Automated Chat Thread Analysis: Untangling the Web

Dr. Sowmya Ramachandran, Randy Jensen,
Oscar Bascara, Tamitha Carpenter
Stottler Henke Associates, Inc.
San Mateo, CA
Sowmya@stottlerhenke.com,
Jensen@stottlerhenke.com,
Bascara@stottlerhenke.com,
Tamitha@stottlerhenke.com

Todd Denning
AFRL/HEA
Nellis AFB, NV
Todd.denning.ctr@nellis.af.mil

Lt Shaun Sucillon
AFRL
Mesa, AZ
Shaun.sucillon@mesa.afmc.af.mil

ABSTRACT

As networked digital communications proliferate in military operational command and control (C2), chat messaging is emerging as a preferred communications method for team coordination. Chat room logs provide a potentially rich source of data for analysis in after-action reviews, affording considerable insight into the decision-making processes among the training audience. The multitasking nature of these types of operations, and the large number of chat channels and participants lead to multiple, parallel threads of dialogs that are tightly intertwined. It is necessary to identify and separate these threads to facilitate analysis of chat communication in support of team performance assessment. This presents a significant challenge as chat is prone to informal language usage, abbreviations, and typos. Techniques for conventional language analysis do not transfer very well. Few inroads have been made in tackling the problem of dialog analysis and topic detection from chat messages. In this paper, we will discuss the application of natural language techniques to automate chat log analysis, using an AOC team training exercise as the source of data. We have found it necessary to enhance these techniques to take into consideration the specific characteristics of chat-based C2 communications. Additionally, our domain of interest provides other data sources besides chat that can be leveraged to improve classification accuracy. We will describe how such considerations have been folded into traditional data analysis techniques to address this problem and discuss their performance. In particular, we explore the problem of automatically detecting content-based coherence between messages. We present techniques to address this problem and analyze their performance in comparison with using distinguishing keywords provided by subject matter experts. We discuss the lessons learned from our results and how it impacts future work.

Distribution A: Approved for public release. Unlimited distribution. As part of 88ABW-2010-3376, 18 June 2010.

ABOUT THE AUTHORS

Dr. Sowmya Ramachandran is a research scientist at Stottler Henke Associates, a small business dedicated to providing innovative Artificial Intelligence solutions to real-world problems. Dr. Ramachandran's interests focus on intelligent training and education technology including intelligent tutoring and intelligent synthetic agents for simulations. She is also interested in issues of motivation and metacognition. Experience with military and private industry gives Dr. Ramachandran a unique perspective on the needs and requirements of the ultimate end-users and their constraints. She contributes expertise in AI, instructional systems, probabilistic reasoning, and knowledge management. She has developed ITSs for a range of topics including reading comprehension, high-school Algebra, helicopter piloting, and healthcare domains. She has participated in workshops organized by the Learning Federation, a division of the Federation of American Scientists, to lay out a roadmap for critical future research and funding in the area of ITSs and virtual patient simulations. She has developed a general-purpose authoring framework for rapid development of ITSs, which is currently being used to develop an intelligent tutor training Navy Tactical Action Officers. She has also developed tools and technologies for training emergency first responders.

Todd Denning is a training research investigator for AOC training research for the Air Force Research Laboratory's Warfighter Readiness Research Division in Mesa, AZ and an instructor/subject matter expert for dynamic and deliberate planning training with the 505th Operations Squadron, Nellis AFB, NV. He has extensive experience in

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE **Automated Chat Thread Analysis: Untangling the Web**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **AFRL/HEA,Nellis AFB,NV,89191**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2010, 29 Nov ? 2 Dec, Orlando, FL

14. ABSTRACT
As networked digital communications proliferate in military operational command and control (C2), chat messaging is emerging as a preferred communications method for team coordination. Chat room logs provide a potentially rich source of data for analysis in after-action reviews, affording considerable insight into the decision-making processes among the training audience. The multitasking nature of these types of operations, and the large number of chat channels and participants lead to multiple, parallel threads of dialogs that are tightly intertwined. It is necessary to identify and separate these threads to facilitate analysis of chat communication in support of team performance assessment. This presents a significant challenge as chat is prone to informal language usage, abbreviations, and typos. Techniques for conventional language analysis do not transfer very well. Few inroads have been made in tackling the problem of dialog analysis and topic detection from chat messages. In this paper, we will discuss the application of natural language techniques to automate chat log analysis, using an AOC team training exercise as the source of data. We have found it necessary to enhance these techniques to take into consideration the specific characteristics of chat-based C2 communications. Additionally, our domain of interest provides other data sources besides chat that can be leveraged to improve classification accuracy. We will describe how such considerations have been folded into traditional data analysis techniques to address this problem and discuss their performance. In particular, we explore the problem of automatically detecting content-based coherence between messages. We present techniques to address this problem and analyze their performance in comparison with using distinguishing keywords provided by subject matter experts. We discuss the lessons learned from our results and how it impacts future work.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

fighter and air operations planning and operations to include combat operations in Southwest Asia and the Pacific region.

Randy Jensen is a group manager at Stottler Henke Associates, Inc., working in training systems since 1993. He has developed numerous Intelligent Tutoring Systems for Stottler Henke, as well as authoring tools, simulation controls, after action review tools, and natural language analysis methods. He is currently leading projects to develop an embedded training Intelligent Tutor for the Army, an after action review toolset for the Air Force, and an authoring tool for virtual training demonstrations for the Army. He holds a B.S. with honors in symbolic systems from Stanford University.

Oscar Bascara is a software engineer at Stottler Henke Associates, Inc. His interests include training systems, authoring tools, and user interface design. He holds an M.Eng. in Electrical Engineering from Cornell University and an M.A. in Mathematics from the University of California at Berkeley.

Lt Shaun Sucillon is a behavioral scientist assigned to the Air Force Research Laboratory's 711th Human Performance Wing, Warfighter Readiness Research Division in Mesa, AZ. He is the government project manager for Stottler Henke Associates, Inc., Small Business Innovative Research effort to develop an automated performance assessment after-action review tool that analyzes chat communications among teams. Lt Sucillon earned his Bachelor of Science degree in Behavioral Sciences from the United States Air Force Academy in Colorado Springs, CO.

Automated Chat Thread Analysis: Untangling the Web

Dr. Sowmya Ramachandran, Randy Jensen,
Oscar Bascara, Tamitha Carpenter
Stottler Henke Associates, Inc.
San Mateo, CA
Sowmya@stottlerhenke.com,
Jensen@stottlerhenke.com,
Bascara@stottlerhenke.com,
Tamitha@stottlerhenke.com

Todd Denning
AFRL/HEA
Nellis AFB, NV
Todd.denning.ctr@nellis.af.mil

Lt Shaun Sucillon
AFRL
Mesa, AZ
Shaun.sucillon@mesa.afmc.af.mil

INTRODUCTION

Ramachandran et. al. 2009 described the need for tools to facilitate the analysis of electronic communications among teams. Communication options like chat and email offer benefits over traditional radio and are becoming a vital part of team interactions. This provides an unprecedented opportunity for team communication analysis. Text-based communications can be logged and analyzed to study the team's performance. If there were failures or undesirable events in an operation, the logs can be examined to determine the contribution of communication failures to the situation.

In a broader context, mining chat and other text message-based communications is going to be of increasing importance in the future. Instant messaging and chat are becoming important tools for team communications. The trend is rapidly towards increasing adoption of chat-based communication even within the military. Automated analysis of chat will grow in value for various purposes such as: 1. Analyzing and improving business communications, 2. Detecting topic trends, 3. Analyzing messages streams for intelligence and counter-intelligence analysis.

Our research to date is focused on the use of chat for military planning operations. The multitasking nature of these types of operations, and the large number of chat channels and participants lead to multiple, parallel threads of dialogs that are tightly intertwined. It is necessary to identify and separate these threads to facilitate analysis of chat communication in support of team performance assessment. This presents a significant challenge as chat is prone to informal language usage, abbreviations, and typos. Techniques for conventional language analysis do not transfer very well.

Despite all the advances in the field of Natural Language Processing (Manning and Schütze, 1999), understanding the semantics of language is still a big challenge for computers. The objective of our research is to explore the extent to which chat-based communications can be analyzed to extract useful information without a deep semantic understanding of the messages. We focus on the use of statistical and rule-based techniques that will analyze messages based on surface features such as word occurrences and correlations.

Ramachandran et. al. discussed a combination of domain-specific and domain-independent techniques to separate chat data into threads of related conversation around a topic. The particular domain of application is the Air and Space Operations Center (AOC) planning operations and as such we were interested in separating out conversation threads relating to different targets or missions. This initial approach consisted of 1. Using Subject Matter Expert (SME)-provided keywords to associate messages with specific mentions, and 2. Using a process of temporal pattern recognition to identify talk-response dyads in the conversations to identify coherent sets of related conversations.

Keyword-based association is a crucial step in the procedure that leads to highly reliable associations since it is provided by experts familiar with the distinguishing characteristic of each mission. However, this human-in-the-loop solution does not achieve our research objective of developing a communication analysis tool that requires minimal human input. To minimize the human effort required to conduct after-action reviews, it is desirable to automate chat analysis to the extent practical.

This paper will discuss techniques for automatically identifying keywords that distinguish the different missions.

BACKGROUND

This work is in support of the research at the Air Force Research Lab at Mesa, AZ aimed at improving team training outcomes and developing exercise visualization and debriefing tools that will help trainees and trainers. As a targeted training domain, the Air Force Research Laboratory's Training Research Exercise (T-REX) provides a controlled research environment to investigate team performance dynamics in an air and space operations center. The environment allows mission-ready warfighters to practice their assigned duties using real-world systems in a scenario designed to test the full spectrum of decisions and coordination required in operational planning. The suite of systems includes collaborative planning tools, including chat rooms. As the warfighters conduct mission duties, researchers collect information on a variety of performance areas, leveraging chat as the complementary real-time communication mode in association with the suite of collaborative tools and shared situation awareness inputs available in an AOC. The research objectives pursued in a T-REX exercise are to: 1) Develop immersive scenarios to stimulate full team participation; 2) Develop tools to capture and validate team performance measures while conducting joint force planning for kinetic and non-kinetic effects; and, 3) Develop a synchronized suite of after-action review displays and tools to effectively communicate performance back to the team immediately after a training session.

The research approach used in determining how to analyze and display information follows the operational planning methodology laid out in joint and USAF doctrine. The initiator for planning is normally a problem statement in the form of intelligence data or operational data reported to the team. The initiating report typically establishes a segregated planning approach to address the problem. The team then examines the problem in sequence with other planning tasks or a sub-team may be tasked to examine the issue in parallel with other team activities. In many cases, planning may be interrupted and take on an interleaved character. When a training session ends, trainees need to be able to see each problem in isolation, as well as in context with other workload. The isolation approach allows the team to review actual process versus doctrine, while the context of workload offers insight into time delays, distractions, errant information sources, and overall cognitive effort.

The most significant challenge to conducting an effective after action review of operational planning is to isolate processes efficiently for consumption by different members or subgroups within the training audience. Problems in operational warfare rarely involve an entire audience, since the team is composed of individuals with unique and non-overlapping areas of expertise. At the leadership level of the team, the decision makers must be able to track and review decisions in full view of the information available at the time to understand how well they acted on it. Planning specialists *involved in a process* will also want to segregate and review information pertaining only to the process in question. The specialists *not involved in a process* will want the review to move quickly enough to get to the next point in time where they are involved. After action review tools must help an instructor to sort and associate information with a unique process and be able to display information cogently to identify key areas that positively or negatively affected team and individual performance. This is true in the general sense, irrespective of the form that exercise data takes. Where chat logs are one of the primary sources of data indicating performance, tools for reviewing multiple chat logs in tandem become critical.

Intelligent Diagnostic Assistant (IDA) is intended as a mixed initiative solution that leverages the strengths of the machine and the human. The strength of the machine lies in data management, organization, filtering, presentation, and automated analysis for simple keyword-based and temporal-based patterns. The strength of the human lies in selecting analysis criteria and performing high-level, big-picture analysis. For example, with the TREX exercises, instructors have expressed a strong need for a tool that will classify the chat data according to missions, further associate chat segments with different phases of a process, and provide complementary visualization that will clarify the communication flow within each process. Rather than supplanting instructional tasks, the goal is to facilitate them, so that instructors will be able to use their expertise efficiently to identify the training points and supporting data they wish to emphasize. Thus, the goal is to develop a tool that serves as a cognitive aid to instructors developing an After-Action Review (AAR).

Our approach to automating chat analysis for the purposes of developing an instructor's tool divides into capabilities to support two primary activities:

1. **Association and filtering**: In order to increase the speed and efficiency of putting together an AAR, automated natural language analysis and pattern recognition techniques produce a preliminary association between chat messages and specific missions of the exercise. This association is the backbone of a filtering capability that instructors

use to narrow the scope of the chat data they will be reviewing as they explore specific lines of inquiry into trainee decisions.

2. **Visualization and browsing**: Even with a filtered set of chat data, it is still a time consuming task to review synchronous conversation streams in multiple chat rooms and develop an understanding of the overall flow to identify performance indicators. This is the motivation for a tailored browsing capability that an instructor can use to review process-specific communications and visualize chronological relationships cross-referenced with missons. Typically, communications regarding a particular target will flow across multiple chat rooms, so synchronous browsing is a key feature. Additionally, the results of associations and filtering can be reflected in the browsing environment as cues during the review process. For example, keywords related to a mission process that were detected in the filtering step will often be of interest to an instructor as highlighted terms while browsing.

The instructor uses these tools to focus on process-specific communications and draw their own conclusions about how the team's communication helped or hindered achieving the mission objectives.

The visualization aspect of this tool was discussed in detail in a previous paper on IDA (Ramachandran et. al., 2009). This paper focuses on IDA's approach for associating chat messages with specific exercise missions.

FIRST PASS: RULE-BASED ASSOCIATIONS

We will describe our initial algorithm and discuss the refinements to it to address its various limitations. The initial rule-based analysis, first described in Ramachandran et. al. applies the following four rules in sequence:

Rule 1: In this domain, each mission/target is assigned a unique identification number (ID). Trainees sometimes, but not always, will refer to this ID while talking about a mission. When they do, this makes it easy to associate the chat messages with a mission/target. IDA makes one pass through the data set to identify those messages that have explicit references. These messages form the core set upon which subsequent rules build.

Rule 2: The next pass uses mission-specific keywords to classify chat messages. The keywords are provided by SMEs in a configuration file prior to analysis. The fact that each mission has a set of unique identifiers (e.g., mission numbers, code names for places or people, target types) is leveraged to tag chat messages. Typically this pass results in a smaller but still significant number of untagged messages.

Rule 3: There are some types of temporal patterns that can be detected with reliable accuracy without the need to understand the content of utterances. An example is recognizing the pattern of a turn-by-turn interaction between two people in the same room (e.g. A says something to B and 3 minutes later B says something to A) and inferring that they belong to the same topic thread. Making an assumption of dialog coherence, one can say with a high degree of confidence that such conversation dyads refer to the same topic thread. The message classifications identified using the keyword-based approach is used as the basis to further identify and tag such pairs of messages.

Rule 4: Finally, locality influence is used to attempt to classify remaining unclassified chat lines. For each such line, IDA examines its neighboring messages and finds the most common mission association, weighted by distance of the neighbor from the line. If the combined influence of all the messages within that window that are associated with this mission is over a threshold, the chat line is also assigned to that mission. Although this rule has been implemented, it has not been analyzed sufficiently to gauge its usefulness. This will be done as a part of future research.

Outcome

Results indicate that the classifications resulting from this approach are moderately accurate. Table 1 shows the classification accuracy of these rules on data from real chat logs from one of the T-REX exercise sessions. We provide three related measures of accuracy. Precision is a measure of the number of data items classified correctly as a fraction of the total number of data that were assigned a classification. Recall is a measure of the number of data items classified correctly as a fraction of the total number data items that actually belong to those classes (as specified by ground truth information). The F-score is harmonic mean of these two measures. All of these measures range between 0 and 1, with 1 signifying the best accuracy.

All of the results reported here use a T-REX data set with 631 chat lines and 20 missions. All accuracies reported in this paper are averages of the precision,

recall, and F-score measures for each mission. The data set was hand labeled with message-mission associations by an SME. This formed the gold standard against which IDA's output was evaluated.

Table 1. Accuracy of Classification Resulting From Initial Rules

Data Set	IDA Classification Accuracy		
	Precision	**Recall**	**F-Score**
T-REX 9.1.3	0.85	0.72	0.76

These rules allow for multiple classifications of the same chat message (i.e. each message can be assigned to multiple classes). This leads to a significant number of false positives (reflected in a lower precision score). For IDA, however, false positives are preferable to false negatives. Filtering data conservatively is better than filtering out messages that are related to the mission of interest. The rules result in a high recall accuracy which means there are few false negatives. However, improving precision will be an ongoing objective.

A more critical limitation of this approach is its reliance on hand coded keywords. The following sections describe our ongoing efforts to eliminate this need.

CLUSTERING

Our first approach to identifying related messages based on statistical analysis was a technique called clustering. This is an established and popular Artificial Intelligence technique for automatically grouping data without human input. Our hypothesis was that this technique would lead to high-value topic based clusters that can be used, in addition to the rules mentioned above, to separate the communications relating to different missions. The clustering approach has the advantage of not requiring that the training data be labeled by hand.

We introduced a clustering step for associating messages with planning processes based on the term frequency–inverse document frequency (TF-IDF) similarity measure presented in (Adams and Martell, 2008). This measure determines similarity by the number of overlapping words between two messages, weighted by the uniqueness of the words. That is, under this scheme, common words such as "at", "the", etc. will have lower weight and therefore smaller influence on the similarity measure. Unique words will contribute more heavily.

The first pass of the modified algorithm is the same as before. It looks for mission IDs to create an initial set of message-mission associations. During the next pass, it uses a nearest neighbor approach to build message clusters based on the TF-IDF similarity measure. Within each cluster, it looks for messages that have been classified to missions based on Step 1. Clusters with messages that have been associated with more than one mission are eliminated as not being relevant to the analysis (i.e. the similarity between the messages in the cluster is not pertinent to the topics of interest). Clusters with at least one message associated with a mission are identified and all the untagged messages in each cluster are assigned to the mission.

Consider an example with 8 chat messages, U1 through U8, and three missions, M1 through M3. The following table shows the process associations after the first pass.

Table 2. Initial Associations between Message and Missions

Message	Process
U1	M1
U2	None
U3	None
U4	None
U5	M2
U6	M3
U7	None
U8	None

Now, suppose the nearest neighbor approach identifies message clusters as shown in the following table.

Table 3. Results of Clustering

Clusters	Messages	Missions
C1	U1, U2, U6, U7	U1<->M1 U6<->M3
C2	U4, U5	M2
C3	U3, U8	None

Cluster C1 is eliminated from any further processing since it has associations with multiple missions. Cluster C3 is also eliminated because it has no mission associations. That leaves C2. In this cluster, U5 is associated with M2 and therefore U4 is also tagged with M2. The resulting process associations after the second pass will be as shown in Table 4.

After the second pass, the remaining passes to make associations based on request-response dyads and

message proximity are made. The rules for these subsequent passes have not changed.

Table 4. Message-Mission Associations after Clustering

Message	Process
U1	M1
U2	None
U3	None
U4	M2
U5	M2
U6	M3
U7	None
U8	None

Additionally, we had to make one modification to the traditional clustering algorithm described above. We observed that while the clusters identified by this method have an identifiably unifying topic, this topic sometimes is tangential to the missions being discussed. For example, the algorithm may cluster together messages which are similar in that they all talk about times on target (TOTs). This is not interesting from the perspective of identifying different mission-related discussions, as there is nothing mission-specific about TOTs. So we included a variation where messages are only compared to others within a specified time window. This improved the relevance of the generated clusters.

Outcome

This approach still needs some messages to be assigned to a mission according to Rule 1 above (i.e. based on references to the actual mission/target IDs). The clustering component uses this seed set to then classify other messages that are in the same clusters as these seeds. While testing the clustering algorithm, we observed that there are chat conversations about missions where the trainees do not ever mention the target IDs. The clustering approach fails under these conditions.

While we have not performed a quantitative evaluation of the utility of the clustering component, indications are that it provides some utility and therefore, we will retain it in the mix. However, this will fail to identify keywords under the conditions mentioned above.

LEVERAGING OTHER RELATED INFORMATION SOURCES

The domain provides another related data source that could be usefully exploited. All trainees use a database system called Joint Automated Deep Operations Coordination System (JADOCS) to record critical information about the various missions, such as target intelligence, operational orders etc. A very common practice is to copy over messages from chat streams to the JADOCS database (DB) as annotations. This results in a set of chat messages stored in the JADOCS DB with definite mission associations that can be mined to learn mission-specific identifiers. However, this data can be sparse and it is an empirical question if it is sufficient to for IDA to learn accurate classifications. .

Our next enhancement was to use Naïve Bayes (Langley 1995) classifiers, described below, that were trained on the message-mission associations found in the JADOCS. This is done in place of Rule 2 of the original approach. The remaining rules are applied as before.

Naïve Bayes Classifiers with Normalization

Our approach uses a separate Naïve Bayes classifier for each mission classification. A normalization process is then applied to the results of these classifiers to obtain the mission classification probabilities for a chat message.

We have a set of mission classifications or classes:

$$C = \{c_1, c_2, \ldots, c_{|C|}\} \qquad (1)$$

These classes are used to label a set of training chat messages or documents:

$$D = \{d_1, d_2, \ldots, d_{|D|}\} \qquad (2)$$

That is, the training document d_j is either associated with class c_i or not associated with class c_i.

The set of all the words contained in the training documents is the vocabulary:

$$V = \{w_1, w_2, \ldots, w_{|V|}\} \qquad (3)$$

Consider a new test document d_α to be classified. Bayes' Rule can be applied to compute the posterior probability $P(c_i \mid d_\alpha)$:

$$P(c_i \mid d_\alpha) = \frac{P(c_i)P(d_\alpha \mid c_i)}{P(d_\alpha)} \quad (4)$$

To compute the prior probability $P(c_i)$, we simply divide the number of training documents associated with class c_i by the total number of training documents:

$$P(c_i) = \frac{\sum_{j=1}^{|D|} P(c_i \mid d_j)}{|D|} \quad (5)$$

where $P(c_i \mid d_j) = 1$ or 0 depending if training document d_i is associated with class c_i.

Next, to compute the likelihood $P(d_\alpha \mid c_i)$, we use the Naïve Bayes assumption that each word in a document is independent of the occurrence of other words to get the following:

$$P(d_\alpha \mid c_i) = \prod_{t=1}^{|V|} [B_{\alpha t} P(w_t \mid c_i) + (1 - B_{\alpha t})(1 - P(w_t \mid c_i))] \quad (6)$$

where $B_{\alpha t} = 1$ or 0 depending if document d_α contains word w_t.

To compute $P(w_t \mid c_i)$, we can divide the number of training documents containing w_t and associated with class c_i by the total number of training document associated with class c_i. But to avoid probabilities of 0 or 1, the division is primed:

$$P(w_t \mid c_i) \approx \frac{1 + \sum_{j=1}^{|D|} B_{jt} P(c_i \mid d_j)}{2 + \sum_{j=1}^{|D|} P(c_i \mid d_j)} \quad (7)$$

Finally, the prior probability $P(d_\alpha)$ in Bayes' Rule is a constant that will cancel during normalization.

Now two normalization steps are performed. First, we normalize between the $P(c_i \mid d_\alpha)$ and $P(\bar{c}_i \mid d_\alpha)$, where $\bar{c}_i$ is the event of not class c_i, to get:

$$P_{norm_1}(c_i \mid d_\alpha) = \frac{P(c_i \mid d_\alpha)}{P(c_i \mid d_\alpha) + P(\bar{c}_i \mid d_\alpha)} = \frac{P(c_i)P(d_\alpha \mid c_i)}{P(c_i)P(d_\alpha \mid c_i) + P(\bar{c}_i)P(d_\alpha \mid \bar{c}_i)} \quad (8)$$

Second, we normalize between all these normalized probabilities to get:

$$P_{norm_2}(c_i \mid d_\alpha) = \frac{P_{norm_1}(c_i \mid d_\alpha)(c_i)}{\sum_{k=1}^{|C|} P_{norm_1}(c_k \mid d_\alpha)} \quad (9)$$

The probability $P_{norm_2}(c_i \mid d_\alpha)$ is used to determine if the test document d_α is associated with class c_i.

Table 5 shows the accuracy of this approach on one data set. For comparison, we have shown the results of applying the rule-based approach described in the Section FIRST PASS: RULE-BASED ASSOCIATIONS, with and without keywords. (Note: all these versions included the clustering step described above). The results reported are for the same data set. Note that the recall accuracy suffers dramatically when no keywords are provided (comparing rows 1 and 2 of the table). The precision score is improved however in the absence of keywords. This indicates that the keywords, while covering more of the true positives, also result in misclassifying more negative examples. The Bayesian classifier (row 3) that replaces the keywords improves the recall accuracy significantly while also improving precision. Thus it is significantly better than using no keywords (row 1) at all, both in terms of identifying more of the true positives and misclassifying fewer of the negatives. Compared to hand-coded keywords (row 2) approach, the Bayesian approach results in better precision but worse recall. Thus it does not misclassify as many negatives but fails to classify as many true positives. Thus the automated approach, while not as effective as using hand-coded keywords, is significantly better than not using any keywords at all. Furthermore, it offers the convenience of not requiring the trainer or a subject matter expert to supply the keywords.

Analyzing IDA's classification on a different data set, we observed that the trainees experienced considerable confusion between two targets and mixed up their target data descriptions several times. As a result, IDA was not able to distinguish between these two missions very accurately.

While this is an issue for IDA, it may also be possible to raise a flag when such confusions are detected as they may indicate useful training points. This possibility will be explored in the future.

Table 5. Accuracy of JADOCS-Based Approach

Classification Method			IDA Accuracy		
Rule-Based	**Hand Coded Keywords**	**Automated Keyword Detection**	**Precision**	**Recall**	**F-Score**
Yes	Yes	No	0.85	0.72	0.76
Yes	No	Yes	0.74	0.58	0.59
Yes	No	No	0.85	0.32	0.43

LESSONS LEARNED

1. Data fusion, i.e. collecting and analyzing data from multiple sources is a powerful way of harnessing complementary information for analysis. IDA is able to tap into data from the JADOCS DB, which, while not extensive, contains crucially salient information that is needed for analyzing the larger database of chat messages.

2. Topics, while separate, are sometimes highly correlated. For example, exercises may have one mission dedicated to a High-Value-Individual (HVI), another to the HVI's location, and another to a planned operation targeting the HVI. Sometimes these distinctions are genuinely necessary; at other times they are just artifacts of a misunderstanding on the part of trainees about communication and bookkeeping protocols. Whatever the reason, this makes topic identification challenging in the absence of a deeper semantic interpretation. Statistical techniques for natural language analysis, such as the ones discussed in this paper, are limited in this respect.

3. For the above-mentioned reason, we have observed that even humans familiar with the details of the exercise find it difficult to associate chat messages to the relevant topic threads (i.e. missions). This makes evaluating the performance of IDA a challenge as there are no accurate ground truth classifications to serve as standards for comparison. We will have to use an alternate approach to perform a robust evaluation of IDA's analysis.

4. Finally, we note that there is room for improvement in IDA's classification approach. The F-Scores for all versions are lower than desired. This is largely due to high numbers of false positives. Improving this will be an important focus in the near future. The discussion points above point out the challenges to achieving this goal. We will work with the AOC trainers to determine how to best approach topic confusions such as the one mentioned above. Trying to differentiate between highly correlated missions is a significant challenge for automated techniques. However, if the existence of such correlations turns out to an artifact of insufficient understanding of the process on the part of trainees, IDA can be engineered to identify such confusions and turn them into training opportunities. This will be an important focus of this research going forward. Finally, we have focused our attention to date largely on automating keyword-based classification. Going forward, we will also analyze the performance of rules 3 and 4, study their contribution to classification accuracy, and tune them.

RELATED WORK

Previous related research involving multi-party dialog analysis has included much work to characterize spoken interactions in multi-party meetings, social structures, and collaborative learning environments. The most relevant work is being done by the "Cognitive Agent that Learns and Organizes" (CALO) project, a joint effort between SRI and Stanford University's Center for the Study of Language and Information. (Zimmermann 2006), and (Tur 2008) describe efforts within the CALO project to support multi-party meetings with transcription, action item extraction, and, in some cases, software control such as document retrieval and display updating. (Niekrasz 2004) describe an architecture in which the spoken conversation between meeting participants is processed using automatic speech recognition techniques, and grounded against the artifact being produced (e.g., a schedule, a budget) and the drawings made on an electronic whiteboard. All of these inputs are used to create an electronic version of the artifact. Although experiments with dialog models from spoken interactions are transferable to research with chat communications, there are also unique challenges with the chat medium.

Much chat-related research has focused on the inherent communication artifacts of the medium, such as the emergence of conventional abbreviations, emoticons, and other common stylistic practices. To a lesser degree, some research has yielded methods and tools to analyze or visualize chat communication patterns. Most require a coding step carried out by a human reader to tag messages or explicitly identify dependencies before analysis takes place in any automated form.

(Cakir 2005) studied methods for assessing team problem solving with a chat environment and shared workspace. Essentially this employed a structure for organizing messages and identifying instances of interactions between two, three, or more participants as well as indices for factors like initiative. This is useful for learning research observations about how level and type of participation contribute to team dynamics and collaboration effectiveness.

(Shi 2006) introduce a conceptual framework for "thread theory," which suggests an approach for sorting out different chat threads based on topic or theme, and for characterizing defining features such as life, intensity, magnitude, and level of participation. (Herring 2006) describes VisualDTA, a tool designed to generate a visualization of a chat conversation that has been manually coded. In this visualization, messages are plotted in a descending tree, with temporal spacing represented on one axis, and semantic divergence represented on the other. The tool also accommodates the possibility of completely new topic threads appearing within the chat stream, resulting in new trees. This is useful for social interaction research, where plots of communication patterns reveal behavioral features.

(Adam and Martell, 2008) used the TF-IDF measure discussed earlier to identify topic threads in chat conversations. Their approach used only clustering whereas we have suite of other techniques to help the process. Whereas they were concerned with detecting topics in general public chat sessions that are not focused on any particular domain, our objectives are narrower. We are concerned primarily with chat conversations that are occur within military team training exercises. This gives us the benefit of leveraging chat protocols, domain-specific vocabulary, and other data sources to help refine our technique.

ACKNOWLEDGEMENTS

The research effort described in this paper is sponsored by the Air Force Research Laboratory.

REFERENCES

Adams, P.H. and Martell, C. H. 2008. Topic Detection and Extraction in chat. *In the Proceedings of the IEEE Conference on Semantic Computing*. Santa Clara, CA.

Cakir, M., Xhafa, F., Zhou, N., and Stahl, G. (2005), Thread-based analysis of patterns of collaborative interaction in chat, paper presented at the *Conference of Artificial Intelligence for Education (AIEd'05),* Amsterdam, Netherlands.

Herring, S. C., & Kurtz, A. J. (2006). "Visualizing dynamic topic analysis", *Proceedings of CHI'06*. New York: ACM Press.

Langley, Pat (1995). *Elements of Machine Learning*.Morgan Kaufman Series in Machine Learning. 1995

Manning, Christopher and Schüteze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press. 1999.

Niekrasz, J., Gruenstein, A., & Cavedon, L. (2004). Multi-human dialog understanding for assisting artifact-producing meetings. In Proceedings of the 20th International Conference on Computational Linguistics (COLING).

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

Shi, S., Mishra, P., Bonk, C. J., Tan, S., & Zhao, Y. (2006). "Thread theory: A framework applied to content analysis of synchronous computer mediated communication data", *International Journal of Instructional Technology and Distance Learning*, *3*(3), 19-38.

Tur, G., Stolcke, A., Voss, L., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Hakkani-Tür, D., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J.,

Peters, S., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., & Yang, F. (2008). "The CALO meeting speech recognition and understanding system", In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*.

Zimmermann, M.; Liu, Y.; Shriberg, E. & Stolcke, A. (2006). "Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings". In *Proceedings of IEEE ICASSP*, Toulouse, France (2006).