Multidimensional Team Communication Modeling for Adaptive Team Training: A Hybrid Deep Learning and Graphical Modeling Framework

Wookhee Min[1], Randall Spain[1], Jason D. Saville[1], Bradford Mott[1], Keith Brawner[2], Joan Johnston[2], and James Lester[1]

[1] North Carolina State University, Raleigh, NC 27695, USA
{wmin, rdspain, jdsavill, bwmott, lester}@ncsu.edu

[2] U.S. Army Combat Capabilities Development Command, Orlando, FL 32826, USA
{ keith.w.brawner.civ, joan.h.johnston.civ}@ mail.mil

Abstract. Team communication modeling offers great potential for adaptive learning environments for team training. However, the complex dynamics of team communication pose significant challenges for team communication modeling. To address these challenges, we present a hybrid framework integrating deep learning and probabilistic graphical models that analyzes team communication utterances with respect to the intent of the utterance and the directional flow of communication within the team. The hybrid framework utilizes conditional random fields (CRFs) that use deep learning-based contextual, distributed language representations extracted from team members' utterances. An evaluation with communication data collected from six teams during a live training exercise indicate that linear-chain CRFs utilizing ELMo utterance embeddings (1) outperform both multi-task and single-task variants of stacked bidirectional long short-term memory networks using the same distributed representations of the utterances, (2) outperform a hybrid approach that uses non-contextual utterance representations for the dialogue classification tasks, and (3) demonstrate promising domain-transfer capabilities. The findings suggest that the hybrid multidimensional team communication analysis framework can accurately recognize speaker intent and model the directional flow of team communication to guide adaptivity in team training environments.

Keywords: Team Communication Analytics, Probabilistic Graphical Models, Deep Learning, Distributed Language Representations, Natural Language Processing.

1 Introduction

There is broad recognition that team training can improve team effectiveness across a wide range of domains [1]. It can improve team knowledge, team coordination, and team leadership behaviors, which can in turn minimize errors, enhance productivity, and help ensure teams are successful. Adaptive team training holds significant potential for providing effective learning experiences by delivering tailored remediation and feedback that support the development of teamwork and taskwork skills and dynamically address a team's training needs [2, 3].

A key challenge posed by team training is developing approaches to reliably assessing and diagnosing team processes in real time. Team training theory and research shows team communication provides a rich source of evidence about team processes that can support team training experiences [3-5]. Team members communicate with one another to develop a shared understanding of goals, tasks, and responsibilities [4], to coordinate actions [6], and to regulate social and cognitive processes associated with team performance [1, 7]. Accurate analyses of team communication can therefore provide deep insight into team cognition, collaboration, and coordination, which can ultimately be used to adaptively support team training needs.

Work on team communication modeling has explored a variety of methods. For instance, latent semantic analysis (LSA) has been used to devise team communication analysis models and assess team discourse using team communication content, sequence, and structure [8]. However, LSA does not adequately account for the dynamically changing dialogue context and semantics of the utterances that could be used for in-depth team discourse analysis. More recently, approaches based on deep neural networks [9] and probabilistic graphical models [10] have demonstrated significant potential for performing fine-grained dialogue analyses using multi-level language data (e.g., characters, words, paragraphs, documents), as well as other discourse and context features (e.g., dialogue sequence, turn taking, task sequences, environmental events). These techniques offer considerable promise for producing more accurate representations of team communication. Thus, a key question is how we can most effectively leverage these recent advances to accurately analyze team discourse, assess team communication, predict team performance and, ultimately, provide adaptive training experiences for learners.

In this paper, we present a hybrid, multidimensional team communication analysis framework supporting adaptive team training (Fig. 1). The framework leverages conditional random fields' structured prediction and deep neural networks' contextual language representation learning capabilities to classify team communication data with respect to the intent of utterances (i.e., speech acts [11]) and how information is conveyed to a team (i.e., team development categories). We investigate the hybrid team communication framework on transcripts of spoken utterances captured from six U.S. Army squads during a live capstone training exercise [12]. We evaluate the predictive performance of the hybrid framework optimized through cross-validation on a held-out test set and compare them to bidirectional long short-term memory networks, which are optimized through multiple configurations of multi-task learning and fusion methods, across the two classification tasks.

2 Related Work

Natural language processing techniques have been used in a wide range of learning analytics tasks to assess student knowledge and competencies, analyze student and teacher dialogue, and provide individualized feedback [13, 14]. Previous work has investigated automated essay scoring [15], short answer grading [16], discourse

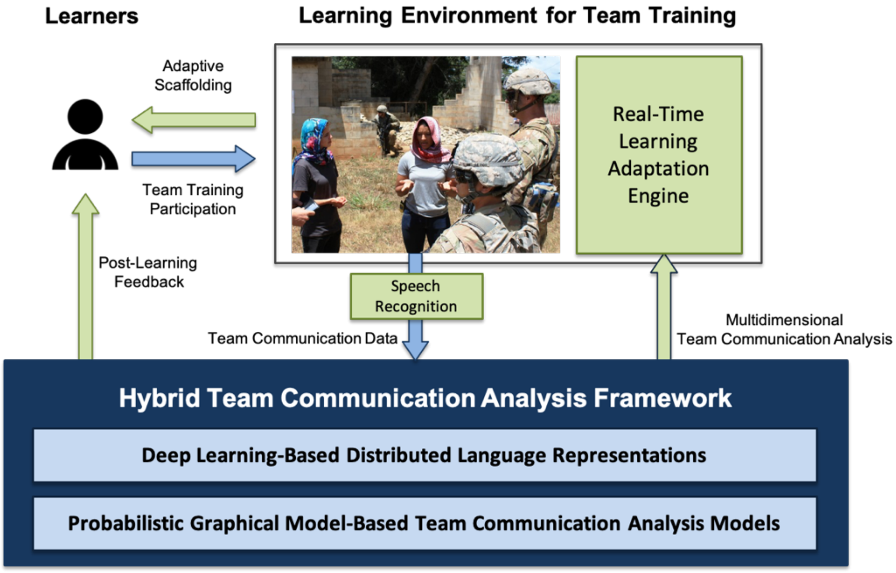
Learners
Learning Environment for Team Training
Adaptive Scaffolding
Team Training Participation
Real-Time Learning Adaptation Engine
Post-Learning Feedback
Speech Recognition
Team Communication Data
Multidimensional Team Communication Analysis
Hybrid Team Communication Analysis Framework
Deep Learning-Based Distributed Language Representations
Probabilistic Graphical Model-Based Team Communication Analysis Models

Fig. 1. Team communication analysis modeling for team training environments.

analysis in classrooms for both students [17] and teachers [18], text difficulty classification [19], and tutorial dialogues [20]. More recently, deep learning-based natural language processing has been explored for learning analytics tasks [e.g., 15, 16, 21], taking advantages of deep neural networks' capabilities on distributed linguistic representation learning [22, 23] and highly accurate modeling in an end-to-end trainable manner [24, 25]. Closely related to team training and performance, deep learning-based methods have been investigated for *computer-supported collaborative learning* (CSCL). In CSCL environments, group members work collaboratively towards a shared goal and solve problems as they learn [26], and deep neural network-based methods have been used in CSCL environments for detecting disruptive talk [27] and off-task behavior [28] with the goal of engaging in dialogues that are most conducive to learning.

While the majority of previous work on natural language processing in learning analytics has focused on tasks centered on individual learners, analyzing team dialogue could offer significant value to support adaptive team training experience and improve team performance. Team communication provides a window into how teams collaborate, coordinate, and distribute information in order to achieve a shared goal during team training and improve team performance [3, 29]. Consequently, many approaches have been investigated for analyzing team dialogue to obtain insight into teamwork, team performance, coordination processes, and training needs, including a growing body of work on computational approaches to team communication analysis [30]. For instance, LSA has been used to detect socio-cognitive roles in multiparty interactions [31] and team communication content analysis [8]. Researchers have also successfully utilized Markov models [32] and support vector machines utilizing multi-party dialogue embeddings [33] to analyze temporal patterns of team communication, as well as hierarchical regression models to investigate relationships between linguistic entrainment and team social outcomes [34].

Our work focuses on computational modeling of sequential communication patterns in actual team dialogue data collected from a set of live capstone training exercises. The hybrid, multidimensional team communication analysis framework shows considerable potential to support creating effective team training environments that adaptively facilitate teamwork and improve team performance.

3 Dataset

We investigate the hybrid team communication framework with transcribed audio logs captured from six U.S. Army squads as they each completed a 45-minute live training scenario (Fig. 1) [12]. The training scenario included a scripted set of training objectives and events (e.g., contacting key local leaders, providing combat casualty care) that were designed to elicit team development behaviors among squad members. Throughout the mission, squad members were required to develop a baseline of advanced situation awareness, identify and report tactical threats, and accomplish mission objectives. Each squad consisted of 10 team members wearing individual microphones, and each team member assumed a designated role and communicated with other key role players to collectively complete the mission.

The audio logs were transcribed and annotated using a coding scheme of 27 speech acts, 18 team development labels, and the speaker's role by domain experts, where speech acts represented the basic purpose of a given utterance, such as requesting information or stating an action being taken, team development labels reflected how different forms of information were being transferred up and down the chain of command in a squad, and speaker roles indicated the role of the team member speaking (six speaker roles including one squad leader and two sub-team leaders). While every utterance was assigned a speech act label, utterances were only assigned team development label when applicable.

Balancing the granularity of dialogue labels, their impact on the predictive accuracy of the models, and the potential utility of their predictions for training, we developed a mapping to reduce the number of speech acts from 27 down to 9 distinct labels consisting of ACKNOWLEDGEMENT, ACTION REQUEST, ACTION STATEMENT, COMMAND, ATTENTION, GREETING, PROVIDE INFORMATION, REQUEST INFORMATION, and OTHER statements. Team development communication behavior labels consisted of 19 labels (e.g., COMMAND COMING FROM THE SQUAD LEADER, PROVIDE INFORMATION UP THE CHAIN OF COMMAND, REQUEST INFORMATION FROM DOWN THE CHAIN OF COMMAND), including one extra label ("N/A") to account for the utterances whose team development labels are not applicable. Overall, the dataset included 4,315 tagged utterances made by the team members from the six squads (Table 1). Frequency analyses showed PROVIDE INFORMATION (*n* = 1,109) was the most prevalent speech act in the dataset, followed by COMMAND (*n* = 805). For team development labels, the most frequent labels were N/A (*n* = 1,978) followed by PROVIDE INFORMATION UP THE CHAIN OF COMMAND (*n* = 550) and COMMAND COMING FROM THE SQUAD LEADER (*n* = 362).

Pearson correlation analyses of the dataset found that squads who provided information statements (*r* = .862, *p* = .027) and issued acknowledgement statements (*r*

= .864, *p* = .027) more frequently received higher ratings of team performance during the training event [35]. Results also showed that ratings of team performance were positively correlated with the number of commands that squad leaders issued during the training event (*r* = .848, *p* = .033). Given the critical role communication plays in team effectiveness, being able to accurately classify team communication content in terms of speech act and team development labels could provide significant value for assessing team performance and developing adaptive training environments for teams.

Table 1. Example utterances and their speech act (SA) and team development (TD) labels.

Speaker	Example Utterances	SA	TD
Team Leader	Where are we moving?	REQUEST INFORMATION	REQUEST INFORMATION FROM UP THE CHAIN OF COMMAND
Team Leader	Hey, we're getting ready to move.	PROVIDE INFORMATION	PASS INFORMATION DOWN THE CHAIN OF COMMAND
Squad Leader	Six four be advised we're going to make contact with Romanov.	ACTION STATEMENT	PROVIDE INFORMATION UP THE CHAIN OF COMMAND
Squad Leader	Hey two alpha, hold right there at those trees.	COMMAND	COMMAND COMING FROM THE SQUAD LEADER

4 Multidimensional Team Communication Analysis Framework

We first devise linear-chain conditional random fields (CRFs) and deep neural network (DNN)-based predictive models that could classify team communication utterances into speech acts and team development labels. CRFs are discriminative models for structured prediction and sequence modeling [36]. CRFs utilize the probabilistic graphical modeling for multivariate data classifications and have been found to be particularly effective for modeling interdependencies in predictive features (e.g., pixels in an image, words in a sentence) along with the class labels associated with the features. While they have proven useful for a variety of tasks, recent work has produced significant advances by incorporating CRFs with deep learning techniques for dialogue act classification [37] and sentiment analysis [38]. These approaches suggest that higher-level features, such as team communication metrics, could be modeled accurately with CRFs.

To effectively model dialogue interactions that have a sequential structure, we investigate linear-chain CRFs (Fig. 2). As shown in Equation 1, the posterior probability of a sequence of classes (y) given a sequence of input feature vectors (x) from time 1 to T is computed using a weighted sum of K feature functions (f) that are parameterized with y at times t and t-1, and x at time t, where Z is an instance-specific

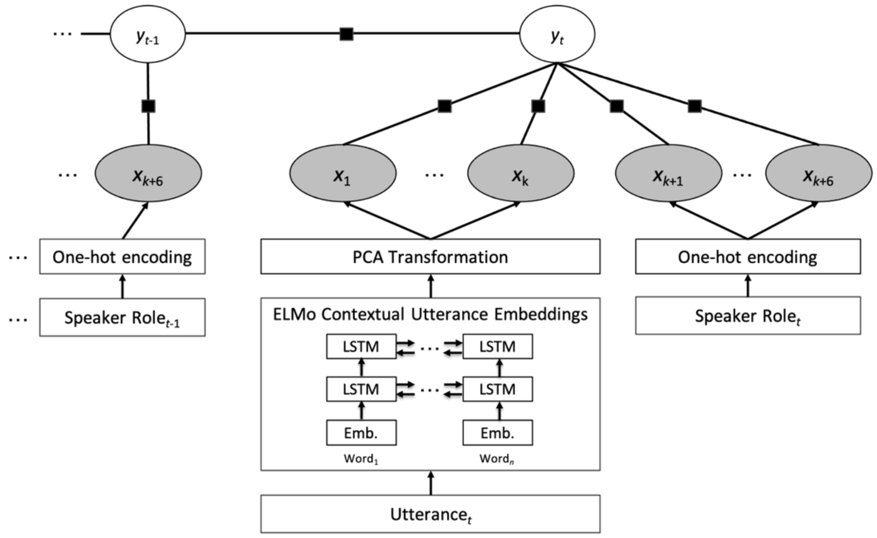

Fig. 2. A factor graph representation of a linear-chain CRF utilizing ELMo contextual utterance embeddings (CRF-ELMo). The gray shaded nodes denote input features (x) that are a concatenation of k-dimensional utterance features and 6 one-hot encoded speaker role features within a time step (t). The white nodes denote a target variable (y) such as speech act. The black shaded boxes indicate factor nodes.

normalization function [39]. To train the model, we sub-sampled sequences using a sliding window of length 100 (i.e., each subsequence with 100 utterances) from each team's communication data, considering both the context to capture from the dialogue sequence and potential data sparsity issues. In this way, we create a set of sub-sampled sequences equal to the number of utterances in each team's communication (for the sequences shorter than 100, we applied zero padding). The outputs y, which are the labels associated with the given input sequence, are generated for both training and testing. We use a block-coordinate Frank-Wolfe optimization technique [40] to train linear-chain CRFs with the maximum iteration number of 100.

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\} \quad (1)$$

To represent the speaker utterance, we employ a DNN-based contextual, distributed representation method using an ELMo language model [23]. In contrast to static, distributed representation methods such as GloVe [41], which provide fixed dictionary-based embeddings, contextual embedding approaches support inducing dynamic representations of text by utilizing a language model that takes as input a sequence of words. Consequently, ELMo-based approaches might be able to generate more accurate representations of utterances included in dialogues. In this work, we use a pre-trained ELMo model to generate utterance-level embeddings with 1,024-dimensional vectors through a mean pooling of all contextualized word representations. This ELMo model was built with stacked bidirectional LSTMs trained on the 1 Billion Word Benchmark, approximately 800M tokens of news crawl data from WMT 2011. Since the 1,024-dimensional vector representation per utterance is prohibitive for models to be effectively trained considering the dataset size examined in this work, we apply principal component analysis (PCA) to reduce the 1,024 dimensions down to one of 32, 64, or 128 dimensions, identifying the optimal reduced dimension through cross-validation. In summary, this hybrid model, referred to as *CRF-ELMo*, takes advantage of CRF's strong structure prediction capacity as well as ELMo's contextual language representation capability.

To identify the best performing CRF-ELMo, we examined two hyperparameters, including the regularization parameter from $\{0.1, 0.5, 1\}$ and the optimizer convergence tolerance from $\{0.01, 0.001\}$. We used PyStruct [42], a Python-based off-the-shelf CRF modeling library, to train the models, while the optimal set of hyperparameters is identified through a cross-validation process.

We also construct bidirectional long short-term memory networks (BLSTMs) [43], deep learning-based sequence model baselines, that use the same ELMo contextual language embeddings (*BLSTM-ELMo*). Specifically, we adopt a two-layer BLSTM architecture, as we anticipated both forward and backward propagations of hidden representations of the input streams would more effectively capture bidirectional, sequential patterns in the streams of speaker role changes and utterances and thus more accurately model dynamics characterized in team communication behaviors. A preliminary analysis conducted with the training set indicated that the stacked BLSTM architecture's speech act classification approach outperformed both single-layer standard LSTMs and two-layer standard LSTMs.

Multi-task neural models offer distinct advantages over single-task variants when performing multiple classification tasks [44]. First, multi-task neural models are more cost-effective for training than single-task models because they use one network architecture with multiple output layers accounting for different classification tasks instead of training multiple models. Second, when multiple tasks are correlated, multi-task models can potentially improve their generalization performance through effective regularization, especially when training data is limited. For this reason, we investigate both multi-task and single-task versions of BLSTM-ELMos in this work.

We also explore two fusion methods, early fusion and late fusion, in terms of the input feature sets (utterance-based feature set and speaker role-based feature set) for optimal BLSTM-ELMo modeling. For early fusion, the PCA-applied, ELMo representations of the speaker utterance and the corresponding speaker role passed through an embedding layer are concatenated into a vector, which is fed into the BLSTM layer. For late fusion, two BLSTMs are created to deal with two input feature sets separately, and the two BLSTM outputs are concatenated to perform classifications in a softmax layer. In both cases, we explore the same set of reduced dimensions (i.e., 32, 64, or 128) by PCA for the utterances as done in the CRF-ELMo.

For the speaker role, we use a trainable embedding layer with the embedding size of 4 to represent the speaker role in a continuous vector space. We use 32 hidden units for the two BLSTM layers with a dropout rate of 0.25 for regularization of the trained models, the softmax activation function for the output layers, and the Adam optimizer [45]. Similar to CRF-ELMo, we set the maximum input sequence length to 100 and the maximum training epochs to 100. Also, we train the models with the same sub-sampled sequential dialogue data and adopt early stopping with the patience duration of 10 epochs using the validation loss computed with 10% of the training set for effective training.

5 Evaluation

To evaluate the hybrid team communication framework, we split the team communication dataset into two sets: one contained data from 5 squads for performing cross-validation and the other data from 1 squad for held-out testing. First, we performed 5-fold cross-validation using data from 1 squad as a test set and data from 4 squads as a training set for each fold. The optimal set of model hyperparameters was identified through cross-validation by choosing the one that achieved the highest average cross-validation accuracy rate. It should be noted that the held-out test data was completely unseen from the cross-validation and its hyperparameter optimization process for fair generalization evaluation across models. The majority class baselines for the 9 speech acts and 19 team development labels were 25.7% and 45.8%, respectively. Table 2 shows the cross-validation results of the speech acts and team development labels. CRF-ELMo uses the format of hyperparameters, {*optimizer regularization parameter*, *optimizer convergence tolerance*, *reduced PCA dimensions*}, and BLSTM-ELMo uses the format of {*task modeling type*, *fusion mode*, *reduced PCA dimensions*}.

Table 2. Averaged cross-validation accuracy rates (%) for CRF-ELMo and BLSTM-ELMo. The highest predictive accuracy rates for speech act (SA) and team development labels (TD) per modeling technique are marked in bold.

CRF-ELMo	SA	TD	BLSTM-ELMo	SA	TD
{0.1, 0.001, 32}	68.80	58.38	{Multi-task, Early, 32}	61.44	**55.79**
{0.1, 0.001, 64}	67.88	56.16	{Multi-task, Early, 64}	**62.07**	53.88
{0.1, 0.001, 128}	64.85	52.67	{Multi-task, Early, 128}	61.97	53.43
{0.1, 0.01, 32}	**68.88**	58.31	{Multi-task, Late, 32}	60.22	53.96
{0.1, 0.01, 64}	67.87	56.20	{Multi-task, Late, 64}	60.19	54.51
{0.1, 0.01, 128}	64.87	52.62	{Multi-task, Late, 128}	59.77	53.22
{1.0, 0.001, 32}	68.76	**58.84**	{Single-task, Early, 32}	62.04	55.13
{1.0, 0.001, 64}	67.45	56.03	{Single-task, Early, 64}	61.16	54.47
{1.0, 0.001, 128}	65.72	53.36	{Single-task, Early, 128}	61.84	53.93
{1.0, 0.01, 32}	68.83	58.77	{Single-task, Late, 32}	61.48	52.64
{1.0, 0.01, 64}	67.45	55.93	{Single-task, Late, 64}	60.53	53.54
{1.0, 0.01, 128}	65.62	53.29	{Single-task, Late, 128}	61.96	50.98

Overall, the CRF-ELMo model achieved higher predictive accuracy compared to BLSTM-ELMo model based on cross-validation results. Both CRF-ELMo and BLSTM-ELMo generally showed higher accuracy when adopting the smallest number of language features (32 dimensions), which could be attributed to model overfitting issues. For BLSTM-ELMo, the early fusion method often outperformed late fusion. Further results showed multi-task learning and single-task learning were competitive, with the highest cross-validation results for both the classification tasks being attained by multi-task learning.

Next, we chose the best performing model hyperparameter configurations for the speech act and team development communication behavior predictions for the CRF-ELMo and BLSTM-ELMo models (marked in bold in Table 2), re-trained the models with the hyperparameters using all available training data (i.e., 5 squad training data), and evaluated the trained models' predictive performance using the held-out test set, which involved a separate squad's communication data. Table 3 reports model test performance across the re-trained models using the best performing hyperparameter configurations identified by cross-validation.

Table 3. Test accuracy rates (%) for best performing CRF-ELMo and BLSTM-ELMo models.

	SA	TD		SA	TD
CRF-ELMo	**69.42**	**64.92**	BLSTM-ELMo	64.61	61.88

The held-out test set-based evaluation results in Table 3 suggest that the hybrid CRF-ELMo approach outperformed the BLSTM-ELMo method with sizable differences for both the classification tasks, as seen in the cross-validation evaluation (Table 2). It is notable that the test accuracy rates are slightly higher than the average cross-validation accuracy rates. The five-fold cross-validation accuracy rates for the best performing CRF-ELMo models vary from 67.24% to 72.41% across the folds for speech act classification (average: 68.88%) and from 56.40% to 64.26% for team development classification (average: 58.84%), and these indicate that the held-out test set evaluation results are in a similar range. Both CRF-ELMo and BLSTM-ELMo models trained with the *entire training data* (i.e., 5 squad communication data rather than 4 in cross-validation) could help capture the test set data distribution thereby exhibiting high generalization performance.

We also trained alternating CRF models using a bag-of-words (BoW)-based static representation for utterances (CRF-BoW). To train the models, we first transformed all of the words that appeared in the training set to lower case and created a dictionary only using the top 80% of the most frequently observed words included in the training set, while treating the remaining 20% of the least commonly occurring words as *unseen* (a special token). This decision was made to effectively deal with an out-of-vocabulary problem (e.g., idiosyncratic words, typographical errors) in the test set. To create a BoW representation for each utterance, we created a vector with the dimension of 1,089, which is *the size of the dictionary* + 1 (the *unseen* special token), and set the word bit to 1 for the words included in the utterance, while setting the *unseen* special token bit to 1 for any undefined words. A CRF-BoW model is trained using the same model architecture used for the best performing CRF-ELMo. This CRF-BoW achieves 59.37% and 56.54% for speech act classification and team development, respectively, for the test set evaluation.

The results indicate that combining CRF's sequence modeling capabilities with ELMo, which uses a deep learning-based contextual, distributed utterance representation learning technique, achieves considerably higher predictive performance for both of the team communication modeling tasks. Together, these results suggest the following: (1) CRF can serve as a high-fidelity, sequence modeling technique for team communication, even with a corpus that is perhaps too small to effectively train LSTMs; and (2) the ELMo deep learning-based contextual language model trained with a large, general natural language dataset can effectively extract context and semantics from team dialogue and improve the predictive accuracy of the CRF models.

To build on these results, we next evaluated the team communication framework's *domain-transfer* capabilities. To facilitate this analysis, we explored how well the models trained with the mission data examined in this work (M_{org}) could classify squad communication that was collected during another training mission (M_{new}) [12]. Results showed that the best performing CRF-ELMo model trained with M_{org} achieved 67.35% predictive accuracy on speech act classification for utterances from M_{new}. These results show promise for developing scalable NLP-based models that can effectively transfer its predictive capacity to data collected from a related training exercise.

6 Conclusion

Adaptive team training is critical for effectively developing teamwork skills, facilitating team processes, and improving team performance. A key challenge posed by creating adaptive training environments is reliably analyzing team communication, which is a crucial source of evidence about team interaction. To address this challenge, we have introduced a hybrid, multidimensional team communication analysis framework incorporating CRF-ELMo, which integrates a high-fidelity, hybrid model that utilizes a probabilistic graphical model with a deep learning-based contextual language representation model. Evaluations conducted with cross-validation followed by a held-out test set showed that CRF-ELMo team communication analysis models achieved the highest predictive accuracy with respect to both speech acts and team development labels by effectively dealing with noisy team communication data captured from a live training exercise, and they significantly outperformed stacked, bidirectional long short-term memory network classifiers as well as majority class baselines. This hybrid approach was also found to have shown promising domain-transfer capabilities when applied to a different training event.

Future research in team communication analytics should investigate other contextual embedding approaches, model architectures, and model optimization and regularization techniques that can support generalizability and further improve the classification accuracy of team communication. Accurately classifying team communication utterances would allow team training researchers to identify if teams are pushing and pulling information at optimal rates and identify if critical pieces of information are being passed to relevant team members. In addition, future research should also conduct error analysis on misclassified instances and investigate the sequential patterns of team communication to facilitate team cognition and team performance. Finally, it will be important for future work to investigate the relationships between team communication and team performance and explore dialogue dynamics that can serve as key team performance indicators with the ultimate goal of creating adaptive team training environments.

Acknowledgements. The research was supported by the U.S. Army Research Laboratory under cooperative agreement #W912CG-19-2-0001. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army.

References

1. Salas, E., DiazGranados, D., Klein, C., Burke, C.S., Stagl, K.C., Goodwin, G.F., Halpin, S.M.: Does team training improve team performance? A meta-analysis. Human Factors, **50**(6), 903-933 (2008).
2. Johnston, J.H., Burke, C.S., Milham, L.A., Ross, W.M., Salas, E.: Challenges and propositions for developing effective team training with adaptive tutors. In: Johnston, J., Sottilare, R., Sinatra, A., Burke, C. (eds.) Building Intelligent Tutoring Systems for Teams, **19**, 75-97. Emerald Publishing Limited (2018).
3. Sottilare, R.A., Burke, C.S., Salas, E., Sinatra, A.M., Johnston, J.H., Gilbert, S.B.: Designing adaptive instruction for teams: A meta-analysis. International Journal of Artificial Intelligence in Education, **28**(2), 225-264 (2018).
4. Smith-Jentsch, K.A., Johnston, J.H., Payne, S.C.: Measuring team-related expertise in complex environments. In: Cannon-Bowers, J.A., Salas, E. (eds.). Making decisions under stress: Implications for individual and team training, 61-87. American Psychological Association (1998).
5. Rousseau, V., Aubé, C., Savoie, A.: Teamwork behaviors: A review and an integration of frameworks. Small Group Research, **37**(5), 540-570 (2006).
6. Marks, M.A., Mathieu, J.E., Zaccaro, S.J.: A temporally based framework and taxonomy of team processes. Academy of Management Review, **26**(3), 356-376 (2001).
7. Stout, R.J., Cannon-Bowers, J.A., Salas, E.: The role of shared mental models in developing team situational awareness: Implications for training. In E. Salas (ed.) Situational Awareness, 287-318. Routledge (2017).
8. Gorman, J.C., Foltz, P.W., Kiekel, P.A., Martin, M.J., Cooke, N.J.: Evaluation of Latent Semantic Analysis-based measures of team communications content. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, **47**(3), 424-428. CA: SAGE Publications (2003).
9. Deng, L., Liu, Y.: Deep learning in natural language processing. Springer (2018).
10. Yu, B., Fan, Z.: A comprehensive review of conditional random fields: variants, hybrids and applications. Artificial Intelligence Review, 1-45 (2019).
11. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, **26**(3), 339-373 (2000).
12. 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.: A team training field research study: extending a theory of team development. Frontiers in Psychology, **10**, 1480 (2019).
13. McNamara, D., Allen, L., Crossley, S., Dascalu, M., Perret, C.A.: Natural language processing and learning analytics. Handbook of Learning Analytics, 93-104 (2017).
14. Litman, D.: Natural language processing for enhancing teaching and learning. In: AAAI Conference on Artificial Intelligence, pp. 4170-4176. AAAI (2016).
15. Kumar, V.S., Boulanger, D.: Automated essay scoring and the deep learning black box: How are rubric scores determined? International Journal of Artificial Intelligence in Education, 1-47 (2020).
16. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: International Conference on Artificial Intelligence in Education, pp. 469-481. Springer, Cham (2019).
17. Clarke, S.N., Resnick, L.B., Rosé, C.P.: Discourse analytics for classroom learning. Learning Analytics in Education, 139 (2018).
18. Jensen, E., Dale, M., Donnelly, P.J., Stone, C., Kelly, S., Godley, A., D'Mello, S.K.: Toward automated feedback on teacher discourse to enhance teacher learning. In: 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-13. ACM (2020).

19. Balyan, R., McCarthy, K.S., McNamara, D.S.: Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. International Journal of Artificial Intelligence in Education, **30**(3), 337-370 (2020).
20. Katz, S., Albacete, P., Chounta, I.A., Jordan, P., McLaren, B.M., Zapata-Rivera, D.: Linking dialogue with student modelling to create an adaptive tutoring system for conceptual physics. International Journal of Artificial Intelligence in Education, 1-49 (2021).
21. Stone, C., Quirk, A., Gardener, M., Hutt, S., Duckworth, A.L., D'Mello, S.K.: Language as thought: Using natural language processing to model noncognitive traits that predict college success. In: International Conference on Learning Analytics & Knowledge, pp. 320-329. ACM (2019).
22. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
23. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018).
24. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science, **349**(6245), 261-266 (2015).
25. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, **13**(3), 55-75 (2018).
26. Sullivan, F.R., Keith, P.K.: Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions. British Journal of Educational Technology, **50**(6), 3047-3063 (2019).
27. Park, K., Sohn, H., Mott, B., Min, W., Saleh, A., Glazewski, K., Hmelo-Silver, C., Lester, J.: Detecting Disruptive Talk in Student Chat-Based Discussion within Collaborative Game-Based Learning Environments. In: International Learning Analytics and Knowledge Conference, pp. 405-415. ACM (2021).
28. Carpenter, D., Emerson, A., Mott, B.W., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., Lester, J.C.: Detecting off-task behavior from student dialogue in game-based collaborative learning. In: International Conference on Artificial Intelligence in Education, pp. 55-66. Springer, Cham (2020).
29. Marlow, S., Lacerenza, C., Paoletti, J., Burke, C., Salas, E.: Does team communication represent a one-size-fits-all approach? A meta-analysis of team communication and performance. Organizational Behavior and Human Decision Processes, **144**, 145-170 (2018).
30. Foltz, P.W.: Automating the assessment of team collaboration through communication analysis. Design recommendations for intelligent tutoring systems, **6**, 179-185 (2018).
31. Dowell, N.M., Nixon, T.M., Graesser, A.C.: Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. Behavior Research Methods, **51**(3), 1007-1041 (2019).
32. Ayala, D.F.M., Balasingam, B., McComb, S., Pattipati, K.R.: Markov modeling and analysis of team communication. IEEE Transactions on Systems, Man, and Cybernetics: Systems, **50**(4), 1230-1241 (2020).
33. Enayet, A., Sukthankar, G.: Analyzing Team Performance with Embeddings from Multiparty Dialogues. arXiv preprint arXiv:2101.09421 (2021).
34. Yu, M., Litman, D., Paletz, S.: Investigating the relationship between multi-party linguistic entrainment, team characteristics, and the perception of team social outcomes. In: International Florida Artificial Intelligence Research Society Conference, pp. 227-232. AAAI (2019).
35. Saville, J.D., Spain, R., Johnston, J., Lester, J.: Exploration of team communication behaviors from a live training event. To appear in the 12th International Conference on Applied Human Factors and Ergonomics (2021).
36. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

37. Kumar, H., Agarwal, A., Dasgupta, R., Joshi, S.: Dialogue act sequence labeling using hierarchical encoder with CRF. In: AAAI Conference on Artificial Intelligence, pp. 3440-3447 (2018).
38. Tran, T.U., Hoang, H.T.T., Huynh, H.X.: Bidirectional independently long short-term memory and conditional random field integrated model for aspect extraction in sentiment analysis. In: Frontiers in Intelligent Computing: Theory and Applications, pp. 131-140. Springer (2020).
39. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, **2**, 93-128 (2006).
40. Lacoste-Julien, S., Jaggi, M., Schmidt, M., Pletscher, P.: Block-coordinate Frank-Wolfe optimization for structural SVMs. In: International Conference on Machine Learning, pp. 53-61. PMLR (2013).
41. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543 (2014).
42. Müller, A.C., Behnke, S.: PyStruct: learning structured prediction in python. J. Mach. Learn. Res., **15**(1), 2055-2060 (2014).
43. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, **18**(5-6), 602-610 (2005).
44. Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017).
45. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).