Automated Team Discourse Annotation and Performance Prediction Using LSA

Melanie J. Martin
Department of Computer Science
New Mexico State University
P.O. Box 30001, MSC CS
Las Cruces, New Mexico 88003-8001
mmartin@cs.nmsu.edu

Peter W. Foltz
Department of Psychology
New Mexico State University
P.O. Box 30001, MSC 3452
Las Cruces, New Mexico 88003-8001
pfoltz@crl.nmsu.edu

Abstract

We describe two approaches to analyzing and tagging team discourse using Latent Semantic Analysis (LSA) to predict team performance. The first approach automatically categorizes the contents of each statement made by each of the three team members using an established set of tags. Performance predicting the tags automatically was 15% below human agreement. These tagged statements are then used to predict team performance. The second approach measures the semantic content of the dialogue of the team as a whole and accurately predicts the team's performance on a simulated military mission.

1 Introduction

The growing complexity of tasks frequently surpasses the cognitive capabilities of individuals and thus, often necessitates a team approach. Teams play an increasingly critical role in complex military operations in which technological and information demands require a multi-operator environment. The ability to automatically predict team performance would be of great value for team training systems.

Verbal communication data from teams provides a rich indication of cognitive processing at both the individual and the team level and can be tied back to both the team's and each individual team member's abilities and knowledge. The current manual analysis of team communication shows promising results, see for example, Bowers et al. (1998). Nevertheless, the analysis is quite costly. Hand coding for content is time consuming and can be highly subjective. Thus, what is required are techniques for automatically analyzing team communications in order to categorize and predict performance.

In the research described in this paper we apply Latent Semantic Analysis (LSA), to measure free-form verbal interactions among team members. Because it can measure and compare the semantic information in these verbal interactions, it can be used to characterize the quality and quantity of information expressed. This 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. In this paper we describe research on developing automated techniques for analyzing the communication and predicting team performance using a corpus of communication of teams performing simulated military missions. We focus on two applications of this approach. The first application is to automatically predict the categories of discourse for each utterance made by team members during a mission. These tagged statements can then be used to predict overall team performance. The second application is to automatically predict the effectiveness of a team based on an analysis of the entire discourse of the team during a mission. We then conclude with a discussion of how these techniques can be applied for automatic communications analysis and integrated into training.

2 Data

Our corpus (UAV-Corpus) consists of 67 transcripts collected from 11 teams, who each completed 7 missions that simulate flight of an Uninhabited Air Vehicle (UAV) in the CERTT (Cognitive Engineering Research on Team Tasks) Lab's synthetic team task environment (CERTT UAV-STE). 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 and 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.

The 67 team-at-mission transcripts in the UAV-Corpus contain approximately 2700 minutes of spoken dialogue, in 20545 separate utterances or turns. There are approximately 232,000 words or 660 KB of text. All communication was manually transcribed.

We were provided with the results of manual annotation of the corpus by three annotators using the Bowers Tag Set (Bowers et al. 1998), which includes tags for: acknowledgement, action, factual, planning, response, uncertainty, and non-task related utterances. The three annotators had each tagged 26 or 27 team-at-missions so that 12 team-at-missions were tagged by two annotators. Inter-coder reliability had been computed using the C-value measure (Schvaneveldt, 1990). The overall C-value for transcripts with two taggers was 0.70. We computed Cohen's Kappa to be 0.62 (see Section 4 and Table 1).

In addition to the moderate level inter-coder agreement, tagging was done at the turn level, where a turn could range from a single word to several utterances by a single speaker, and the number of tags that taggers assigned to a given turn might not agree. We hope to address these limitations in the data set with a more thorough annotation study in the near future.

3 Latent Semantic Analysis

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

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), and automated text analysis (Foltz, 1996). In addition, LSA has been applied to a number of NLP tasks, such as text segmentation (Choi et al., 2001). More recently Serafin et al. (2003) used LSA for dialogue act classification, finding that LSA can effectively be used for such classification and that adding features to LSA showed promise.

To train LSA we added 2257 documents to the corpus UAV transcripts. These documents consisted of training documents and pre- and post-training interviews related to UAVs, resulting in a total of 22802 documents in the final corpus. For the UAV-Corpus we used a 300 dimensional semantic space.

4 Automatic Discourse Tagging

Our goal was to use semantic content of team dialogues to better understand and predict team performance. The approach we focus on here is to study the dialogue on the turn level. Working within the limitations of the manual annotations, we developed an algorithm to tag transcripts automatically, resulting in some decrease in performance, but a significant savings in time and resources.

We established a lower bounds tagging performance of 0.27 by computing the tag frequency in the 12 transcripts tagged by two taggers. If all utterances were tagged with the most frequent tag, the percentage of turns tagged correctly would be 27%.

Automatic Annotation with LSA. In order to test our algorithm to automatically annotate the data, we computed a "corrected tag" for all 2916 turns in the 12 team-at-mission transcripts tagged by two taggers. This was necessary due to the only moderate agreement between the taggers. We used the union of the sets of tags assigned by the taggers as the "corrected tag".

The union, rather than the intersection, was used since taggers sometimes missed relevant tags within a turn. The union of tags assigned by multiple taggers better captures all likely tag types within the 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 team-at-missions. For a given turn, T, the algorithm proceeds as follows:

Find the turns in the other 11 team-at-mission transcripts, whose vectors in the semantic space have the largest cosines, when compared with T's vector in the semantic space. We choose either the ones with the top n (usually top 10) cosines, or the ones whose cosines are above a certain threshold (usually 0.6). The corrected tags for these "most similar" turns are retrieved. The sum of the cosines for each tag that appears is computed and normalized to give a probability that the tag is the corrected tag. Finally, we determine the predicted tag by applying a cutoff (0.3 and 0.4 seem to produce the best results): all of the tags above the cutoff are chosen as the predicted tag. If no tag has a probability above the cutoff, them the single tag with the maximum probability is chosen as the predicted tag.

We also computed the average cosine similarity of T to its 10 closest tags as a measure of certainty of categorization. For example, if T is not similar to any previously categorized turns, then it would have a low certainty. This permits the flagging of turns that the algorithm is not likely to tag as reliability.

In order to improve our results, we considered ways to incorporate simple discourse elements into our predictions. We added two discourse features to our algorithm: for any turn with a question mark, "?", we increased to probability that uncertainty, "U", would be one of the tags in its predicted tag; and for any turn following a turn with a question mark, "?", we increased to probability that response, "R", would be one of the tags in its predicted tag.

We refer to our original algorithm as “LSA” and our algorithm with the two discourse features added as “LSA+”. Using LSA+ with our two methods now performs only 11% and 15% below human-human agreement (see Table 1).

We realize that training our system on tags where humans had only moderate agreement is not ideal. Our failure analyses indicated that the distinctions our algorithm has difficulty making are the same distinctions that the humans found difficult to make, so we believe that improved agreement among human annotators would result in similar improvements for our algorithm.

The results suggest that we can automatically annotate team transcripts with tags. While the approach is not quite as accurate as human taggers, 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.

Measuring Agreement. The C-value measures the proportion of inter-coder agreement, but does not take into account agreement by chance. In order to adjust for chance agreement we computed Cohen’s Kappa (Cohen 1960), as shown in Table 1.

CODERS-AGREEMENT	C-VALUE	KAPPA
Human-Human	0.70	0.62
LSA-Human	0.59	0.48
LSA+-Human	0.63	0.53

Table 1. Kappa and C-Values.

5 Predicting Overall Team Performance

Throughout the CERTT Lab UAV-STE missions a 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 0 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 the tag frequencies with the scores and by correlating entire mission transcripts with one another.

Team Performance Based on Tags. We computed correlations between the team performance score and tag frequencies in each team-at-mission transcript.

The tags for all 20545 utterances 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 preliminary results indicate that frequency of certain types of utterances correlate with team performance. The correlations for tags predicted by computer are shown in Table 2.

TAG	PEARSON CORRELATION	SIG. 2-TAILED
Acknowledgement	0.335	0.006
Fact	0.320	0.008
Response	-0.321	0.008
Uncertainty	-0.460	0.000

Table 2. Tag to Performance Correlations.

Table 2 shows 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.

Team Performance Based on Whole Transcripts. Another approach to measuring content in team discourse is to analyze the transcript as a whole. Using a method similar to that used to score essays with LSA (Landauer et al. 1998), we used the transcripts to predict the team performance score. We generate the predicted team performance scores was 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 UAV-Corpus 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). Our results indicated that the LSA estimated performance scores correlated strongly with the actual team performance scores ($r = 0.76$, $p < 0.01$), as shown in Figure 1. 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 an analysis of their transcript from the mission.

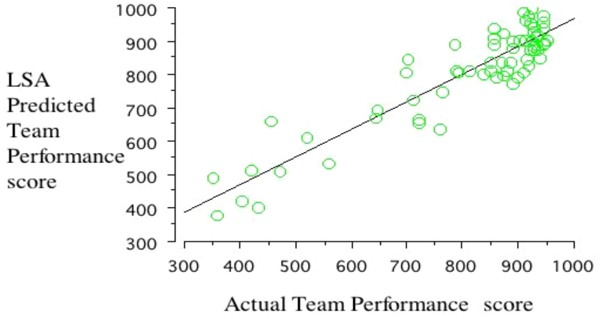

Figure 1. Correlation: Predicted and Actual Team Performance.

6 Conclusions and Future Work

Overall, the results of the study show that LSA can be used for tagging content as well as predicting team performance based on team dialogues. Given the limitations of the manual annotations, the results from the tagging portion of the study are still 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. We plan to conduct a more rigorous manual annotation study. We expect that improved human inter-coder reliability would eliminate the need for "corrected tags" and allow for sequential analysis of tags within turns. It is also anticipated that incorporating additional methods that account for syntax and discourse turns should further improve the overall performance, see also Serafin et al. (2003).

Even with the limitations of the discourse tagging, our LSA-based approach demonstrates it can be applied as a method for doing automated measurement of team performance. Using automatic methods we were able to duplicate some of the results of Bowers, and colleagues, (1998) who analyzed the sequence of content categories occurring in communication in a flight simulator task. They found that high team effectiveness was associated with consistent responding to uncertainty, planning, and fact statements with acknowledgments and responses.

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. We are currently exploring two promising avenues to predict performance in real time: integration of speech recognition technology, and inter-turn tag sequences.

Research into team discourse is a new but growing area. However, up to recently, the large amounts of transcript data have limited researchers from performing analyses of team discourse. The results of this study show that applying NLP techniques to team discourse can provide accurate predictions of performance. These automated tools can help inform theories of 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, Steven Shope, Preston Kiekel, Jamie Gorman and Susan Smith. This work was supported by Office of Naval Research and Army Research Laboratory grants.

References

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

F. Y. Y. Choi, P. Wiemer-Hastings, and J. D. Moore. 2001. Latent Semantic Analysis for Text Segmentation. *In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing,* Lillian Lee and Donna Harman (Eds.), 109—117.

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

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

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

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

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

R. Serafin, B. Di Eugenio, and M. Glass. 2003. Latent Semantic Analysis for dialogue act classification. *HLT-NAACL03, 2003 Human Language Technology Conference*, Edmonton, Canada, May (Short Paper)

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