15

Automated Communication Analysis of Teams

Peter W. Foltz and Melanie J. Martin

With the advent of advanced communication technology, individuals are better able to work as teams in complex and geographically distributed situations. Teams provide effective means to solve problems in these complex task environments. In such domains as military, civil emergency response, business planning, and medicine, tasks can often exceed the capacity of individual performance. This necessitates that individuals work in teams, with each individual providing some part of the overall solution. One critical difference between problem solving as an individual and in teams is that team members must communicate with each other, providing information that can permit a more effective solution than any one individual working alone. This communication can be a rich indicator of teamwork, coordination, learning, knowledge, collaboration, situation awareness, stress, and workload. Therefore, analysis of such communication can be used to generate measures of team performance and can provide a better understanding of team processes.

While such an analysis can provide useful characterizations of team performance and processes, communication analysis can be very time consuming, often requiring large amounts of tedious hand-coding of the data. Even when the analyses are performed, it can still be hard to relate the results to models of performance in a way that may be generalized to other teams and domains. Thus, what is required are ways of quickly analyzing communication and automatically deriving models of performance. This chapter focuses on automated approaches to characterizing team performance through team communication data. It examines a number of methods that analyze both the pattern of interactions as well as the content of what is said by team members. The chapter takes a multidisciplinary approach, incorporating methods from computational linguistics, machine learning, artificial intelligence, as well as engineering psychology approaches to modeling teams. Highly effective measures of team performance can be derived using this approach, and these measures can

be used for both improving the modeling of teams and within applications for team training and monitoring.

Does Communication Predict Performance?

Communication is a very rich source of data on team interactions. When tasks are performed by individuals, it is difficult to measure what a person is thinking. In contrast, tasks that require team members to communicate with each other force the team members to transmit information that, in turn, reveals parts of their cognitive states. These states can include information about individual and common knowledge, situation awareness, degree of uncertainty present, and plans and strategies. Viewed in this manner, team communication may provide something akin to a verbal protocol analysis (e.g., Ericsson & Simon, 1984). Unlike many other measures taken during team tasks (e.g., tests of knowledge, situation awareness; see Cooke, Salas, Cannon-Bowers, & Stout, 2000), it is not invasive and is a natural byproduct of team interaction.

Subjectively, by listening to teams performing tasks and with reasonable accuracy, how well a team is performing can be characterized. Illustrating this is the fact that subject matter experts (SMEs) often monitor teams by purely listening to their communication. For example, in military situations, commanders may listen to radio communication and, based on its content, flow, and speed, may be able to assess their unit's performance. Communication serves as a noninvasive, yet important, proxy for measuring cognition reflected in both team process and performance.

A wide range of studies have shown that hand-coded analyses of communication in teams can predict performance (see Harris & Sherblom, 2002 for a review). These studies have looked at the frequency, patterns, and content of communication. The frequency of types of communications has often been quantified to measure performance. Bowers, Jentsch, Salas, and Braun (1998) analyzed communication sequences of aircrews in flight simulation experiments with a goal of providing better team training and of reducing crew-generated errors. They developed a tag set to annotate the team discourse, and the results of their manual analyses showed promise for further automated investigation of team communication patterns. For example, they found that by examining individual statements, poorer-performing teams had a higher proportion of nontask-related communications. An analysis of the communication patterns, revealed significant differences between successful and unsuccessful crews; generally good teams were more likely to follow statements of uncertainty, fact, planning,

or action with acknowledgments or responses (Bowers et al., 1998; see also Oser, Prince, Morgan, & Simpson, 1991).

In another study of communication patterns, Xiao, Seagull, Mackenzie, Ziegert, and Klein (2003) asked experts to annotate videotapes of surgical teams doing trauma resuscitations for initiator and target team members. These communication diagrams were then analyzed for particular patterns. Using this technique, they were able to quantitatively differentiate high versus low task urgency, high versus low team experience, and leadership.

Computing the frequency of communication has also been used to characterize performance, although with varied results. In some research, high-performing teams communicate with higher overall frequency than low-performing teams (Mosier & Chidester, 1991; Orasanu, 1990), but in other cases, this finding has not been supported (e.g., Thornton, 1992). Communication frequency can be affected by such factors as the level of team workload, task difficulty, and team and individual expertise. In some studies frequency is reduced with high workload (Kleinman & Serfaty, 1989; Oser et al., 1991), whereas in other studies it is increased with high workload (e.g., Stout, 1995). Although not all findings are consistent, taken as a whole, results suggest that the manual analysis of communication (based on the frequency, pattern, and content) may be useful in characterizing aspects of team performance.

Automatically Analyzing Communication

While manual analysis of communication can prove useful, it can be quite expensive and time consuming. A single team, such as a team in a command and control center, could generate many hours of data in a typical one-day task. Despite the large volume of communication generated during team tasks, very small amounts of the data are typically collected or analyzed because the amount of communication can easily exceed the capacities of the people who need to study it. Emmert and Barker (1989, p. 244) cited an example of a study requiring 28 hours of transcription and encoding for each hour of communication. Automating the study of communication has the potential to permit such analyses to be done in near real time. The claim made in this chapter is that by applying computational approaches to modeling language, methods to model team communication can be derived that provide an automated and highly accurate approach for generating team performance measures.

Building Communication Models

The remainder of the chapter lays out an approach to automated analysis of communication using a range of cross-disciplinary modeling methods. With the description of each method, the chapter also details studies that test the approach in a number of domains. Finally, a description is given of the utility of this approach to improving modeling of teams as well as developing applications for team monitoring and training.

There are two primary approaches to developing models of human performance. The first, which we call a *theory-driven approach*, is to start with a cognitive, social, or communication theory that posits something about task communication. The researcher then decides the key factors to examine and devises an approach to test these factors. After running an experiment, the experimenter concludes how well the model accounts for the factors. The second approach is a *model-building approach*, which we use in this chapter. In this approach, we start with a set of human-derived or objective performance measures for teams. Human-derived measures are ratings of team performance by SMEs, often including intuitive or holistic ratings of communication, situation awareness critical events, team errors, or classifications of utterances that are indicative of performance (e.g., uncertainty vs. planning). Objective performance measures can include time on task, objectives completed, resource utilization, kills, or communication failures. Given that a set of communication outputs from the team and some performance measures exist, machine-learning techniques are then used to infer the relationship between the performance measures and the communication data. Essentially, the system is discovering relationships between communication and performance. The theory-driven and model-driven approaches differ in that the latter does not initially posit a particular relationship between communication and performance but instead tries to find out if, and the degree to which, such a relationship exists. The validity of the relationship can then be tested on new or held-out data.

Of course, to find these relationships one must have capable computational models that can perform this inferential step. These computational models must accurately measure features in communication that would relate to measures of team cognition. To create such a model, recent advances in the fields of computational cognitive models (e.g., latent semantic analysis, or LSA; Landauer, Foltz, & Laham, 1998); computational linguistics (e.g., Jurafsky & Martin, 2000); social network analysis (see Carley & Prietula, 2001, and Chapter 16 in this volume); machine-learning techniques that employ hill climbing, clustering, classification, and generalization methods; and automated speech recognition can be leveraged. By combining these techniques and applying them to team communication and performance data, predictive models of performance can be derived.

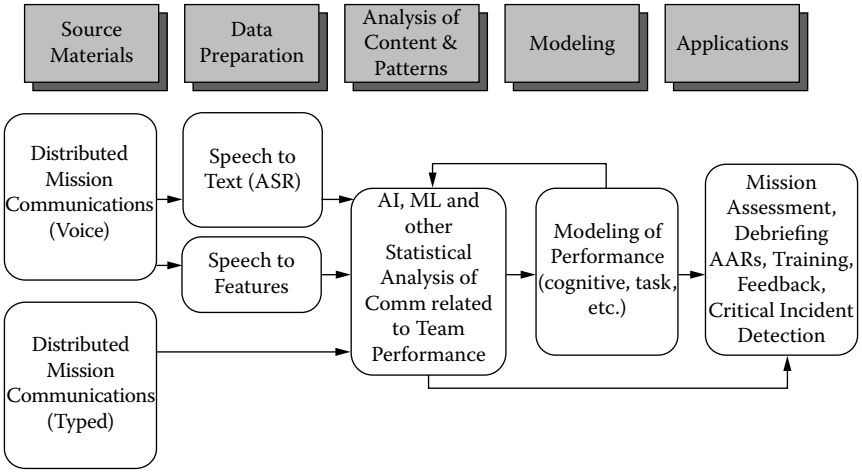

FIGURE 15.1
Communications analysis pipeline.

The goal is a communication analysis system that can turn communication into performance metrics. Figure 15.1 shows the outline of such a system. A communication analysis pipeline should be able to take input from voice data (or written chat or e-mail). Speech is automatically converted to text, which is then analyzed through computational linguistics and machine-learning statistical processes. The output can then be incorporated with other cognitive and task models of performance. In addition, the final output can provide indications of performance that can be used within training, feedback, or monitoring of teams.

Automated Communication Analysis Methods

To perform automated communication analysis, we need to distinguish the types of communication that can be analyzed. Communication data can be separated into two distinct types. First, pattern data describe the physical pattern of interactions among team members. This type of data includes who talks to who, when, and how much. Second, content focuses on what was actually said, including the content of the whole team's discourse, individual utterances, and the classification of these utterances. The remainder of the chapter covers these approaches with an emphasis on the analysis of content.

Pattern of Communication

The pattern of communication provides information about the type and duration of interactions among team members. This information can be turned into such measures as the duration of communication among team members, a characterization of the patterns of interactions, or frequency counts of team members' contributions. Analyses of the network of dynamic patterns can provide information about the social networks within the team (see Monge and Contractor, 2003, and Chapter 16 in this volume on social network analysis).

Interaction pattern data are often readily available from team tasks. These data can be obtained by, for example, recording the time and duration of communication events such as microphones, telephone calls, e-mail, or instant message use. This makes collection and analysis relatively straightforward and can permit a range of measures addressing team quality and performance, situation awareness, social structure, and adaptability of the team network.

The communication patterns can be related to social theories of communication and analyzed as a complex networked system. Such analyses can be used to measure the communication patterns over time by using lag sequential or Markov chains, time series modeling, Fourier analysis (Watt & VanLear, 1996, p. 12), or related methods that reveal the changes in the communication patterns over time (Sanderson & Fisher, 1994). The patterns can also be linked to social theories, thereby providing characterization of how well a set of communication patterns matches particular expected social patterns. Communication patterns can be analyzed as frequency counts of the categories or as a series of events (called *interaction analysis*; see Emmert, 1989 for discussion and Poole, Holmes, Watson, & DeSanctis, 1993 for an example) or by using lag sequential analysis, which examines sequences of communication patterns over different time lags. For example, in a simulated unmanned air vehicle (UAV) environment with three team members, Kiekel, Cooke, Foltz, Gorman, and Martin (2002) analyzed turn-taking sequences and dominance of team members using a communications log (CommLog) that records the quantity of communication: who is speaking to whom and the duration of the speech. Using procedural networks (ProNet; Cooke, Neville, & Rowe, 1996), which perform an automated sequential analysis using the Pathfinder (Schvaneveldt, 1990) network modeling tool, they were able to derive network path-length variables to measure a team's consistency and turn-taking behavior. These variables correlated with team performance, particularly in the skill acquisition phase of team training (Kiekel et al., 2002), and successfully identified communication glitches, where the communication channel between two team members fails during a mission (Kiekel, Gorman, & Cooke, 2004).

Kiekel et al. (2002) also developed Clustering Hypothesized Underlying Models in Sequence (CHUMS), a clustering method to determine pattern shifts in sequences of data of who talked with whom. This measure of the stability of team communications correlated with team performance during the skill acquisition phase and with situation awareness after skills had been acquired. This measure was also used to study the effects of a communication glitch on colocated and distributed teams.

Taking a different approach, but also using a simulated military framework, Carley and colleagues (Carley, Moon, Schneider, & Shigiltchoff, 2005; Moon, Carley, Schneider, & Shigiltchoff, 2005) analyzed a large amount of interaction data from America's Army, an on-line multiparty first-person-shooter game. Two types of team communication analysis were conducted on this data: (1) who talks after whom (all communications are from an individual to the whole team); and (2) type of communication. The two types of communication considered were "normal" and "in report." In-report communication occurs when a player presses a special "hot" key that broadcasts his or her location to other team members, whereas normal communication consists of the selection of predefined phrases or typed messages to broadcast to the team. Their work is automated, first creating a relational database to organize, mine, and then perform statistical analyses. Results indicated that high-frequency in-reporting was essential for winning games. Among findings for who talked after whom, a communication structure with a high sequential edge count and high network level can reduce the damage a team received (Carley et al., 2005).

In subsequent work Moon, Carley, Schneider, and Shigiltchoff (2005) used the same data with location information added to the log records, a who-was-close-to-whom social network was created. Results suggested that dense networks with two subgroups performed best. In communication network analysis, two dominant communication networks—star-shaped and long-chained—were found. Long-chained networks minimized the need for excess communication and were better (e.g., higher-performing teams communicated more than poor-performing teams; Moon et al., 2005). Some of the tools used in the work by Moon et al. (2005) just discussed have been applied to studying team situation awareness and mental models of teams during a simulated task of planning to rescue personnel from an island in the midst of war (Weil, Carley, Diesner, Freeman, & Cooke, 2006).

Analysis of the Content of Communication

While analysis of patterns of interaction among team members provides information about *who* is talking and *when* information was passed, it does

not provide information on *what* information was passed. By focusing on the communication content, one can monitor the exact words used to determine an individual's and a team's level of knowledge, situation awareness, errors in process, and workload and potentially to predict future performance problems. Thus, analysis of the content of the communication provides a much greater wealth of information about the performance of the team than pattern-based analysis alone.

Nevertheless, content analyses of verbal interactions have been hindered by a lack of effective tools. While some methods described earlier rely on tedious hand-coding of verbal interactions, automated analyses through computational linguistic and knowledge representation techniques provide the promise of real-time assessment of teams' and users' mental and performance states. A number of artificial intelligence, 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 a team's discourse.

In the remainder of this chapter, several computational approaches to content analysis are described, with a focus primarily on the approach of using latent semantic analysis, a cognitive discourse modeling technique. LSA is a fully automatic corpus-based statistical modeling method for extracting and inferring relations of expected contextual usage of words in discourse (Landauer et al., 1998). In LSA a training text is represented as a matrix, where each row represents a unique word in the text and each column represents a text passage or other unit of context. The entries in this matrix are the (possibly weighted) 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. One effect of the creation of the semantic space is that semantically similar words in the corpus are represented close to each other in the semantic space. The meaning of any passage is the average of the vectors of the words in the passage (Landauer, Laham, Rehder, & Schreiner, 1997). Words, utterances, and whole documents can then be compared with 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. Additional details are not covered here, but information on the theory behind LSA and its application can be found in papers about information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), automated essay scoring (Landauer Laham, & Foltz, 2000), and automated text analysis (Foltz, 1996, 2007).

To apply LSA to communication analysis, we generate predictive models to measure the 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.

There are two primary approaches to which the LSA-based content analysis has been applied. The first approach is to generate predictive models of performance. These models predict such measures as SME ratings of situation awareness, solution quality, planning, and objective measures of performance. The second approach is to tag (annotate) the communication for features that are predictive of performance. For example, one might want to identify all utterances where a team member does planning or expresses uncertainty (e.g., Bowers et al., 1998). In both cases we use the model-building approach. The computer generates variables from computational analyses of the communication and then uses these variables to predict objective or subjective measures of performance. The following section describes these two approaches.

Generating Predictive Models of Performance

Generating a predictive model of performance is based on the notion that team performance is reflected in the team's communication. The goal of creating such a model is to train an algorithm to extract features from the communication predictive of the team's performance. In a sense, the system learns to mimic the ability of humans to observe the communication of teams and generate a score for the team. A wide range of potential scores of performance or team process can be used to evaluate team performance and process: ratings of team performance by SMEs; holistic ratings of communication; situation awareness; identification of critical events; team errors; and objective performance measures such as time on task, number of objectives completed, kills, or communication failures. The system must infer the relationship between given scores and automatically extracted communication features.

The communication features used are a series of LSA-based measures as well as other computational linguistic features, including syntactic features and statistical features of the language (see Jurafsky & Martin, 2000 for examples of typical features used in computational linguistics). The features include measures that examine how semantically similar a team transcript is to other transcripts of known quality, measures of the semantic coherence of one team member's utterance to the next, the overall cohesiveness of the dialogue, characterizations of the quantity and quality of information provided by team members, and measures of the types of

words chosen by the team members. Hill-climbing methods (e.g., stepwise regression, random forests, support vector machines) are then used to select a subset of the features that best predict performance variables (see Witten & Eibe, 1999). Typically, the derived model has three to five features that together best predict the team performance score. The quality of the model is then tested by using cross-validation or hold-out procedures in which the model is derived on a subset of the data and tested on the remaining data. Additional details of the measures and technical information on this approach applied to military and simulated military team communication can be found in Foltz, Martin, Abdelali, Rosenstein, and Oberbreckling (2006), Gorman, Foltz, Kiekel, Martin, and Cooke (2003), and Kiekel et al. (2002).

Predictive models have been built using a number of team communication data sets and for a number of different performance measures within those data sets. Table 15.1 shows a summary of results of using derived models to predict different team performance measures across a number of data sets. For each one, every transcript was associated with one or more objective performance scores or SME ratings for the team's mission. Each of the models was developed to be specific to that particular data set and performance measure.

The data sets are as follows:

1. CERTT-UAV: Typed transcripts of teams of three people who performed UAV missions in a synthetic task environment with an objective measure of their overall team performance (see Gorman et al., 2003).
2. Tactical Decision Making Under Stress (TADMUS): Typed transcripts collected at the Surface Warfare Officer's School (SWOS) (see Johnston, Poirer, & Smith-Jentsch, 1998). In the scenario, a ship's air defense warfare (ADW) team performed the detect-to-engage (DTE) sequence on aircraft in the vicinity of the battle group and reported it to the tactical action officer and bridge. Associated with the transcripts were a series of SME-rated performance measures.
3. Air Force Research Laboratory (AFRL) F16: Automatic speech recognition generated transcripts of teams of four F-16s and an airborne warning and control system (AWACS) controller in AFRL Mesa's Distributed Mission Training Simulator with SME ratings of a range of team performance variables.
4. Office of Naval Research (ONR) noncombatant evacuation operation (NEO): Typed transcripts from teams of undergraduates who performed planning for a simulated noncombatant extraction operation with SME ratings of overall team performance.

TABLE 15.1

Predictions of Performance for Different Team Data Sets

Data Source	Number of Transcripts	Team Performance Measure	Objective/ Subjective	Corr. to Measure
CERTT-UAV AF1	67	Composite score of objective measures	O	0.76
CERTT-UAV AF3	85	Composite score of objective measures	O	0.72
Navy TADMUS	64	Leadership	S	0.73
Navy TADMUS	64	Completeness of reports	S	0.63
Navy TADMUS	64	Providing/requesting backup	S	0.62
Navy TADMUS	64	Error correction	S	0.57
Navy TADMUS	64	Information exchange	S	0.50
AFRL F-16 DMT	229	Planning operations	S	0.58
AFRL F-16 DMT	229	Situation awareness	S	0.54
AFRL F-16 DMT	229	Overall engagement quality	S	0.44
AFRL F-16 DMT	229	Number of prior missions in simulator	O	0.67
ONR NEO	16	Rating of overall team performance	O	0.90
ARL SASO	480	Aggregate score for correct team actions	O	0.61

5. Army Research Laboratory (ARL) stability and support operations (SASO): Typed transcripts from undergraduates who performed intelligence decision making during simulated stability and support operations.

The correlations presented are Pearson correlations using a hold-out procedure in which each team transcript score is predicted by deriving a model based on all remaining transcripts. This approach provides a conservative estimate of prediction ability and generalizability.

Overall, the results from Table 15.1 show that the technique can predict many different objective and SME-rated metrics of team performance. These metrics include quality of communication, leadership, information passing, providing and requesting assistance, error correction, planning, situation awareness, and engagement quality. These metrics are aspects not only of communication but also of general cognition, knowledge, and skills involved in performing team tasks. Although predicted

performance varied across domains, all were highly significant and, as such, suggest ways to improve performance in those areas. Human interrater reliability was not assessed for any of the subjective measures; however, the correlations presented in Table 15.1 are likely close to the level of agreement among humans. For example, a prior study on similar AFRL human subject ratings data found SME agreement (using alpha) of 0.42 (see Krusmark, Schreiber, & Bennett, 2004). Thus, it is likely that the computer-based methods were correlating at near the maximum level of human–human intercorrelation.

The results show that by measuring language variables from a team's transcript as a whole, one can accurately characterize the quality of the team. While the models used LSA as a critical component for measuring content in the discourse and it accounted for the largest amount of variance, additional language variables significantly improved the predictions. These variables included the complexity of language, the frequency of usage, and the choice of words.

While these studies all focus on spoken or written speech acts, it should be noted that communication does not just have to represent actual speech acts. Communication can be any information shared among team members, such as documents. Hill, Dong, and Agogino (2002) studied levels of shared understanding and team cohesiveness in engineering design teams by applying LSA to design documents generated in the collaborative design process (e.g., mission statements, concept selection rationale, prototype description, test plans, design evaluation). Assessments generated by their automated methods had about 80% agreement with the assessments of human experts.

Automated Discourse Annotation of Content

As described already, analyzing networks based on who speaks to whom and the content of a team's whole transcript can provide a large amount of information about team processes, situation awareness, and performance. However, a more complete picture requires that we also analyze the content of the individual utterances or dialogue acts within the team communication dialogue. This provides refined information about what any individual or group of individuals are saying at any point in time. For instance, one would want to know when an individual is planning versus expressing uncertainty, since this provides information about the individual's situation awareness and their performance as a team member contributing to the overall team performance. Similarly, if a particular person has passed information to his or her commander, or if a leader has sent the appropriate commands at the appropriate time, this may enable real-time performance evaluation during training and monitoring of teams.

Most analysis of discourse content at the utterance level has required hand annotation of the discourse (e.g., Bowers et al., 1998). Manual annotation is expensive and time-consuming and can introduce subjectivity or bias. To remedy the situation, extensive recent work in the computational linguistics community has been performed to develop automatic annotation techniques using primarily statistical and machine learning tools. Essentially, the problem of annotating is a problem of classification. A computer or human annotator needs to examine a part of the communication and assign it to a particular category. This work is now surveyed to give the reader an idea of the range of methods that can be applied.

Classifiers

Dialogue act annotation or tagging can be viewed as a classification problem: Given a finite set of tags and a set of features (attributes) of the dialogue act (utterance), the goal is to assign the correct (or most probable) tag to the utterance guided by the values of the features. Viewed this way, the problem is well suited to supervised machine-learning approaches to build a classifier. Machine learning in this context is generally supervised because the classifier needs to learn to tag the dialogue acts based on some amount of manually tagged data.

To better comprehend the classifier, it is noted that the set of possible tags is predefined and may be specific to the discourse genre being tagged. For example, a tag set for annotating military mission transcripts may be smaller than would be necessary for unrestricted general conversations. A key element in building a classifier is finding a good set of features that can be efficiently and automatically extracted from the dialogue. Features generally contain information about the syntax, semantics, or context of the utterance. Each utterance can be represented by a vector containing the values for each feature on that utterance. The vector can then be used by the classifier to assign an appropriate tag to the utterance.

LSA-Based Classifiers

LSA lends itself well as a feature for classifiers because it provides information about the semantic content of any utterance. Martin and Foltz (2004) and Foltz et al. (2006) used LSA to classify utterances from the CERTT-UAV corpus using the Bowers et al. (1998) tag set. In their classifier, the main features were the semantic similarity between a given utterance and utterances whose tags are known, augmented with some additional syntactic features. The semantic similarity between utterances is measured by the cosine of the angle between the vectors representing the utterances in the LSA semantic space. The concept behind the approach is that if a subset of the utterances was already tagged by a human, the computer could then learn to tag in the same way by comparing a new utterance to

utterances that had already been tagged. The results showed that the system could tag within 15% of the accuracy of human taggers, as measured by human–human agreement versus Human–computer agreement (see Foltz et al., 2006). In addition, the system was able to tag an hour of team discourse in under a minute, whereas it took human taggers about 45 minutes per hour of dialogue. A similar approach was taken by Serafin and Di Eugenio (2004) to classify dialogue acts in tutoring conversations.

Additional Tagging Methods

Many of the other methods that have been applied to discourse annotation or dialogue act tagging have been successfully used in part-of-speech tagging, where context features play a larger role. Prominent among these methods are decision trees, Markov chains (n-gram models), and hidden Markov models (HMMs).

A decision-tree classifier is constructed by recursively partitioning the training data (an already classified set of utterances) based on statistical features extracted from the utterances. At each step the feature is selected that most reduces the uncertainty about the class in each partition of the data. Once the decision tree is constructed, it can be used to assign the most probable class to a new utterance, based on its features. In relatively early work, studies by Mast, Niemann, Nöth, and Schukat-Talamazzini (1996) and Core (1998) used decision trees in dialogue act classification and concluded that n-gram or HMMs seemed more promising.

N-grams or Markov chains estimate the probability of a given tag under the assumption that the probability of the tag depends only on the previous n-tags (local context) and that it is stable over time. A Markov chain can be represented as a state diagram, where the states are tags and the edges are transitions between the states; when a probability is assigned to each edge, we have a Markov model. An HMM has an additional layer of broader categories or hidden relationships learned by the classifier. Promising results in the area of dialogue act tagging were obtained by Chu-Carroll (1998), Stolcke et al. (2000), and Venkataraman, Stolcke, and Shirberg (2002). For example, Stolcke et al. was able to predict the tags assigned to discourse within 15% of the accuracy of trained human annotators in conversational speech.

As automatic dialogue act tagging continues to grow as a research area, the range of tag sets, corpora, and methodologies has grown. In more recent work, Clark and Popescu-Belis (2004) explored the use of multilayered maximum entropy classifiers on multiparty meeting corpora. Their work includes the definition of a new tag set and discussion of some issues of tag-set design, including theoretical soundness, empirical validation, and mapping to existing tag sets. Also working in the area of multiparty meetings and using a maximum entropy classifier, Ang, Liu, and Shirberg (2005) found that both the segmentation of spoken dialogue and classification

of dialogue acts are difficult for a fully automatic system in this domain. Challenges include system performance degradation due to word recognition errors, multiple speakers with frequent overlap, and interruption. Ji and Bilmes (2005), who also worked in this area, provided a full analysis of how graphical models—in particular generative and conditional dynamic Bayesian networks—can be adapted to dialogue act tagging.

In the area of e-mail dialogue act classification, Carvahlo and Cohen (2005) developed a dependency-network-based collective classification algorithm using maximum entropy classifiers that provides modest but statistically significant improvement in some cases. Also providing a new approach, combining natural language processing (NLP) analysis with information retrieval (IR) techniques, Feng, Shaw, Kim, and Hovy (2006) were able to detect conversation focus in threaded discussions.

Overall, the area of dialogue act tagging is an emerging area of proven usefulness in many settings, including automated tutoring systems and information retrieval. Work discussed herein shows reasonable success with LSA-based methods and HMMs. Improved results will aid in our ability to understand team communication and, hence, processes, mental models, situation awareness, and performance. Thus, although this area is growing quite quickly in computational linguistics, the techniques are quite applicable to those who want to apply them to team analyses.

Conclusions

Communication represents a rich resource for monitoring and assessing teams. It provides a natural form of data that reveals cognitive and social aspects of individual and team functioning. Recent research has started to examine the role of communication in team cognition, both from the point of view of understanding how communication affects teams and how communication can reveal the functioning of the teams. However, until recently, the large amounts of transcript data and the difficulty in having reliable coding have limited researchers from performing effective analyses of team discourse. Team performance measurement requires understanding theories and techniques from a range of fields. Most typically these fields have included human factors, cognitive psychology, educational measurement, and communications. This chapter posits that more attention to computational language fields can further improve measurement of team process and performance. With the advances in artificial intelligence, computational cognitive modeling, and computational linguistics, as well as sufficiently fast computers, it has become possible to perform automated analyses of team discourse.

The methods described in this chapter primarily took a model-driven approach, deriving the team performance model from collected data rather than generating a theory-based model and then testing against the data. Both approaches are valid ways of performing team analysis; however, the model-driven approach lends itself well to applying computational linguistic and machine-learning techniques to large amounts of team communication data. The two approaches can still be used on the same data and may help support each other.

There is a range of different types of communication that can be analyzed in teams, including analysis of patterns, communication content of the team as a whole, and individual utterances. Within each of those, a range of computational techniques can be applied. Therefore, before applying such analyses, team researchers will need to determine what aspects of communication they want to measure and what performance metrics they want to derive from the analysis. For example, a pattern analysis may be more suitable for addressing questions of social structure within a team, whereas automatically deriving team performance scores from the content may be more suitable for analysis of team and individual situation awareness. Nevertheless, these methods can be combined. One can use the methods together to examine the pattern of the flow of content. For instance, to trace a commander's intent, one could use content-based tagging to identify utterances associated with the commander's intent and then could use pattern analysis to measure how that particular information has moved through the network of team members. Indeed, a hybrid approach in which multiple automated measures are used can help provide converging evidence, higher reliability, and novel methods of tracing content across teams.

The methods described in this chapter can yield information on how communication can be turned into metrics that are valid, reliable, and useful to the assessment and understanding of team performance and cognition. The measures can address both individual- and group-level performance and can provide metrics to quantify aspects such as the quality of team planning, decision making, performance, communication, and process. The metrics can further be used to identify sources of failures and successes within teams, which can be used for both monitoring and feedback to teams. Thus, such metrics are 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 such streamlined methods for assessing team communication.

A number of applications have been developed to perform automated analyses. Foltz, Laham, and Derr (2003) and Foltz et al. (2006) showed that they could take audio communication from F-16 pilots in simulators,

convert it through automated speech recognition, and produce overall team performance predictions in almost real time. In an application focusing more on monitoring teams of learners, LaVoie et al. (in press), Lochbaum, Streeter, and Psotka (2002), and Streeter, Lochbaum, and LaVoie (2007) developed and tested a collaborative learning environment called knowledge post. The application consisted of an off-the-shelf threaded discussion group that has been substantially augmented with LSA-based functionality to evaluate and support individual and team contributions. Tests on the system at the Army War College and the U.S. Air Force Academy showed that it was able to automatically notify the instructor when discussion went off-topic, to insert expert comments and library article interjections into the discussion in appropriate places by automatically monitoring the discussion activity, and to enhance the overall quality of the discussion and consequent learning level of the participants when compared to more standard threaded discussion applications. Thus, the team analysis pipeline shown in Figure 15.1 can be completed. The techniques described in this chapter can be implemented within tools to automatically monitor teams and have effective measures and feedback.

This approach suggests a range of potential applications for assessing teams. These applications can include systems to detect critical incidents, to monitor for poor performance, to generate automated after-action reviews, to detect workload, and to provide feedback to teams and individuals on such aspects as communication and process quality, knowledge, and situation awareness failures. While helping provide new applications, the approach also helps inform theories and models of team cognition and communication. This opens new frontiers in research in which we can improve our modeling of teams through applying computational modeling. As we are better able to analyze the wealth of communication data generated by teams, we will be better able to understand and develop better theories of how teams perform.

Acknowledgments

This research was has benefited from collaboration work with the team members from Pearson Knowledge Technologies (Robert Oberbreckling, Mark Rosenstein, Noelle LaVoie, and Marcia Derr), the Cognitive Engineering Research on Team Tasks (CERTT) laboratory (Nancy Cooke, Preston Kiekel, Jamie Gorman, and Susan Smith); the Computing Research Laboratory at New Mexico State (Ahmed Abdelali and David Farwell); and from data provided by Joan Johnston at the Naval Air Systems Command (NAVAIR), Norm Warner at the Office of Naval Research (ONR),

and Winston Bennett at the Air Force Research Laboratory (AFRL). This work was supported by the ONR, the Defense Advanced Research Projects Agency (DARPA), Army Research Laboratory, and AFRL.

References

Ang, J., Liu, Y., & Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 1061–1064). Philadelphia.

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

Carley, K. M. & Prietula, M. (Eds.). (2001). *Computational organization theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carley, K. M., Moon, I., Schneider, M., & Shigiltchoff, O. (2005*). Detailed analysis of factors affecting team success and failure in the America's Army game.* Technical Report CMU-ISRI-05-120. Pittsburgh, PA: CASOS, Carnegie Mellon University.

Carvalho, V. & Cohen, W. W. (2005). On the collective classification of e-mail "speech acts." In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 345–352. New York, NY: ACM Press.

Chu-Carroll, J. (1998). A statistical model for discourse act recognition in dialogue interactions. In J. Chu-Charroll & N. Green (Eds.), *Applying machine learning to discourse processing. Papers from the 1998 AAAI spring symposium*. Technical Report SS-98-01 (pp. 12–17). Menlo Park, CA: AAAI Press.

Clark, A. & Popescu-Belis, A. (2004). Multi-level dialogue act tags. *Proceedings of SIGDIAL '04 5th SIGDIAL workshop on discourse and dialog*, 163–170. East Strandsburg, PA: Association of Computational Linguistics.

Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996). Procedural network representations of sequential data. *Human-Computer Interaction, 11,* 29–68.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42*, 151–173.

Core, M. (1998). Analyzing and predicting patterns of DAMSL utterance tags. In J. Chu-Charroll & N. Green (Eds.), *Applying machine learning to discourse processing. Papers from the 1998 AAAI spring symposium*. Technical Report SS-98-01 (pp. 18–24). Menlo Park, CA: AAAI Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T.K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Emmert, V. J. (1989). Interaction analysis. In P. Emmert & L. L. Barker (Eds.), *Measurement of communication behavior (*pp. 218–248). White Plains, NY: Longman, Inc.

Emmert, P. & Barker, L. L. (1989). *Measurement of communication behavior.* White Plains, NY: Longman, Inc.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis.* Cambridge, MA: MIT Press.

Feng, D., Shaw, E., Kim, J., & Hovy, E. (2006). Learning to detect conversation focus of threaded discussions. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 208–215. East Standsburg, PA: Association for Computational Linguistics.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computer, 28*, 197–202.

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K Landauer, W. Kintsch, D. McNamara, & S. Dennis (Eds.), *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Publishing.

Foltz, P. W., Laham, R. D., & Derr, M. (2003). Automated speech recognition for modeling team performance. In *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting*, 673–677. Santa Monica, CA: HFES.

Foltz, P. W., Martin, M. J., Abdelali, A., Rosenstein, M. B., & Oberbreckling, R. J. (2006). Automated team discourse modeling: Test of performance and generalization. *Proceedings of the 28th Annual Cognitive Science Conference,* 1317–1322. Mahwah, NJ: Lawrence Erlbaum Associates.

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. J., & Cooke, N. J. (2003). Evaluation of latent semantic analysis-based measures of communications content. *Proceedings of the 47th annual human factors and ergonomic society meeting*, 424–428. Santa Monica, CA: HFES.

Harris, T. E. & Sherblom, J. C. (2002). *Small group and team communication.* New York: Allyn & Bacon.

Hill, A. W., Dong, A., & Agogino, A. M. (2002). Towards computational tools for supporting the reflective team. In J. Gero (Ed.), *Artificial intelligence in design* (pp.305–325). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ji, G. & Bilmes, J. (2005). Dialog act tagging using graphical models. *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing* (pp. 33–36). Piscataway, NJ: IEEE.

Johnston, J. H., Poirier, J., & Smith-Jentsch, K. A. (1998) Decision making under stress: Creating a research methodology. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individuals and teams* (pp. 39–59). Washington, DC: APA.

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J., & Martin, M. J. (2002). Some promising results of communication-based automatic measures of team cognition. *Proceedings of the Human Factors and Ergonomics Society 46th annual meeting*, 298–302. Santa Monica, CA: HFES.

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.

Kleinman, D. L. & Serfaty, D. (1989). Team performance assessment in distributed decision making. In R. Gilson, J. P. Kincaid, & B. Godiez (Eds.), *Proceedings of the Interactive Networked Simulation for Training Conference* (pp. 22–27). Orlando, FL: Institute for Simulation and Training.

Krusmark, M., Schreiber, B., & Bennett, W. (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.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412–417). Mahwah, NJ: Erlbaum.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T., Laham, D., & Foltz, P. (2002). The intelligent essay assessor. *IEEE Intelligent Systems 15*(5), 27–31. Piscataway, NJ: IEEE.

LaVoie, N., Streeter, L., Lochbaum, K., Boyce, L., Krupnick, C., & Psotka, J. (in press). Automating expertise in collaborative learning environments. *International Journal of Computer-Supported Collaborative Learning.*

Lochbaum, K., Streeter, L., & Psotka, J. (2002). *Exploiting technology to harness the power of peers.* Paper presented at the Interservice/Industry Training, Simulation and Education Conference. Arlington, VA: NTSA.

Martin, M. J. & Foltz, P. W. (2004). Automated team discourse annotation and performance prediction using LSA. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL),* Boston, MA.

Mast, M., Niemann, H., Nöth, E., & Schukat-Talamazzini, E. G. (1996). Automatic classification of dialog acts with semantic classification trees and polygrams. In S. Wermter, E. Riloff, & G. Scheler (Eds.), *Connectionist, statistical, and symbolic approaches to learning for natural language processing.* Heidelberg: Springer.

Monge, P. R. & Contractor, N. S. (2003). *Theories of communication networks.* Oxford, England: Oxford University Press.

Moon, I., Carley, K. M., Schneider, M., & Shigiltchoff, O., (2005). Detailed analysis of team movement and communication affecting team performance in the America s Army Game. Technical Report CMU-ISRI-05-129, Carnegie Mellon University.

Mosier, K. L. & Chidester, T. R. (1991). Situation assessment and situation awareness in a team setting. In Y. Quéinnec & F. Daniellou (Eds.), *Designing for everyone: Proceedings of the 11th Congress of the International Ergonomics Association* (pp. 798–800). London: Taylor & Francis.

Orasanu, J. (1990). *Shared mental models and crew performance* (Report CSLTR-46). Princeton, NJ: Princeton University.

Oser, R. L., Prince, C., Morgan, B. B., Jr., & Simpson, S. (1991). *An analysis of aircrew communication patterns and content* (NTSC Tech. Rep. 90-009). Orlando, FL: Naval Training Systems Center.

Poole, M.S., Holmes, M., Watson, R., & DeSanctis, G. (1993). Group decision support systems and group communication: A comparison of decision making in computer-supported and nonsupported groups. *Communication Research, 20*, 176–213.

Sanderson, P. M. & Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction, 9*, 251–317.

Serafin, R. & Di Eugenio, B. (2004). *FLSA: Extending latent semantic analysis with features for dialogue act classification.* In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, 692–699. East Strandsburg, PA: Association for Computational Linguistics.

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

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., et al. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics, 26*, 339–373.

Stout, R. J. (1995). Planning effects on communication strategies: A shared mental model perspective. In *Proceedings of the Human Factors Society 39th annual meeting*, 1278–1282.

Streeter, L. A., Lochbaum, K. E., & LaVoie, N. (2007). Automated tools for collaborative learning environments. In T. K. Landauer, W. Kintsch, D. McNamara & S. Dennis. (Eds.) *LSA: A Road to Meaning*, (pp. 279–292). Mahwah, NJ: Lawrence Earlbaum Associates.

Thornton, R. C. (1992). *The effects of automation and task difficulty on crew coordination, workload, and performance.* Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

Venkataraman, A., Stolcke, A., & Shirberg, E. (2002). Automatic dialog act labeling with minimal supervision. In *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, 70–73. Canberra City, Australia: ASSTA.

Watt, J. H. & VanLear, A. C. (1996). *Dynamic patterns in communication processes.* Thousand Oaks, CA: Sage Publications.

Weil, S. A., Carley, K. M., Diesner, J., Freeman, J., & Cooke, N. J. (2006). *Measuring situational awareness through analysis of communications: A preliminary exercise.* The Command and Control Research and Technology Symposium, San Diego, CA.

Witten, I. A. & Eibe, F. (1999). *Data mining: Practical machine learning tools and techniques with Java implementations.* New York: Morgan Kaufmann.

Xiao, Y., Seagull, F. J., Mackenzie, C. F., Ziegert, J. C., & Klein, K. J. (2003). Team communication patterns as measures of team processes: Exploring the effects of task urgency and shared team experience. In *Proceedings of the Human Factors and Ergonomics Society 47th annual meeting*, Santa Monica, CA, 1502–1506.