Can Dialogue Features Help Predict Team Performance?

Kallirroi Georgila, Carla Gordon, Anton Leuski, Ron Artstein, David Traum
University of Southern California Institute for Creative Technologies
Los Angeles, California
{kgeorgila, cgordon, leuski, artstein, traum}@ict.usc.edu

ABSTRACT

This paper explores the question of whether team performance scores can be automatically predicted from team dialogue features. We analyze data from U.S. Navy military training exercises designed to improve decision-making under stress. These exercises were scored by subject matter experts on 11 team performance indicators, e.g., situation updates, error correction, brevity, clarity. We compute multiple dialogue features from transcriptions of the intercom messages from the participants. These features include number of speakers, number of turns, average number of words per turn, number of occurrences of specific dialogue acts, and others. Some of these features are based on manual annotations of the transcripts, while others are calculated automatically. To enhance our models with more informative features, we develop a novel annotation scheme which handles lower-level task coordination, marking the initiation and resolution points for events (commands, suggestions, and requests). Then using these features and regression we train automatic performance prediction models for each of the 11 team performance indicators, and report results varying the dialogue features and the type of regression used (Linear Ridge Regression, Support Vector Regression, Gaussian Process Regression). Our results are promising and in most cases the prediction errors (Root Mean Square Error values) fall within one standard deviation from the mean. Our models also consistently outperform the baseline that always predicts a neutral score of 3. However, more data are needed to draw stronger conclusions and compute better and more robust team performance predictions. Our work advances the state of natural language dialogue processing as a means to understand and predict team performance.

ABOUT THE AUTHORS

Kallirroi Georgila, PhD is a Research Associate Professor at the USC Institute for Creative Technologies and at USC's Computer Science Department. Her research focuses on machine learning for spoken dialogue processing. She is a past Vice President of SIGdial, and has chaired and served on many conference program committees. She is currently serving as an Associate Editor of the Dialogue and Discourse journal and an Action Editor of the Transactions of the Association for Computational Linguistics journal.

Carla Gordon for the last 8 years has been the Data Management Specialist for the Natural Language Dialogue group at the USC Institute for Creative Technologies. She is an expert on language data annotations.

Anton Leuski, PhD is a Research Scientist at the USC Institute for Creative Technologies (ICT). His research interests center around interactive information access, human-computer interaction, and machine learning. He has developed the NPCEditor toolkit used in many ICT dialogue systems for Army research and applications.

Ron Artstein, PhD is a Research Scientist at the USC Institute for Creative Technologies. He is an expert on the collection, annotation and management of linguistic data for spoken dialogue systems, and on the evaluation of implemented dialogue systems. He has led the data acquisition efforts for large-scale, public-facing spoken dialogue systems and has worked extensively in the military domain.

David Traum, PhD is the Director for Natural Language Research at the USC Institute for Creative Technologies and Research Professor at USC's Computer Science Department. His research focuses on Dialogue Communication between Human and Artificial Agents. He is a founding editor and current Editor in Chief of the Dialogue and Discourse journal, has chaired and served on many conference program committees, and is a past President of SIGdial.

Can Dialogue Features Help Predict Team Performance?

Kallirroi Georgila, Carla Gordon, Anton Leuski, Ron Artstein, David Traum
University of Southern California Institute for Creative Technologies
Los Angeles, California
{kgeorgila, cgordon, leuski, artstein, traum}@ict.usc.edu

INTRODUCTION

Communication is an important part of teamwork, but it is an open research question how different patterns of communication affect team success. In this paper, we study team dialogue and explore the question of whether team performance scores can be automatically predicted from dialogue features. We use a corpus, TADMUS (Tactical Decision-Making Under Stress; Smith et al., 2004), in which U.S. Navy trainees work together as a team to accomplish complex tasks, and are scored by subject matter experts (SMEs) on a variety of indicators of team performance (e.g., passing information, proper phraseology, brevity, clarity, error correction). In our previous work (Georgila et al., 2024) we annotated this corpus with information about content and meaning (dialogue acts) and dialogue structure (transactions). Here, we continue this work and focus on automatic team performance prediction based on manually and automatically extracted dialogue features. To enhance our models with more informative features than dialogue acts, we develop a novel annotation scheme which handles lower-level task coordination, marking the initiation and resolution points for events (commands, suggestions, and requests). There can be cases where initiating one event can trigger the initiation of another sub-event. These nested events can show how issuing commands, suggestions, and requests follows the chain of command downwards (from higher levels to lower levels) and then their resolution follows the chain of command upwards (from lower levels to higher levels). We use Linear Ridge Regression, Support Vector Regression, and Gaussian Process Regression to build team performance prediction models for each of the 11 team performance indicators, and report results varying the dialogue features used.

Our contributions can be summarized as follows: (1) We develop a novel annotation scheme annotating complex dialogue structure, namely, initiation and resolution points for commands, suggestions, and requests. (2) We provide new results (compared to Georgila et al. (2024)) based on large language models (LLMs) for automatically tagging dialogue acts. (3) We use 3 types of regression methods to build models for predicting team performance using both manually annotated and automatically extracted dialogue features (varying the dialogue features used). (4) Our experiments are performed on a real-world corpus recording military training exercises. (5) Our work advances the state of natural language dialogue processing as a means to understand and predict team performance.

RELATED WORK

Teams are small groups that work together to achieve joint goals (Cohen & Levesque, 1991). Teams collaborate on activities such as construction, resource production, maintenance, transportation, and reconnaissance, and sometimes compete against other teams (e.g., in sports or games). Team activities include joint action and full-team dialogues, but also allocation of tasks to individual team members, dialogues among subsets of team members, dialogues between team and non-team members, team formation and maintenance, and creation and updating of common ground across the team (Bell et al., 2004; Remolina et al., 2005; Priest & Stader, 2012; Brown et al., 2021). Some teams consist of only two members, or only (dyadic) conversation episodes between two members. However, many teams involve more members contributing to team tasks, thus it is important to be able to understand and analyze communication in multiparty dialogues. There has been limited work on studying team communication and analyzing team performance using natural language dialogue processing. Below we discuss some of this work.

Spain et al. (2019) explored techniques to develop a team communication analysis toolkit that can perform real-time end-to-end natural language analysis on team spoken dialogue and generate team dialogue analytics. Spain et al. (2021) used basic linguistic features such as n-grams and found that low-performing teams generated fewer unique unigrams, bigrams, and trigrams than high-performing teams. Saville et al. (2022) compared behaviors of high and

low-performing teams and found significant interaction effects between time and performance group for the overall speech frequency and the number of given commands. More recently, Spain et al. (2025) used LLMs (particularly GPT-4) for dialogue act classification using team dialogue data.

Rahimi & Litman (2020) developed a method for learning entrainment embeddings to predict team performance using the Teams Corpus (Litman et al., 2016). Enayet & Sukthankar (2021) also used the Teams Corpus to learn embeddings from multiparty dialogues so that teams with similar conflict scores are closer in the vector space. These embeddings were extracted from dialogue acts, sentiment polarity, and syntactic entrainment. Enayet & Sukthankar (2021) found that the teamwork phase affected the utility of each embedding type. Shibani et al. (2017) designed an automated assessment system for providing students with feedback on their teamwork competency. They extracted features from text, such as unigrams and bigrams, and compared a rule-based approach vs. supervised machine learning methods for classifying coordination, mutual performance monitoring, team decision-making, constructive conflict, team emotional support, and team commitment.

In our previous work (Georgila et al., 2024) we used transcriptions from two military training exercises, TADMUS (U.S. Navy) and Squad Overmatch (U.S. Army), which were designed to improve team decision-making under stress. These exercises were scored by SMEs on a variety of indicators of team performance. We annotated part of the TADMUS and Squad Overmatch datasets with information about dialogue participation, content and meaning, and dialogue structure. Also, we annotated Squad Overmatch with dialogue actions relevant to team development (TD) and advanced situation awareness (ASA). We built machine learning models for automatic dialogue act labeling, and used both manually annotated and automatically extracted dialogue features to calculate correlations between indicators of team effectiveness and dialogue features. We found that requesting and providing information were strongly correlated with how teams were rated on TD and ASA, and identifying and describing threats were correlated with ratings on TD (but not ASA, probably due to data sparsity). Additionally, for each indicator of team performance, there were some dialogue acts that exhibited strong correlation with that indicator.

DATA

TADMUS (Tactical Decision-Making Under Stress) is an empirical decision support system (DSS) developed at the Space and Naval Warfare Systems Center in San Diego to mitigate the limitations of human cognition in the following 3 areas: perception, attention, and memory (Smith et al., 2004). TADMUS DSS was used as a decision aid tool in U.S. Navy team training exercises. Ninety-six U.S. Navy officers were randomly assigned to 16 teams. Each team had 6 members playing the roles of decision makers in a medium fidelity combat simulation. The task of the team was to defend their ship from attacking aircraft. The roles of the team members were Commanding Officer (CO, at the top of the chain of command), Tactical Action Officer (TAO, reporting to CO), Electronic Warfare Supervisor (EWS, reporting to TAO), Anti-Air Warfare Coordinator (AAWC, reporting to TAO), Tactical Information Coordinator (TIC, reporting to AAWC), and Identification Supervisor (IDS, reporting to TIC). Figure 1 shows the team hierarchy. The participants wore headsets and microphones and communicated using an intercom. There were also an Airborne Warning And Control System (AWACS) with call sign "RAINBOW", and other external entities with call signs "GW" and "GB".

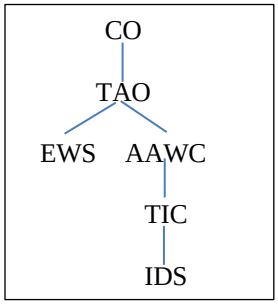

Figure 1. Team Hierarchy

Each of the 16 teams completed 4 scenarios (Bravo, Charlie, Delta, India), simulating peace-keeping missions with a very high number of ambiguous targets to deal with in a short period of time. Thus the TADMUS corpus includes 64 dialogues from the above scenarios plus a few more dialogues from additional scenarios (85 dialogues in total). The TADMUS dialogues are quite long, about 250 turns per dialogue on average. The dialogues are manually transcribed and annotated with speaker and timing information. An excerpt of a TADMUS dialogue can be seen in Table 1, along with dialogue annotations described below.

The original TADMUS corpus also includes team performance scores. Each team exercise (dialogue) was scored by SMEs on a variety of team effectiveness indicators (general-purpose and domain-specific). Whenever there was disagreement between the two SMEs, it was resolved by having a more senior SME provide the final score. For our experiments we use 11 such general-purpose indicators (score types), all ranging from 1 (lowest) to 5 (highest).

Table 1. Excerpt from a TADMUS Team Dialogue

Trans	Speaker	Transcript	Dialogue Act	Initiate Event	Ack/Resolve Event
29	RAINBOW	PANTHER forward, this is RAINBOW, interrogative, do you have comms with DESERT EAGLE 101 102 over?	request-info	init-request(23)	
30	TIC	EW/TIC you have anything bearing 023?	request-info	init-request(24)	
29	GW	GW that's negative over.	negative		resolve-request(23)
29	RAINBOW	GW, this is RAINBOW, I have poor comms with DESERT EAGLE 101 102.	inform		
29	RAINBOW	Can you contact them this circuit over?	command	init-command(25)	
30	EWS	Negative.	negative		resolve-request(24)
30	TAO	I got it at 500ft doing 80knts. So it is a possible helo.	inform		
30	TAO	So you might go out with a level 1 query on that one.	suggest	init-suggest(26)	
30	TIC	I copy that TAO.	ack		ack-suggest(26)
29	GW	This is GW, roger over.	ack		ackwilco-command(25)
30	IDS	Unidentified aircraft bearing 275 ... identify yourself and state your intentions over.	warning		resolve-suggest-action(26)
30	EWS	I see that you are looking at track 7031.	confirm-info	init-request(27)	
30	TAO	That's correct.	affirmative		resolve-request(27)
30	TAO	Track 7014 just dropped about 20 thousand ft range about 39m bearing 302.	inform		
30	IDS	TAO I issue level 1 query on 7014.	confirm-action		resolve-suggest(26)

The team performance indicators are:

- **Seeking Sources:** Proactively asking for information from multiple (internal or external) sources to accurately assess the situation.
- **Passing Information:** Anticipating another team member's need for information and passing it to an individual or group of individuals without having to be asked.
- **Situation Updates:** An update given by a team member either to the entire team or a subset of the team (or to others outside the team) which provides an overall summary of the big picture as they see it.
- **Proper Phraseology:** Use of standard terms or vocabulary when sending a report.
- **Complete Reports:** Following standard procedures that indicate which pieces of information are to be included in a particular type of report and in what order.
- **Brevity:** The degree to which team members avoid excess chatter, stammering and long winded reports which tie up communication lines.
- **Clarity:** The degree to which a message sent by a team member is audible (e.g., loud enough, not garbled, not too fast).
- **Error Correction:** Instances where a team member points out that an error has been made and either corrects it themselves or sees that it is corrected by another team member.
- **Provide/Request Backup/Assistance:** Instances where a team member either requests assistance or notices that another team member is overloaded or having difficulty performing a task, and provides assistance to them by actually taking on some of their workload.

- **Providing Guidance:** Instances where a team member directs or suggests that another team member take some action or instructs them on how to perform a task.
- **Stating Priorities:** Instances where a team member specifies, either to the team as a whole or to an individual team member, the priority ordering of multiple tasks.

In our previous work (Georgila et al., 2024) we annotated TADMUS with transactions (Sinclair & Coulthard, 1975; Carletta et al., 1997; Kawano et al., 2023) and dialogue acts (Bunt et al., 2012; 2020). Transactions represent sub-dialogues that together are part of attempting to achieve the same task purpose. Transactions are indicated with an integer, and utterances that are part of the same transaction will have the same integer. Transactions can be interleaved and examples are shown in the first column of Table 1. Dialogue acts indicate the main purpose of each utterance, e.g., requesting information ("request"), confirming information ("confirm-info"), providing a suggestion ("suggest"), confirming an action ("confirm-action"), issuing a command ("command"), etc. Examples of dialogue act annotations are shown in the fourth column of Table 1. Our full dialogue act taxonomy is presented in our previous work (Georgila et al., 2024).

NEW ANNOTATION SCHEME

We created a novel annotation scheme to handle lower-level task coordination, marking the initiation and resolution points for events (commands, suggestions, and requests). These are shown in the last two columns of Table 1, with the full taxonomy presented in Table 2. The initiations and resolutions are accompanied by numbers in parentheses, indicating the association between individual initiations and their resolution. For example, during transaction 30 (see Table 1), we can see that a suggestion is initiated by TAO ("init-suggest(26)"), acknowledged by TIC ("ack-suggest(26)"), and resolved with an action (issuing a warning) by IDS ("resolve-suggest-action(26)"). Then IDS fully resolves the suggestion by confirming the warning action ("resolve-suggest(26)"). There can also be cases where initiating one event can trigger the initiation of another sub-event. These nested events can show how issuing commands, suggestions, and requests follows the chain of command downwards (from higher levels to lower levels) and then their resolution follows the chain of command upwards (from lower levels to higher levels). Note that one transaction may involve multiple initiation/resolution pairs. It is also possible but less frequent that a single utterance (depending on its complexity) triggers multiple initiations and/or resolutions of events.

A TADMUS dialogue, annotated by three annotators, was used for measuring inter-annotator reliability of the initiation/resolution scheme. Agreement was measured separately on initiation and resolution, since these are distinct annotation fields. Since both initiation and resolution are fairly sparse annotations (most utterances are not an initiation or resolution), we first calculated agreement just on whether an utterance was marked as an initiation or resolution, without consideration of what type of initiation or resolution it was. Krippendorff's α among all three raters was 0.77 for initiation and 0.66 for resolution (raw agreement 92% and 88%, respectively); pairwise agreements among the annotators were 0.67, 0.74, and 0.90 for initiation and 0.58, 0.66, and 0.72 for resolution (raw agreement 89%, 91%, 97% and 85%, 89%, 89%, respectively). All the agreement figures were somewhat lower when taking into account the type (label) of initiation or resolution: Krippendorff's α among all three raters was 0.73 for initiation and 0.56 for resolution (raw agreement 90% and 83%, respectively); pairwise agreements among the annotators were 0.64, 0.68, and 0.87 for initiation and 0.49, 0.54, and 0.64 for resolution (raw agreement 86%, 89%, 95% and 80%, 85%, 85%, respectively).

AUTOMATIC DIALOGUE ACT TAGGING

We built automatic dialogue act classifiers using 21 dialogues. The reason for using automatically annotated tags in addition to manually annotated tags is because for future work we envision an automatic pipeline for analyzing team communication and predicting team performance. For training our classifiers, we used the MLTextClassifier library from Apple[1], which can generate 5 types of models: a conditional random field model (CRF), a maximum entropy model (MaxEnt), a static transfer learning model (tl.static), and two dynamic transfer learning models (tl.bert and tl.elmo). We also developed another model based on LLMs by fine-tuning GPT-4o Mini.

[1] https://developer.apple.com/documentation/createml/mltextclassifier/modelalgorithmtype

Table 2. List of Initiation and Resolution Tags

Initiate/Resolve Tag	Description
INITIATE EVENT	
init-request	initiate a new request
init-request-repeat	initiate again a request
init-suggest	initiate a new suggestion
init-suggest-repeat	initiate again a suggestion
init-command	initiate a new command
init-command-repeat	initiate again a command
ACKNOWLEDGE EVENT	
ack-request	acknowledge a request
ackwilco-request	acknowledge a request committing to carry it out
ack-suggest	acknowledge a suggestion
ackwilco-suggest	acknowledge a suggestion committing to carry it out
ack-command	acknowledge a command
ackwilco-command	acknowledge a command committing to carry it out
RESOLVE EVENT (provide information or confirm that the event has been resolved)	
resolve-request	resolve a request
resolve-suggest	resolve a suggestion
resolve-command	resolve a command
INITIATE ACTION (start performing an action)	
init-request-action	start carrying out a request (start performing an action)
init-suggest-action	start carrying out a suggestion (start performing an action)
init-command-action	start carrying out a command (start performing an actions)
RESOLVE ACTION (finish performing an action)	
resolve-request-action	resolve a request by performing an action
resolve-suggest-action	resolve a suggestion by performing an action
resolve-command-action	resolve a command by performing an action

Table 3. Dialogue Act Tagging Results – Precision, Recall, and F1-score – Best result per metric in bold

Model	Macro			Micro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
GPT-4o Mini fine-tuned	74.10	**69.92**	**75.01**	**84.23**	**83.70**	**83.55**
CRF	**75.07**	54.28	65.70	76.93	75.83	75.38
MaxEnt	66.93	58.39	65.99	75.27	75.34	75.03
tl.bert	63.21	56.52	65.55	76.15	75.94	76.22
tl.elmo	68.58	55.41	67.69	76.67	76.23	76.68
tl.static	69.97	46.98	62.84	72.11	71.95	71.88

For training and evaluating all classifiers we used 10-fold cross-validation. Compared to our previous work (Georgila et al., 2024), here we present results with additional transfer learning models and the fine-tuned GPT-4o Mini model. Results for macro precision, recall, and F1-score (with equal weights to each class) and micro precision, recall, and F1-score (with equal weights to each utterance, i.e., weighted by class size) are shown in Table 3. Our classifiers only assign one dialogue act per utterance (it is very rare that one utterance is annotated with more than one label). The best model is fine-tuned GPT-4o Mini and thus we use its annotations for the score prediction experiments presented below.

We have only manually annotated a relatively small portion of our data with dialogue act tags so there is certainly room for improvement. Also, currently our models do not use context from previous utterances, which is another consideration for future work.

Table 4. Performance Prediction (16 dialogues) – Best RMSEs per performance indicator in bold black, best model per group (manual dialogue acts vs. automatic dialogue acts) in bold italic and in a different color per group (red and blue)

Performance Indicators	Root Mean Square Error (RMSE)							Score Mean	Score Std Dev
	Manual Dialogue Acts			Automatic Dialogue Acts			Base-line		
	Linear	SVR	GPR	Linear	SVR	GPR			
Seeking sources	0.71	0.70	***0.68***	0.89	0.74	***0.69***	0.79	3.13	0.81
Passing info	***0.68***	0.71	0.71	***0.62***	0.67	0.70	0.90	3.44	0.81
Situation updates	0.79	***0.75***	0.78	***0.78***	0.81	0.80	0.87	2.88	0.89
Proper phraseology	***0.45***	0.51	0.49	***0.45***	0.49	0.48	0.83	3.69	0.48
Complete reports	***0.59***	0.63	0.61	0.76	0.69	***0.62***	0.75	3.06	0.77
Brevity	0.66	***0.54***	0.59	***0.46***	0.60	0.60	0.71	3.25	0.68
Clarity	1.12	0.88	***0.82***	0.89	***0.80***	***0.80***	1.35	3.94	1.00
Error correction	***0.47***	0.81	0.80	0.90	0.82	***0.81***	0.87	3.25	0.86
Backup/assistance	0.75	***0.62***	0.66	0.79	***0.58***	0.65	0.90	3.56	0.73
Providing guidance	0.79	0.73	***0.68***	0.80	0.70	***0.67***	1.09	3.81	0.75
Stating priorities	***0.92***	0.97	0.95	1.14	1.04	***0.96***	1.00	3.25	1.00

PERFORMANCE PREDICTION EXPERIMENTS

From each dialogue we extracted the following features: number of transactions, number of speakers, average number of turns per speaker, number of turns, number of words, average number of words per turn, number of occurrences of each dialogue act (e.g., num-request-info, num-request-confirm, num-ack, etc.), number of occurrences of each initiation or resolution tag (e.g., num-init-request, num-init-suggest, num-ack-request, num-init-request-action, num-resolve-command-action, etc.), and percentages of resolved requests, commands, and suggestions. Sums of the above numbers are also used as additional features. For example, num-resolve-action-all is the sum of num-resolve-request-action, num-resolve-command-action, and num-resolve-suggest-action.

Similarly to Georgila et al. (2019b; 2020) and Georgila (2022; 2024), we built models using different types of regression because we do not have many data points for data-hungry methods such as neural networks. In particular, we used Linear Ridge Regression (i.e., linear regression with L2 regularization), Support Vector Regression (SVR), and Gaussian Process Regression (GPR) employing the scikit-learn[2] toolkit. For SVR we used the RBF kernel and for GPR the Matérn kernel. For all our experiments we used leave-one-out cross-validation. Prediction results in terms of Root Mean Square Error (RMSE) are shown in Tables 4 and 5 (based on 16 and 8 dialogues, respectively). The lower the RMSE value the better and RMSEs range from 0 to 4 (given that scores were on a scale from 1 to 5). We can also see the mean and standard deviation for each score type in the data, and the RMSE values for a baseline model always predicting a neutral score of 3.

In Table 4 we show performance prediction results with features including manually annotated dialogue acts and automatically annotated dialogue acts using 16 dialogues, all from the same scenario (Bravo). There is no clear winner in terms of regression method. Interestingly, often models based on features including automatically annotated dialogue acts outperform models based on features including manually annotated dialogue acts. But this is not very surprising given that our best dialogue act tagging model (fine-tuned GPT-4o Mini), which generated the automatically annotated dialogue acts, performs quite well. For all score types our best models always outperform the baseline and produce RMSE values falling within one standard deviation from the mean. This is also true for most of our models (not just the best performing ones).

[2] https://scikit-learn.org/stable/

Table 5. Performance Prediction (8 dialogues) – Best RMSEs per performance indicator in bold black, best model per group (manual dialogue acts vs. automatic dialogue acts vs. manual dialogue acts plus percentages of resolved events vs. manual dialogue acts plus manual initiation/resolution tags plus percentages of resolved events) in bold italic and in a different color per group (red, blue, green, and purple)

Performance Indicators	Root Mean Square Error (RMSE)												Score Mean	Score Std Dev
	Manual Dialogue Acts			Automatic Dialogue Acts			Manual + % resolved			Manual + init/res tags + % resolved			Base line		
	LR	SVR	GPR	LR	SVR	GPR	LR	SVR	GPR	LR	SVR	GPR			
Seeking sources	***0.43***	0.76	0.75	***0.68***	0.74	0.74	***0.47***	0.72	0.74	0.80	0.79	***0.77***	0.79	2.88	0.84
Passing info	0.88	0.71	***0.67***	0.75	***0.66***	***0.66***	0.75	0.69	***0.67***	0.71	***0.61***	0.65	1.00	3.75	0.71
Situation updates	***0.94***	1.02	0.99	1.10	1.02	***1.00***	***0.89***	1.01	0.99	***0.99***	1.05	1.00	0.94	2.88	0.99
Proper phraseology	0.70	0.60	***0.57***	0.57	***0.55***	0.56	0.61	0.58	***0.57***	0.52	***0.50***	0.54	0.79	3.63	0.52
Complete reports	***0.64***	0.74	0.75	***0.59***	0.73	0.75	***0.59***	0.75	0.75	***0.73***	0.77	0.76	0.79	2.88	0.84
Brevity	***0.53***	0.65	0.60	***0.56***	0.64	0.59	0.64	0.65	***0.61***	***0.56***	0.63	0.60	0.71	3.00	0.76
Clarity	***1.03***	1.13	1.11	***1.12***	1.13	***1.12***	1.16	1.15	***1.13***	1.42	1.19	***1.15***	1.28	3.63	1.19
Error correction	***0.84***	1.08	1.08	***1.03***	1.06	1.08	***0.86***	1.06	1.07	***1.00***	1.07	1.09	1.00	3.25	1.04
Backup/ assistance	1.30	***0.91***	0.95	1.34	***0.89***	0.94	1.11	***0.89***	0.94	1.09	***0.90***	0.94	0.94	3.38	0.92
Providing guidance	1.05	0.82	***0.75***	0.93	0.76	***0.73***	0.92	0.78	***0.74***	0.95	***0.72***	0.73	0.94	3.63	0.74
Stating priorities	1.04	0.97	***0.88***	1.15	0.95	***0.87***	0.90	0.97	***0.88***	***0.87***	0.95	0.88	***0.87***	3.00	0.93

Due to the complexity of the initiation/resolution annotation scheme, we only annotated 8 dialogues with initiation and resolution events (a subset of the 16 dialogues). Thus Table 5 shows results for predicting scores based on 8 dialogues. Note that for the initiation and resolution events we only have manual annotations. For future work we intend to build models for automatic annotation of initiation and resolution events. For each score type, the first and second groups show results using the same features as in Table 4 (but for 8 dialogues only). The third group shows results using features including manually annotated dialogue acts and percentages of resolved events (based on the manually annotated initiation/resolution tags). The fourth group shows results using features including manually annotated dialogue acts, manually annotated initiation/resolution tags, and percentages of resolved events (based on the manually annotated initiation/resolution tags). Basically the features used in the fourth group are a superset of the features used in the third group.

Again in Table 5, similarly to Table 4, there is no clear winner in terms of regression method. Sometimes models based on features including automatically annotated dialogue acts outperform models based on features including manually annotated dialogue acts. For some score types (passing info, situation updates, proper phraseology, providing guidance), using information from the initiation/resolution tags helps. For all score types our best models always produce RMSE values falling within one standard deviation from the mean. Also, for all score types, except for "stating priorities", our best models outperform the baseline. For "stating priorities" the performance of some of our models is the same as the performance of the baseline. Note that for "stating priorities" the mean value of this score in the data is 3 which is what the baseline predicts. Thus it is not surprising in this case that the baseline performs as well as our best models.

Overall, our results are promising and, as mentioned above, in most cases the prediction errors (RMSE values) fall within one standard deviation from the mean. Our models also consistently outperform the baseline that always predicts a neutral score of 3. However, more data are needed to draw stronger conclusions and compute better and more robust team performance predictions. For our experiments we used data from only one scenario (Bravo), while there are overall 4 scenarios (Bravo, Charlie, Delta, India). In the future we would like to use data from more than one scenario and investigate whether results from one scenario generalize to other unseen scenarios.

CONCLUSION

We explored the question of whether team performance scores can be automatically predicted from team dialogue features. We used data from U.S. Navy military training exercises designed to improve decision-making under stress. These exercises were scored by SMEs on 11 team performance indicators, e.g., situation updates, error correction, brevity, clarity. We computed multiple dialogue features from transcriptions of the intercom messages from the participants. These features include number of speakers, number of turns, average number of words per turn, number of occurrences of specific dialogue acts, and others. Some of these features are based on manual annotations of the transcripts, while others are calculated automatically. To enhance our models with more informative features, we developed a novel annotation scheme which handles lower-level task coordination, marking the initiation and resolution points for events (commands, suggestions, and requests). Then using these features and regression we trained automatic performance prediction models which outperform baselines for each of the 11 team performance indicators, and reported results varying the dialogue features and the type of regression used (Linear Ridge Regression, Support Vector Regression, Gaussian Process Regression).

We have shown that it is possible to predict team performance based on a variety of dialogue features. Performance of course depends on the availability of team dialogue data and annotations of dialogue structure. It is interesting that more complex annotations of dialogue structure did not always result in performance gains but this could be due to lack of adequate data. It is also encouraging that using automatic dialogue act annotations resulted in performance similar to relying on manually annotated dialogue acts. With the continual advancement of LLMs we expect in the future to also see progress in automatic annotations of more complex dialogue structure. Our work advances the state of natural language dialogue processing as a means to understand and predict team performance. For future work we would also like to explore how dialogue features can be combined with other aspects related to team performance, for example, time to detect an event or threat, time to engage, etc.

Our ultimate goal is to build an automatic pipeline for analyzing team communication, predicting team performance, and providing feedback to individual team members and the team as a whole, preferably in real time. Automatically generated real-time feedback could potentially be provided with as few disruptions in the team exercises as possible, and the type and timing of feedback could be controlled to maximize efficiency, something that may not be possible with feedback generated by human instructors. Such a process would revolutionize team training in military settings and beyond. This is a very challenging task and there is still much work to be done to achieve this goal but our work is an important step forward.

Looking beyond team training, our work also has important implications for human-machine interaction, particularly with machines acting as teammates. Machines that act as teammates, must go beyond the current focus on dyadic communication (Georgila et al., 2019a) and engage in multiparty interactions (Traum et al., 2008; Xiao & Georgila, 2018; Gu et al., 2021), ideally adopting behaviors of good human teammates, contributing to team success in a range of mission types.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Army Research Office under Cooperative Agreement Numbers W911NF-20-2-0053 and W911NF-25-2-0040, and the U.S. Army under Contract Number W911NF-14-D-0005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Office or the U.S. Government. We thank Dr. Joan Johnston for providing access to the data and answering our questions.

REFERENCES

Bell, B., Johnston, J., Freeman, J., & Rody, F. (2004). STRATA: DARWARS for deployable, on-demand aircrew training. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Brown, O., Power, N., & Conchie, S. M. (2021). Communication and coordination across event phases: A multi-team system emergency response. *Journal of Occupational and Organizational Psychology*, 94(3):591–615.

Bunt, H., Alexandersson, A., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., & Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 430–437, Istanbul, Turkey.

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., & Prévot, L. (2020). The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 549–558, Marseille, France.

Carletta, C., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Cohen, P. R., & Levesque, H. J. (1991). Teamwork. *Noûs*, 25(4):487–512.

Enayet, A., & Sukthankar, G. (2021). Analyzing team performance with embeddings for multiparty dialogues. In *Proceedings of the IEEE International Conference on Semantic Computing*.

Georgila, K., Core, M. G., Nye, B. D., Karumbaiah, S., Auerbach, D., & Ram, M. (2019a). Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 737–745, Montreal, Canada.

Georgila, K., Gordon, C., Choi, H., Boberg, J., Jeon, H., & Traum, D. (2019b). Toward low-cost automated evaluation metrics for Internet of Things dialogues. In *Proceedings of the 9th International Workshop on Spoken Dialogue Systems Technology (IWSDS), Lecture Notes in Electrical Engineering*, Vol. 579, pages 161–175, Springer Singapore.

Georgila, K., Gordon, G., Yanov, Y., & Traum, D. (2020). Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 726–734, Marseille, France.

Georgila, K. (2022). Comparing regression methods for dialogue system evaluation on a richly annotated corpus. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-DubDial)*, pages 81–93, Dublin, Ireland.

Georgila, K., Gordon, C., Leuski, A., Artstein, R., & Traum, D. (2024). Studying team effectiveness via dialogue analysis. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Georgila, K. (2024). Comparing pre-trained embeddings and domain-independent features for regression-based evaluation of task-oriented dialogue systems. In *Proceedings of the 25th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 610–623, Kyoto, Japan.

Gu, J.-C., Tao, C., Ling, Z.-H., Xu, C., Geng, X., & Jiang, D. (2021). MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 3682–3692.

Johnston, J., Gamble, K., Patton, D., Fitzhugh, S., Townsend, L., Milham, L., Riddle, D., Phillips, H., Smith, K., Ross, W., Butler, P., Evan, M., & Wolf, R. (2016). Squad Overmatch for Tactical Combat Casualty Care: Phase II Initial Findings Report. Orlando, FL: *Program Executive Office Simulation, Training and Instrumentation*.

Johnston, J. (2018). Team performance and assessment in GIFT – Research recommendations based on lessons learned from the Squad Overmatch research program. In *Proceedings of the Sixth Annual GIFT Users Symposium*, vol. 6, pages 175–187.

Johnston, J. H., Phillips, H. L., Milham, L. M., Riddle, D. L., Townsend, L. N., DeCostanza, A. H., Patton, D. J., Cox, K. R., & Fitzhugh, S. M. (2019). A team training field research study: Extending a theory of team development. *Frontiers in Psychology*, 10.

Kawano, S., Yoshino, K., Traum, D., & Nakamura, S. (2023). End-to-end dialogue structure parsing on multi-floor dialogue based on multi-task learning. *Frontiers in Robotics and AI*, 10.

Litman, D., Paletz, S., Rahimi, Z., Allegretti, S., & Rice, C. (2016). The Teams Corpus and Entrainment in Multi-Party Spoken Dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1421–1431, Austin, Texas, USA.

Priest, H. A., & Stader, S. (2012). A framework for developing synthetic agents as pedagogical teammates: Applying what we already know. In *Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society*, pages 2552–2556.

Rahimi, Z., & Litman, D. (2020). Entrainment2vec: Embedding entrainment for multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8681–8688.

Remolina, E., Li, J., & Johnston, A. E. (2005). Team training with simulated teammates. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Saville, J., Spain, R., Johnston, J., & Lester, J. (2022). An analysis of squad communication behaviors during a field-training exercise to support tactical decision making. In *Proceedings of the 13th International Conference on Applied Human Factors and Ergonomics*, pages 109–116.

Shibani, A., Koh, E., Lai, V., & Shim, K. J. (2017). Assessing the language of chat for teamwork dialogue. *Educational Technology & Society*, 20(2):224–237.

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press.

Smith, C. A. P., Johnston, J., & Paris, C. (2004). Decision support for air warfare: Detection of deceptive threats. *Group Decision and Negotiation*, 13:129–148.

Spain, R., Geden, M., Min, W., Mott, B., & Lester, J. (2019). Toward computational models of team effectiveness with natural language processing. In *Team Tutoring Workshop in conjunction with the Artificial Intelligence in Education Conference (AIED)*, pages 30–39, Chicago, Illinois, USA.

Spain, R., Min, W., Saville, J., Brawner, K., Mott, B., & Lester, J. (2021). Automated assessment of teamwork competencies using evidence-centered design-based natural language processing approach. In *Proceedings of the 9th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium*.

Spain, R., Min, W., Kumaran, V., Pande, J., Saville, J., & Lester, J. (2025). Applying large language models to enhance dialogue and communication analysis for adaptive team training. *International Journal of Artificial Intelligence in Education.*

Traum, D., Marsella, S. C., Gratch, J., Lee, J., & Hartholt, A. (2008). Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proceedings of the International Workshop on Intelligent Virtual Agents (IVA), Lecture Notes in Computer Science (LNAI, volume 5208)*, Springer, pages 117–13.

Xiao, G., & Georgila, K. (2018). A comparison of reinforcement learning methodologies in two-party and three-party negotiation dialogue. In *Proceedings of the 31st International FLAIRS Conference*, pages 217–220, Melbourne, Florida, USA.