Automated Team Discourse Modeling: Test of Performance and Generalization

Peter W. Foltz[1,3] (pfoltz@crl.nmsu.edu), Melanie J. Martin[2] (mmartin@cs.csustan.edu), Ahmed Abdelali [1](ahmed@crl.nmsu.edu), Mark Rosenstein[3] (mbr@pearsonkt.com), Rob Oberbreckling[3] (rjo@pearsonkt.com)

1. Computing Research Laboratory and Department of Psychology
New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003

2. Department of Computer Science
California State University, Stanislaus, 801 West Monte Vista Avenue, Turlock, CA 95382

3. Pearson Knowledge Technologies
4940 Pearl East Circle, Suite 200, Boulder, CO, 80305

Abstract

Team communication provides a rich source of discourse, which can be analyzed and tied to measures of team performance. Our goal is to better understand and model the relationship between team communication and team performance to improve team process, develop collaboration aids, and improve the training of teams. In the present work, we use Latent Semantic Analysis (LSA) for automating the analysis and annotation of team discourse. We describe two approaches to modeling team performance. The first measures the semantic content of a team's dialogue as a whole to predict the team's performance. The second categorizes each team member's statements using an established set of discourse tags and uses them to predict team performance. In three experimental settings we demonstrate the ability of these approaches to model performance and generalize to new datasets.

Introduction

Evaluating teams of decision-makers in complex problem-solving environments requires effective models of individual and team performance. Without these models, current methods of assessing team and group performance rely largely on global outcome metrics, which often lack information rich enough to diagnose failures or suggest improvements in team cognition or process. Modeling and assessment of teams has been hindered by the fact that the richest source of data, the verbal communication among team members, is difficult to collect and analyze. Prior attempts at annotating the content of communication have relied on tedious manual techniques or have only employed limited data such as frequencies, pattern analyses and durations of communications. With the advent of NLP, AI and machine-learning techniques that can measure the semantic content of communication discourse, novel methods for the analysis and modeling of team communication can be applied.

A team's verbal communication data provides a rich indication of cognitive processing at both the individual and the team level. This can be tied back to both the team's and each individual team member's abilities and knowledge. The manual analysis of team communication has shown promising results, see for example, Bowers et al. (1998). This analysis, however, is quite costly, where coding for content can take upwards of 28 hours per 1 hour of tape (Emmert, 1989) and can be subjective. Thus, there is a need for techniques to automatically analyze team communications to categorize and predict performance.

The ultimate goal is to be able to automatically analyze a team's communication and incorporate that analysis into models that can provide feedback during or after training, as well as predict performance. This work extends approaches to computational analysis of language, while also providing techniques to improve modelling of teams and measuring performance.

A number of AI, statistical and machine learning techniques have been applied to discourse modeling, generally for the purpose of improving speech recognition and dialogue systems. However, few have focused directly on just the content of the discourse of teams. Recent methods include decision trees (Core 1998), statistical modelling based on current utterance and discourse history (Chu-Carroll 1998), and hidden Markov models. For example, in the work by Stolcke et al., (2000), they were able to predict the tags assigned to discourse within 15% of the accuracy of trained human annotators, while Kiekel et al., (2004) developed markov models of communication patterns among team members that predicted overall performance.

In this work we generate predictive models using Latent Semantic Analysis (LSA) to measure free-form verbal interactions among team members. Because LSA can measure and compare the semantic information in these verbal interactions, it can be used to characterize the quality and quantity of information expressed. LSA analysis can be used to determine the semantic content of any utterance made by a team member as well as to measure the semantic similarity of an entire team's communication to another team.

This paper extends research on automated techniques for analyzing the communication and predicting team performance using corpora of communication of teams performing simulated military missions (see Kiekel et al., 2002; Martin & Foltz, 2004). We focus on two applications of this approach in order to test how well

these methods generalize to different data sets and parameter conditions. The first application predicts team effectiveness based on an analysis of the entire discourse of the team during a mission. The second application predicts categories of discourse for each utterance made by team members and uses the tags to predict performance. Testing generalizability permits characterizations of the robustness of the methods. We look at two levels of generality. At the lower level, we consider the range of situations for which a given communications model can produce meaningful predictions. At the higher level, we consider the range of team communication scenarios that LSA based techniques can successfully model. We illustrate the applicability of the approach across a number of datasets. We conclude with a discussion of how these techniques can account for the role of communications in teams. Overall we illustrate how NLP, AI, and machine learning techniques can be used to automatically model and predict team performance using realistically sized data sets of team communication.

Latent Semantic Analysis

LSA is a fully automatic corpus-based statistical modeling method for extracting and inferring relations of expected contextual usage of words in discourse (Landauer, Foltz and Laham, 1998).

In LSA a training text is represented as a matrix, where each row represents a unique word in the text and each columns represents a text passage or other unit of context. The entries in this matrix are the frequency of the word in the context. A singular value decomposition (SVD) of the matrix results in a 100-500 dimensional "semantic space", where the original words and passages are represented as vectors. The meaning of any passage is the average of the vectors of the words in the passage (Landauer et al., 1997). Words, utterances, and whole documents can then be compared against each other by computing the cosine between the vectors representing any two texts. This provides a measure of the semantic similarity of those two texts, even if they do not contain words in common.

LSA has been used for a wide range of applications and for simulating knowledge representation, discourse and psycholinguistic phenomena. These approaches have included: information retrieval (Deerwester et al., 1990), automated essay scoring (Landauer et al., 2001), and automated text analysis (Foltz, 1996). More recently Serafin and Di Eugenio (2004) used LSA for dialogue act classification, finding that LSA can effectively be used for such classification and that adding features to LSA showed promise. In addition, Martin and Foltz (2004) found that LSA predicted overall team performance scores as well as effectively classified speech acts.

Experiment 1

Three corpora of team transcripts were collected as a result of three different experiments that simulate operation of an Uninhabited Air Vehicle (UAV) in the CERTT (Cognitive Engineering Research on Team Tasks) Lab's synthetic team task environment (CERTT UAV-STE) (see http://www.certt.com). CERTT's UAV-STE is a three-team-member task in which each team member is provided with distinct, though overlapping, training; has unique, yet interdependent roles; and is presented with different but overlapping information during the mission. The overall goal is to fly the UAV to designated target-areas and to take acceptable photos at these areas. To complete the mission, the three-team members need to share information with one another and work in a coordinated fashion. Most communication is done via microphones and headsets, although some involves computer messaging.

Table 1. Corpora Statistics

Corpus	Transcripts	Teams	Missions	Utterances
AF1	67	11	7	20245
AF3	85	21	7	22418
AF4	85	20	5	22107

The three corpora are labelled by experiment name: AF1, AF3, and AF4. Each corpus consists of a number of team-at-mission transcripts, where mission duration is approximately 40 minutes. Some statistics are shown in Table 1. All communication was manually transcribed.

To train LSA, we added 2257 documents to the transcripts of all of the corpora to create a semantic space. These documents consisted of training documents and pre- and post-training interviews related to UAVs. Unless otherwise noted all results reported were computed using 300 dimensions, although results were robust across a range of dimensions.

Predicting Team Performance

Throughout the CERTT UAV-STE experiments an objective performance measure was calculated to determine each team's effectiveness at completing the mission. The performance score was a composite of objective measures including: amount of fuel/film used, number/type of photographic errors, time spent in warning and alarm states, and un-visited waypoints. This composite score ranged from -200 to 1000. The score is highly predictive of how well a team succeeded in accomplishing their mission. We used two approaches to predict these overall team performance scores: correlating entire mission transcripts with one another and by correlating tag frequencies with the scores.

Prediction Using Whole Transcripts Our first approach to measuring content in team discourse is to analyze the transcript as a whole. Using a k-nearest neighbor method that has been highly successful for scoring essays with LSA (Landauer et al., 1998), we used whole transcripts to predict the team performance score. We generate the predicted team performance scores as follows: Given a subset of transcripts, S, with known performance scores, and a transcript, t, with unknown performance score, we can estimate the performance score for t by computing its

similarity to each transcript in S. The similarity between any two transcripts is measured by the cosine between the transcript vectors in the semantic space. To compute the estimated score for t, we take the average of the performance scores of the 10 closest transcripts in S, weighted by cosines. A holdout procedure was used in which the score for a team's transcript was predicted based on the transcripts and scores of all other teams (i.e. a team's score was only predicted by the similarity to other teams). Tests on the AF1 corpus showed that the LSA estimated performance scores correlated strongly with the actual team performance scores ($r = 0.76$, $p < 0.01$, $r=0.63$, $p<.01$ when correcting for the repeated measure structure (see Martin & Foltz, 2004 for additional details on the approach). Thus, the results indicate that we can accurately predict the overall performance of the team (i.e. how well they fly and complete their mission) just based on our automatic analysis of their mission transcripts.

Generalization of Team Performance scores for Different Corpora While the results were successful for the AF1 corpus, it is important to determine if similar results hold for the other two corpora. In addition, it is important to determine if the technique can operate successfully by training the algorithm on the performance scores of one corpus in order to predict performance scores on another corpus. This approach is equivalent to having collected N transcripts from teams flying UAVs on a set of particular missions and then trying to predict a new set of teams performing a different set of missions. Thus, delimiting the bounds of generalization reveals how robust such a system could be in more realistic contexts where different teams may have to fly entirely novel missions.

We tested the generalization for the AF3 set of transcripts, by training our algorithm on the performance scores of the AF3 experiment and predicting the performance scores from the other experiment (AF4). Using the 10 closest transcripts, as before, the LSA estimated scores strongly correlated with the actual scores or AF3, showing only a four percent degradation in performance ($r=0.72$ to $r=0.66$). Thus, there was a high level of generalization from one training corpus to predicting the performance scores of another.

Automated Discourse Tagging

Another approach to utilizing the semantic content of team dialogues to measure performance is to focus on the dialogue in the transcripts at the turn level. We designed an algorithm to learn from human coded content of the communication data and then apply the tool to code new communication data.

We used a tag set developed by Bowers et al. (1998) to analyze airplane cockpit team communication. The set consists of tags for *acknowledgement, action, fact, planning, response, uncertainty, and non-task* communication. The frequency of the occurrence of these tags in team discourse has been shown to be predictive of team performance.

The 67 transcripts in AF1 were manually coded. Of these 12 were coded by two humans independently. Working within the limitations of the manual annotations, we developed an algorithm to code transcripts automatically, resulting in some decrease in accuracy, but at significant savings in time and resources.

We established a lower bound for coding performance of 0.27 by computing the tag frequency in the 12 AF1 transcripts coded by two annotators. If all utterances were coded with the most frequent tag, the percentage coded correctly would be 27%.

Algorithm for Automatic Annotation We tested our algorithm to automatically annotate the data, by computing a "corrected tag" for all turns in the 12 transcripts coded by two human annotators. This was necessary due to the only moderate agreement between the humans. We used the union of the sets of tags assigned as the "corrected tag". The union better captures all likely code types within a turn. A disadvantage to using "corrected tags" is the loss of sequential tag information within individual turns. However the focus of this study was on identifying the existence of relevant discourse, not on its order within the turn.

Then, for each of the 12 team-at-mission transcripts, we automatically assigned "most probable" tags to each turn, based on the corrected tags of the "most similar" turns in the other 11 transcripts, using Martin and Foltz (2004)algorithms: LSA and LSA+.

The LSA algorithm uses only LSA to measure semantic similarity to predict the most probable codes, while the LSA+ algorithm adds two syntactic features (Martin & Foltz 2004). Using LSA+ the performance was only 10% and 15% below human-human agreement, depending on which agreement measure is used. (see Table 2).

We used two measures of agreement: C-value (Schvaneveldt, 1990) measures the proportion of inter-coder agreement, without taking into account agreement by chance and Cohen's Kappa, does accounts for chance agreement, as shown in Table 2.

Table 2. Kappa and C-Values.

Annotators-Agreement	C-Value	Kappa
Human-Human	0.70	0.62
LSA-Human	0.59	0.48
LSA+-Human	0.63	0.53

The results suggest that we can automatically annotate team transcripts with tags. While the approach is not quite as accurate as human coders, LSA is able to tag an hour of transcripts in under a minute. As a comparison, it can take half an hour or longer for a trained tagger to do the same task.

Predicting Performance from Tags While demonstrating that the method is able to tag, it is critical to determine the relationship of the type of utterances to performance. We computed correlations between the

team performance score and tag frequencies in each team-at-mission transcript.

The tags for all utterances in the 67 AF1 transcripts were first generated using the LSA+ method. The tag frequencies for each team-at-mission transcript were then computed by counting the number of times each individual tag appeared in the transcript and dividing by the total number of individual tags occurring in the transcript.

Our results indicate that frequency of certain types of utterances correlate with team performance. The correlations for tags predicted by computer are shown in Table 3.

Table 3. Tag to Performance Correlations.

TAG	PEARSON CORR.	SIG. 2-Tailed
Acknowledgement	0.335	0.006
Fact	0.320	0.008
Response	-0.321	0.008
Uncertainty	-0.460	0.001

We see that the automated tagging provides useful results that can be interpreted in terms of team processes. Teams that tend to state more facts and acknowledge other team members more tend to perform better. Those that express more uncertainty and need to make more responses to each other tend to perform worse. These results are consistent with those found in Bowers et al. (1998), but were generated automatically rather than by the hand-coding done by Bowers. The results therefore illustrate a direct link between the types of language used and team performance.

Generalization Experiments

While the above results show that the approach can provide accurate predictions of team performance, it is important to test the generalization to different team tasks. In this section, we present results from two additional data sets: Navy TADMUS and Air Force Research Labs WCAS.

We obtained 64 transcribed missions from the Navy Tactical Decision Making Under Stress (TADMUS) dataset collected at the Surface Warfare Officer's School (SWOS) (See Johnston, Poirer and Smith-Jentsch, 1998). In the scenario, a ship's air defense warfare (ADW) team performs the detect-to-engage (DTE) sequence on aircraft in the vicinity of the battle group, and report it to the Tactical Action Officer and Bridge. Associated with the transcripts were a series of Subject Matter Expert (SME)-rated performance measures (see Table 4).

As in Experiment 1, we developed a semantic space of information relevant to the domain and then computed predicted scores for each of the SME ratings. Unlike Experiment 1, we combined the LSA k-nearest measure with a series of other natural language measures that have been successfully used in characterizing essay quality and account for effects of word order (See Landauer, Laham & Foltz, 2001). Using a stepwise regression model, the system built a model that incorporated, typically the 4 to 6 best measures. The correlation of the model's predicted scores to the SME scores using a Jack-knife correlation which predicts each score for each transcript on the basis of the other 63 transcripts is shown in Table 4.

Table 4. Model correlations to SME ratings

SME measure	Jack-knife Corr. to SME ratings
Leadership (guidance + priority)	0.73
Brevity	0.58
Clarity	0.68
Completeness of reports	0.63
Passing information	0.61
Proper phraseology	0.78
Seeking sources	0.57
Situation update	0.45
Providing and requesting backup/assistance	0.62
Error correction	0.57
Providing guidance	0.61
Stating priorities	0.73
Anti-Air Warfare Team Observation	0.61
Communication	0.59
Information Exchange	0.50
Supporting Behavior	0.61

The results show significant correlations to all SME ratings and they show that the approach to predicting team performance generalizes to highly different types of team communication data.

In another experiment in cooperation with the Air Force Research Lab (AFRL) in Mesa, Arizona, we analyzed engagement transcripts recorded from F-16 simulators. We analyzed communication transcribed from radio audio using automatic speech-to-text software. All engagements were air-to-air scenarios against an enemy threat involving a team of four F-16s and an airborne warning and control system aircraft (AWACS). AFRL provided subjective SME team performance ratings corresponding to 230 engagement transcripts.

One SME evaluates an engagement, but different SMEs were used over the time of our data. SME agreement data was not available, although previous research on similar data found SME agreement of 0.42. (See Krusmark, Schreiber & Bennett, 2004)

As before, we created a semantic space of relevant text. For team performance prediction, we used the same language measures as the TADMUS experiment augmented with speaker frequency counts and speaker frequency ratios. Because we had utterance time stamps, we also included aggregate time measures of average time between utterances and the engagement time duration. Predictive models were created in the same manner as in the TADMUS experiment. Table 5 shows the resulting correlations between our predictions of team performance and the original SME ratings of team performance and Table 6 shows correlations to objective measures.

Table 5. Correlations to subjective SME ratings

SME measure	Corr
Planning operations	0.58
Rate the comm.. about the targeting/sorting/shots	0.50
Aggregate score for overall situational awareness	0.54
Aggregate score for overall communications	0.42
Aggregate score for overall engagement	0.44

Table 6 Correlations to objective indicators

Objective indicator	Corr
Team identification number	0.80
Mission identification number	0.67
Day of week of engagement	0.63

These result show we could predict/identify which team was performing the task based solely on our communications analysis. The teams' tasks vary in a known way over the course of the week of the team's training and the results show we can often determine the day of the week the engagement was performed using our communications analysis alone.

As with the TADMUS results, these AFRL results show significant correlations ($p < 0.0001$) with both the SME and the objective indicators again demonstrating that this approach generalizes to disparate team communication data sets.

Conclusions

Overall, the results of the studies show that LSA-based algorithms can be used to predict team performance both through modeling team communications directly and via tagging content. These results show that the approach generalizes beyond specific corpora and training sets. The results from the tagging portion of the study are comparable to other efforts of automatic discourse tagging using different methods and different corpora (Stolcke et al., 2000), which found performance within 15% of the performance of human taggers. While LSA relies only on a semantic model, ignoring word order and other syntactic and discourse factors, Experiment 2 shows that it can also be incorporated with additional statistical measures of language (see also Serafin & Di Eugenio 2004).

In addition to being able to use the LSA-based approach to discourse tagging, this study demonstrates how it can be applied as a method for doing automated measurement of team performance. The LSA-predicted team performance scores correlated strongly with the actual team performance measures. This demonstrates that analyses of discourse can automatically measure how well a team is performing on a mission. This has implications both for automatically determining what discourse characterizes good and poor teams as well as developing systems for monitoring team performance in near real-time.

For example, we can now locate utterances in the semantic space that correspond to places where teams received high or low team scores. These can provide indications of the types of language that are strongly correlated with good and poor performance. It can further identify potential knowledge gaps in teams. Because of the highly interactive nature of the task, there are certain pieces of knowledge that must flow between team members at critical points. These techniques can identify if this information has been conveyed. Typically, modelling these tasks have heretofore relied on large amounts of hand-coding and SME knowledge in order to determine the types of knowledge and team processes involved at any part of a large team interaction. The results suggest that the current approach can be used in combination with other modelling efforts of teams.

In terms of applying this research to team dialogues, the automated modeling provide cost-effective and efficient approaches for analyzing communications data. The experiments reported here show that the LSA techniques work with both transcribed and automated speech recognition (ASR) output confirming, Foltz, Laham & Derr (2003) which showed that LSA predictions derived from ASR output was highly robust. With 40 percent word error rates, LSA's prediction ability decreased by only 10-15%. Thus, these methods can yield information on team communication patterns that are valid, reliable, and useful to the assessment and understanding of team performance and cognition--necessary prerequisites to the development of team training programs and the design of technologies that facilitate team performance. In particular, application domains that are communications-intensive and that require a high degree of team coordination can especially benefit from these streamlined methods for assessing team communication.

Research into modelling team cognition based on team discourse is a new but growing area. However, until recently, the large amounts of transcript data have limited researchers from performing analyses of team discourse. The results of this study show that applying AI and NLP techniques to team discourse can provide accurate predictions of performance. These automated tools can help inform theories of the nature of communication in team performance and also aid in the development of more effective automated team training systems.

Acknowledgements

This research was completed in collaboration with the CERTT laboratory including, Nancy Cooke, Preston Kiekel, Jamie Gorman, Susan Smith, and David Farwell from the Computing Research Laboratory and with data provided by Joan Johnston, NAVAIR and Winston Bennett, AFRL. This work was supported by the Office of Naval Research, DARPA, Army Research Laboratory and Air Force Research Laboratory.

References

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

Chu-Carroll, J. (1998). A Statistical Model for Discourse Act Recognition in Dialogue Interactions. Papers from the 1998 AAAI Spring Symposium. Jennifer Chu-Carroll and Nancy Green, Program Cochairs. 2001. *Technical Report SS-98-01. Published by The AAAI Press*, Menlo Park, California. Pp. 12-17.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 34-46.

Core, M. (1998). Analyzing and Predicting Patterns of DAMSL Utterance Tags. Papers from the 1998 AAAI Spring Symposium, Jennifer Chu-Carroll and Nancy Green, Program Cochairs, *Technical Report SS-98-01, Published by The AAAI Press*, Menlo Park, California. Pp. 18-24.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Emmert, V. (1989). Interaction analysis. In P. Emmert and L. L. Barker (Eds.), *Measurement of Communication Behavior*, pp. 218-248. White Plains, NY: Longman, Inc.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers. 28(2)*, 197-202.

Foltz, P. W. Laham, D., & Derr, M. (2003). Using Automatic Speech Recognition for Team Communication Analysis. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*.

Johnston, J.H., Poirer, J., and Smith-Jentsch, K.A. (1998). "Decision Making Under Stress: Creating a Research Methodology". In, Cannon-Bowers, J. & Salas, E. (Eds.), *Making Decisions Under Stress: Implications for Individual and Team Training,* pp. 39-5). Washington, DC: APA,

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

Kiekel, P., Gorman, J., & Cooke, N. (2004). Measuring Speech Flow of Co-located and Distributed Command and Control Teams During a Communication Channel Glitch. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 683-687.

M. Krusmark, B. Schreiber, & W. Bennet (2004). The Effectiveness of a Traditional Gradesheet for Measuring Air Combat Team Performance in Simulated Distributed Mission Operations. (AFRL-HE-AZ-TR-2004-0090). Air Force Research Laboratory, Warfighter Readiness Research Division, May 2004.

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

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

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

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

Serafin, R., & Di Eugenio, B. (2004). FLSA: Extending Latent Semantic Analysis with features for dialogue act classification. *ACL04, 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July.

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