Predicting Situation Awareness from Team Communications

Cheryl A. Bolstad[1], Peter Foltz[2], Marita Franzke[2], Haydee M. Cuevas[1], Mark Rosenstein[2], & Anthony M. Costello[1]

[1]SA Technologies
Marietta, GA

[2]Pearson Knowledge Technologies
Boulder, CO

Given the importance of Situation Awareness (SA) in military operations, there is a critical need for a real-time, unobtrusive tool that objectively and reliably measures warfighters' SA in both training and operations. Just as the requirement for improved access to SA measures has become vital, it is now commonplace for military team communications to be mediated by technology, hence easily captured and available for analysis. We believe that team communications can be used to derive SA measures. To address this issue, we are developing the Automated Communications Analysis of Situation Awareness (ACASA) system. ACASA combines the explanatory capacity of the SA construct with the predictive and computational power of TeamPrints, to assess team and shared SA as well as other cognitive processes. TeamPrints is a system that combines computational linguistics and machine learning techniques coupled with Latent Semantic Analysis (LSA) to analyze team communication. In this paper, we present the findings from an exploratory evaluation of how well TeamPrints predicts SA from the team communications arising during a military training exercise.

INTRODUCTION

Situation awareness (SA) is critical for teams to function effectively and to work resourcefully together. SA provides a foundation for decision-making and action. Not surprisingly, the SA construct has received considerable research attention in recent years. There is an urgent need for not only better methods to enhance and train SA, but also for better methods to assess SA among team members. As part of an ongoing research project sponsored by the Office of Naval Research, we are combining the explanatory capacity of the SA construct with the predictive and computational power of TeamPrints. TeamPrints is a system that uses computational linguistics and machine learning techniques coupled with Latent Semantic Analysis (LSA) to analyze team communications.

The result is the Automated Communications Analysis of Situation Awareness (ACASA) system, which addresses the need for a real-time, unobtrusive tool that objectively and reliably measures warfighters' SA in both training and operations. The ACASA system leverages the TeamPrints methodology to predict team SA by analyzing naturally occurring team communications. In this paper, we first discuss the concepts underlying the ACASA system and then present findings from a preliminary evaluation of TeamPrint's usefulness for predicting SA.

Situation Awareness

While many definitions of SA exist, we follow Endsley's (1995) definition, which views SA as "…the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future " (p. 36). This definition encompasses several concepts that are important to understanding the SA construct.

First, SA is comprised of three levels: perception, comprehension and projection. Level 1 SA, perception, involves perceiving critical information from the environment. Level 2 SA, comprehension, involves integrating and comprehending the information in working memory (Salas, Prince, Baker, & Shrestha, 1995) to understand how the information will impact the individual's goals and objectives. This process involves combining individual pieces of information to form a comprehensive picture of the world, or of that portion of the world of concern to the individual. Level 3 SA involves extrapolating this information forward in time to determine how it will affect future states of the operating environment (Endsley, 1988; Endsley, 1993). Level 3 SA combines what the individual knows about the current situation with their mental models or schemata of similar events to predict what might happen next.

TeamPrints

TeamPrints is a system that uses computational linguistics and machine learning techniques coupled with Latent Semantic Analysis (LSA) to analyze team communications. From this analysis, TeamPrints can make predictions of team performance, predict other team characteristics, such as the quality of leadership, or automatically tag communications for relevant discourse events (Foltz, Martin, Abdelali, Rosenstein & Oberbreckling, 2006). Its computational core is based on LSA, which models the semantic relationships of language (Landauer, Foltz & Laham, 1998).

TeamPrints can process an incoming stream of free-form communication data and evaluate it in near real-time. It finds patterns of similarity between communication elements and

correlates these patterns to external team performance measures. These correlations allow making reliable predictions based on analyses of communication data alone. The underlying technique has proven successful in the context of essay scoring (Landauer, Laham & Foltz, 2001) as well as for predicting operator team performance in a number of different military training tasks (Kiekel, Cooke, Foltz, Gorman & Martin, 2002; Martin, & Foltz, 2004, Foltz et al., 2006). In addition to predicting external performance measures, TeamPrints has been successfully used to automatically classify communication events from the Bowers, Jentsch, Salas, & Braun (1998) tagging set.

Using TeamPrints to Predict Situation Awareness

The goal of the work described here was to determine if TeamPrints can make predictions about a range of SA metrics as well as objectively measured team performance. Starting with a training set of communication elements and their associated human rated SA levels, TeamPrints generates a model relating communication and SA level. This model is then used to predict SA levels for new team communications. If successful, we will be able to automatically tag dynamically occurring team communication data for team SA. In addition, the automatically assessed team SA can be used to predict team performance. A candidate set of communication and performance data was selected in order to test the TeamPrints model.

METHOD

The NEO Mission Scenario Data Set

The data set used to evaluate TeamPrints consisted of the expanded Noncombatant Evacuation Operation (NEO) Mission Scenario (Warner, Wroblewski, & Shuck, 2003). The data was collected as part of an experiment that was part of the Collaborative Knowledge in Asynchronous Collaboration (CASC) Phase II Project) and kindly provided by Norm Warner of the Naval Air Systems Command. The scenario was developed utilizing the expertise from operational personnel (Navy Seal, Marine, and Army aircrew). Data was collected from 32 teams of college students with each team consisting of three team members in each scenario. Each team was given one hour to generate a plan, as specified in the following mission statement read to the students:

> *The time is 2:00am, January 15. Your mission is to rescue 3 stranded Red Cross workers from a church basement, on a remote island, caught in the middle of guerilla warfare, within 24 hours. The situation is described in the next few pages along with the assets of US forces that are available to rescue the workers. You need to work together and develop a course of action (using ANY assets available to you), which includes a plan for getting to the church, a plan for evacuating the workers, and a plan for the return to the Army base or aircraft carrier. The course of action solution can be an Army, Marine, Navy Seals solution, or a combination of the assets of the three. You want to choose the optimal and most efficient solution. You want to minimize damage to the village and villagers; you want to avoid contact with enemy if possible, and to rescue the workers safely. However, the rules of engagement are that any forces will defend themselves if needed. Good Luck!*

The experiment consisted of a two-factorial design. First, subjects were working either in a *face-to-face* condition where all communications were synchronous and oral, or a *distributed* condition where communications were asynchronously written notes, mediated through the EWall communications environment (Warner et al., 2003). Second, the scenario provided to the teams either remained the same throughout the whole planning session (*static condition*), or changed at a standard time, making it necessary to adjust the emerging plans (*dynamic condition*). Each team was scored by independent scorers on a 100-point scale as to their team performance. The average score across all 32 teams was 83.8, with a standard deviation of 7.2.

The very different nature of face-to-face communications compared with communications mediated by EWall caused us to limit our analysis in this preliminary study to the face-to-face data. Team communications for the face-to-face data were recorded, transcribed and time-stamped, and made available to us. Seven of the transcripts from the Face-To-Face condition of this experiment were randomly selected for our analyses; four of these were from the dynamic and three from the static condition.

Given that the NEO dataset was collected for related but different purposes, no traditional online process measures (e.g., Situation Awareness Global Assessment Technique – SAGAT) or post-experimental measures (e.g., Participant Subjective Situation Awareness Questionnaire – PSAQ) indicating the actual or self-perceived SA of the team were collected. This led us to experiment with a new way of assessing the teams' SA based on their recorded communications. We had human experts rate the level of SA evidenced by groupings of utterances in the team communications, and then used TeamPrints to model these human ratings. Finally, we used that model to predict SA.

Procedure

We manually rated groupings of utterances from seven of the team communication protocols for indications of SA level as defined by Endsley (1995). We first created a goal directed cognitive task analysis (GDTA) for the NEO domain (see illustrated example in Figure 1). The GDTA seeks to uncover the goals operators have in a particular domain, the decisions that must be made to achieve these goals, and the dynamic information requirements needed to support these decisions grouped by SA level (for more information on GDTA, see Endsley, Bolte, & Jones, 2003).

1. **Develop Plan to get to Church**
   What is the best route?
   1.1. Projected time on route - 3
   1.2. Distance of route -2
      1.2.1. Passable routes by land -2
         1.2.1.1. Danger areas -2
         1.2.1.2. Projected impact of weather on roads -2
            1.2.1.2.1. Projected weather (winds/rain) - 3
            1.2.1.2.2. Terrain -1
            1.2.1.2.3. Road location/type -1

Figure 1. Goal 1 from the NEO GDTA with decisions, information requirements and SA levels.

We developed a protocol for grouping communication exchanges into meaningful, coherent units that provided a larger unit of analysis, approximately paragraph sized, rather than an isolated word or sentence. Two researchers determined the grouping boundaries of the communications and assigned an SA element and SA level to each grouping (see Figure 2).

Subject	Content	GDTA Code	SA Element	SA Level
W	First think I think we have to think about is when we want to do that, and it said that, um, it's really foggy during the morning time and then it gets completely clear by noon, so we probably want to do it somewhere between, you know, 2 am and when it gets real clear during the day so that we're not easily detected, or whatever.	1.2.3.10.1	projected visibility to enemy	3
I	OK			
E	So you say it's going to be easier to do it during, like, nighttime?			
W	Well, if we don't want to be seen and we know, like, we know where we're going, and we can get there, when it's like foggy outside so nobody will see...			
E	Oh, ok. I see what you mean.			

Figure 2. Example of a level 3 SA grouping taken from the NEO data.

To develop the protocol for scoring SA, three researchers jointly rated a single transcript. The scored transcript was then shared with other members of the project team for feedback. Based on these discussions, the transcript was rescored and the protocol revised. Six additional transcripts were then rated by two raters, with two of the transcripts being scored by both raters. There were noticeable differences in the number of conversational groupings assigned to SA level between the two scorers in the double-scored transcripts. All groupings and ratings were later revised by a third, more experienced rater. All analyses reported below are based on the final ratings of the third rater; therefore, no standard inter-rater reliability statistics can be reported. It is important to stress the exploratory nature of this work. If our initial findings turn out promising, further research could introduce more standardized tagging methods and investigate generalizability of these methods by examining inter-rater reliability between independent scorers.

RESULTS AND DISCUSSION

Using TeamPrints to Predict SA Level

The next step in this exploration was to show if TeamPrints can be used to automatically tag communication streams for SA level (eliminating the need for time-consuming human tagging), and whether these automatically generated tags can also predict team performance. The TeamPrints software was used to model the language of the transcripts and the human SA level assignment. The experiment used a hold-one-out procedure, where TeamPrints was iteratively trained with 6 transcripts and the resulting model was used to predict the SA levels of the groupings in the 7th transcript.

Table 1 shows the results of these predictions, both in terms of percent exact agreement, (perfect agreement on SA level) adjacent agreement (disagreement by one or less levels), and correlation between the actual scores and the predicted scores. While this performance is better than a random assignment (which in this case would predict an exact agreement of 33.3), in other applications of this type of labeling team communication data, TeamPrints has been substantially more reliable (e.g., Foltz, et al, 2006). However, given the relatively small training set of only 7 transcripts, these findings are promising.

Table 1
Results of TeamPrints Predicting Human SA Rating of a Held Out Transcript

Transcript Held-Out	% Exact Agreement	% Adjacent Agreement	Correlation
Team-1-F2F-S	47.8	98.9	0.38
Team-2-F2F-D	47.3	99.5	0.29
Team-2-F2F-S	54.3	98.4	0.36
Team-4-F2F-D	51.9	99.2	0.42
Team-4-F2F-S	41.5	97.6	0.37
Team-7-F2F-D	53.4	97.7	0.35
Team-9-F2F-D	61.7	100.0	0.44
Mean	51.1	98.8	0.37
Standard Dev	6.4	0.9	0.05

Using Team Prints to Model Human Performance Predictions

Given the promise of automatically tagging communications data using TeamPrints, we now wanted to determine whether the automatic SA level tags would be able to predict team performance in a manner similar to the tags determined by our human expert. To approach this issue, we first have to describe how the human scores can be used to predict team performance scores. Team performance scores were provided to us in the original dataset and consisted of a holistic performance score that varied between 81 and 90 points for the seven missions described above.

Table 2 below shows for each mission, what proportion of tags was assigned by the human expert for each SA level. Visual inspection of this table suggests that lower proportions of SA level 1 and higher proportions of SA level 3 per mission appear to be associated with higher performance scores.

Table 2
Proportion of SA Level Tags Assigned by Human Expert

Team	Percent of SA 1	Percent of SA 2	Percent of SA 3	Performance score
Team-1-F2F-S	0.37	0.41	0.22	87.0
Team-2-F2F-D	0.45	0.41	0.14	81.0
Team-2-F2F-S	0.41	0.43	0.17	90.0
Team-4-F2F-D	0.47	0.34	0.19	86.0
Team-4-F2F-S	0.43	0.27	0.29	82.5
Team-7-F2F-D	0.53	0.34	0.14	83.0
Team-9-F2F-D	0.46	0.47	0.07	81.5

Correlations using the percentages of SA level tags to predict the performance scores confirm this finding, although we did not have enough data points to reach statistical significance for these correlations. Table 3 shows the correlations of SA level percentages with performance for both the human SA ratings and the TeamPrints predicted SA ratings. The strongest relationship appears to manifest itself between the proportion of SA level 1 tags and performance: the more a team is absorbed with perception of information alone, the lower its performance score. The reverse appears to emerge for SA level 3 tags: the more team communications provide evidence for SA level 3, the higher the performance score. We see no obvious explanation between the discrepancy in relationship between the human and predicted scores for SA level 2, and exploration of that issue awaits further research.

Table 3
SA Tag Performance Predictions

	SA 1 Percentage	SA 2 Percentage	SA 3 Percentage
Human	-0.48 (p=.27)	0.12 (p=.80)	0.24 (p=.61)
TeamPrints	-0.38 (p=.40)	0.51 (p=.24)	0.01 (p=.98)

A final step in this exploration was to directly model performance using TeamPrints. In this case, we were not limited to the 7 transcripts and used the entire set of 16 face-to-face transcripts. We used a jackknife predicted correlation (similar to a hold-one-out method) and had an R of .77 (p<.01) between TeamPrints predicted performance and the actual holistic performance score. It is not surprising that using a larger data set and modeling performance directly captures more of the variance of the relationship between communications and performance. It is exactly one of the strengths of SA to provide diagnostic value beyond that of simple performance and one reason why we are interested in accurately predicting it.

Our data analyses had several goals. The first was to show that TeamPrints analyses could be used to automatically tag communication data using an SA level tagging scheme. The data presented in Table 1 suggests that ours is indeed a promising new approach to measuring team SA. On average, TeamPrints was able to predict the exact SA level of a communication segment with 51 percent accuracy. Our second goal was to show that the automatically derived SA tags would predict performance in a way similar to tags assigned by human scorers also seems obtainable. Both human and machine tagging schemes make some of the same predictions, namely that high proportions of SA level 1 in communication data are indicative of low performing teams. It is important to note that our results fell short of statistical significance, but the overall pattern of the results is consistent and suggestive. Further research using larger datasets and more developed coding protocols might provide an extended foundation for this newly established analysis method.

CONCLUSION

By way of introducing a new way of tagging communication according to SA level and showing that these tags were predictive of performance, we also provided evidence for our assumption that conversation streams indeed express the joint building of the situation awareness and decision making that underlies team performance. Analyses of team communications can, therefore, be seen as a valid measure of this construct. We also demonstrated that TeamPrints can be used to tag SA level automatically, and that these automatically derived tags should in principle predict performance accurately and reliably. The results of our preliminary analyses point into the right direction, further investigations would confirm the feasibility of our approach.

Findings from this preliminary investigation will feed into the further development and refinement of our ACASA system. Future work will focus on identifying and selecting additional valid and reliable metrics and techniques that can be incorporated into the ACASA system to unobtrusively and objectively assess team and shared SA as well as other cognitive processes such as shared understanding and team knowledge building.

Our ultimate goal is to create a system that will be highly useful to system integrators by providing a method for evaluating the quality and effectiveness of system design solutions in terms of their degree of supporting operator SA. In addition, ACASA may provide military organizations with a capability they never had before – the ability to determine the SA and performance of warfighter teams in situ, in any training or operational setting. This information can be very important for determining training needs and providing training intervention mechanisms.

ACKNOWLEDGEMENTS

Work on this paper was partially supported by a Phase II Small Business Innovative Research Contract (N00014-05-C-0438) awarded to the first and second authors from the Office of Naval Research (ONR). The views and conclusions contained herein, however, are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR, U. S. Navy, Department of Defense, U. S. Government, or the organizations with which the authors are affiliated. Correspondence concerning this paper should be addressed to Cheryl A. Bolstad, Ph.D., SA Technologies, 76 Lillian Court, Forest Hill, MD 21050, email: cheryl@satechnologies.com.

REFERENCES

Bowers, C., Jentsch, F., Salas, E., & Braun, C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors*, 40, 672-679.

Endsley, M. R. (1988). *Design and evaluation for situation awareness enhancement*. Paper presented at the Human Factors Society 32nd Annual Meeting, Santa Monica, CA.

Endsley, M. R. (1993). A Survey of situation awareness requirements in air-to-air combat fighters. *International Journal of Aviation Psychology, 3*(2), 157-168.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M. R., Bolte, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to human-centered design*. New York, NY: Talyor & Francis.

Foltz, P.W., Martin, M.J., Abdelali, A., Rosenstein, M., & Oberbreckling, R. (2006). Automated Team Discourse Modeling: Test of Performance and Generalization. In Proceedings of the Cognitive Science Annual Meeting.

Kiekel, P.A., Cooke, N.J., Foltz, P.W.. Gorman, J., and Martin, M. (2002). Some Promising Results of Communication-Based Automatic Measures of Team Cognition. In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Landauer, T.K. Laham, D. and Foltz, P.W (2001). Automated essay scoring. IEEE Intelligent Systems. September/October 2001.

Martin, M.J. , and Foltz, P.W. (2004). Automated Team Discourse Annotation and Performance Prediction using LSA. Proceedings of the Human Language Technology and North American Association for Computational. Linguistics Conference (HLT/NAACL), May 2-7, 2004, Boston, Massachusetts.

Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (1995). Situation awareness in team performance: Implications for measurement and training. *Human Factors, 37*(1), 123-136.

Warner, N., Wroblewski, E., & Shuck, K. (2003). Noncombatant Evacuation Operation Scenario, Naval Air Systems Command, Human Systems Department (4.6), Patuxent River, Maryland