Toward Automated Detection of Phase Changes in Team Collaboration

Julie L. Harrison (julieharrison@gatech.edu)[1], Sona A. Jain[2], Terri Dunbar[1], Jamie C. Gorman[1] & Sashank Varma[1,2]

[1]School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

[2]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Team science research heavily relies on communication data—that is, data derived from audio, video, or text-chat communication streams between team members. Between transcription and content analysis, significant overhead is required to work with these data. Recent developments in natural language processing (NLP) may help ameliorate time constraints in this domain. Using transcript data, the present study, presented as a proof-of-concept, assesses how the BERT NLP model performs in a team communication categorization task, in comparison to ground truth measures. This work builds upon past work that relied on human-coded transcripts to identify phase transitions in team collaboration. Results suggest BERT's capabilities at phase change detection are promising for experienced teams, though further iteration is needed on the methods in the current study. Applications of this work extend to real-time collaboration with an artificial agent, as this requires the real-time semantic processing of human communication data.

Keywords: NLP; BERT; teams; planning; collaborative problem solving; phase changes

Introduction

Increasingly, everyday tasks—whether they be at work, at home or at play—rely on teams of actors to complete. Methods are needed for studying team coordination that rely less on resource-intensive human assessment and more on automated assessment. Here, we present a proof-of-concept of a method for automatically detecting transitions in team problem solving using the natural language processing (NLP) tool BERT (Devlin et al., 2018). This method relies on the analysis of communication data and falls under the theoretical purview of Interactive Team Cognition (ITC; Cooke, et al., 2013), which is introduced below.

Historically, approaches to team cognition have posited that static knowledge structures exist in the minds of individual team members (e.g., shared mental models; Salas & Fiore, 2004). Team cognition then becomes the study of the overlap in these knowledge structures across team members. Under this approach, it is challenging to truly observe team cognition because it is reduced to latent and static cognitive products. In contrast, ITC posits that team cognition is: (1) an activity; (2) something only understood at the team level; and (3) embedded in the context of the task environment (Cooke et al., 2013). In this framework, team cognition becomes observable. Moreover, through direct observation of team cognition via measures such as team communication, we can assess how cognition unfolds over time. Cognition, then, is not a static product, but rather a dynamic state analyzable at the team level, while team members interact, where *time* is a central variable.

Team collaboration is the process by which parties with distinct information and roles interact to search for solutions to a problem (Gray, 1989). The focus of the present study is the identification of phase changes. Phases are conceptualized as "qualitatively different subperiods within a total continuous period of interaction in which a group proceeds from initiation to completion of a problem involving group decision" (Bales & Strodbeck, 1951, p. 485). Teams move through and revisit different phases to accomplish their collective tasks (Wiltshire et al., 2017). The shifts between phases can be, but are not always, sudden and are referred to as *phase transitions* (Kelso, 2009). There is a dearth of studies that identify these changes empirically; more work is needed to create bottom-up measures of team collaboration phase changes (Gorman et al., 2012).

The present study builds upon work by Wiltshire and colleagues (2017) applying the ITC approach to recognize collaborative problem solving (CPS) phase transitions in a dyadic task. They developed a method of recognizing phase transitions in team collaboration using task transcripts. Such CPS phases included *knowledge construction*, *team problem model*, *team consensus,* and *evaluation/revision*, and are understood to be important to successful teamwork (Fiore et al., 2010; Wiltshire et al., 2017). Qualitative content codes were applied to transcripts of teams completing a task by human raters. These annotations noted the content of utterances as they applied to the four different CPS phases listed above. Through a sliding window entropy measure, Wiltshire and colleagues identified peaks in communication *in*stability, represented by peaks in entropy, which indicated that the current collaboration phase is becoming unstable. These entropy peaks aligned with CPS phase transitions.

The present study adopts a similar approach. However, the goal is to develop a method of working with communication data for phase change detection and phase identification that does not rely on human raters. The innovation here is to apply the NLP model BERT (Devlin et al., 2018) to automatically (1) detect phase transition points and (2) identify the content within a given phase.

Our approach follows that of Gorman and colleagues (2016) who used an earlier model of word meaning, Latent Semantic Analysis (LSA), to analyze team communication during CPS. They computed successive cosine similarity ratings of utterances in a task transcript. This measure estimates the degree of *semantic* similarity between any two pieces of discourse. Their prediction, that semantic coherence

would break down at phase transitions was not supported. We borrow their approach of generating successive cosine similarity ratings, but instead using the more modern and much larger BERT model. We chose BERT because, unlike most other NLP models, it examines utterances bidirectionally (i.e., both from left to right and right to left). Thus, a higher level of contextual understanding is built (Devlin et al., 2018).

The first research goal is to use BERT to automatically detect phase transitions from semantic content. We expected it to outperform LSA in this regard. The second research goal is to use BERT to automatically detect the subgoal that is being discussed in each phase of the teams' transcripts.

The present study analyzes data collected as part of a study by Dunbar and Gorman (2020). Here, we analyze the transcripts of eight dyads completing the Non-combatant Evacuation Operation planning task (NEO task; Warner et al., 2003). Dyads were required to work together to plan a rescue mission of Red Cross workers trapped in a church on a remote island, while avoiding rebel forces and other environmental hazards on the island. The task required the development of plans around three subgoals: 1) getting rescuers to the shore of the island, 2) arriving at the church and 3) returning to safety. These are the three phases of team collaboration we sought to automatically identify using our methods. Task transcripts are particularly rich in semantic content pertaining to each of the three subgoals (i.e., phases), as each member of the dyad received different information pertaining to the evacuation plan. Thus, both individuals had to discuss the task in detail.

From our two research goals followed the following predictions and hypotheses. First, we predicted phase transitions would coincide with valleys of successive cosine similarity scores generated using BERT, as semantic content across phases should be less related than semantic content within a phase (Hypothesis 1; see Figure 1). We also predicted that BERT would be able to accurately categorize utterances as relevant to a given subgoal, thus automatically

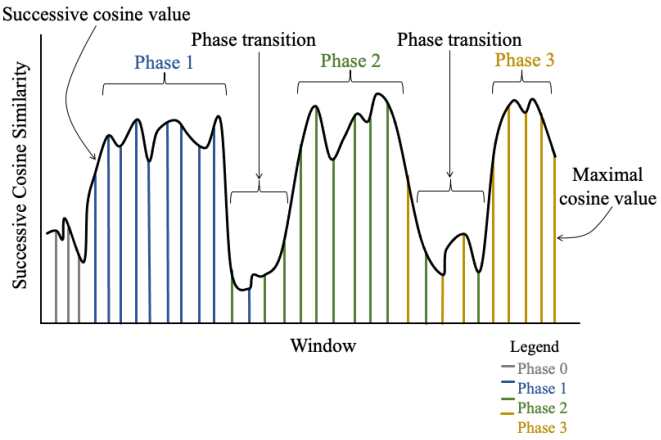

Figure 1: A visualization of our hypotheses, wherein phase transitions will correspond to low values of successive cosine similarity.

recognizing the purpose of each phase (Hypothesis 2). The benchmark against which we compared the BERT model was the coding of a human rater.

Method

Participants

Forty-six participants (23 dyads) were recruited by Dunbar and Gorman (2020) from the Georgia Tech psychology participant pool. The present study assessed transcripts from four of the six teams who completed the NEO task (numbered Teams 3, 12, 20 and 23). Dunbar and Gorman (2020) reported average participant age as $M = 19.70$ ($SD = 1.77$) and a 71.74% male sample. Participants were compensated with course credit. The protocol was approved by the local IRB.

Procedure

After obtaining informed consent, dyads had 15 minutes to familiarize themselves with the NEO task materials. They then had 15 minutes to collaborate with their interaction partner to complete the NEO task. They then completed the task a second time with a different interaction partner. In one task sessions, they worked with another novice participant to plan the rescue mission. In the other, they worked with an experimenter to plan the rescue mission. The experimenter-participant groups are considered more experienced teams in the present study, as the experimenter had performed the task numerous times and was familiar with the material given to both team members. Though both participants in the novice dyads completed the task with the experimenter, for the current analysis, we only assessed the transcripts of the two novice participants completing the task together and Participant 1 completing the task with the experimenter. This equated the sample size between the two types of dyads.

Members of the dyads worked in separate rooms, such that they only had verbal interactions over a push-to-talk microphone and headset channel. Participants and experimenters used computers with dual monitors to complete the NEO task. They wore EEG headsets as they completed the experimental task, which did not affect communication behaviors assessed in the present study. Upon completion of the task sessions, participants completed demographic questionnaires, were debriefed and then dismissed. The sessions lasted two hours.

Experimental Task

The NEO task was adapted from a three-person task developed by the Navy to a dyadic task by Dunbar and Gorman (2020). The mission was based on a hypothetical scenario that required the coordination of a rescue mission of Red Cross workers on a remote island. Subgoals of the task required participants to consider 1) which aircraft(s) and/or watercraft(s) they would use to reach shore; 2) how they would arrive to the church from the shore; and 3) how they would return the rescuers and Red Cross workers to safety within a 24-hour period.

While both teammates were apprised of the general requirements needed to rescue the workers within 24 hours and general information on the island, only one teammate received detailed information on weapons resources, while the other received information on intelligence resources and the island's environment. This interdependency between the teammates' roles required them to communicate to accomplish each subgoal. Participants were instructed to record the timeline of their hour-by-hour plans in a shared document.

Measures of Communication Content

A human rater first annotated the eight transcripts for teams' transitions into and out of the three subgoals. The human coding of the transcripts served as the ground-truth demarcation of CPS phase transitions in the NEO task. Below, we describe the criteria used by the human rater and BERT for detecting phase transitions.

Human Rater Coding Scheme A human rater reviewed the eight transcripts—four from novice participant teams and four from experimenter-participant teams. A nominal code associated with one of the three subgoals was ascribed to each relevant utterance in the transcripts. If an utterance pertained to other matter—experimental procedures, making notes in the shared document, or minor difficulties with the push-to-talk system—then a code of 0 was assigned and, as such, it was not ascribed to a particular subgoal. Short utterances (e.g., "yeah", "OK", etc.) were ascribed to the subgoal for which they were responding. In other words, the context of the surrounding utterances was taken into consideration when assigning the nominal codes. If an utterance pertained to more than one subgoal, the code most prominent in either the utterance and/or the context immediately surrounding it was chosen. This practice ensured each utterance applied to at most one subgoal.

Example utterances, one from each of the three subgoals, include, "*Um I think the helicopter will work, because we don't really need to carry more than fifteen people, because there are only three workers*" (subgoal 1); "*Yeah it says the local military has three trucks, bikes, and donkeys*" (subgoal 2); and "*Um, so on the way back we should probably not have them walking, right? So we maybe put them in the tank*" (subgoal 3).

BERT Phase Change Recognition Criteria We used the *nli-bert-large* model which, as the name suggests is pre-trained on the Natural Language Inference (NLI) data (Reimers, n.d.). NLI is used in supervised transfer tasks, and the utterance data is analyzed for how well subsequent ideas build on each other (Choi, 2021), making this BERT model particularly appropriate to quantify successive cosine similarity values. Devlin and colleagues' (2018) found that using a larger BERT model tends to be beneficial for analyzing task-specific text data. Because the present study assesses transcripts tied to a specific task, we employ a large BERT model here.

First, consider how we used BERT to identify phase transitions. Using the model described above, a pair of utterances from the task transcripts was passed in, encoded, and the cosine similarity between the two encoded utterances calculated. Here, similarity quantifies how well the utterances in one window build on the ideas in another window. A sliding window approach was used to assess utterance similarity. Using a window size of five utterances generates, on the first iteration, the utterance pairs of (1,2), (1,3), …, (1,5), (2,3), …, (4,5). For the second iteration, the pairs were defined over utterances 2-6, and so on. For each iteration, the similarities between all pairs of utterances were averaged to arrive at a single similarity measure for that window. These successive cosine similarity calculations were used to determine the points at which a phase transition might have occurred, wherein valleys of cosine similarity, recognized visually as local minima, were predicted to align with phase transitions.

We compared the effectiveness of window sizes of two, five, and ten (Figure 2). A window size of two produced graphs that made it more difficult to visually interpret phase transitions, as there are frequent sudden spikes and valleys. A window size of ten produced graphs that were smoothed too much, such that peaks and valleys were not prominent and could not be visually identified as phase transitions. A window size of five produced peaks and valleys that were still pronounced but occurred at a frequency aligned with that of the NEO task structure, and was therefore used.

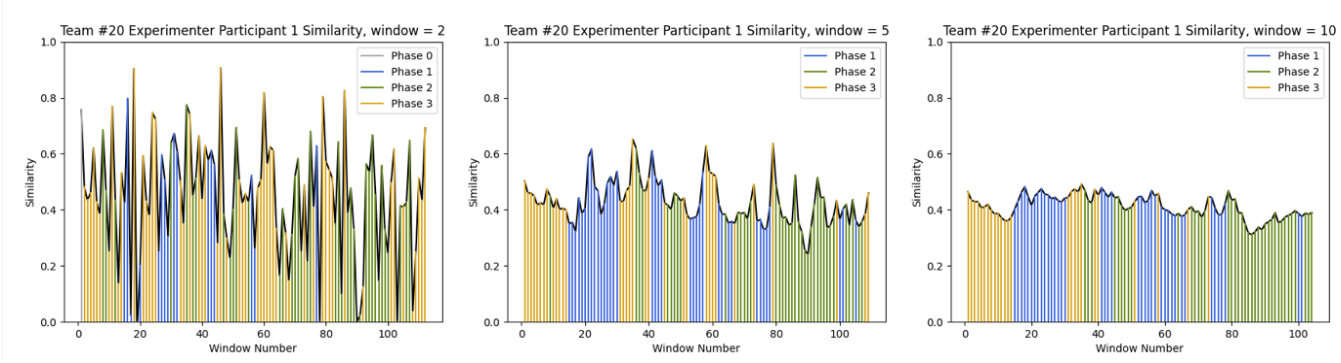

Figure 2: Example from a dyad in Team 20 demonstrating the effect of window size (2, 5, 10) on cosine similarity scores.

Next, consider how we used BERT to identify the phase of problem solving. The vertical lines below the black successive cosine similarity line, as seen in Figure 1, are color-coded to indicate the most likely phase for each window. This was determined by first combining the five utterances in a window into a single vector. The cosine similarity between this vector and a vector encoding the key words of the textual description of the first phase (i.e., subgoal) was then computed. The process was repeated for the second and third phases (i.e., subgoals). The maximal similarity value gives BERT's identification of the current phase.

Thus, the successive cosine similarity metric was calculated to designate where phase transitions occur, while the maximal cosine similarity values were calculated to determine what was being discussed within each window of neighboring utterances.

Lastly, we note that at the beginning of the task, some groups engaged in off-task conversation for a few utterances to get oriented to the directions and materials given to them. Phase 0 is associated with this off-task or task logistics conversation toward the beginning of the transcripts. Once BERT detected a phase other than Phase 0, the algorithm no longer compared the subsequent utterances against the Phase 0 bag of words.

Results

For this proof of concept, we randomly selected four teams to evaluate the hypotheses. Teams 3 and 20 were randomly selected to assess Hypothesis 1. Teams 12 and 23 were randomly selected to assess Hypothesis 2.

Hypothesis 1: Automatic Phase Change Detection

As Figure 1 depicts, phase transitions as detected by the BERT model were predicted to be marked by low successive cosine similarity values, as this would indicate prior utterances did not relate strongly to subsequent utterances, consistent with a shift in subgoal topic. Maximal cosine values were used to categorize utterances to a respective phase. They were predicted to be most variable at valleys of successive cosine similarity; in other words, at successive cosine similarity valleys categorization would be noisy. In Figure 3 we see the results of these metrics from Teams 3 and 20. The green circles in

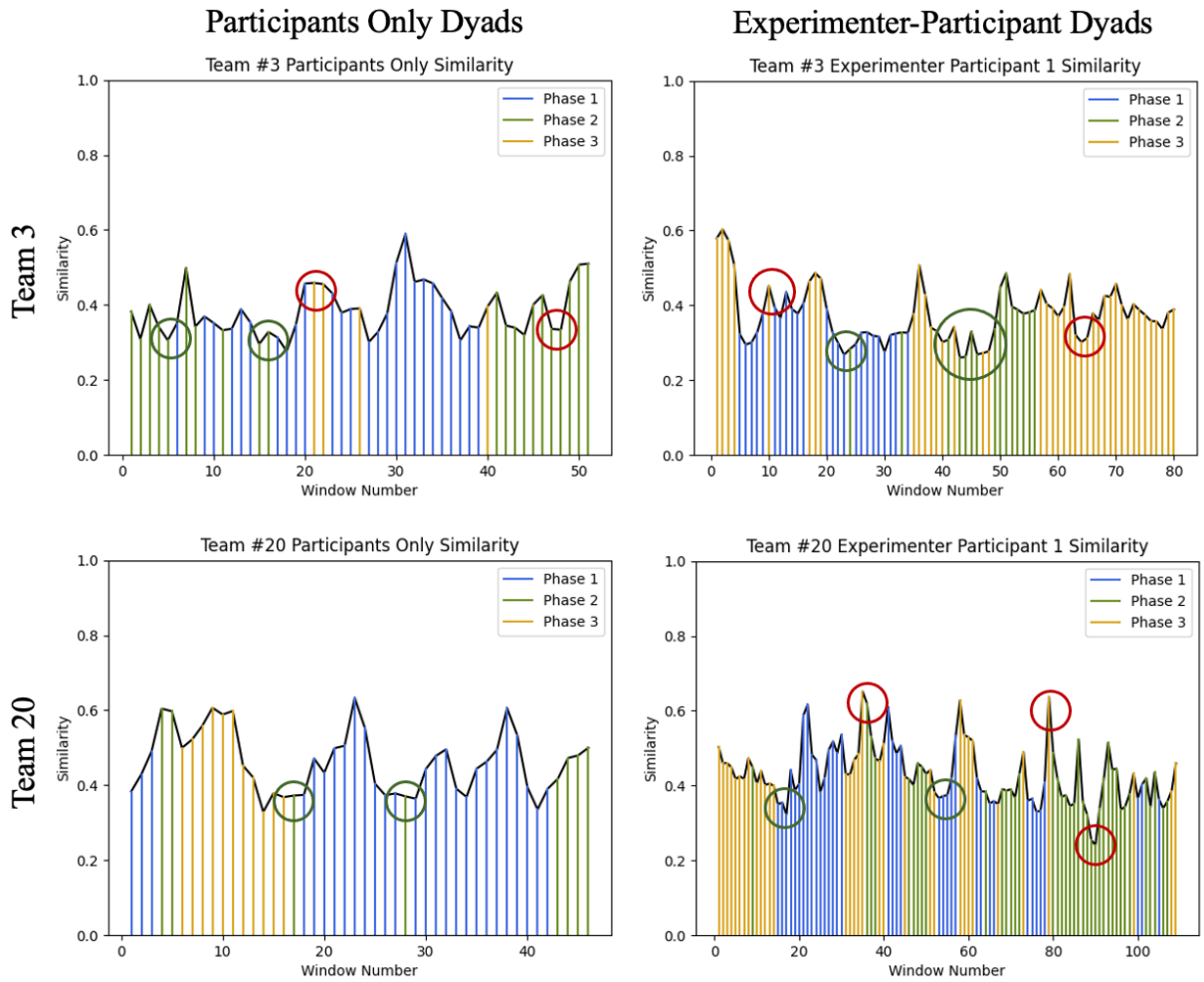

Figure 3: Successive and maximal cosine values plotted for Teams 3 and 20. Green circle indicate, results consistent with Hypothesis 1, whereas red circles indicate results in violation of this hypothesis.

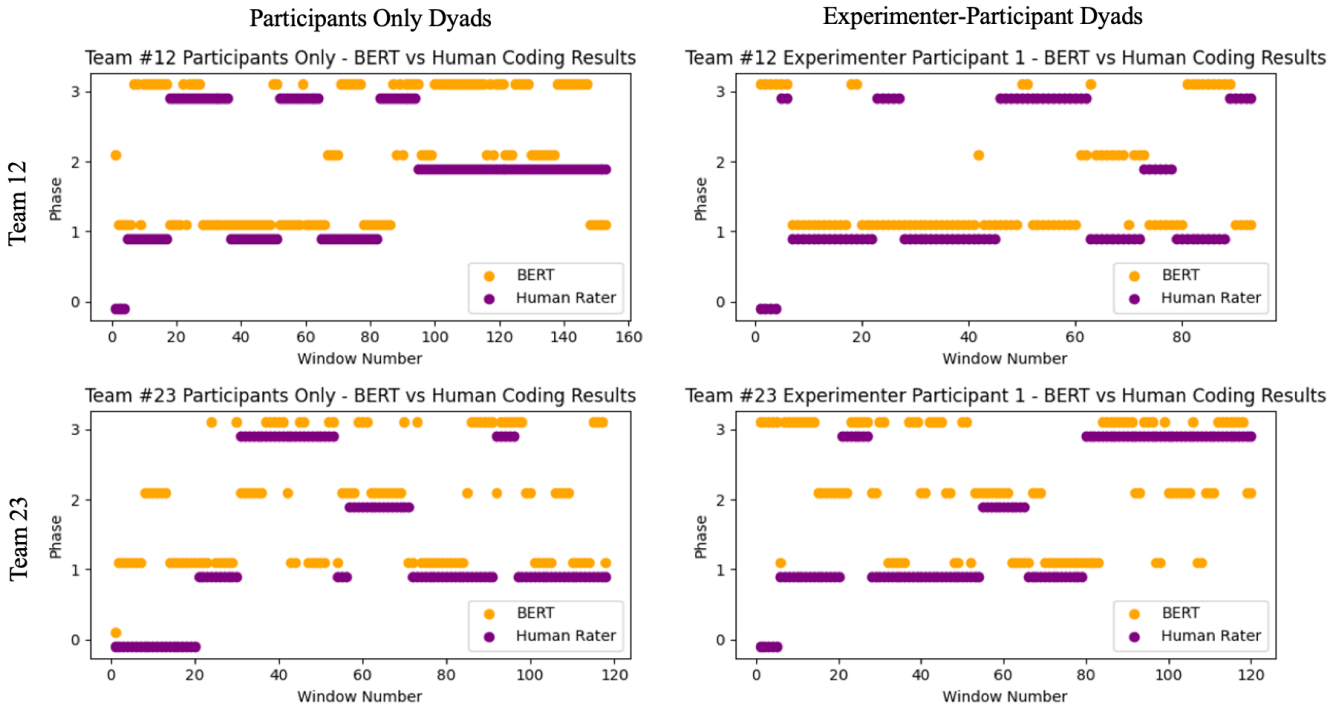

Figure 4: Phase identifications made by BERT and by the human rater.

this figure indicate results that are consistent with Hypothesis 1, whereas the red circles indicate results that are in violation of this hypothesis. There does not appear to be a systematic relationship between successive cosine similarity and phase transitions as detected by the BERT model. Phase transitions do not consistently occur at valleys—and only at valleys. Moreover, variability in the categorization of the current phase (i.e., subgoal) is present at both valleys, as predicted, but also at peaks. Thus, successive cosine similarity does not appear to capture phase transitions in the NEO planning task. The failure of the BERT model in detecting phase transitions parallels the failure of LSA observed in earlier work (Gorman et al., 2016).

Hypothesis 2: Automatic Phase Identification

To evaluate the understand how successful BERT is at categorizing utterances, we assessed its performance at phase transition detection and identification against that of a human rater. Hypothesis 2 predicted that the maximal cosine values used to assign each window of utterances to a specific phase (i.e., subgoal) would align with the human categorization. In Figure 4, we see that although BERT tends to be more "jittery" than the human rater, the two detect phase transitions at roughly the same locations, and assign roughly the same phase (i.e., subgoal) labels.

Overall, BERT had approximately a 40% success rate for the four teams, where success is defined as agreement with the human rater. The Participant-Only teams had rates of 33.99% and 47.46%, while the Experimenter-Participant teams had rates of 43.01% and 43.33% (for Teams 12 and 23, respectively). Moreover, BERT's agreement rate was relatively consistent across all four teams. The differences in these rates for the two Participant-Only teams may be driven by the preponderance of longer utterances in Team 23. With more words in the vector, BERT had more semantic content to consider when making its categorization. Many of the shorter utterances found in Team 12's dialogue did not contain semantic content, which may have contributed to noise in BERT's categorization.

Examination of the Experimenter-Participant dyads shows that Team 12 had a more organized flow of communication through the three subgoals. Team 23, on the other hand had disjointed dialogue. In particular, their dialogue did not involve an initial information share, and thus they frequently had to check in on resources available to them instead of planning how to use those resources. Despite these differences, BERT agreed with the human rater to the same degree. BERT may be more reliable at detecting phase changes in experienced teams as their conversations contain more semantic content that is relevant to task subgoals. This result supports findings from Gorman and colleagues (2016), wherein submarine crews' transcripts could be clustered by experience level using a cosine measure of semantic relatedness. Thus, BERT's detection of phase transitions may be more appropriate for experienced teams.

Discussion

Our hypotheses were not fully supported by the results of the present study. Nevertheless, automated detection of team collaboration phase transitions is important for the advancement of team science, and we believe the application of BERT and other NLP models is worth exploring further (Gorman et al., 2012; Wiltshire et al., 2017).

Hypothesis 1 predicted phase changes to be marked by valleys in the cosine similarity graphs. The results depicted a more complicated picture wherein phase changes occur at both valleys and peaks in successive utterance cosine similarity ratings, and, at times, randomly in the similarity distribution. Thus, we conclude that there is no systematic relationship between phase transition behavior and successive cosine similarity values. This finding is consistent with Gorman and colleagues' (2016) conclusion that successive cosine values generated by LSA did not identify task transitions.

Hypothesis 2 predicted BERT's maximal cosine similarity values would perform similarly to a human rater in identifying the phase (i.e., subgoal) of an utterance. This approach to phase transition detection was more successful. There was an approximately 40% success rate across the four team transcripts between BERT's categorizations of utterances and those of the human rater. Ideas for future iterations and improvement upon this approach are provided below.

Future Directions

Different NLP Models The current study used the nli-bert-large BERT model. It is possible that a different BERT model might have produced better results. We explored another model within this famil, stsb-bert-large, but the results were not promising. Future research should also explore other NLP models of word semantics such as word2vec (Mikolov et al., 2013a; 2013b) and Global Vectors (GloVe; Pennington et al., 2014), which have had some success in modeling cognitive science data.

Utilizing the BERT Toolbox We calculated (1) the average cosine similarity for automatic phase transition detection and (2) the maximal cosine similarity (for automatic phase identification. We did so within a range of content specified by a window size of five utterances. This windowed approach enables a role for context. However, a strength of BERT is that it enables the use of context directly. For example, BERT has a tool called Next Sentence Prediction (NSP) that enables it to predict what the next sentence/utterance will be based on prior context (Devlin et al. 2018). Assume that BERT takes on the role of both participants talking back and forth. Comparing the BERT vs. human utterances may give some insight into BERT's ability to capture how the two group members interact. These results could give more insight into the effectiveness of BERT as an analysis tool for investigating how teams work together as a team to accomplish some common goal.

Limitations

Ground Truth Ratings It is worth noting that this initial effort took the human rater's categorizations as ground truth. These categorizations may have been biased in this proof-of-concept study because the human rater was not blind to the task, and may have expected three distinct sequential phases, i.e., that the first *x* utterances belong to Phase 1, the next *y* utterances align with Phase 2, and the rest are assigned to Phase 3, with some Phase 0 assignments scattered intermittently. The present study employed only a single rater due to time constraints. However, the continuation of this work will ensure interrater agreement across raters who are blind to hypotheses. This limitation points to an interesting opportunity: Due to unavoidable bias and noise in human ratings—even with multiple raters—NLP methods should continue to be explored as a solution to transcript annotation, since they do not embody expectations of task structure.

Phases 1 & 3 are Semantically Similar Based on the BERT results, Phase 3 is frequently detected near the beginning of each transcript, though it was not discussed until the end of the task. Additionally, Phase 1 is occasionally detected near the end. This is because the descriptions of the two phases are semantically similar to each other, even if the associated subgoals are quite different. This ambiguity may be an inherent barrier in applying NLP models to automatic phase transition detection and phase (i.e., subgoal) identification. If so, future studies might explore the utility of ascribing utterances to multiple phases. This approach would perhaps enable the study of ambiguity in collaborative problem solving, or parallel pursuit of multiple goals, as a single utterance may contain information relevant to multiple parts of the solution.

Conclusion

Team science research continues to rely on human raters for transcript analysis. This practice can be costly, time intensive and unreliable. The further development of NLP tools and continued iteration on the methods described here may lead to more automatic and accurate semantic processing of task transcripts, lifting the current time constraints when processing communication data. Word embedding tools like BERT, in conjunction with automatic speech recognition (ASR) used for transcription (e.g., Dale et al., 2022), may reduce the time burden of working with communication data.

Additionally, real time team metrics of team collaboration are of increasing interest. With the development of automated semantic coding of communication data to the ASR methods currently employed by the field, it is possible to imagine the realization of real-time collaboration with an artificial agent. For example, Pugh and colleagues (2021) have employed real-time semantic analysis methods using ASR and BERT to identify CPS skills, with the envisioned application of providing immediate post-task feedback to students working on CPS tasks. Other current real-time communication processing methods rely on communication flow, i.e., who is speaking and when (Gorman et al., 2020). Iterations on the methods employed in the present study may extend to real-time analysis of communication content as well.

Acknowledgements

This research is supported by the NSF National AI Institute for Student-AI Teaming (iSAT; subaward 1559732 from NSF Grant DRL2019805). Any opinions, findings and

conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

References

Bales, R. F., & Strodtbeck, F. L. (1951). Phases in group problem-solving. *The Journal of Abnormal and Social Psychology*, 46(4), 485.

Choi, H., Kim, J., Joe, S., & Gwon, Y. (2021). Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. a*rXiv preprint at arXiv: 2101.10642*

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive science*, *37*(2), 255-285.

Dale, M. E., Godley, A. J., Capello, S. A., Donnelly, P. J., D'Mello, S. K., & Kelly, S. P. (2022). Toward the automated analysis of teacher talk in secondary ELA classrooms. *Teaching and Teacher Education*, *110*, 103584.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dunbar, T. A., & Gorman, J. C. (2020). Using communication to modulate neural synchronization in teams. *Frontiers in Human Neuroscience*, *14*.

Fiore, S. M., Smith-Jentsch, K. A., Salas, E., Warner, N., & Letsky, M. (2010). Towards an understanding of macrocognition in teams: developing and defining complex collaborative processes and products. *Theoretical Issues in Ergonomics Science*, *11*(4), 250-271.

Gorman, J. C., Cooke, N. J., Amazeen, P. G., & Fouse, S. (2012). Measuring patterns in team interaction sequences using a discrete recurrence approach. *Human Factors*, *54*(4), 503-517.

Gorman, J.C., Grimm, D.A., Stevens, R.H., Galloway, T., Willemsen-Dunlap, A.M., & Halpin, D.J. (2020). Measuring real-time team cognition during team training. *Human Factors, 62,* 825-860.

Gorman, J. C., Martin, M. J., Dunbar, T. A., Stevens, R. H., Galloway, T., Amazeen, P. G., & Likens, A. D. (2016). Cross-level effects between neurophysiology and communication during team training. *Human Factors, 58,* 181-199.

Gray, B. (1989). *Negotiations: Arenas for reconstructing meaning*. Unpublished working paper, Pennsylvania State University, Center for Research in Conflict and Negotiation, University Park, PA.

Kelso, J. A. (2009). Coordination dynamics. In R. A. Myers (Ed.), *Encyclopedia of complexity and systems science* (pp. 1537–1565). New York: Springer.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013), 1–12.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS, 1–9

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543

Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., & D'Mello, S. K. (2021). Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. *International Educational Data Mining Society*.

Reimers, Neil. (n.d.). *SentenceTransformers Pre-trained Models.* Google Sheets. https://docs.google.com/spreadsheets/d/14QplCdTCDwEmTqrn1LH4yrbKvdogK4oQvYO1K1aPR5M/edit#gid=0

Salas, E. E., & Fiore, S. M. (2004). *Team cognition: Understanding the factors that drive process and performance.* Washington, DC: American Psychological Association.

Wiltshire, T. J., Butner, J. E., & Fiore, S. M. (2017). Problem-solving phase transitions during team collaboration. *Cognitive science*, *42*(1), 129-167.