Chat Analysis for After Action Review

Andrew Duchon, Cullen Jackson
Aptima, Inc.,
Woburn, MA
aduchon@aptima.com, cjackson@aptima.com

ABSTRACT

The after action review (AAR) is an essential component to training. To better focus AARs on critical points of team performance, Aptima previously developed a software tool called SPOTLITE that incorporates Mission Essential Competencies[SM] and guides observer/instructors easily through assessment questions during live-virtual-constructive training. However, even with such tools to systematize the assessment and debrief process, performance measurement is very labor intensive. Methods are needed to automate this process, and at least approximate the final results in order to guide the AAR. One form of data that is available digitally, and thus could be a basis for such an automated process, is Internet Relay Chat. Chat allows an operator to monitor more than one channel (or chat room) at once, is persistent so operators can trace back the information and decision process, yet is as instant as radio communications. Chat is used extensively in command and control operations, and during training is itself a major point of focus for AARs. We previously presented (I/ITSEC 2009) measures of communications performance based on chat that changed over the course of a week's training at an Air Force Training Research Exercise (TREX) which models the Dynamic Targeting Cell of an Air and Space Operations Center. Since observer-based performance data also was recorded during these exercises, we investigated predicting these scores using communication measures applied to chat. Results indicate that these measures can provide an adequate ($r^2 > 0.6$) proxy for the average of the manual observer-based performance assessment of the team's performance; thus, chat analysis could provide a mechanism for missions to be ranked automatically for AAR.

ABOUT THE AUTHORS

Andrew Duchon, Ph.D., is a Lead Scientist on the Communications Dynamics Team in Aptima's Analytics, Modeling and Simulation Division. He applies statistical natural language processing to projects requiring the analysis of text or verbal information to gain better understanding of a situation. Dr. Duchon holds a Ph.D. in Cognitive Science from Brown University, and a B.A. in Psychology and the Integrated Science Program from Northwestern University. He is a member of the Association for Computational Linguistics.

Cullen Jackson, Ph.D., has expertise in human performance measurement and assessment, human visual perception and cognition, and knowledge elicitation. Dr. Jackson applies his knowledge of human perceptual and cognitive capabilities to designing systems and techniques for measuring the impact of technology on human performance effectiveness. Dr. Jackson holds a Ph.D. in Experimental Psychology from Brown University, and a B.S. in Computer Science and a B.A. in Psychology from Trinity University. He is a member of the Cognitive Science Society, the Association for Computing Machinery, and the Institute of Electrical and Electronics Engineers Computer Society.

Chat Analysis for After Action Review

Andrew Duchon, Cullen Jackson
Aptima, Inc.,
Woburn, MA
aduchon@aptima.com, cjackson@aptima.com

INTRODUCTION

The after-action review (AAR) is an essential component to training. AARs are facilitated discussions of individual and team performance, generally conducted soon after the conclusion of a training event, with the goal of improving subsequent performance, by helping both trainees and instructors discover and diagnose performance successes and failures in the context of the event (Wiese, Freeman, Salter, Stelzer & Jackson, 2008). To better focus AARs on critical points of performance, competency modeling is used to define and validate higher-order individual and team competencies that a fully-prepared individual and team requires for successful mission completion under adverse conditions (Alliger, Colegrove & Bennett, 2003; Tossell, Wiese, Garrity, Denning & Alliger, 2006). Aptima previously developed a software tool called SPOTLITE (Scenario-based Performance Observation Tool for Learning In Team Environments) that incorporates these competencies and guides observer/instructors easily through assessment questions during live-virtual-constructive training (MacMillan, Entin, Morley & Bennett, in press). In particular, a competency model for the Air Force's Air and Space Operations Center (AOC) was used as a foundation for developing appropriate performance measures for assessing individual and team training for four AOC cells/teams (Jackson, Wiese, Garcia, Wolfe & Stephens, 2008); in this paper, we specifically discuss the measures developed for the Dynamic Targeting Cell (DTC).

Other tools (e.g., CAOC Performance Assessment System) exist to automatically collect data off of some of the primary systems (e.g., JADOCS) used by the DTC (Case, Koterba, Conrad, Ockerman & Vanderberry, 2006). However, these tools do not capture behaviors that exist outside of interactions with those tools (e.g., face-to-face communications; viewing target imagery), which is why observer-based measurement is still so valuable. Unfortunately, this type of measurement is very labor intensive. To systematically assess the performance of the DTC during a training event, observers must rate approximately 150 behaviors in order to cover all of their processes and interactions for one target.

Given the value, but labor-intensive nature, of observer-based measurement, methods are needed to automate this process, or at least approximate the final results. The benefits of this automation will be numerous. For instance, fewer observers will be needed for data collection and/or can be repurposed as exercise controllers. Furthermore, to the extent that performance data is collected in real-time, observers will be able to concentrate on assessing team behaviors across all the various performance measurement systems (rather than just one) during the training event, which significantly will shorten the amount of time between the end of the event and the beginning of the AAR. Shortening this time period will enable trainees and instructors to more easily link assessments with actions during training events, which in turn will enable more contextual diagnoses of performance successes and failures.

One form of data that is available digitally, and thus could be a basis for such an automated process, is Internet Relay Chat. Chat allows an operator to monitor more than one channel (or chat room) at once, is persistent so operators can trace back the information and decision process, yet is as instant as radio communications. Chat is used extensively in operational command and control (AFDC, 2007), and we have observed chat as a major focus of AARs during training events.

Here we report our initial results correlating automated measures of communications in chat data with observer-based measures of performance on a per-target basis. We first give an overview of the data and the scenario from which it was obtained. We then describe the measures of chat we employed and briefly how chat messages about specific targets were disentangled, regardless of explicit mention of the target within a message. We then present the results of using chat communication content measures to predict observer-based performance scores. We conclude with ideas of how these communications measures are able to be utilized in this fashion and how they could be applied in an online, real-time operational environment in addition to the training environment where they can

be used to speed and enhance after action review of performance on a per-target basis.

MILITARY CHAT

Computer-based messaging systems such as email, Internet Relay Chat (IRC), online forums and Web 2.0 social media systems like Facebook and Twitter are critical for many people's social and business lives. They are also playing a more important role in today's military (Heacox, Moore, Morrison & Yturralde, 2004). While much of this use is similar to that in the civilian world in terms of its social uses and business-like planning and coordination, in one realm such messages are time-critical and have life-and-death consequences.

One such arena where this is true is the Air and Space Operations Center (AOC), the operational-level, warfighting command center for joint aerospace forces. It is the senior element of the Theater Air Control System and provides centralized planning, direction, control, and coordination for all joint aerospace operations. The Dynamic Targeting Cell (DTC), one of a dozen teams within the Combat Operations Division of the AOC, is responsible for directing the prosecution of time-sensitive, dynamic, and emerging targets (Time Sensitive Targets, or TSTs). They are often given less than 30 minutes to determine what exactly the TST is, prioritize it, determine the best approach to the target, find an asset that can appropriately prosecute the target, and then assess the results.

All members of the AOC must communicate effectively and coordinate their actions efficiently. They must know with whom (among perhaps hundreds of people) to communicate during a given task, at what times those communications should occur, and the best communications medium in which to have the conversation; sometimes radio, phone, or even face-to-face interaction is necessary. However, text chat is one of the primary media for information dissemination in the DTC and throughout the rest of the AOC (AFDC, 2007; Eovito, 2006).

Joint Automated Deep Operations Coordination System (JADOCS)

Chat is of course not the only tool that operators have in the DTC. The main coordination software is the Joint Automated Deep Operations Coordination System (JADOCS) which integrates a number of tools and maps with which the operators perform their specific jobs such as nominating and vetting potential targets, and once nominated, coordinating strikes and other effects.

Intelligence sources typically provide the first information about a TST in JADOCS, but targets must be approved by at least half a dozen different operators who bring different perspectives. These operators, both within and outside the DTC, must collect a great deal of information to appropriately approve and accurately prosecute the TST.

Some information can be discovered and decisions made independently, for example by looking at role-specific information sources or examining the information in JADOCS itself. However, other information must be obtained from other operators. The primary means of obtaining this information is through IRC.

Chat in the DTC

It is not uncommon for ten targets to be active simultaneously, each with a specified priority and each at a different stage of completion. Information about one target may appear in multiple chat rooms, and discussion of multiple targets may be interleaved within a single chat room.

While chat has been called the "Wild West," the Armed Forces are beginning to develop protocols for IRC (ALSA, 2009). These protocols currently exist at a fairly high level, describing how chat rooms should be named, and what types of information should be conveyed in each room and by whom. The protocol does not currently go into details of message formatting, except to say that messages should be addressed to the intended recipient.

A protocol is difficult because while chat in the AOC is tactically oriented, communication need not be as brief as, for example, the 3-1 Brevity Codes (AFDC, 2001) used by fighter pilots. The nature of the targets in the AOC are diverse and information required to understand them varied, so the language in chat (and why chat is preferred; Eovito, 2006) must capture a wide variety of circumstances. But as this tactical chat approaches normal language use, it lends itself to the same ambiguity, misinterpretation, and confusion, not to mention the complexity of its abbreviated syntax and hastily-entered orthography.

During training and in real-time operations, much of the "action" is taking place within these chat rooms, but it is difficult for leadership (and trainers) to follow these communications which, if improper, could lead to poor results which can often only be discovered when a target has been mishandled, with dire consequences. Thus, an automated assessment of these activities is desirable.

Numerous groups are tackling this issue within the armed forces (e.g., Salter, Duchon & Weil, 2009; La Voie, Foltz, Rosenstein, Oberbreckling, Chatham & Psotka, 2008; Ramachandran, Jensen, Bascara, Carpenter, Denning & Sucillon, 2009) and are even using communications analyses to predict performance (e.g., Foltz, Bolstad, Cuevas, Franzke & Costello, 2008). These studies often apply a large number of undisclosed analyses that lead to performance predictions. Here we describe our analyses in some detail in order to facilitate understanding of how automatic communications analysis can reveal performance.

DATA

The data used in this study came from a Training Research Exercise (TREX 09-2) conducted at AFRL/RHAS in Mesa, AZ in their Part Task Trainer (PTT) AOC Test Bed (Wolfe, Garcia & Denning, 2008). The exercises were conducted over a period of 4 days with six exercise periods of two hours each. The exercises simulate a Dynamic Effects Cell (DEC) within an (AOC). The DEC is an extension of the DTC, which is only concerned with "kinetic" effects (KE). In modern warfare however, destruction of the target may not be the most advantageous approach and the concept of the DEC is being developed to enable other effects to be effectively administered.

The figure below shows the organization of the 11 major players in the DEC at TREX. The major decision makers are the SIDO (Senior Intelligence Duty Officer) who must nominate and validate targets, and the DECC (DEC Chief) who must approve the tasking.

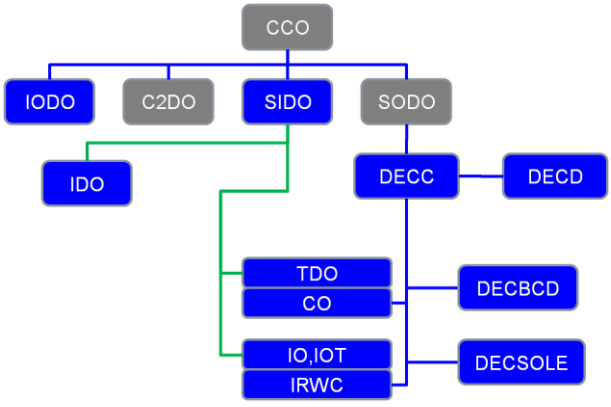

FIGURE 1. ORGANIZATION OF THE DEC IN TREX (GRAY ROLES ARE SIMULATED).

In the TREX, the rest of the AOC is composed of "White Force" players who play numerous roles, mostly for providing intelligence inputs to the DEC and assessments of their decisions. Time-sensitive targets are "injected" into the scenario at various times based on the Master Scenario Event List (MSEL). As these injects occur, the targets must be vetted and prioritized, the appropriate effect decided upon, the effect coordinated, tasked, and finally assessed. Three to five dynamic targets are injected per hour, which is sufficient to keep most participants very occupied. The main goal is to have all targets tasked appropriately and "moved off the floor" as quickly as possible.

SPOTLITE DTC Questions

Observer-based performance scores were obtained via SPOTLITE DTC for the final 3 days of the four day exercise. Each day, two observers graded the performance on two targets, for a total of 12 targets. For each target, we used the average aggregate observer score on those measures with a Likert rating (1-5, with 5 being the best score). This aggregate was composed of 7 to 39 measures per target. We also took an average score from a subset of nine SPOTLITE DTC measures which were oriented towards communications. This aggregate was comprised of 1 to 8 measures per target. These measures were related to data requests, risk analysis, joint coordination, and monitoring engagement progress. With more data (i.e., more observer-assessed targets), it may be feasible to determine which questions in particular may be best predicted by the communications measures.

METHODS

The goal of this study was to use measures of chat communications to predict observer-based performance scores. Because observer scores were recorded for each target individually, the threads of chat messages about the different targets first needed to be "disentangled." Once done, measures of the content of just those messages about a particular target were applied to make predictions about that target.

Message Threading

The first task was to disentangle the threads of communications about the various targets that might be active at one time. The problem of conversation "disentanglement," "threading," "de-threading," "de-interleaving" or "conversation extraction" is beginning to be addressed in the academic literature. The basic problem is that of clustering the messages using a variety of features by which two messages can be similar. Elsner and Charniak (2008) used features that were: a) chat-specific: the time between messages, the same sender, or mentioning other senders, b) discourse-based: using questions and greetings, and c) content-based: looking at shared words between the messages.

Wang et al. (2008) expanded the content-based similarity by using a large corpus to first create a semantic space in which word similarity could be defined. Wang & Oard (2009) expanded this representation to include content from messages from the same author (the idea being that the same author is likely to continue speaking about the same thread), the content of messages from those explicitly addressed, and the content of messages nearby in time.

These techniques performed well at agreeing with human judgments of which messages were in a thread when only looking three adjacent messages at a time, typically around 75%. However, when trying to extract entire threads, or conversations, they agreed only about 50% with humans (for whom the task is also difficult).

Semi-supervised kernel k-means clustering

With TREX chat data, there is a subset of the messages related to a particular mission which explicitly mentions the mission ID number (e.g., "JA0013"). As a result, we can use *semi-supervised clustering* to take advantage of these explicit mentions in order to perform the threading function and extract all the messages about a particular mission/target.

In contrast to the unsupervised clustering algorithms used in previous work, semi-supervised clustering algorithms aim to make use of a limited amount of background knowledge to improve the overall clustering solution. In semi-supervised clustering, prior knowledge is expressed as constraints regarding whether two data points should be in the same cluster. *Must-link* constraints require that pairs of data points must appear in the same cluster, while *can't-link* constraints indicate that data points should not.

In semi-supervised clustering, the objective function to be maximized includes these must-link and can't-link constraints. Basu, Bilenko and Mooney (2004) proposed an objective function based on weighted kernel *k*-means clustering. Kulis, Basu, Dhillon and Mooney (2009) proposed two modifications: a reward for constraint satisfaction if points are in the same cluster, and weighted penalties based on cluster size. We employ a modified version of this algorithm to cluster messages related to the same mission. First, we create must-link constraints between all messages that mention the same mission ID, and can't-link constraints between each message that mentions a mission ID and each other message that mentions a different mission ID.

To obtain mentions of a mission, we first look at the mission IDs themselves, finding exact and inexact matches ("O" for zero, or zeros missing, etc.). We then find all the entries in the JADOCS table that are tied to that mission ID (and the number of these for a given exercise period provide the *k* in the *k*-means clustering). We extract text from seven fields holding text data from JADOCS which the operators and controllers enter. Terms which appear for multiple missions are removed. Other general terms are also removed (e.g., "site"). An additional cluster is created to account for "Overhead" discussions and contains terms such as "test" or "chat check." Any message that has either explicit mentions, JADOCS-based mentions, or Overhead mentions is included in the must-link and cant-link designations.

The initial similarity matrix is created from a weighted kernel matrix of similarities. These similarities, like those mentioned above, are composed of a number of features of the data. Here, we investigate four such features, each of which is conditioned on the pair of messages appearing in the same chat room. *Temporal Similarity* relates two messages with a normal probability distribution $N(0,60)$, based on the number of seconds between the two messages. *Sequential Similarity* used a logistic function of the sequential difference (regardless of time) between two messages. *Jaccard Similarity* uses the extracted features of the two messages (minus any addressees): the number of features in common divided by the total number of unique features in the two messages. Finally, the *Sender-Addressee Similarity* uses the extracted addressees of a message to relate it to all messages sent by those addressees.

These similarity measures are weighted uniformly to give the initial similarity space. After the initial clustering the similarity measures are re-weighted based on how many must-link and can't-link constraints they violate (similar to Bilenko et al., 2004). The entire algorithm is repeated until the clusters stop changing.

Previous Clustering Results

This technique was applied to a "gold standard" of message threads created by a subject matter expert (SME) for a previous TREX exercise (TREX 09-1). Each message was manually coded with a mission ID, or "Overhead" if it concerned the mechanics of the exercise, or "Orphan" if it could not be determined which mission ID the message concerned. Three of the six exercise periods were manually coded.

Table 1 gives a summary of the input data. Explicit mentions of a mission ID (even accounting for typos) appear in fewer than 30% of the messages. The SME was able to find messages about a particular target in up to 83% of the messages. Our goal was to accurately tag messages with the same mission ID as the SME, but we did not count as incorrect if we tagged a mission ID

to a message that he had labeled either as overhead or orphan.

Table 1: Message statistics from TREX 09-1.

	Period 0	Period 2	Period 5
Total Messages	581	625	714
Explicit Mission ID (with typos)	110 (19%)	131 (21%)	187 (26%)
Manual Mission IDs	327 (56%)	442 (71%)	591 (83%)

Using semi-supervised clustering with only the must-link and can't-link constraints gave precision of 53-65% on the three periods (recall values were always within a couple of percentage points). Various combinations of the measures were tested and a four-way ANOVA was conducted with the three periods as repeated measures. Only the Temporal ($F(1,2) = 39.1$, $p < 0.03$) and Sequential Similarity ($F(1,2) = 30.88$, $p<0.04$) had significant main effects. Two-way interactions indicated that both Jaccard and Sender-Addressee actually reduced performance if either of the other two measures were present, but Temporal and Sequential Similarity together can slightly improve over each alone ($F(1,2)=93.5$, $p<0.02$), resulting in precision of 77-85%.

This threading system was applied to TREX 09-2 data without change to produce the "CIFTS Threads." In addition, the messages from TREX 09-2 were hand-tagged by three subject matter experts from AFRL/HEA for these three days of the exercise. We will call these the "SME threads."

Content Measures

For the communications analysis, we use measures from Aptima's CIFTS software (Communication and Information Flow Tracking System, e.g., Salter, Duchon & Weil, 2009). This software automatically categorizes each message into a variety of types. We report here the use of Dialogue Acts.

Due to the uniqueness of the domain, and the lack of annotated data for the dialogue acts necessary for machine learning (e.g., Stolcke, 2000; Ivanovic, 2005), we apply a rule-based approach using term and regular expression matching, examples of which are shown in Table 2. These techniques are similar to those for information extraction when data are extremely "noisy" like chat and not as clear, grammatical, or spelled correctly as, e.g., *Wall Street Journal*, on which more formal methods have been developed (for discussion see e.g., Creswell, Schwartzmyer, and Srihari, 2007). Other groups have taken a similar approach to military chat such as Berube, Hitzeman, Holland, Anapol & Moore (2007) who use a "Military Language Pre-Processor" to pull out mentions of latitude, longitude, call signs, and other semi-structured data types. Budlong, Walter & Yilmazel (2009) perform a similar analysis to that described here to search for military chat messages indicating "Urgency" or "Uncertainty" though they also combine this with a learned model. As we were not concerned with accuracy of the labeling on a per message basis, so much as the preponderance of the label types in a mission thread, our rule-based approach seemed adequate for now. Production of an annotated dataset and application of machine learning methods is left to future work.

Table 2: Dialogue Act Examples.

Category	Description	Examples
Positive Reply	Positive answers	Yes, affirm, as fragged, concur
Negative Reply	Negative answers	no, unable, unknown
Politeness	Unnecessary polite terms	Please, sir, thanks, sorry
Acknowledgment	Ack. of the receipt of information or command	Roger, c, copy, wilco
Anticipation	Asking to be asked	Do you want ___
Question	Indications of a question being asked	?, anything, do we, I need, is there
Correction	Explicit changes	Correction, I meant, adjust
Confirmation	Uncertainty or need of confirmation	Are you sure, confirm, repeat, resend
Proposal	Polite commands	We should, advise, if you
Ambiguity	Indications or recognition of ambiguity	Which convoy, talking about, believe, think

RESULTS

The basic approach was to perform a forward stepwise regression of communications measures against the observer scores. We correlated the SME-threaded and CIFTS-threaded communications measures against the overall aggregate and communications-oriented observer scores.

Message Threading

We applied the message threading techniques designed for TREX 09-1 data to the TREX 09-2 data without modification. Table 3 shows the precision and recall of the CIFTS message threading when compared to the SME message threading for all messages. The results are shown for all missions, and just those for which assessed by observers. These results are generally much worse than those for the TREX 09-1 data which generally had precision and recall around 80%. Further investigation revealed a number of issues with the data, both from the CIFTS side and the SME side.

Table 3: Message Threading Precision and Recall for TREX 09-2

	All Missions		Observer-assessed Missions	
	Prec.	**Recall**	**Prec.**	**Recall**
Day 2	0.46	0.51	0.62	0.53
Day 3	0.45	0.57	0.77	0.52
Day 4	0.18	0.31	0.77	0.35

The CIFTS threading relies on information in the JADOCS table where operators enter information about a target, such as the description, target number, and aircraft call signs. The JADOCS data for TREX 09-2 contained information for a number of missions (3-20 per day) which were not active on a given day. This adds confusion to the algorithm and it will assign messages to a mission thread that does not actually exist. Also, the algorithm confused targets that had nearly identical descriptions. These problems can be addressed in the future, e.g., by using the entry time in JADOCS, and by requiring at least one explicit mention of a target ID in chat in order to have it be a potential thread. There is other information in the JADOCS tables which could also be exploited for these purposes.

In general, inter-rater reliability and coding guidelines are essential to create a gold standard to which a machine learning algorithm will be compared. The TREX 09-1 data were threaded by a single SME, but the TREX 09-2 data were threaded by three SMEs. It is likely that guidelines were not well established beforehand, which is understandable the time and difficulty of doing so. In any case, for the TREX 09-2 data, non-mission IDs were often used to assign a thread (e.g., not JF0001, but "Safehouse" was used as the assignment). Thus, these assignments could not be matched by the algorithm (15 occurrences over all).

Nevertheless, the precision numbers are generally high (77% on Day 3 and Day 4) for the observer-assessed targets that we are interested in. This means that 77% of the messages that CIFTS threading determined were about a target, were actually about that target (according to the SMEs). If these data are a reasonable sample of all the chat about the target, then the analyses of these messages can still be used to predict the observer score about that target.

Observer-based Performance Prediction

Number of Messages

Initial observations indicated that the total number of messages for a target was also somewhat predictive of the score, so this was always the first predictor used. Table 4 shows the results of correlating the number of messages about a target to the target's observer score.

Table 4: R-squared values (*p*-value) for predicting observer scores with the number of messages, using all chat rooms

		Threading Source	
		SME	**CIFTS**
Observer Scores	**Aggregate**	0.1274 (0.138)	0.1825 (0.092)
	Comms-oriented	0.2732 (0.046)	0.3582 (0.023)

As an example, for the bottom right cell, predicting the communications-oriented observer scores given CIFTS threading, the model is: score = 3.18 + 0.011*count, that is, for every additional message, the average communications-oriented observer score is predicted to rise by 0.011 from an initial value of 3.18.

Communications Measures

Starting from this high baseline using just the number of messages, we then sought to improve the model by adding CIFTS categorizations of those messages. We did this by using forward stepwise regression. Given the basic model of the count of messages, we added information about the ratio of messages about a target having a certain Dialogue Act assignment. All the measures were tested, and the one that improved the model most (resulted in the lowest *p*-value) was added to the model. The next step then looked at all the remaining measures to see which one further improved the model. This was repeated until the next step did not

improve the model. This process was repeated 12 times, each time leaving out one of the targets.

The top four predictors that appeared in these twelve models most, either using CIFTS or SME threads, were then used to build a single model. Given only these categories for the two types of observer scores, we re-ran the stepwise regression, the results of which are shown in Table 5.

Table 5: Regression model results.

Predictor	Aggregate Observer Score		Comms-Oriented Observer Score	
	SME	**CIFTS**	**SME**	**CIFTS**
Intercept	3.57	3.50	3.66	3.72
# Messages	0.002	0.009	0.002	0.01
Correction	-10.68	-8.17	6.99	8.31
Anticipation	-21.28	-21.12		
Neg. Reply	4.62			
Confirmation			-7.10	-9.56
Ambiguity				-4.80
Politeness			4.89	
r^2 *p*	0.64 0.02	0.68 0.006	0.84 0.001	0.93 <0.0001

DISCUSSION

For both sets of models, the weightings of those factors which both models used were similar for SME and CIFTS threaded messages. For example, the parameter values for Correction were similar (-10.68 vs. -8.17). Thus, despite the low recall for CIFTS and the noise added to both the CIFTS and the SME labels, similar models were found.

The SME threads had more messages for a target, thus the weight of the # Messages was lower for the SME threads. While there were some differences in the final model found using the two threading sources, it seems reasonable to say that the CIFTS threading is capturing enough of the messages about a target to make a similar conclusion about the score.

The models found for the communications-oriented observer scores were generally better than those for the aggregate scores. This is as expected since many of the SPOTLITE DTC questions are concerned with aspects of the mission and team for which a great deal of domain knowledge would be required in order to assess the behavior from communications alone. For example, consider the question: "Does the Target Duty Officer (TDO) make a collection request that includes all the pertinent information (EEI, timeline, resolution of image or product, target surveillance)?" While this is a type of communications (a "collection request"), understanding the specific "pertinent information" given the situation would require a great deal of domain knowledge and automated reasoning to assess the TDO's performance. In addition, the communications so far have not been broken out by operator and related to operator-specific SPOTLITE DTC questions.

It may, however, be second, third or fourth-order effects of this performance that can be captured in the communications as they are analyzed here, e.g., the Intelligence Duty Officer (IDO) sends a message back to the TDO that says something like "resend with higher resolution" (a Confirmation Dialogue Act). Thus, what these communications analyses are measuring is not performance per se, but the team's reactions to poor performance. The factors that best related to the observer scores are those that measure indications that something has gone wrong.

Confirmations are mostly requests for someone to repeat information, confirm that what was sent is true, or confirm that what was sent was received correctly. These requests would only be made if the original information was not good enough. *Corrections* are indications that the operator has realized the original information sent was not good enough. *Anticipations* may indicate that one operator realizes another operator does not have the right information. *Ambiguities* are indications that one is unsure of the information that one is giving or receiving. Occurrences of all of these Dialogue Acts lower the eventual observer score given by the observer. Interestingly, *Corrections* actually has a positive correlation with the communications-oriented scores. This might be because when judging coordination itself, it is a good thing to transmit corrections before they become mistakes, but overall, the more corrections one has to make, the more likely a problem will arise or has arisen.

We should note that the observer using SPOTLITE DTC is not recording these Dialogue Acts themselves, but is mostly likely aware of the informational challenge that the operators are reacting to when these types of messages are sent. The observer is not lowering the team's score because they made a confirmation request. Rather, it suggests that the team is itself aware of a problem (perhaps one already made and scored) and is trying to recover.

CONCLUSION

These analyses suggest that at least some observer-based performance scores can be reliably predicted from communications data. The fact that the automatic CIFTS-threading gave results similar to the SME-threading of the messages suggests that a fully automated system is feasible for guiding or prioritizing after action review. This will reduce the burden on instructors and lessen the time between training and review. One can imagine a real-time system that

1) monitors the chat stream,
2) labels which messages are about which target,
3) gives an initial baseline score for the target that generally increases as more discussion takes places, but
4) the score is decremented or incremented as the proportion of messages are labeled with these various Dialogue Acts.

This real-time score could be provided to trainers for real-time support or AAR learning points, or to the DTC/DEC Chief for better management, or to the team for improved focus, or even to the observers themselves to be able to change focus onto those targets that appear to have the most issues in order to provide more focused information for the AAR.

However, clearly more data is required. These results are based on data from only 12 targets in one exercise. The true generality of these results is unknown. Other environments for instance, might not have the same rate of messaging so the message count alone may no longer be a good predictor. Or, the language used to indicate confirmation requests may be different and not covered by the current CIFTS data structures. In addition, these results are based only on the analysis of chat data. Much of the actual communication (in this team in particular) was face to face. If the distributions of message types are different between these two data sources, then the chat may not be a good sample of the types of messages actually being sent about a target. Overall though, these results are a promising first step.

ACKNOWLEDGEMENTS

This work was supported by AFOSR contract #FA9550-07-C-0134 and L-3 Communications contract #PO-DS11223. We greatly appreciate the time and effort of all those at AFRL/RHAS, especially Todd Denning, Oscar Garcia, and Cameron Barnes.

REFERENCES

Air Force Doctrine Center (2001). Air Force Tactics, Techniques, and Procedures 3-1.1 Operational Brevity Standards. AFDC, Maxwell AFB, AL.

Air Force Doctrine Center (2007). Air Force Tactics, Techniques, and Procedures 3-3.AOC Operational Employment-Air and Space Operations Center. AFDC, Maxwell AFB, AL.

Air Land Sea Application Center (2009). Multi-Service Tactics, Techniques, and Procedures for Internet Relay Chat for Command and Control. ALSA Center, Langley AFB, VA.

Air Land Sea Application Center (2003). Multi-Service Brevity Codes. ALSA Center, Langley AFB, VA.

Alliger, G., Colegrove, C.M., & Bennett, Jr., W. (2003). Mission Essential Competencies: New method for defining operational readiness and training requirements. Paper presented at the Thirteenth International Occupational Analyst Workshop, San Antonio, TX.

Basu, S., Bilenko, M., & Mooney, R.J. (2004) A Probabilistic Framework for Semi-Supervised Clustering. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, Seattle, Washington.

Berube, C.D., Hitzeman, J.M., Holland, R.J., Anapol, R.L., & Moore, S.R. (2007) Supporting Chat Exploitation in DoD Enterprises. The 12th International Command and Control Research and Technology Symposium. Newport, Rhode Island.

Bilenko, M., Basu, S., & Mooney, R.J. (2004) Integrating Constraints and Metric Learning in Semi-Supervised Clustering. In: *Proceedings of the 21st International Conference on Machine Learning*.

Boiney, L.G., Goodman, B., Gaimari, R., Zarrella, J., Berube, C. and Hitzeman, J. (2008). Taming Multiple Chat Room Collaboration: Real-Time Visual Cues to Social Network and Emerging Threads. *Proceedings of the 5th International ISCRAM Conference.*

Budlong, E., Walter, S.M., and Yilmazel, O. (2009). Connotative meaning of military chat communications. Proceedings of SPIE, 7347, 73470G.

Carley, K., Diesner, J., Reminga, J. & Tsvetovat, M. (2007). Toward an Interoperable Dynamic Network Analysis Toolkit. *Decision Support Systems*, *Special Issue on Cyberinfrastructure for Homeland Security: Advances in Information Sharing, Data Mining, and Collaboration Systems*, 43, 1324 - 1347.

Case, F. T., Koterba, N., Conrad, G., Ockerman, J., & Vanderberry, R. (2006). An instrumentation capability for dynamic targeting. Command and

Control Research and Technology Symposium, San Diego, CA, June 20-22, 2006.

Cowell, A.J., Gregory, M.L., Bruce, J., Haack, J., Love, D., Rose, S., and Andrew, A.H. (2006). Understanding the dynamics of collaborative multi-party discourse. *Information Visualization*, 5, 250-259.

Creswell, C., Schwartzmeyer, N. and Srihari, R.K. (2007). Information extraction for multi-participant, task-oriented, synchronous, computer-mediated communication: a corpus study of chat data. IJCAI Workshop on Analytics for Noisy and Unstructured Text Data, pp. 131-138.

Elsner, M. & Charniak, M. (2008) You talking to me? A Corpus and Algorithm for Conversation Disentanglement. *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics*, pages 834-842, Columbus, OH, USA.

Eovito, B. A. (2006). An Assessment of Joint Chat Requirements from Current Usage Patterns. Naval Postgraduate School, Monterey, CA. DTIC: ADA451327.

Foltz, P. W., Bolstad C. A., Cuevas H. M., Franzke M. and Costello A. M. (2008). Measuring Situation Awareness through Automated Communications Analysis. In M. Letsky, N. Warner, S. Fiore & CAP Smith, (eds.). *Macrocognition in Teams*, Elsevier.

Heacox, N., Moore, R., Morrison, J., Yturralde, R. (2004). Real-time Online Communications: 'Chat' User in Navy Operations. *Proceedings of Command and Control Research and Technology Symposium.* San Diego, CA.

Ivanovic, E. (2005). Dialogue Act Tagging for Instant Messaging Chat Sessions. Proceedings of the ACL Student Research Workshop, pages 79–84.

Jackson, C., Wiese, E., Garcia, O., Wolfe, A., and Stephens, N. (2008). Developing performance measures for Air and Space Operations Center training environments: Triumphs, challenges, and next steps. In *Proceedings of the 2008 Simulation Interoperability Workshop (online)*, Orlando, FL.

Kulis, B., Basu, S., Dhillon, I., and Mooney, R. (2009). Semi-supervised Graph Clustering: A Kernel Approach. *Machine Learning*, Vol. 74, No. 1, pp. 1–22.

La Voie, N., Foltz, P., Rosenstein, M., Oberbreckling, R., Chatham, R. and Psotka, J. (2008) Automated Support for AARs: Exploiting Communication to Assess Team Performance. Interservice/Industry Training, Simulation, and Education Conference.

MacMillan, J., Entin, E.B., Morley, R., & Bennett, Jr., W. (in press). Measuring team performance in complex dynamic environments: The SPOTLITE method. *Military Psychology*.

Meyer, B. (1998). Self-organizing graphs—a neural network perspective of graph layout. In *Graph Drawing. 6th International Symposium, GD'98. Proceedings*, pages 246–62, Berlin, Germany. Springer-Verlag.

Ramachandran, S., Jensen, R., Bascara, O., Carpenter, T., Denning, T., & Sucillon, S. (2009). After Action Review Tools For Team Training with Chat Communications. Interservice/Industry Training, Simulation, and Education Conference.

Salter, W.J., Duchon, A. and Weil, S. (2009) Communications Assessment at an Air Force Exercise: Specific Results, Larger Implications. Interservice/Industry Training, Simulation, and Education Conference.

Shen, D., Yang, Q., Sun, J.T., and Chen, Z. (2006) Thread detection in dynamic text message streams. Proceedings of the 29th ACM International Conference on Research and Development in Information Retrieval.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R.., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C., Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3):339–373.

Tossell, C., Wiese, E. E., Garrity, M. J., Denning, T., & Alliger, G. M. (2006). Developing command and control performance-based training criteria in a network centric environment. Proceedings of the 11th International Command and Control Research and Technology Symposium, Cambridge, UK.

Wang, L., Jia, Y., & Chen, Y. (2008) Conversation Extraction in Dynamic Text Message Stream. *Journal of Computers*, 3(10), p. 86-93.

Wang, L. & Oard, D. (2009) Context-based Message Expansion for Disentanglement of Interleaved Text Conversations. *The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, CO, p. 200–208. Association for Computational Linguistics.

Wiese, E.E., Freeman, J., Salter, W.J., Stelzer, E.M., and Jackson, C. (2008). Distributed After-Action Review in simulation-based training. In D. Vincenzi, P. Hancock, M. Mouloua, and J. Wise (Eds.), *Human Factors in Simulation and Training*. Boca Raton, FL: CRC Press.

Wolfe, A., Garcia, O., & Denning, P. C. (2008). AOC Training Research Exercise (T-REX) Hits New Heights. *Fight's On!*, *7*(2).