EVALUATION OF LATENT SEMANTIC ANALYSIS-BASED MEASURES OF TEAM COMMUNICATIONS CONTENT

Jamie C. Gorman Peter W. Foltz Preston A. Kiekel Melanie J. Martin
New Mexico State University
Las Cruces, NM

Nancy J. Cooke
Arizona State University East
Mesa, AZ

Team process is thought to mediate team member inputs and team performance. Among the team behaviors identified as process variables, team communications have been widely studied. We view team communications as a team behavior and also as team information processing, or team cognition. Within the context of a Predator Uninhabited Air Vehicle (UAV) synthetic task, we have developed several methods of communications content assessment based on Latent Semantic Analysis (LSA). These methods include: Communications Density (CD) which is the average task relevance of a team's communications, Lag Coherence (LC) which measures task-relevant topic shifting over UAV missions, and Automatic Tagging (AT) which categorizes team communications. Each method is described in detail. CD and LC are related to UAV team performance. AT-human is comparable to human-human agreement on content coding. The results are promising for the assessment of teams based on LSA applied to communication content.

INTRODUCTION

Team process is thought to mediate team member inputs and team performance. Among the team behaviors identified as process variables, team communications has been widely studied. Communications can be modeled and analyzed in different ways, including content coding and code sequence analysis (Bowers, Jentsch, Salas, and Braun, 1998), hidden Markov models (Stolcke, Ries, Coccaro, Shriberg, Bates, Jurafsky, Taylor, Martin, Van Ess-Dykema, and Meteer, 2000), and classification along linguistic dimensions and word counts (Sexton & Helmreich, 1999).

Over the past five years, scientists in the Cognitive Engineering Research on Team Tasks Laboratory (the CERTT Lab; Cooke & Shope, 2002) have studied teams in the context of a synthetic Predator Uninhabited Air Vehicle (UAV) reconnaissance mission. The CERTT Lab is fully customized and records a variety of mission data including performance, communication flow, and digital audio recordings of team communications. The CERTT Lab and the synthetic Predator task were designed for the express purpose of studying and developing measures of team cognition. This work has included the development and evaluation of methods for team communication analysis.

A major concern in the CERTT Lab is how teams process information and how this relates to their performance. Team communications modeling in the CERTT Lab falls within the traditional purview of team process modeling. However, we also view team communications as team information processing or team cognition (Kiekel, Cooke, Foltz, Gorman, and Martin, 2002). Developing methods for analyzing team communications have therefore been paramount in many of our efforts. Two general classes of communication data have been addressed by CERTT researchers: communications flow and communications content. Flow can more generally be described as who is talking to whom, when, and for how long. Content can be thought of as the meaning of what was said relative to the task being performed. Several methodologies under development in the CERTT Lab concerning the latter have been discussed previously including several methods for analyzing communications flow and LSA-based measures of performance and efficiency (Kiekel et al., 2002). This paper extends that discussion with more detailed results and new methods under development.

LSA and Team Communications Content

Latent Semantic Analysis (LSA) is a fully automatic method for representing and analyzing semantic information within a domain (Landauer, Foltz, and Laham, 1998). Running LSA on a corpus of UAV relevant material provides a semantic space (308 factors) based on UAV knowledge for use with communications

in the Predator task context. The "LSA approach" to evaluating team communications content has three big advantages over more traditional approaches. First, content assessment can be automated, resulting in a great time savings, second, the analysis is done at a semantic level, rather than at a keyword matching level, and third, assessment is empirical and internally consistent. Two metrics derived from a geometric interpretation of the semantic space have been employed for the purposes of analyzing team communications content- vector length and cosine. *Vector length* of a statement (or utterance; e.g., "AVO, I need you to be at 3000 feet") is the length of the summed word (= sentence) vector when plotted in the semantic space. In LSA, the length of a statement vector indicates the amount of information the statement carries concerning the modeled domain of discourse. The *cosine* between statement vectors provides a measure of how two statements are semantically related in the semantic space.

METHOD

Data Collection and Performance Measure

Transcribed mission communications from two Predator experiments in the CERTT Lab serve as units of analysis. The first experiment consisted of 11 3-person teams flying 10 missions each. Each team member had a specialized role, distinguished as AVO (air vehicle operator), PLO (payload operator), or DEMPC (navigator). These teams had to coordinate their efforts orally over microphones and headsets in order to maneuver their UAV into a position from which acceptable photographs of specified targets (critical waypoints) could be taken. Mission performance scores were based on a weighted sum of several mission variables, including: number of photos taken, total mission time, fuel/film used, and time in alarm state. The weights for this composite score were based on the instructed relative importance of a particular component to the team task. The basis for the first study was to examine the nature of team skill acquisition over time. The second experiment consisted of 20 3-person teams flying 7 Predator missions each. This study was conducted in order to identify differences between co-located for distributed teams and low (missions 1-4) and high (missions 5-7) workload missions. Sixty-seven mission transcripts from the first experiment were analyzed, 41 from the second were analyzed. Remaining missions from the second experiment are now in the process of being transcribed and analyzed in the semantic space.

LSA-Based Content Measures

Several methods developed for analyzing LSA team communications content include: 1.) Communication density, 2.) Lag coherence, and 3.) Automatic tagging. Individual mission transcripts in textual format, comprised of all statements among three team members made during the course of a given mission, are the units of analysis for each of these methods.

Communication density is motivated by the concept that for team communication to be effective, information should be conveyed in a concise manner. It is based on the formula for average velocity,

$$\text{Rate} = \frac{\text{Distance}}{\text{Time}}$$

But in this case, rate is rate of meaningful discourse. Distance becomes meaningfulness and time becomes number of words spoken. Therefore the density formula is:

$$\text{Density} = \frac{\text{Meaningfulness}}{\text{Words Spoken}}$$

Operationally, density is the average task-relevance of a team's communications, which is measured by the ratio of LSA vector length, summed over all statements, to the number of words spoken in a given mission.

Lag coherence is used to measure task-relevant topic shifting over speech turns, or utterances, by team members during Predator missions. In team discourse, presumably there is an appropriate level of relatedness of one utterance to the next. If this relatedness is too low, then each team member is communicating on entirely different topics. If it is too high, then each team member is simply repeating information previously conveyed to them. Lag coherence is computed as average LSA cosine between a statement and other statements over varying utterance lags (e.g., 2 utterances away, 3 utterances away, etc.). We average statement cosines over a 36 lag moving window. The decision to use a 36 lag window was not empirical, we simply feel that we can safely capture any interesting mission event by at most 36 speech turns (e.g., we have estimated that photographing a target is minimally a four-turn event). Once these averaged cosines are collected, log lag is used to predict log cosine in a linear regression equation. The estimated slope of this relationship is the final measure of topic shifting, or lag coherence, for a given mission.

Automatic tagging Along with measuring team communication, LSA can be applied to categorize team communication according to a set of content codes. Tagging of content is achieved by comparing statements from a mission transcript to a tag database, and retrieving the most similar tagged statements. The tag database consists of the 67 mission transcripts from Experiment 1 with each statement individually tagged by a human. In addition, for comparisons between human-human tagging agreement and LSA-human tagging agreement, we have a tag database consisting of 12 human-coded transcripts from Experiment 1 that were independently coded by at least two coders. Our human coders used the same tags developed by Bowers et al. (1998) to categorize statements (e.g., factual statement, acknowledgment, uncertainty statement, etc.). For the automatic tagging procedure, each line in the un-coded transcript is automatically tagged by retrieving the most similar statements in the human-tagged database. "Most similar" is defined as the top n (e.g., 10) statements in terms of cosine or above threshold statements (e.g., cosine > .6). Then for each unique tag retrieved, the associated cosine(s) are summed and normalized. The tag with the highest probability from this distribution is assigned and a confidence measure, the average cosine of the assigned tag within the retrieved set, is also provided. This automatic tagging then permits characterizations of the types (categories) of utterances being used by team members during missions.

RESULTS AND DISCUSSION

It should be noted analyses are ongoing. Specifically, team performance regression models with the various communication metrics as predictors are continually re-evaluated as new transcripts become available.

Communication Density

Following Experiment 1, a linear regression performance model indicated that more task-specific communications are not always associated with top mission performance (linear effect 8.54; t (47) = .48, p = .68). A negative quadratic effect was more appropriate for describing the density-performance relationship (quadratic effect –25.4; t (47) = -2.49, p = .016), which can be interpreted as "optimal" communication density. In Experiment 2, this optimum level was not detectably different under the low workload (t (10) = 1.33, p = .21) or co-located (t (7) = .48, p = .646) experimental conditions, which were similar to those experienced by teams in Experiment 1. Although a stronger than expected positive linear estimate was obtained under both conditions.

Communication density under Experiment 2's high workload condition did not provide evidence of any "optimal" rate of information transfer (the quadratic estimator was near zero at .93). However the linear estimator surfaced as being important, though not significantly given the model developed following Experiment 1 (b = 17.52; t (8) = .64, p = .73). These results suggest the need for further study under high workload conditions in which densely informative communications may indeed be optimal.

Results for communications density under distributed task conditions were less clear. The linear effect was negative (-50.49) as was, expectedly, the "optimal", or quadratic effect (-65.44). However both of these estimators had relatively high sampling error due to little variation in the observed communication densities among our distributed teams. Observation of communication densities over a wider array of distributed teams is required before drawing further conclusions.

Interestingly, in each of our experiments, as teams gained skill in performing the Predator task their performance increased monotonically from mission 1 until asymptote around mission 4. During this period of skill acquisition co-located teams started out communicating rather inefficiently, then became overly task-specific, until task-relevant communications were used most effectively -- optimal communications density (refer to missions 4-6 in Figure 1).

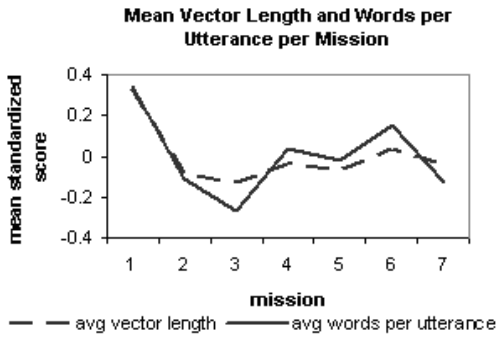

Figure 1. Composition of communication density by mission for Experiment 1

In both experiments the composite measure, communication density, predicted performance better than either average words or vector length alone. The results described here suggest that effective communications, in terms of UAV-STE performance, are a function of both what was said and the number of words used to say it. Presumably then, effective

communications are not overly terse with task-specific nomenclature.

Lag Coherence

Lag coherence performance models have not yet undergone validation in Experiment 2, but initial results (from Experiment 1) indicate that remote statements are more positively correlated in high performing missions than in low (t (55) = 2.2, p = .02). These results suggest that topic shifting within the 36 statement moving window is most frequent in low performing missions before teams have reached asymptotic levels of performance (i.e., missions 1-3; see Figure 2).

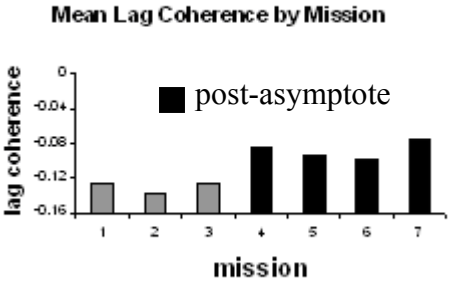

Figure 2. Pre- and post-performance asymptote levels of lag coherence for Experiment 1

The tendency for inexperienced teams to quickly cover a wide array of topics implies that they are working with a much broader palette of UAV semantics. In the process of reaching asymptote-level performance, team members may also be converging on a more constrained team lexicon, or common ground, upon which subsequent task-relevant communications can be based.

Automatic Tagging

We assessed inter-coder agreement between human-human and LSA-human for the tag database from Experiment 1. For assessing agreement, we used the C-value measure (Schvaneveldt, 1990). C-value, which ranges from 0 to 1, was chosen for its ability to handle arbitrarily long sequences of tags for a given turn and cases where taggers assigned sequences of tags of different lengths to a given turn. The average C-value among the 12 human-coded transcripts in the tag database was .71. This C-value served as the benchmark for LSA-human agreement. Each transcript in the 12-transcript tag database was then automatically tagged by LSA, using our entire UAV corpus, according to similarity criteria. Using the top n cosine (n = 10) criteria and excluding retrieved cosines < .3, average LSA-human C-value was .57, or about 20% below human-human agreement. Using the threshold cosine > .6 criteria, and excluding cosines < .4, LSA-human C-value was .61, about 14% below human-human. Finally, we included a syntactic feature along with cosines for retrieving the most similar statements. When the syntactic feature "?" is added to the cosine > .6 criteria and excluding cosines < .4, average LSA-human C-value is .64, about 9% below human-human agreement.

CONCLUSIONS

This research provides two key contributions. First, it provides a diverse set of methods for analyzing team communication. These methods can be thought of as components of a general toolbox for automated analyses. Second, the results have clear implications for team communication, which translate into team cognition. For example, the results on communication density and lag coherence show that we can determine optimal levels of task-relevant information that are communicated by successful teams. Because results like these are correlational, they might best be thought of as drivers for hypotheses regarding causal factors of effective team performance. Although these methods can be used to assess the quality of team performance, a deeper understanding of the causal factors underlying that performance is necessary for effective intervention through training and/or design.

In the near term, the methods of communication content analysis discussed in this paper will continue to be refined and validated for consistency. Collecting mission transcripts is currently cumbersome, requiring several typists and an inordinate amount of time. However an initial study regarding LSA performance using speech recognition software was promising (Foltz, unpublished), supporting the eventual possibility of complete automation.

Future applications of the analysis methods discussed in this paper might include automated team process monitoring agents. The invocation of an LSA-based artificial agent for team process monitoring assumes that knowledge will be transferred either verbally or textually, and that knowledge transfer among team members is necessary for completing the task. Essentially an agent would record communication, evaluate it against some semantic space, and compare its vector properties to an optimal content distribution. *Unusual* statements would be flagged and perhaps automatically tagged with a corresponding degree of tag-certainty. For example, for a 3-member team consisting of the members AVO, PLO, and DEMPC, the agent might log:

1. PLO: "AVO, I need you to be at 3000 feet"
log: |optimal|

2. AVO: "PLO, why aren't we circling?"
log: |alert-incoherent response .47|

3. DEMPC: "O.K. guys, what's happening?"
log: |alert-incoherent non-salient uncertainty .85|

A running log of this sort could provide feedback to team members and team coordinators, possibly in the form of real-time process feedback. Such a log would also allow researchers and team coordinators to quickly pinpoint shortcomings in team information processing. For the example log above, the researcher would be able to quickly make notes such as:

"Statement 2 looks like loss of team SA"
"Statement 3 looks like poor communications"

Finally, with additional research based on notes like those from the sample log above, specific patterns in communication may be associated with a particular diagnosis of a team dysfunction that can then be targeted for intervention.

ACKNOWLEDGEMENTS

This work was supported by ONR Grant No. N00014-00-1-0818 and greatly benefited from the contributions of Steven Shope, Susan Stevens, and Jessica Cox.

REFERENCES

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

Cooke, N. J., & Shope, S. M. (2002). The CERTT-UAV Task: A Synthetic Task Environment to Facilitate Team Research. *Proceedings of the Advanced Simulation Technologies Conference: Military, Government, and Aerospace Simulation Symposium*, pp. 25-30. San Diego, CA: The Society for Modeling and Simulation International.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J. C., and Martin, M. J. (2002). Some promising results of communication-based automatic measures of team cognition. *Proceedings of the Human Factors and Ergonomic Society 46th Annual Meeting*, 298-302.

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

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Sexton, J. B. and Helmreich, R. L. (1999). Analyzing cockpit communication: the links between language, performance, error, and workload. In *Proceedings of the Tenth International Symposium on Aviation Psychology* (pp. 689-695). Columbus, OH: The Ohio State University.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000) Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech, *Computational Linguistics 26(3)*, 339-373, 2000.