, Location, Unit, Time and Equipment. The military user is used to giving information in a structure way, therefore, information entry is structured but the vocabulary is reasonably varied, an example report is "Size is three insurgents, Activity is transporting weapons." These reports are tagged by WIRE with GPS position and time of filing. The report can be sent in real-time over 802.11 or radio link or downloaded on return to base and viewed on a C2 Interface. WIRE will allow for increased amounts of digitized intelligence that can be correlated in space and time to predict adverse events. In addition, pre and post-patrol briefings will be more efficient, accurate and complete. Additionally, if reports are transmitted in real time, they have the potential to improve situational awareness in the field.

This paper discusses the challenges of taking spoken language understanding technology out of the laboratory and into the hands of dismounted warfighters. We also discuss usability tests and results from an initial test with Army Reservists.

2 System Overview

WIRE is a spoken language understanding system that has a plug-and-play architecture (Figure 1) that allows for easy technology refresh of the different components. These components pass events to each other via an event bus. The speech is collected by an audio server and passed to the Automatic Speech Recognizer (ASR) server, which is responsible for converting the audio waveform into an N-best list. The Natural Language (NL) under

Bridging the Gap: Academic and Industrial Research in Dialog Technologies Workshop Proceedings, pages 84–88,
NAACL-HLT, Rochester, NY, April 2007.

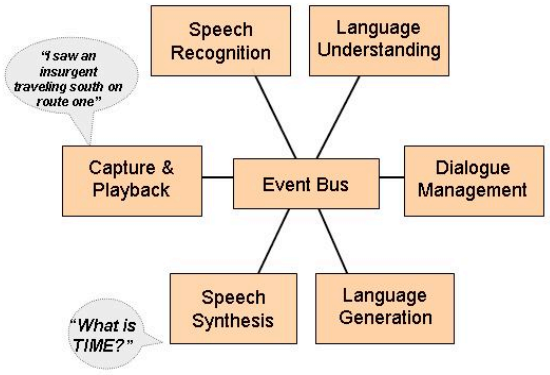

Figure 1. WIRE System Architecture

standing component executes a named-entity tagger to tag and retain key text elements within the each candidate N-best list element. The sets of tagged entities are then parsed using a bottom-up chart parser. The chart parser validates each named entity tag sequence and generates a syntactic parse tree. A heuristic is then applied to select the best parse tree from the N-best list as the representative spoken text. After a parse tree is selected, a semantic parser is used to prune the parse tree and produce a semantic frame—a data structure that represents the user's spoken text. The semantic frame is then passed through a rule-based filter that translates text as necessary for processing, e.g., converting text numbers to digits.

The semantic frame is then passed to the Dialogue Manager which decides what action to take based on the most recent utterance and its context. If the system is to speak a reply, the natural language generation component generates a string of text that is spoken by the Text-To-Speech engine (TTS).

The WIRE spoken language understanding system was fully developed by the authors with the exception of the ASR, called Dynaspeak™, which was developed by SRI International (Franco et al., 2002) and the TTS engine from Loquendo S.p.A. Grammars for the ASR and NL have to be written for each new domain and report type.

In order for the system to adapt to the user's environment, there are two modes of operation. *Interactive* mode explicitly confirms what the user says and allows the user to ask the system to read back certain fields or the whole report. Alternatively, in *stealth* mode, the user simply speaks the report and WIRE files it immediately. In both cases, audio is recorded as a back-up for report accuracy.

3 Challenges of Deployment to Dismounted Warfighters

The goal of WIRE is to provide a means of reporting using an interface that is conceptually easy to use through natural language. This is particularly challenging given the fluid nature of war and the constant emergence of new concepts such as different types of Improvised Explosive Devices (IEDs) or groups of insurgents. Another challenge is that each unit has its own idiosyncrasies, call signs and manner of speaking. Because WIRE is a limited-domain system and it is not possible to incorporate all of this variability, we found training to be a key factor in user and system performance and acceptance.

A new challenge that phone-based or desk-top systems have yet to face is the need for a *mobile* spoken language understanding system that can be worn by the user. From a software perspective, WIRE has to have a small footprint. From a hardware perspective, the system has to be lightweight, robust, and rugged and must integrate with existing equipment. Wearable computing is constantly evolving and eventually WIRE will be able to run on a system as small as a button. We have also been working with various companies to create a USB noise-canceling microphone similar to what the military user is accustomed to.

4 Experiment Design

Fifteen Army Reservists and three former Marines participated in WIRE usability tests in a laboratory environment. The Reservists predominately provide drill-instructor support for Army basic training groups. The session began with a brief introduction to the WIRE system. Following that, participants reviewed a series of self-paced training slides. They then completed two sets of four scenarios, with one set completed in stealth mode and the other in interactive mode. A total of 523 utterances were collected. Participants were asked to complete five-question surveys at the end of each set of scenarios. For the regression model described below, we averaged User Satisfaction scores for both types of interaction modes.

We adopted the PARADISE evaluation method (Walker et al., 1997). PARADISE is a "decision-theoretic framework to specify the relative contribution of various factors to a system's overall performance." Figure 2 shows the PARADISE model which defines system performance as a weighted function of task-based success measures and dialogue-based cost measures. Dialogue costs are further divided into dialogue efficiency measures and qualitative measures. Weights are calculated by correlating User Satisfaction with performance.

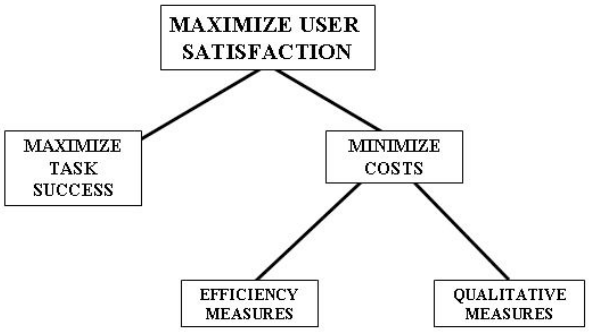

Figure 2. PARADISE Model (Walker et al., 1997)

The set of metrics that were collected are:

- **Dialogue Efficiency Measures:** User Turns, Average Length of Utterance, Average Response Latency and Platform.
- **Dialogue Quality Measures:** Word Accuracy.
- **Task Success Measures:** Report Accuracy, Field Correctness for Size, Activity, Location, Unit, Time and Equipment.
- **User Satisfaction:** Average of User Expertise, User Confidence, System Trust, Task Ease, Future Use.

User Satisfaction is the average of responses from a survey of five questions on a five-point Likert scale with five being the highest rating. These questions include:

- **Q1**: I knew what I could say at any point (User Expertise).
- **Q2**: I knew what I was doing at any point in the dialog (User Confidence).
- **Q3**: I trusted that WIRE accurately captured my report information (System Trust).
- **Q4**: I felt like I could create and file a report quickly (Task Ease).
- **Q5**: I would recommend that this system be fielded (Future Use).

These questions are modified from the more traditional User Satisfaction questions (Walker et al., 2001) that include *TTS Performance* and *Expected Behavior*. TTS Performance was substituted because the voice is of such a high quality that it sounds just like a human; therefore, the question is no longer relevant. Expected Behavior was substituted for this study because WIRE is mostly user initiative for the reporting domain.

The Task Success metric was captured by Report Accuracy. This was calculated by averaging the correctness of each field over the number of fields attempted. Field correctness was scored manually as either 1 or 0, depending on whether the report field was filled out completely correctly based on user's intent. Partial credit was not given.

Various platforms were used in the experiment, including laptops, tablet PCs and wearable computers. The Platform metric reflects the processing power with 0 being the highest processing power and 1 the less powerful wearable computers.

5 Experimental Results

We applied the PARADISE model using the metrics described above by performing multiple linear regression using a backward coefficient selection method that iteratively removes coefficients that do not help prediction. The best model takes into account 68% of the variance of User Satisfaction (p=.01). Table 1 gives the metrics in the model with their coefficients and p values. Note that the data set is quite small (N=18, df=17), which most likely affected the results.

Table 1. Predictive Power and Significance of Metrics

Metric	Standardized β Coefficients	*p* value
User Turns	-0.633	0.01
Unit Field Correctness	0.735	0.00
Platform	-0.24	0.141

Results show an average User Satisfaction of 3.9 that is broken down into 4.09 for interactive mode and 3.73 for stealth. The lowest medium user satisfaction score was for System Trust (3.5), the highest for Task Ease (4.5).

Speech recognition word accuracy is 79%, however, Report Accuracy, which is after the speech has been processed by the NL, is 84%. Individual field correctness scores varied from 93% for Activity to 75% for Location. From previous tests, we have found that word accuracy increases through user training and experience up to 95%.

6 Interpretation and Discussion

These initial results show that the User Turns metric is negatively predictive of User Satisfaction. This is intuitive as the more user turns it takes to complete a report the less satisfied the user. (Walker et al., 2001) have similar findings for the Communicator data where Task Duration is negatively predictive of User Satisfaction in their model (coefficient -0.15).

Secondly, Unit Field Correctness is predictive of User Satisfaction. Given this model and the limited data set, this metric may represent task completion better than overall Report Accuracy. During the test, the user can visually see the report before it is sent. If there are mistakes then this too will affect User Satisfaction. This is similar to findings by (Walker et al., 2001) who found that Task Completion was positively predictive of User Satisfaction (coefficient 0.45).

Finally, Platform is negatively predictive, in other words: the higher the processing power (scored 0) the higher the User Satisfaction and the lower the processing power (scored 1) the lower the User Satisfaction. Not surprisingly, users prefer the system when it runs on a faster computer. This means that the success of the system is likely dependent on an advanced wearable computer. There have been recent advances in this field since this experiment. These systems are now available with faster Intel processors and acceptable form factor and battery life.

The User Satisfaction results show that areas of improvement include increasing the trust in the user (Q3). This challenge has been discussed previously for military applications in (Miksch et al., 2004) and may reflect tentativeness of military personnel to accept new technology. Trust in the system can be improved by putting the system in "interactive" mode, which explicitly confirms each utterance and allows the user to have the system read back the report before sending it. A Wilcoxon signed-rank test ($Z = 2.12$, $p < .05$) indicated that scores for this question were significantly higher for interactive mode (M = 3.93) than stealth mode (M=3.27).

Our current evaluation model uses User Satisfaction as a response variable in line with previous PARADISE evaluations (Walker et al., 2001). However, User Satisfaction may not be the most appropriate metric for military applications. Unlike commercial applications, the goal of a military system is not to please the user but rather to complete a mission in a highly effective and safe manner. Therefore, a metric such as mission effectiveness may be more appropriate. Similarly, (Forbes-Riley and Litman, 2006) use the domain-specific response variable, of student learning in their evaluation model.

An obvious extension to this study is to test in more realistic environments where the users may be experiencing stress in noisy environments. Initial studies have been performed whereby users are physically exerted. These studies did not show a degradation in performance. In addition, initial tests outside in noisy and windy environments emphasize the need for a high quality noise canceling microphone. Further, more extensive tests of this type are needed.

In summary, we have presented the WIRE spoken language understanding system for intelligence reporting, and we have discussed initial evaluations using the PARADISE methods. Through advances in spoken language understanding, hardware and microphones, this technology will soon transition out of the laboratory and into the field to benefit warfighters and improve security in conflict regions.

Acknowledgments

Thanks to the Army Reservist 1/417th Regt, 1st BDE 98th Div (IT).

References

Forbes-Riley, K. and Litman, D.J. "Modeling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters." *HLT-NAACL*, 2006.

Franco, H., Zheng, J., Butzberger, J., Cesari, F., Frandsen, M., Arnold, J., Rao, R., Stolcke, A., and Abrash, V. "Dynaspeak™: SRI International's scalable speech recognizer for embedded and mobile systems." *HLT*, 2002.

Miksch, D., Daniels, J.J., and Hastie, H. (2004). "Establishing Trust in a Deployed Spoken Language System for Military Domains." *In Proc. of AAAI Workshop*, 2004.

Walker, M.A., Litman, D., Kamm, C. and Abella, A. "PARADISE: A Framework for Evaluating Spoken Dialogue Agents." *ACL*, 1997.

Walker, M.A., Passonneau, R., and Boland, J.E. "Quantitative and Qualitative Evaluation of DARPA Communicator Spoken Dialogue Systems." *ACL*, 2001.