Intelligent After-Action Review Support Tools for Large Team Training

Sowmya Ramachandran and Randy Jensen

Stottler Henke Associates, San Mateo CA 94010, USA
sowmya@stottlerhenke.com

Abstract. The challenges of intelligent tutoring are greatly amplified when applied to team training. One avenue for near-term progress is to explore highly utilitarian solutions that automate certain instructional tasks, e.g. tools that support human instructors and extend their abilities without the features of a full ITS. Human instructors bring a wealth of expertise and experience that are difficult to replicate or replace. However, their attention can be overwhelmed by the volume of data generated during simulation-based training. Intelligent tools to digest the data into intelligible and actionable forms can greatly enhance the capabilities of instructors in a team setting. This paper describes two tools designed to support team training simulation exercises. They facilitate after-action review by analyzing the simulation interaction data for patterns and events that are interesting from the perspective of assessment and feedback. While instructors are still in charge of training, these tools provide important automated assessment support. Such tools can serve as stepping stones on the path to realizing intelligent tutoring capabilities for team training.

Keywords: simulation-based team training, after-action review, natural language processing

The challenges of intelligent tutoring are greatly amplified when applied to team training. The problem of inferring an individual's knowledge state, intentions, and motivations now transforms into one of inferring the hidden states of many individuals. The uncertainties are magnified because attribution must be spread across multiple participants, groups or sub-groups. The challenge of communicating with one grows into the challenge of communicating with many. Furthermore, there is an added requirement of being able to process and understand the communications amongst the trainees. Add to this the need for the tutor to assess not just task performance but also team performance, and the complexity explodes for developing a team training ITS.

Given these challenges, it is worthwhile to explore less automated, but highly utilitarian solutions, e.g. tools that support human instructors and amplify their abilities. While addressing practical needs, the solutions so developed can contribute to the ultimate objective of developing team ITSs.

Team training is an area that is very well-suited for the "Stupid Tutoring, Intelligent Humans" idea advocated in [1]. When humans conduct large team training exercises,

their experience and domain knowledge give them an edge in assessing performance and skill gaps, and in providing appropriate feedback. However, the amount of data generated in such simulation-based trainers can quickly overwhelm human instructors' attention and processing powers. Intelligent tools to digest the data into intelligible and actionable forms can greatly enhance the capabilities of instructors in the team setting.

This paper explores two case studies involving tools for supporting team training by automating or facilitating instructor after-action review (AAR) tasks: (i) a tool tailored for U.S. Marine Corps combined arms training and (ii) IDA (Intelligent Diagnostic Assistant). Both were designed to assist AAR by analyzing the interaction data generated during simulation-based exercises for patterns and events that are interesting from the perspective of assessment and feedback. In the combined arms training application, rules are used to detect significant events during an exercise and correlate these conditions with the communications amongst the team members. IDA approaches communication analysis from a different angle and focuses on filtering, organizing and visualizing communication streams to enhance human assessment and feedback.

Both tools were developed in response to stated user needs. As such, they represent steps toward addressing the kinds of practical problems that instructors face with team training and providing utilities to help augment their effectives. Intelligent solutions to analyze large volumes of data into actionable intelligence that instructors can use to inform their assessment and feedback are key steps on the path to building intelligent tutoring systems for team training.

1 AAR Tool for Combined Arms Team Training

Combined arms operations involve teams of teams coordinating maneuver with multiple supporting arms, including direct fire, indirect fire, and fixed- and rotary-wing aviation. Teamwork is required both within elements (e.g., within an artillery battery) and across elements (e.g., between an artillery battery and forward observers, or also between the artillery battery and commanders). Task-specific skills like adjusting artillery fire onto a target are a part of the training. But a key objective is to practice team skills, like the successful employment of indirect fire through coordination between observers, approval authorities, and liaisons with air and other elements of the battle. Training exercises may involve 100 or more participants at various stations, carrying out their respective operational responsibilities.

An intelligent AAR tool was developed for USMC combined arms training, which aims to help instructors not only to identify examples of good or bad performance from the exercise, but also to put together an effective debriefing that can convey information about the context of a training point and key participants involved. The data sources available for automated analysis include the simulation event stream, voice communications, and also human-in-the-loop inputs in command and control (C2) tools used in the exercise environment. For example, the data artifacts associated with an artillery mission include radio communications during planning, C2 tool inputs specifying the timing and parameters for events to be triggered in the simulation, communications

again for final approval before execution, and finally the simulation events when executed. In addition to the collection of raw data from the different sources, the system performs automated speech recognition on radio communications, to diagnose causal factors for possible errors. The combination of multiple data streams for automated analysis allows the system to identify training points related to individual task-related decisions and skills, team performance factors, and instances involving both.

1.1 Applying Team Behavioral Markers to Combined Arms Training

The process of constructing team performance measures for the combined arms domain involved an initial practical step of identifying the behavioral markers that are possible to collect and reason about from the available data streams, overlaid with consideration of which dimensions are most important for the domain in terms of their relationships to training objectives. The study of teamwork competencies has evolved over time, producing a variety of different models with common features, collectively providing a theoretical basis for team diagnostic measures [2, 3, 4]. In a recent meta-analysis [5], findings from key publications were combined to suggest behavioral markers to be used for intelligent team tutoring, organized into five factors: trust, collective efficacy, cohesion, communication, and conflict management. For example, markers for collective efficacy include a combination of actions like backup behaviors, and communication artifacts like affirmative comments about the team's ability to complete tasks.

In addition to recent efforts to formalize the implementation of team tutors with concrete domains (e.g., [6]), a significant development is the availability of increasingly sophisticated natural language processing technology that can be used to mine communications data for instances of behavioral markers. For some applications, challenges remain where certain team interactions take place outside of the instrumented environment (e.g., an exercise environment may be constructed with digital communications infrastructure, but team members may still occasionally call out across the room or tap someone on the shoulder), or where subtleties embedded in interactions (e.g., inflection, voice level) are simply not among the collected data. For many markers the best available methods for collection still involve the use of human observers applying subjective interpretations. Yet a growing focus of research is to identify areas where automated systems can be used to assess elements of team performance in a concrete manner from available exercise data streams, and help offload instructors.

In combined arms, some of the most critical behavioral markers relating to effective team operations involve communications and especially information sharing (timely, complete, clear, concise, acknowledged). For example, during a coordinated assault, maneuver units should periodically report their position, roughly every 500 meters of movement. In the training environment, the tactical networks can be easily monitored for verbal position reports, which can be compared against the actual positions of the units in the simulation for accuracy and frequency. Similarly scout teams should be reporting the location and movement of enemy units within their line of sight. A dedicated surveillance radio network can be monitored for reports from the scout teams,

and compared against ground truth from the simulation.

Another example is the Call For Fire (CFF), which initiates the events that ultimately cause a fire series to be executed. There are at least three key parties directly involved in a CFF – the requester, the deciding authority, and the battery that would fire the series. But any approval of requests for missions and movement should be disseminated beyond the directly involved parties; for example, the maneuver companies need to be aware of the upcoming fire missions in their area because it supports their tactical activities and may also create new danger areas. Whether a CFF is approved, denied, or approved with modified parameters, this is important information to communicate.

In addition to *what* is communicated, a system can look for markers related to *how* content is communicated within or across teams. In combined arms operations there is a very specific vocabulary and syntax that should be used. The spirit of this from an operational teamwork perspective is to ensure that other team members can understand the content, especially in noisy environments with the potential for distractions and degraded signal. This also means that software tools can be constructed to look for verbal utterances that conform to the vocabulary and syntax. For example, in combined arms communications, mission approval and denial should be clearly indicated with the phrases "is approved" and "is denied", respectively. In particular, the phrase "not approved" should not be used because of the potential confusion with “is approved,” and such negative examples can be easily detected with automated software.

Backup behaviors relating to the factors of trust and collective efficacy also play a significant role in combined arms team performance, especially in the avoidance of battlespace conflicts. The most important example arises when there is an incorrect clearance of a fire mission or air strike which violates safety constraints for a friendly unit. When conflicts arise, often they can be attributed to an error that a team member could have corrected. One simple example relates to the fact that artillery calculations are done in metric units, but altitude guidelines are conventionally reported to the relevant aircraft in feet. Other examples can often be traced to a change in the timing or location of the mission or friendly unit involved in the conflict, where such a change may not have been noted by a particular team member. In such cases, typically other team members monitoring the same communications networks should be aware of the correct or most recent information, at least in approximate terms. If the aircraft “stay above” for an artillery mission is conveyed with a significant unit conversion error, this should be recognized by other team members. And fire missions approved with timings that haven’t been updated should also draw attention from other team members. Automated measures can be implemented with triggers that start from the detection of initial errors (e.g., an approval of a mission that leads to battlespace geometry conflicts), and then look for any backup behaviors among the team.

1.2 Combined Arms Example Scenario

This example walks through a vignette from a combined arms training event, where automated assessment mechanisms detect behavioral markers from the data streams,

resulting in feedback produced in after action review. This was implemented as part of an operational demonstration prototype, using a training scenario based on those used for existing training. The scenario contains examples of both individual and team performance errors that occur in the course of a combined arms offensive operation, where a Close Air Support (CAS) attack against red force tanks is planned, as blue forces on the ground move to contact. Participants occupy roles in either the Fire Support Coordination Center (FSCC) or the distributed elements in communications with the FSCC. Key FSCC decision-makers are responsible for oversight of ground movements, indirect fire suppression missions, and the CAS mission against the target.

Early in the scenario, a tank platoon begins a movement toward the enemy target, following the convention of giving position reports at every 500 meters as they approach their objective position. According to the plan, the tanks are to move to a position just outside the danger area for the planned CAS attack on the target and then halt and report position. Instead, the tank platoon leader makes the error of moving his unit past the halt position and into the danger area, and also fails to send position reports at either the planned position just outside of this area, or the current halted position inside the danger area. This results in a conflict later when the CAS mission proceeds and the tanks are inside the danger area at the time of CAS ordnance detonation.

In this sequence of events, the tank platoon leader's error is compounded by a team error in that the FSCC staff failed to anticipate the tank platoon's position and request a position report when none had come. This is a reasonable and common form of backup behavior, as the maneuver units are expected to follow patterns established as part of a coordinated operation. Relevant data include simulation data such as tank platoon positions over time and CAS attack and detonation events; communications data such as tank platoon position reports (and lack of elicited reports) and CAS final approval; and C2 tool injections such as the CAS attack triggering to simulation events.

The battlespace geometry conflict (tanks inside the active danger area from the CAS mission) is detected by assessment mechanisms in two forms. First, it is identified in a predictive manner when the CAS attack is approved and input with C2 tools. In an analysis of the attack as entered (and before it takes place in the simulation), the danger area associated with the attack ordnance and the current position of the tank platoon already constitute a predicted conflict. Later when the CAS attack takes place, the conflict becomes realized as an actual event in the simulation. This in turn triggers analytics that look for markers relating to individual and team errors, especially from communications collected on the tactical networks. The implementation uses a combination of speech recognition methods such as keyword spotting with a simple grammar to characterize radio transmissions.

In this case, the analytics need to identify the absence of certain communications. From a team performance metrics standpoint, there is a communication deficiency in terms of information sharing when the position report is not proactively provided. There is also a deficiency in terms of collective efficacy and trust when no markers can be found for backup behavior from the FSCC staff, in failing to ask for the position report. Furthermore, within the FSCC itself, there are several staff roles beyond the individual

who interacts with the tank platoon leader, who also should have been aware of the situation and also expecting confirmation of ground unit positions before giving the final clearance to aircraft on dropping ordnance. Any of them could have initiated prompting for the request of an updated position report.

Once concrete markers are identified for team performance factors in specific exercise events, the system must convey to the team and its individuals what problems occurred and how teamwork needs to be improved. In this domain, teamwork is primarily carried out through radio communications, so these transmissions are very important and therefore prominent in the debrief. Figure 1 below shows the automatically generated information for a debrief training point, with time-synchronized timelines for communications and simulation events.

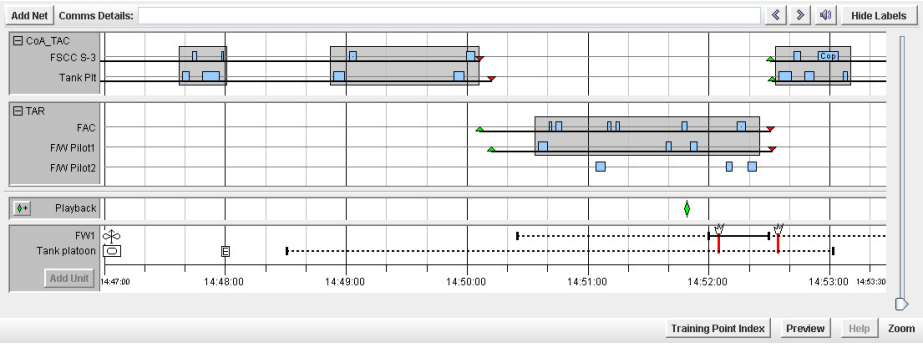

Fig. 1. Training point timeline display

In this case, two communications timelines are relevant: position report communications on the CoA_TAC net, and also air communications on the TAR net. These combine visual information about transmissions and dialogs that took place, along with speech recognition results for their content. Simulation events are depicted in the bottom timeline, with domain-standard symbology for the CAS mission (FW1) and the tank platoon maneuver, and markings for conflicts. These timelines are provided as supplementary information for the debrief, to accompany text generated for situations like a conflict, and any individual or team performance markers of note. Training points are automatically constructed with pre-configured 3D battlespace playback using logged simulation data, and then the human instructors select information to include from the timelines, such as key communications associated with behavioral markers (e.g., the last tank platoon position report, or the CAS final approval).

Similar mappings are constructed for other combined arms related conditions to be identified and debriefed. While it is a detailed process to define the rules for these measures, and also to implement the mechanisms that can analyze exercise data from multiple streams to identify behavioral markers for team performance, the benefits come in offloading the burden on instructors. Particularly in large team training events,

the operational tempo is such that instructors need to meet a rapid turnaround in preparing AAR debriefing materials, and intelligent tools facilitate this data-rich task.

2 Chat Analysis Tool for Large Team Training Exercises

Intelligent Diagnostic Assistant (IDA) is an AAR tool to enhance after-action review via visualization and analysis of inter-team communications during training exercises. Chat-based communications are becoming convenient and common channels for teams. As such, their role will become increasingly central in large military operations. In anticipation of this development, the US Air Force was interested in investigating the role of chat-based communications in air operations exercises and their impact on performance assessment and feedback. IDA is primarily a tool to help team trainers visualize chat streams from multiple perspectives so they can perform an informed and detailed assessment of team performance. Its analysis capabilities are in support of visualization.

The research approach used in determining how to analyze and display information follows the operational planning methodology laid out in joint and USAF doctrine. The initiator for planning is normally a problem statement in the form of intelligence data or operational data reported to the team. The initiating report typically establishes a segregated planning approach to address the problem. The team then examines the problem in sequence with other planning tasks or a sub-team may be tasked to examine the issue in parallel with other team activities. In many cases, planning may be interrupted and take on an interleaved character. When a training session ends, trainees need to be able to see each problem in isolation, as well as in context with other workload. The isolation approach allows the team to review actual process versus doctrine, while the context of workload offers insight into time delays, distractions, errant information sources, and overall cognitive effort. After action review tools must help an instructor to sort and associate information with a unique process and be able to display information cogently to identify key areas that positively or negatively affected team and individual performance. Where chat logs are one of the primary sources of data indicating performance, tools for reviewing multiple chat logs in tandem become critical.

Based on a requirement analysis, we determined that classification of chat data according to missions (or processes) is an important capability for IDA. The objective of our research was to explore the extent to which this data can be analyzed to extract useful information without a deep semantic understanding of the messages. We focused on the use of statistical and rule-based techniques that analyze messages based on surface features such as word occurrences and correlations. The goal was to de-clutter the communication streams so that instructors can focus on meaningful threads to assess and discuss during AAR.

IDA supports two primary activities: (i) association and filtering, and (ii) visualization and browsing.

2.1 Association and Filtering

The objective of the associative mapping is to identify topics on the same thread, where a thread is defined as a mission. IDA first starts out with an untagged set of chat messages sorted in a chronological order. It incrementally tags the messages with associated missions using a combination of heuristic and machine learning approaches. Multiple passes are made over the chat data to successively refine the associations. It is possible for a message to be associated with multiple missions. IDA performs the following two types of analyses to recognize associations. First it performs keyword-based associations over two passes through the data. An important feature of this domain is that each process or mission is associated with a specific identifier that is sometimes referenced in chat utterances. While only a small fraction of the chat messages includes such mentions, these messages nonetheless form seed data from which other associations may be gleaned. Once this seed set has been identified, IDA uses supervised machine learning techniques to learn classifiers to map chat utterances to a mission. The second pass through the data is then performed to classify the remaining messages using these classifiers. This still typically leaves a number of unclassified messages. IDA uses temporal pattern-matching to classify these. First, a turn-by-turn interaction between two people in the same chat room within a time window (e.g. A says something to B and 3 minutes later B says something to A) is associated with a common mission. Making an assumption of dialog coherence, it is likely that such conversation dyads refer to the same topic thread. Finally, the remaining messages are clustered according to the distribution of tags in the neighborhood of each message. A window around each message is analyzed, and the message is tagged with the most commonly tagged mission.

2.2 Visualization and Browsing

Even with a filtered set of chat data, it is still a time-consuming task to review synchronous conversation streams in multiple chat rooms and develop an understanding of the overall flow to identify performance indicators. This is the motivation for a tailored browsing capability that an instructor can use to review process-specific communications and visualize chronological relationships cross-referenced with exercise states. Typically, communications regarding a particular process will flow across multiple chat rooms, so synchronous browsing is a key feature. Additionally, the results of associations and filtering can be reflected in the browsing environment as cues during the review process. For example, keywords detected by the supervised learning algorithm will often be of interest to an instructor as highlighted terms while browsing.

Figure 2 shows the primary visualization view implemented with the prototype. IDA enables simultaneous, synchronous browsing of multiple chat rooms, while preserving chronology, thereby making it possible to follow the communications across time and across chat rooms. With synchronous scrolling, the user can browse through the chat data exactly as it unfolded in the exercise. IDA has simple rules for automatically configuring the windows based on the phase of the exercise, the mission under consideration, and the density of chat traffic in the rooms.

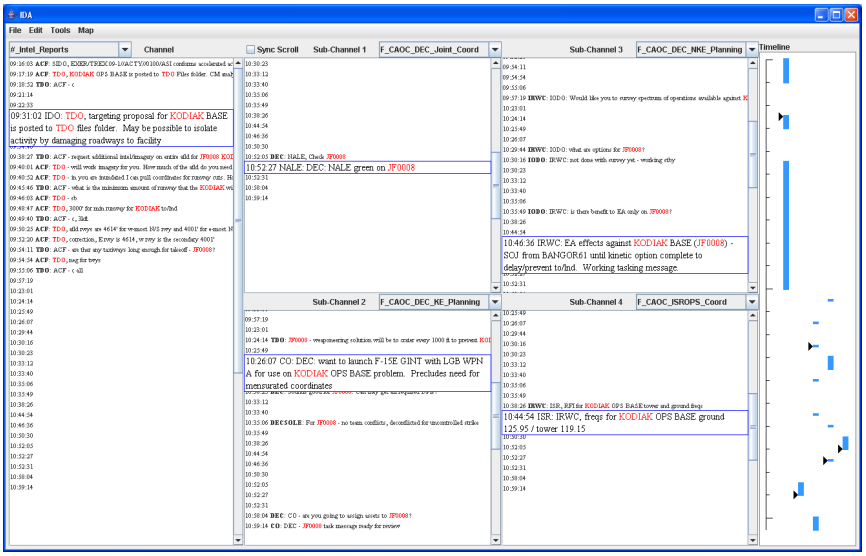

Fig. 2. The IDA visualization tool

The IDA chat visualizer provides information at three levels of detail. A timeline, shown along the right side of Figure 2, provides a birds-eye view of the distribution of communication in the selected chat rooms, without providing the details of communications. Its purpose is to establish the overall temporal context in a manner consistent with other tools utilized on adjacent screens during the after-action review process. Individual channels are represented in the timeline, with independent markers to show the current temporal location of selected chat lines in each channel. The chat channels provide the next level of detail, each presented in a dedicated panel for a particular chat room. Each chat line for a channel is shown with the time stamp, the sender, and the first line of the message content as a scannable summary. Channels can be scrolled independently or synchronously with chat lines aligned by time. A movable magnifying lens within the channel display provides the third level of detail. It shows the entire contents of a selected chat line in larger text, using multiple lines if necessary.

The combination of analysis, filtering and visualization is designed to facilitate rapid assessments of team performance markers by instructors. The birds-eye view of the communications about a process/mission allows instructors to assess whether information is flowing through the right channels at the expected times. Filtering and synchronous browsing allows looking for behavioral markers such closed-loop communications. Instructors can also see patterns and tempo of communications during different phases of a process to inform their AAR. Within each topic thread, instructors can observe how critical keywords (e.g. “approved”, “denied”) are propagated through different parts of the team.

3 Conclusions

The two systems discussed represent different points on the spectrum of intelligent automation in support of team training. Whereas IDA aimed provide a tool to help instructor-derived assessments and feedback, the AAR tool for the combined arms training exercises went a step further by automatically assessing mission events, correlating them with team communications, and tying the analysis to team performance behavioral markers. The users of these solutions were not seeking a fully automated tutor to imitate the capabilities of their instructors. What they desired were intelligent tools to serve as the "eyes and the ears" of the instructors, amplifying their capacities to process the data from training exercises and construct tailored feedback. Incremental development of such intelligent training support tools is one promising path towards ultimately developing advanced intelligent tutoring capabilities.

4 Acknowledgment

Portions of this research were funded under a contract with the Air Force Research Laboratory, Wright-Patterson AFB. We are grateful for this support.

References

1. Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614. https://doi.org/10.1007/s40593-016-0105-0
2. Johnston, J. H., Smith-Jentsch, K. A., & Cannon-Bowers, J. A. (1997). Performance measurement tools for enhancing team decision-making training. In M. T. Brannick, Salas, E., & Prince, C. (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 311-327). Mahwah, NJ: Erlbaum.
3. Salas, E., Rosen, M. A., Burke, C. S., & Goodwin, G. F. (2009). The wisdom of collectives in organizations: An update of the teamwork competencies. *Team effectiveness in complex organizations. cross-disciplinary perspectives and approaches*, 39-79.
4. Salas, E., Shuffler, M. L., Thayer, A. L., Bedwell, W. L., & Lazzara, E. H. (2015). Understanding and improving teamwork in organizations: a scientifically based practical guide. *Human Resource Management*, *54*(4), 599–622. https://doi.org/10.1002/hrm.21628
5. Sottilare, R. A., Shawn Burke, C., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2018). Designing adaptive instruction for teams: a meta-analysis. *International Journal of Artificial Intelligence in Education*, *28*(2), 225–264. https://doi.org/10.1007/s40593-017-0146-z
6. Gilbert, S.B., Slavina, A., Dorneich, M.C., Sinatra, A. M., Bonner, D., Johnston, J., Holub, J., MacAllister, A., & Winer, E. (2018). Creating a Team Tutor Using GIFT. *International Journal of Artificial Intelligence in Education*, *28*(2), 286-313. https://doi.org/10.1007/s40593-017-0151-2