Studying Team Effectiveness via Dialogue Analysis

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

In this paper we use natural language dialogue processing as a means to understand and assess team effectiveness. In particular, we explore the question of what dialogue-related aspects contribute to the success of a team. We use 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 subject matter experts on a variety of indicators of team effectiveness, e.g., team development (TD), advanced situational awareness (ASA), situation updates, stating priorities, error correction, brevity, and clarity. We annotate part of the TADMUS and Squad Overmatch datasets with information about dialogue participation (addressees), content and meaning (dialogue acts), and dialogue structure (transactions). Also, we annotate Squad Overmatch with dialogue actions relevant to TD, e.g., providing information up and down the chain of command, and ASA, e.g., identifying and describing threats. We build machine learning models for automatic dialogue act labeling, and use both manually annotated and automatically extracted dialogue-related features to calculate correlations between indicators of team effectiveness and dialogue-related features. Our annotations show that requesting and providing information are strongly correlated with how teams were rated on TD and ASA, and identifying and describing threats is correlated with ratings on TD (but not ASA, probably due to data sparsity). Additionally, for each indicator of team effectiveness, there are some dialogue acts that exhibit strong correlation with that indicator. We conclude with a discussion on how our work can be extended and applied to automatically analyzing team communication and assessing team effectiveness.

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 has served as Vice President of SIGdial, and 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 7 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 of the Dialogue and Discourse journal, has chaired and served on many conference program committees, and is a past President of SIGdial.

Studying Team Effectiveness via Dialogue Analysis

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 what specific kinds of messages (or patterns of communication) are used by successful and problematic teams? In this paper, we study team dialogue and are particularly interested in understanding what dialogue-related aspects contribute to the success of a team. We use two corpora, TADMUS (Smith et al., 2004) and Squad Overmatch (Johnston et al., 2016), in which military trainees work together as a team to accomplish a task, and are scored by subject matter experts (SMEs) on a variety of indicators of team effectiveness (e.g., team development, advanced situational awareness, stating priorities, error correction, brevity, clarity). We annotate these corpora with information about dialogue participation (addressees), content and meaning (dialogue acts), and dialogue structure (transactions).

Our contributions are as follows: To annotate our datasets we develop 3 novel annotation schemes at the utterance level, marking dialogue acts, team development (TD) aspects, and advanced situational awareness (ASA) aspects. We also annotate transactions (Sinclair & Coulthard, 1975; Carletta et al., 1997). We build machine learning models for automatic dialogue act labeling, and use both manually annotated and automatically extracted dialogue-related features to calculate correlations between team performance scores and dialogue features. Our annotations show that requesting and providing information are strongly correlated with how teams were rated on TD and ASA, and identifying and describing threats is correlated with ratings on TD (but not ASA, probably due to data sparsity). Additionally, for each indicator of team effectiveness, there are some dialogue acts that exhibit strong correlation with that indicator. Using natural language dialogue processing as a means to understand and assess team communication is an understudied topic, and our work is an important step toward understanding the factors that affect team performance.

RELATED WORK

People are often organized into teams, i.e., 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). Teams have team goals, which are often distinct from the individual goals of their members. Team goals may be conveyed by one team member, but often originate outside the team, e.g., an order from a commander outside the team. Team members often try to manage potential conflicts between individual goals (safety of self, impression management) and team goals (completing the mission, sharing necessary information). Teams may also negotiate over goals and the best ways to achieve them, which may involve (re-)allocation of roles or tasks. 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 using Squad Overmatch data. They found significant interaction effects between time and performance group for the overall speech frequency and the number of given commands. 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.

DATA

We investigate team communication in two datasets, TADMUS and Squad Overmatch, both involving small group teams engaged in joint military training missions, with distinguished roles for team members. However, there are also notable differences, in terms of the type of mission (Army squad vs. Navy ship), the type of communication (mixed face to face and radio vs. radio), and topics (urban encounters vs. aircraft identification).

TADMUS Dataset

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 US Navy team training exercises. Ninety US Navy officers were randomly assigned to 15 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 with call sign “GOLF WISKEY” (GW), reporting to TAO), Tactical Information Coordinator (TIC, reporting to AAWC), and Identification Supervisor (IDS, reporting to TIC). The participants wore headsets and microphones and communicated using an intercom. There was also an Airborne Warning And Control System (AWACS) with call sign “RAINBOW”.

The TADMUS corpus, recording the aforementioned team training exercises, includes 85 team dialogues in 4 different scenarios (variations of the same task). 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):

- **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.

Table 1. Excerpt from a TADMUS Team Dialogue

Trans	Speaker	Transcript	Dialogue Act
29	RAINBOW	PANTHER forward, this is RAINBOW, interrogative, do you have comms with DESERT EAGLE 101 102 over?	request-info
30	TIC	EW/TIC you have anything bearing 023?	request-info
29	GW	GW that's negative over.	negative
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
30	EWS	Negative.	negative
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
30	TIC	I copy that TAO.	ack
29	GW	This is GW, roger over.	ack
30	IDS	Unidentified aircraft bearing 275 ... identify yourself and state your intentions over.	warning
30	EWS	I see that you are looking at track 7031.	confirm-info
30	TAO	That's correct.	affirmative
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

- **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.

Squad Overmatch Dataset

The Squad Overmatch research objective is to improve dismounted squad decision making under stress (Johnston et al., 2016; Johnston, 2018; Johnston et al., 2019). Seventy-one US Army squad members participated in the final evaluation event which included 6 squads completing simulation-based training exercises and live training exercises in a controlled and safe environment. Each squad consisted of 10 members, divided into two teams, with a squad leader and two team leaders in addition to other squad members. There were also some non-team members in the scenario that the team interacted with at times.

Note that the number of turns per dialogue in Squad Overmatch is much larger than in TADMUS, about 1000 turns per dialogue on average. Squad Overmatch dialogues were also more complex, involving noise and much overlapping speech. An excerpt from a Squad Overmatch dialogue can be seen in Table 2. Similar to TADMUS, these dialogues were also assessed by SMEs but with respect to Team Development (TD) and Advanced Situational Awareness (ASA).

Table 2. Excerpt from a Squad Overmatch Team Dialogue

Trans	Speaker	Transcript	Dialogue Act	TD	ASA
1	SQL	be advised we have S P time now	inform	provide-info-up	
1	RADIO	good copy three one S P time now	ack-repeat		
2	A_TMLDR	three one three one alpha <unintelligible>	inform		
2	SQL	roger just waiting for three one bravo to be set	ack, inform	provide-info-down	
3	A_TMLDR	three one three one alpha	hail		
3	SQL	three one send it	ack		
3	A_TMLDR	roger I got Father Romanov in the market	inform	provide-info-up	inform-potential-threat
3	SQL	roger that keep eyes on still waiting for three one bravo over	ack, command, inform	command-middle	
4	B_TMLDR	<name> that way go	command	command-bottom	

ANNOTATION SCHEMES

We used the following dialogue annotation schemes on the corpora presented in top-down order.

Transactions (Sinclair & Coulthard, 1975; Carletta et al., 1997) represent sub-dialogues that together are part of attempting to achieve the same task purpose, such as moving something from point A to point B. Transactions may involve combinations of requests, feedback, and information about task status. Transactions are indicated with an integer, and utterances that are part of the same transaction will have the same integer. Examples can be seen in the first columns of Tables 1 & 2. As we can see, transactions can be interleaved.

Dialogue acts indicate the main purpose of each utterance. Rather than use an established general-purpose scheme for dialogue acts, such as the ISO standard (Bunt et al., 2010; Bunt et al., 2012; Bunt et al., 2020), we focused specifically on team-related communications, oriented to exchange of information or coordination of action. When we started analyzing our data it became clear that existing general-purpose dialogue act schemes could not capture various team-related interaction phenomena that often occur in a military setting, such as tracking unsolicited vs. solicited information or targets of confirmation, at least not without extensive modifications. Nevertheless some of the dialogue acts in our scheme are similar to dialogue acts in the ISO standard, such as “request”, “suggestion”, “inform”, and “confirm”. However, the dialogue acts in our scheme tend to be more specific, a deliberate design decision given that we aim to extract informative features for assessing team effectiveness. We distinguish between “inform” and “provide-info” in the sense that the latter is always used in response to a request for information. A confirmation can also appear in two forms, information confirmation (“confirm-info”) vs. action confirmation (“confirm-action”). There are also specific dialogue acts for requesting, granting, or denying permission, as well as dialogue acts for requesting confirmations and responses, all essential communicative actions in a military setting where establishing common ground before any actions are performed is especially important. Some examples of dialogue act annotations can be seen in the fourth columns of Tables 1 and 2. Our full dialogue act taxonomy is shown in Table 3.

Finally we developed 2 new schemes for annotating dialogue actions relevant for **Team Development (TD)** and **Advanced Situational Awareness (ASA)**. The full taxonomies are shown in Table 4. The TD tags are designed to capture how information is relayed up and down the chain of command. They are classified into 4 categories in which a squad member provides new information, passes information to another squad member, gives a command, and makes a request. The ASA tags aim to encode information about threats or potential threats. **Note the distinction between TD and ASA tags, and TD and ASA scores.**

Table 3. List of Dialogue Acts (for both TADMUS and Squad Overmatch)

Dialogue Act	Description
INFORMATION	
inform	provide unsolicited information to another party
request-info	one user asks another for information
provide-info	response to request-info which provides information
confirm-info	one user repeats info back to confirm
ack	acknowledges message is heard
ack-repeat	an acknowledgment that repeats back info
neg-ack	negative acknowledgment (basically “no copy”)
hail	just getting someone’s attention (“alpha two this is bravo two”)
ACTION	
commit-action	someone says they are going to perform some action
command	one user issues a command to another
request-repeat	one user requests another to repeat last request/command
suggest	one user suggests a course of action
request-permission	one user requests permission to perform a certain action
grant-permission	grants permission to perform requested action
deny-permission	denies permission to perform requested action
confirm-action	confirm an action has been done
request-response	one user calls out to another user to respond
request-confirm	one user makes sure another user has heard a certain message
OTHER	
request-other	other kinds of requests
affirmative	simply agreeing with a statement
negative	simply disagreeing with a statement
greeting	greetings of any kind
pleasantries	small talk in general
warning	warnings issued by the team to outsiders
unintelligible	poor transcription due to noise in the recordings
other	anything that cannot be classified to any of the previous categories

Inter-annotator Agreement

To measure inter-annotator reliability for the dialogue act scheme, two annotators annotated one TADMUS dialogue with dialogue acts; Krippendorff's alpha was found to be 0.92 (93% raw agreement). To assess the reliability of the Squad Overmatch annotations, two annotators annotated one team dialogue with dialogue acts, TD, and ASA attributes. Chance-corrected inter-annotator agreement (Krippendorff's alpha) was 0.68 for dialogue acts, 0.71 for TD, and 0.41 for ASA (raw agreement was 71.5%, 81.7%, and 95.4%, respectively). TD and especially ASA are fairly sparse annotations. This causes high expected agreement, which lowers the chance-corrected agreement substantially. The biggest confusions for dialogue acts were inform vs. provide-info, and for TD, provide-info vs. pass-info (for both up and down). For TD and ASA, there were also a number of cases where one annotator labeled an utterance with a category and the other did not.

AUTOMATIC DIALOGUE ACT TAGGING

We built automatic dialogue act and TD classifiers using 20 dialogues from TADMUS (5060 utterances) and 4 dialogues from Squad Overmatch (3665 utterances). Because of data sparsity we did not build a classifier for ASA. For TD we added a ‘no-value’ class because of the sparseness of this tag (57% of the samples had no value). 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.

Table 4. List of Team Development (TD) and Advanced Situation Awareness (ASA) Tags

TD-related Tag	**Description**
PROVIDE-INFO	a squad member provides new information
provide-info-down	new info passed down chain of command (COC)
provide-info-up	new info passed up COC
provide-info-lateral	new info passed to same level of COC
provide-info-all	new info passed to team with mixed COC
correction	a squad member corrects another
PASS-INFO	a squad member relays information from one squad member to another
pass-info-down	relaying info down COC
pass-info-up	relaying info up COC
pass-info-lateral	relaying info to same level of COC
pass-info-all	relaying info to team with mixed COC
COMMAND	a squad member issues a command to another
command-top	a command coming from the top of the COC
command-middle	a command coming from the middle of the COC
command-bottom	a command coming from the bottom of the COC
REQUEST	a squad member makes a request for something
request-backup	calling for backup
request-info-up	lower in command requests info from higher in command
request-info-down	higher in command requests info from lower in command
request-info-all	asking for information in general from an audience with mixed authority
request-info-lateral	asking for information from someone with the same level of authority
request-guidance	lower in command requests guidance on what action should be taken
ASA-related Tag	**Description**
inform-potential-threat	informing someone of a potential threat (person, object or location)
describe-potential-threat	describing a potential threat
potential-threat-behavior	verbalizing nonverbal behaviors of potential threat
inform-threat	informing someone of a clear threat (person, object or location)
describe-threat	describing a clear threat
threat-behavior	verbalizing nonverbal behaviors of a confirmed threat
other-sense	an utterance related to smell, taste or touch

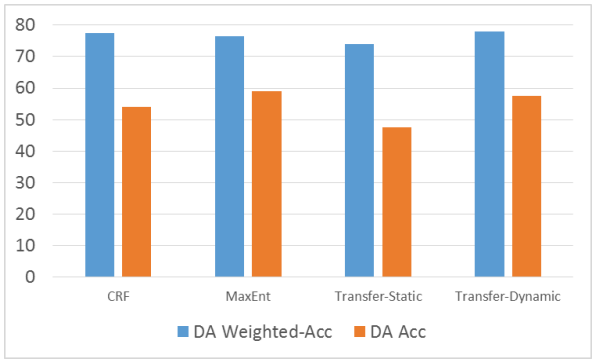

Figure 1. Classification Results for TADMUS

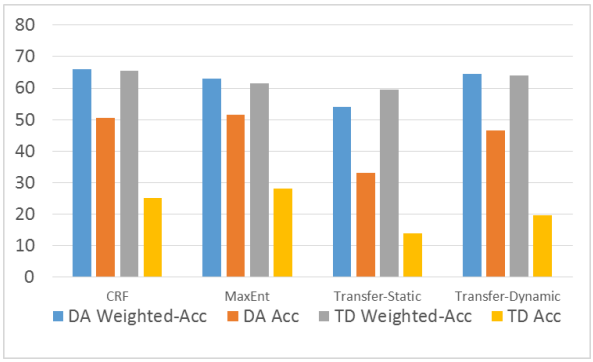

Figure 2. Classification Results for Squad Overmatch

For training our classifiers, we used the MLTextClassifier library from Apple[1], which can generate 4 types of models: a conditional random field model, a maximum entropy model, a static transfer learning model, and a dynamic transfer learning model. For dialogue act classification on TADMUS we used 10-fold cross-validation, and for dialogue act and TD classification on Squad Overmatch we used leave-one-out cross-validation.

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

Results for weighted accuracy (taking into account tag frequencies) and accuracy are shown in Figures 1 and 2 (for TADMUS and Squad Overmatch respectively). Our classifiers only assign one dialogue act per utterance (it is very rare that one utterance is annotated with more than one label) and we always assign a value ('no-value' is one of the labels). Thus accuracy, precision, recall, and F1-score are all equal. As expected, accuracy is higher for TADMUS given the larger amount of data. For TD, weighted accuracy is much higher than accuracy because it takes into account the distribution of TD tags, which as mentioned above is skewed. For the correlation experiments described below we used the outputs of the maximum entropy models and automatically annotated tags for TADMUS only. This is because classification performance on Squad Overmatch was not as good as performance on TADMUS.

Our results are promising, and as shown in the correlation experiments below, specific dialogue acts are strongly and significantly correlated with specific team performance scores. The higher the accuracy of our dialogue act classifiers, the more accurate these correlations will be. We have only annotated a relatively small portion of our data with dialogue act tags so there is certainly room for improvement. Another consideration for future work is using pre-trained large language models and investigating whether they can help with automatic tagging.

CORRELATION EXPERIMENTS

From each TADMUS 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, and number of occurrences of each dialogue act (e.g., num-request-info, num-request-confirm, num-ack, etc.).

Similar to Georgila et al. (2020) and Georgila (2022) we calculated correlations between scores and features. Table 5 shows correlations between scores and features for TADMUS using the feature set including the manually annotated dialogue acts and the automatically annotated dialogue acts, using 16 dialogues, all from the same scenario. Only statistically significant correlations are shown (***: $p<0.001$, **: $p<0.01$, *: $p<0.05$). There were no significant correlations for proper phraseology and provide/request backup/assistance.

From each Squad Overmatch dialogue we extracted the following features: 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, etc.), number of occurrences of each Team Development annotation type (e.g., num-provide-info-down, etc.), and number of occurrences of each Advanced Situational Awareness annotation type (e.g., num-inform-potential-threat, etc.). Table 6 shows correlations between TD and ASA scores and features for Squad Overmatch with the feature set including the manually annotated dialogue acts, the manually annotated TD tags, and the manually annotated ASA tags, using 4 dialogues. Again, only statistically significant correlations are shown (***: $p<0.001$, **: $p<0.01$, *: $p<0.05$).

Table 5. Pearson's Correlations between Performance Scores and Features (TADMUS, Manual and Automatic Dialogue Act Tags, 16 dialogues)

Score	Feature (Pearson's r)
Seeking sources	num-inform (manual -0.58*, auto -0.53*), num-ack-repeat (manual -0.56*)
Passing info	avg-num-words-per-turn (0.66**), num-request-permission (manual 0.53*), num-grant-permission (manual 0.60*), num-other (auto 0.53*)
Situation updates	num-speakers (0.64**), avg-num-words-per-turn (0.52*)
Complete reports	num-speakers (0.66**), num-suggest (manual 0.55*)
Brevity	num-request-confirm (manual 0.54*)
Clarity	num-speakers (0.55*)
Error correction	num-request-response (manual -0.71**)
Providing guidance	num-speakers (0.54*)
Stating priorities	num-speakers (0.64**), num-command (manual 0.54*)

Table 6. Pearson's Correlations between TD and ASA Scores and Manually Annotated Dialogue Act, TD, and ASA Tags (Squad Overmatch, 4 dialogues)

Score	Feature (Pearson's r)
TD	num-request-info (0.73*), num-request-confirm (0.93***), num-inform (0.95***), num-ack (0.83**), num-ack-repeat (0.97***), num-inform-potential-threat (0.95***), num-describe-potential-threat (0.73*), num-confirm-threat (0.82*), num-provide-info-up (0.71*), num-pass-info-up (0.85**), num-pass-info-lateral (0.96***), num-command-middle (0.82*), num-request-info-up (0.73*), num-request-info-down (0.79*)
ASA	num-request-confirm (0.95***), num-ack-repeat (0.88**), num-provide-info-down (0.75*), num-pass-info-lateral (0.77*), num-correction (-0.77*)

As mentioned above, our ultimate goal is to build an automatic pipeline for analyzing team behavior and predicting team performance. Preliminary experiments on automatic prediction of team performance have shown that the features depicted in Tables 5 and 6, which are strongly and significantly correlated with team performance scores, are indeed good predictors of team effectiveness, but again more annotated data is needed. For our correlation experiments on TADMUS we used data from only one scenario, while there are overall 4 scenarios. 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 used two corpora (TADMUS and Squad Overmatch) in which military trainees work together as a team to accomplish a task, and are scored by SMEs on a variety of indicators of team effectiveness, e.g., TD, ASA, situation updates, stating priorities, error correction, brevity, and clarity. We annotated part of the TADMUS and Squad Overmatch datasets with information about dialogue participation (addressees), content and meaning (dialogue acts), and dialogue structure (transactions). Also, we annotated Squad Overmatch with dialogue actions relevant to TD, e.g., providing information up and down the chain of command, and ASA, e.g., identifying and describing threats. We built machine learning models for automatic dialogue act labeling, and used both manually annotated and automatically extracted dialogue-related features to calculate correlations between indicators of team effectiveness and dialogue-related features. Our annotations show that requesting and providing information are strongly correlated with how teams were rated on TD and ASA, and identifying and describing threats is correlated with ratings on TD (but not ASA, probably due to data sparsity). Additionally, for each indicator of team effectiveness, there are some dialogue acts that exhibit strong correlation with that indicator.

We have shown that there is important information in team dialogue structure, which can be harvested via natural language processing and machine learning techniques, with the goal of calculating correlations between team performance scores and dialogue-related features. Then these correlations can be useful for training team performance prediction models, which is part of our planned future work. Using information extracted from natural language and dialogue structure to understand what communication aspects contribute to team success is an understudied topic, and our work is one of a few studies in this research area.

Our goal is to extend this work and ultimately 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. Of course, 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., 2019) and engage in multiparty interactions (Traum et al., 2008; Gu et al., 2021), ideally adopting behaviors of good human teammates, contributing to TD and ASA, and ultimately team success in a range of mission types.

ACKNOWLEDGMENTS

This work was supported by the U.S. Army under Cooperative Agreement Number W911NF-20-2-0053 and Contract Number W911NF-14-D-0005. Statements and opinions expressed and content included are those of the authors and do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. 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, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., & Traum, D. (2010). Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 2548–2555, Valletta, Malta.

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. (2019). 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, 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.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.

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*.

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 117–130.