AUTOMATING EXPERTISE IN COLLABORATIVE LEARNING ENVIRONMENTS

Noelle LaVoie
Parallel Consulting
Lynn Streeter, Karen Lochbaum and David Wroblewski
Pearson Knowledge Technologies
Lisa Boyce
U.S. Air Force Academy
Charles Krupnick
U.S. Army War College
Joseph Psotka
U.S. Army Research Institute

ABSTRACT

We have developed a set of tools for improving online collaborative learning including an automated expert that monitors and moderates discussions, and additional tools to evaluate contributions, semantically search all posted comments, access a library of hundreds of digital books and provide reports to instructors. The technology behind these tools is Latent Semantic Analysis (LSA), a machine learning technology that understands the meaning of words and text in ways that agree highly with human judgments. These tools were evaluated in a series of studies with the U.S. Army War College and U.S. Air Force Academy. At the Army War College, we found that the automated monitor was as accurate at identifying discussion groups in trouble as trained human instructors, and has the potential to effectively reduce the amount of time instructors spend monitoring distance learning courses. At the Air Force Academy, the expert moderator significantly improved the quality of cadets' discussion comments in a collaborative learning environment.

Key words: asynchronous learning networks, automated software agents, collaborative learning, computer supported learning.

INTRODUCTION

We have developed a set of distance learning tools that can monitor, moderate, and assess distance learning courses automatically, greatly reducing instructor workload and increasing student performance. These tools were used at both the U.S. Air Force Academy and the U.S. Army War College. At the Army War College we monitored and assessed an online course activity. We were able to demonstrate the effectiveness of the monitoring tool for providing instructors with current information about how their students' discussions were proceeding, and were able to assess final products from the course activities with high reliability. In the experiment conducted at the Air Force Academy, students participated in one of three discussion conditions: (1) online and moderated by our automated moderator, (2) online with no automated moderator, and (3) face-to-face in a classroom. The quality of the students' discussions was higher when they discussed online with the automated moderator.

The paper begins with a review of relevant literature. Then the moderation and monitoring tool, Knowledge Post®, is described in detail followed by a description of the technology underlying the tool, Latent Semantic Analysis (LSA). The research conducted at the Army War College and the Air Force Academy follows. Next we present a description of how the technology could be made available to integrate with commercial distance learning products. The paper ends with a discussion of the implications that can be drawn from our research, and suggestions for future research directions.

A. Distance Learning

With the growth in popularity of distance learning courses, additional difficulties in knowledge management and assessment of student performance need to be addressed by technological rather than human capital. In 2003 an estimated 11 million post-secondary students participated in at least one distance learning course [1]. Nearly all of these distance learning courses used technology that included asynchronous computer-based instruction, and typically used online discussions. Other technologies, such as video conferencing, have become less popular.

Distance learning has also increased in elementary and secondary schools. During the 2002-2003 school year, over one third of public school districts had students enrolled in distance learning courses [2]. More recently (April, 2005), the U.S. Department of Education has recommended easing limitations on institutions to increase delivery of distance learning programs, arguing that online courses help students access a college education [3]. In military situations, communicating and learning at a distance is a given, hence the early and large emphasis on high quality online learning alternatives at schools including the US Army War College and US Air Force Academy.

The beneficial effects of properly constructed and administered online learning courses are well documented. A meta-analysis of 19 independent case and controlled studies found that when asynchronous discussion groups are instructor moderated, and engage both students and instructors, learning effectiveness is as high as face-to-face instruction and in some instances better [4]. Learning outcomes were measured differently in different studies, but included objective measures such as exam and course grades, quality of work, and participation rates, as well as more subjective measures, such as the overall amount of student learning, level of motivation, and frequency of access to instructors.

Some researchers prefer measures other than learning outcome as indicators of success in computer-supported collaborative learning (CSCL). [5] for example, argues that CSCL research should be concerned with understanding the process of meaning-making, not with the acquisition of knowledge. In essence, Koschmann proposes that learning outcomes are too separate from the actual processes involved in collaborative learning to provide a useful measure of success. Rather, an investigation of the interactions between students during learning, for example the language used, may provide a more appropriate assessment. Using a more process oriented technique to evaluate learning, [6] did a content analysis of four face-to-face classroom discussions and four asynchronous learning discussions with the same questions and instructor. They coded three different cognitive processes going from lowest to highest:

1. Exploration (e.g., rote factual responses, information exchange)
2. Analysis (simple and deep clarification)
3. Integration (connecting ideas, inference, judgment, resolution)

Their most interesting finding was that while there were about two times as many coded instances of cognitive processes in the face-to-face discussion, they were mostly in the lowest category—exploration. In the Asynchronous Learning Network discussions there were more instances of the two higher level processes in absolute as well as relative terms, further supporting the idea that asynchronous discussion groups show evidence of equal or better, and to some degree qualitatively different learning.

An important prerequisite of running a successful online course is having a moderator. A skilled moderator prevents many of the problems that can plague online courses, such as log-in lags that lead to a lack of continuity in discussions, too much student time spent coordinating, and problems using the technology [7], [8]. Regular monitoring allows instructors to identify these problems and provide intervention before the problems escalate and disrupt the quality of the online course.

In order to keep an online discussion flowing smoothly, a moderator should check in on a discussion at least daily to detect any problems as early as possible. The moderator also needs to provide feedback to students to increase learning and motivation. If a discussion gets off-track, a moderator often needs to provide guidance to get students back on topic. Good moderation involves a lot of time and effort on the part of faculty. As a result, many faculty report that teaching online courses is actually more work than teaching traditional face-to-face courses [9], [10], and that they often spend several small chunks of time monitoring online courses, further disrupting their work schedules [11].

In an ideal world, instructors would have tools to manage their limited time and allow students to control the pace of their learning. Online course software would be self-monitoring, alerting instructors as needed. The software would keep track of participation and the ratio of on-topic to off-topic comments to determine when students need help. The software would provide students with feedback about their comments and overall performance during the discussion, as well as aid instructors in assessing the relative contributions made by each student.

Other researchers have attempted to address some of these issues. For example, [12] report on their efforts to use features of student discussions, such as dialogue acts and student roles, to predict the quality of a discussion and the appropriate intervention. Their system was unable to handle natural discourse, and instead relied on sentence openers that students selected to identify dialogue acts, e.g. “Do you think…” and “To elaborate...” Their automated learning companion, which behaved as a peer to the student discussants, was able to identify and intervene when students were confused. The learning companion drew attention to a student’s confusion by adding comments to the discussion reiterating the student’s question. However, the agent was not able to provide explanations or add information to the discussion. Furthermore, the agent missed occasions when it should have intervened, and frequently intervened when it was not necessary, both of which can disrupt collaboration among students.

Unfortunately, automated moderation software is not yet commercially available. However, these tools do exist in prototype form. The tools to be described have been integrated into the Knowledge Post online discussion environment and empirically evaluated in several military user communities. This report outlines the findings and a scheme for integrating these tools into disparate collaborative environments.

B. Knowledge Post

The capabilities of Knowledge Post as it exists today include the abilities:

To find material in the discussion or electronic library that is similar in meaning to a given posting (see Figure 1).

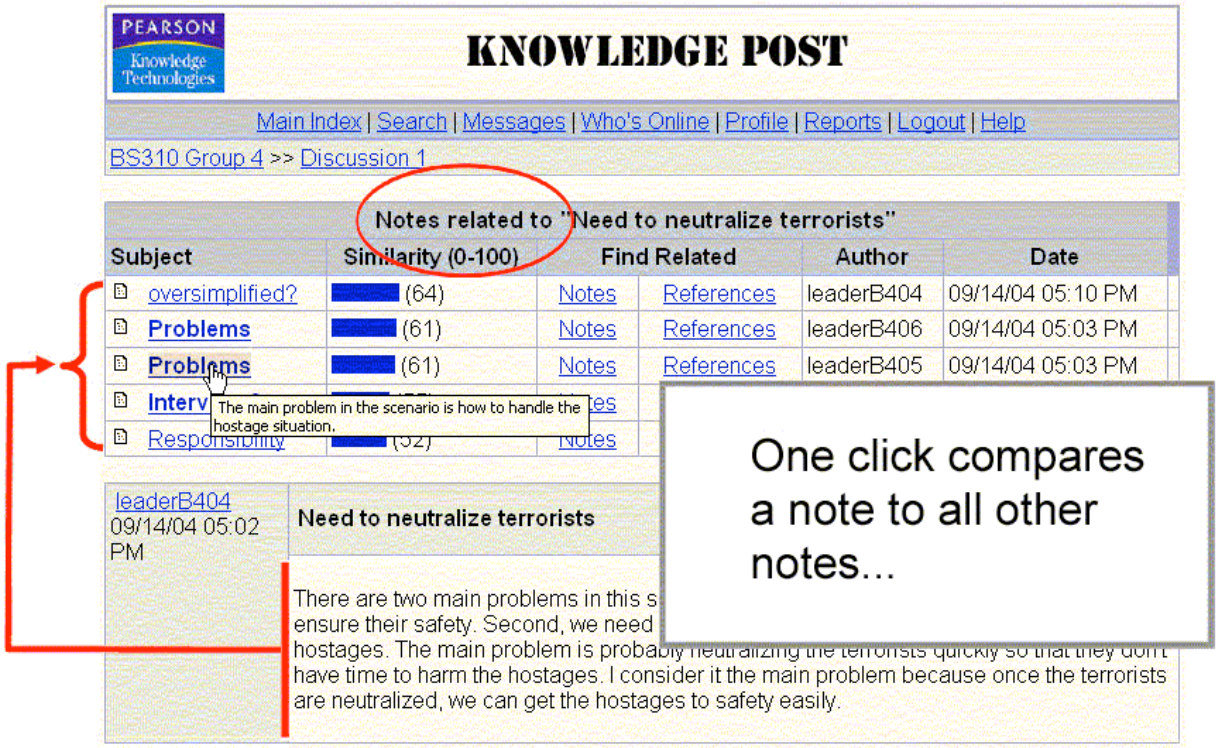

Figure 1. A screenshot from Knowledge Post showing the results of a search for notes similar to the note "Need to neutralize terrorists."

To have contributions automatically summarized by hovering the mouse over the subject of the note (see Figure 2).

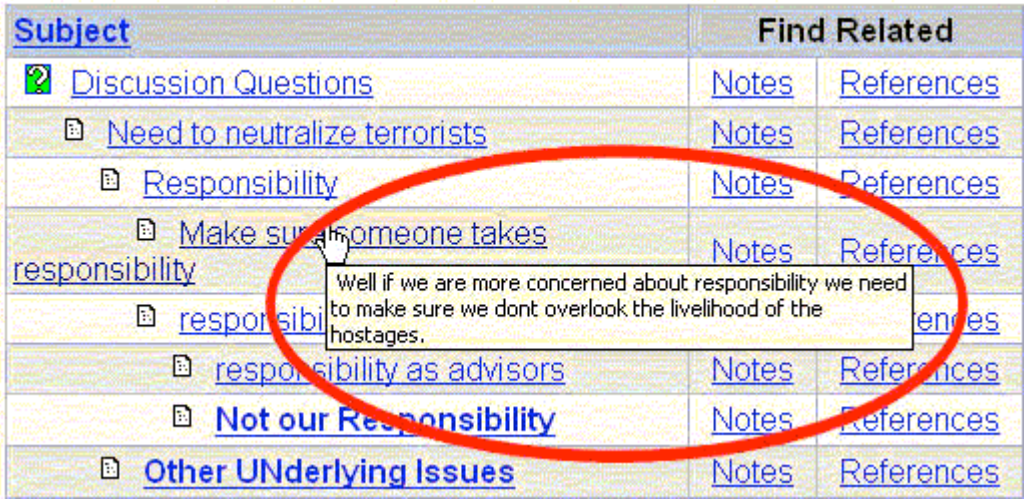

Figure 2. A screenshot from Knowledge Post showing the an automatic summary of the note "Make sure someone takes responsibility" generated by holding the mouse over the subject line.

To enhance the overall quality of the discussion and consequent learning level of the participants.

To have expert comments or library articles interjected into the discussion in appropriate places by automatically monitoring the discussion board activity (see Figure 3).

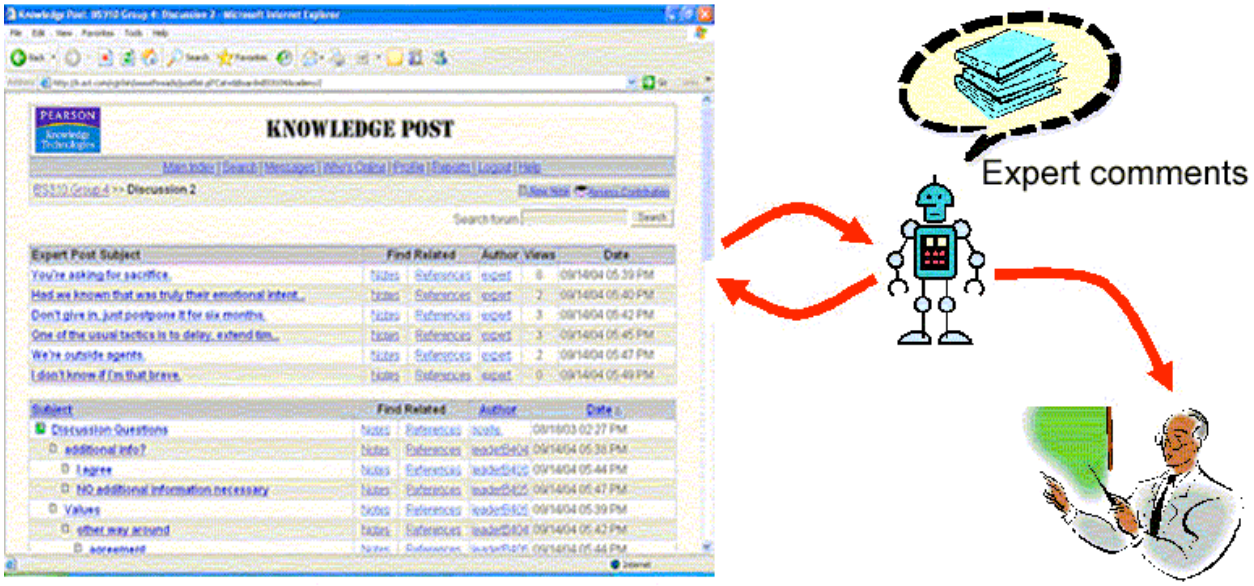

Figure 3. The automated moderator in Knowledge Post.

To automatically notify the instructor when the discussion goes off track.

The first three functions have been part of Knowledge Post for some time, and have been evaluated [13]. Several hundred Army officers, including Lieutenants, Captains, Majors, and Lieutenant Colonels have discussed military scenarios of this sort either face-to-face or using *Knowledge Post*®. Across a broad range of managerial and military scenarios, officers learn more using the threaded discussion tool than they do in face-to-face discussions. The scenarios used in these experiments were either "Think Like a Commander" (TLAC) scenarios developed at Ft. Leavenworth or "Tacit Knowledge of Military Leadership" (TKML) scenarios developed jointly by the Army Research Institute and Yale University [14]. The TLAC scenarios were developed to teach tactical and strategic thinking skills. The TKML scenarios are based on a carefully developed set of representative scenarios of challenging interpersonal leadership situations that are commonly encountered by Army officers, along with sets of alternative actions that a leader might take.

Thirty-eight officers discussed the scenarios face-to-face and wrote their responses using pencil and paper, while the twenty-eight typed their thoughts and eventual "solutions" into the online discussion environment. The electronic discussion group entered an initial response and then a final response after an online synchronous discussion. All responses were randomly sorted and the rank of the officers removed before they were then "blind" graded by two military leadership experts. The grading was based on their expert assessment of the quality of the proposed actions and comments. The results are shown in Figure 4 for all of the TLAC scenarios for one rater. (The same pattern of results has been found for the TKML scenarios and the other raters.)

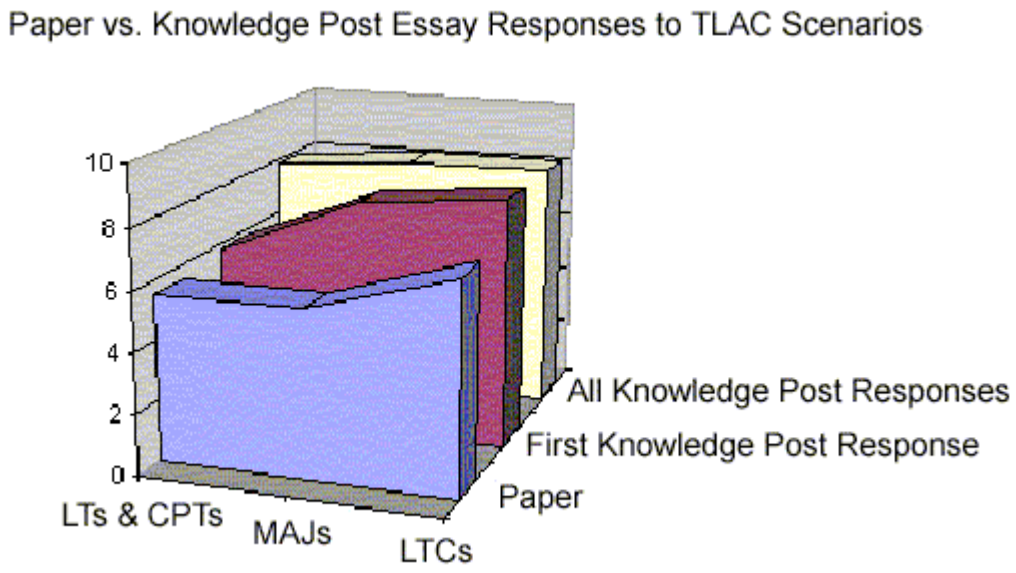

Figure 4. LTs, CPTs and MAJs all wrote better scenario responses using Knowledge Post than using pencil and paper [13].

The rated quality of responses, on a 10 point scale, is plotted for each of the four officer categories as a function of whether they were in a face-to-face discussion group and summarized their views using pencil and paper or whether they were in the electronic discussion group. Whether the discussion took place face-to-face or electronically made a large difference—those who used the electronic discussion group contributed much higher quality initial responses (shown as First *Knowledge Post*® in Figure 4) than those in the paper and pencil group. In addition, the lower ranking officers (Lieutenants and Captains) learned more using the electronic discussion group than did the same ranks of face-to-face participants.

Although senior officers (Majors and Lieutenant Colonels) had a slight superiority in the paper-based version, all officer groups improved through the *Knowledge Post*® discussion. Even the first response in the online format was superior to the final responses in the face-to-face discussion. Several factors may contribute to the better discussion and learning in the electronic medium. It is likely that the tendency for participants in online discussions to be more thoughtful, and engage in deeper cognitive processes [6], accounts for the advantage of the online responses. The anonymity that contributing to an online discussion provides may be another reason. The endurance of contributions made in an online environment may have also increased the officers' motivation to provide a better response. An additional factor is the parallel nature of the discussion—members of an electronic discussion can contribute simultaneously, thereby making more effective use of the time available. This is not possible in face-to-face discussions. The potential for deeper cognition, coupled with greater equality of participation and anonymity, results in a richer set of ideas generated by a greater number of people.

The other two features of Knowledge Post, automatic interjections of expert comments, and automatic notification of instructors when discussion go off track, have been evaluated more recently and are the subject of this paper.

C. The Technology behind the Tools

The tools that were developed do meaning-based search, summarize and assess postings, and moderate discussion groups by gleaning the meaning of the posted text automatically. The critical technology that performs these functions is Latent Semantic Analysis (LSA), a machine learning algorithm that understands the meaning of words and text in such a way that frequently matches the judgments of

humans. For example, LSA infers similarity between words from the contexts in which they occur. Table 1 shows the similarity of five words. Similarity is measured by the cosine of the words in a high dimensional LSA semantic space. The cosine ranges from 1 (perfect similarity and a zero degree angle between the two vectors) to -1 (180 degree angle between the two vectors). On average, two unrelated words will have a cosine of about zero.

	doctor	physician	surgeon	lawyer	attorney
doctor	1				
physician	0.61	1			
surgeon	0.64	0.65	1		
lawyer	0.06	0.06	0.13	1	
attorney	0.03	0.05	0.09	0.73	1

Table 1. LSA similarity between 5 words.

LSA uses a fully automatic mathematical technique to extract and infer meaning relations from the contextual usage of words in large collections of natural discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed ontologies, dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies [15].

LSA takes as input large quantities of raw text parsed into words and separated into meaningful passages such as sentences or paragraphs. Although the technique is based on the statistics of how words are used in ordinary language, its analysis is much deeper and more powerful than the simple frequency, co-occurrence, or keyword counting and matching techniques that have sometimes been assumed to be the only purely computational alternatives to traditional Natural Language Processing. LSA learns about the relations between words and documents by machine analyzing large collections of text—currently handling a billion words of input. The output of this analysis is a several hundred dimensional semantic space in which every word and every document is represented by a vector of real numbers—one for each dimension. Semantic similarity is measured by the cosine of the angles between the vectors representing two words or documents. LSA is not keyword matching. For instance, in LSA similarity space the meaning of the two sentences listed below is extremely close, yet there are no words in common between them:

The doctor operates on the patient.

The physician is in surgery.

LSA is also able to deal with issue of polysemy, or words that have multiple meanings. LSA is able to determine that a word is similar to all of its meanings. For example, the word “fly” has several meanings, including an insect, and to travel through the air. In LSA similarity space “travel through the air” and “fly” are equally close in meaning as are “ two-winged insect” and “fly.” Further, LSA can disambiguate a word with multiple meanings by relying on the context that the word occurs in.

LSA simulates important practical aspects of human meaning to a very useful level of approximation [16]. And it does so using a robust, domain-independent process that allows it to effectively perform tasks

that, when performed by a human, clearly depend on understanding the meaning of textual language. LSA does require a sufficient amount of appropriate text to perform optimally. There are countless applications of LSA including the ability to grade text essays and assign grades that are indistinguishable from a skilled human grader [17], as well as the ability to perform meaning-based search.

Alternative technologies have been used to support computer-supported collaborative learning, such as TagHelper, which was designed to analyze the collaborative learning process through automatic corpus analysis [18]. TagHelper is a technology tool designed to partially automate the tagging of content in a discussion for purposes of process analysis. The authors propose that TagHelper may be used to support moderators of on-line discussions. [12] have also applied a variety of NLP techniques, including hidden Markov models and neural networks, with modest success.

We have taken a different approach using LSA to solve similar analysis and support problems with on-line collaborative learning. With LSA, there is no need for time consuming coding of content, which is often a subjective process that fails to generalize to new course content. Further, LSA has been shown to be more effective than other automatic techniques for modeling the semantics of human language in retrieval and indexing, [19], [20]. However, when text is very brief (less than a sentence) LSA may be less able to accurately assess meaning, and an approach more like TagHelper where content is tagged could be more appropriate.

THE AUTOMATED MONITOR: RESEARCH WITH THE ARMY WAR COLLEGE

The U.S. Army War College's distance education program is of the highest caliber in terms of course construction, faculty moderation, and the students who attend, typically high potential Lieutenant Colonels and Colonels. The Army War College has a large distance learning program with online discussions actively monitored by course instructors. We participated in a study with the Army War College in 2004, where we applied our monitoring technology to an online course activity, and returned live feedback to instructors. The monitoring tool was able to detect important shifts in discussions, and alert instructors in near real-time to potential problems. These initial results indicated that the automated monitor could benefit distance learning instructors by minimizing the amount of time spent monitoring online discussions.

A. Method

The automated monitor was developed and tested during an asynchronous distance learning activity offered to several hundred senior officers in January 2004 over a ten day period. There were 20 separate discussion groups with 12 to 15 participants per group. Participants were U.S. government personnel from all over the world—from Kuwait to Kenya to Kansas. The activity was titled "Interagency Process Simulation" (IPS) and dealt with U.S. foreign and security policy and the future of NATO [21]. Students addressed the following issues during the simulation:

Continued U.S. engagement in Europe through NATO

Russian membership in NATO

European Union security and defense development

Participants were given unique roles, such as Deputy Secretary of State, and assigned to departments and committees, e.g., State Department, Policy Coordinating Committee, Deputies Committee. The departments and committees were organized according to the structure shown in Figure 5, with members of the Policy Coordinating Committee (PCC) also belonging to a department.

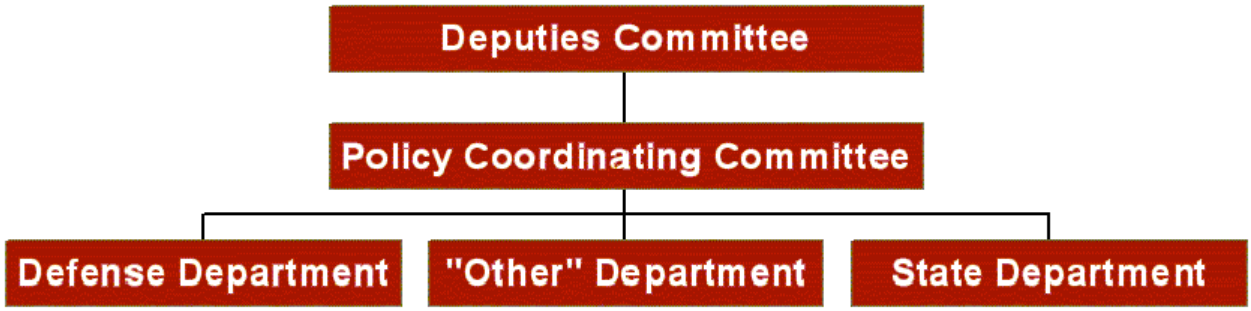

Figure 5. The organization of the Interagency Process Simulation activity.

During the first five days of the activity, the students who were part of the PCC discussed the three issues first in their departments, and then together as the PCC. At the end of the four days, these students completed a 450 word policy alternatives document that they sent to the members of the Deputies Committee (DC). Over the following five days, the students on the Deputies Committee discussed the issues and the recommendations made by the PCC. At the end of their deliberations, they submitted a 450 word final policy recommendation.

The discussion generated over six megabytes of text with a few hundred comments per group, for a total of 1829 comments. Comments were quite long and well-thought out, averaging around 150 words, indicating that students were heavily engaged in the simulation.

There are several ways to tell that a discussion group is "in trouble." The most obvious one is a low participation rate, and detecting this is not a technological challenge, although it provides useful information to instructors. Another indication of trouble is spending far too much time coordinating the group's work, i.e. figuring out who is going to do what, or straying too often from the topic at hand. Automatically detecting coordination difficulties and off-topic remarks was the technological challenge we addressed.

We reasoned that both coordination difficulties and off-topic remarks should be signaled by a series or a high proportion of "administrative" comments. At the other end of the spectrum a group that was "cooking" would have few administrative comments and a large number of comments related to the subject matter of the course. It is reasonable to expect that a higher number of planning and coordination (administrative) comments would occur at the very beginning of the course activity, as students make decisions about how to divide work, and discuss any technological or coordination issues.

B. Automatically separating administrative from content comments

The method developed automatically separated administrative comments from content comments. Separating the two types of comments can be viewed as a signal detection problem—a threshold is needed that identifies the greatest number of hits with the fewest number of false alarms. We used two techniques to score the comments, and then chose thresholds to maximize the accuracy of the separation of the comments. The two techniques are discussed in more detail below.

We were given text from an earlier run of the IPS activity in 2003, including the full text of the discussion with 1605 comments, and the 40 policy papers produced by the students. For the first scoring technique, we had four human graders (one military expert, two political science graduate students, and one non subject matter expert) assign a grade to the policy papers generated by each group yielding a total of 40 papers. The four scorers used the same rubric to score the papers. The reliabilities between the graders were modest with $r(40) = 0.43$, $p = .005$ as the best inter-grader reliability, which is on the low side of inter-rater reliabilities for such exercises. We then used a commercially available essay grading product, the Intelligent Essay Assessor (IEA), [17] based on LSA and other statistical language measures to predict the average of the four human grades for all of the group essays. We chose to model the average of the

human graders, rather than any individual grader, because the average was a more stable estimate of the essay quality. The auto essay grader creates the best fitting statistical model to the grades, which can then be used to score new essays as accurately as human graders. In this case, the model correlated with the average human scores r (40) = 0.56, p = .005 and consisted of two LSA content measures and one readability measure. Using the IEA model produced for the final essays, we were able to score each comment. Because the model was trained on the course content, higher scores indicated that a comment was more similar to a high scoring policy paper, so included more content while lower scores indicated that the comment was more similar to a low scoring policy paper, and had less content.

The second technique we developed using the 2003 IPS course data did not require human scores. Instead, we conducted a principal components analysis on the set of LSA vectors representing the meaning of each paper. This method does not require any training on the human scores of the comments. The intuition behind this method is that a smart human grader, but one who is not necessarily a subject matter expert, could take a pile of essays on the same topic, read them all, compare them and in the end have formed some opinions of which were the good essays and which were the poor essays, and these opinions would be reasonably accurate. This method produced a score for each paper, indicating the amount of content in the paper, with higher scoring papers having more content and lower scoring papers having less. The correlation between this unsupervised learning technique's scores and the average human scores was significant, r (40) = .51, p < .001.

The goal was to use these two types of scores to accurately separate the comments into two categories: administrative and content. A comment is defined as the entire text of a student's post. A human coded the comments as administrative or content, coding 449 of the comments as administrative, and 1156 of comments as content. A second human rater coded a randomly selected 10% of the comments (162). The correlation between the two human scorers was acceptably high (r (162) = .852, p < .001).

Separating the two types of comments can be viewed as a signal detection problem – a threshold is needed that identifies the greatest number of hits with the fewest number of false alarms. Using the first comment scoring method, an appropriate threshold was selected for separating the comments based on the IEA scores. The threshold was selected based on the IEA scores of the comments in such a way that the most comments, both administrative and content, were correctly classified. All of the comments above the threshold were classified as content, while those below the threshold were classified as administrative. This classification was compared to the human coding, and then hits and false alarms were calculated for various thresholds. The threshold that maximized hits and minimized false alarms was selected. Table 2 shows the classification based on the threshold selected.

	Classified Admin	Classified Content	TOTAL
Admin	331	118	449
Content	62	1094	1156

Table 2. The classification of comments from the IPS course.

This model correlated significantly with the human scores, r (1605) = 0.72, p < .001. Note that only 11 % of the comments were misclassified. Using the second method, the unsupervised learning algorithm, the comments were all scored, a threshold was selected in the same way and the comments were classified. This model also correlated significantly with the human scores r (1605) = .84, p < .001, and only 4.9 % of the comments were misclassified. Below are examples of the comment types:

ADMINISTRATIVE: I just want to say 'Thank You' to each of you in our group. It's been great teamwork. Look forward to meeting you all in June. God bless. Jim

CONTENT: The EU military capability needs to be more than a collective security (defensive)

organization, it also needs to have a power projection capability to react to global contingencies that affect Europe. If the concerns can be overcome with effective cooperation and coordination between NATO and the EU, then the U.S. should support the EU's ESDP.

CONTENT MISCLASSIFIED AS ADMINISTRATIVE: Mark and Al: Greetings from San Antonio. The paper here says that the Germans and the French do not like our President's straight forward, Texas way. Well I guess they probably did not like LBJ in the 60's. Mark: I read your paper and then I read the Kennedy's comments. Pretty good stuff. I agree with Gary that the State may water down what comes out of the PCC in order not to offend. This is what Steve pointed out during his AAR. I am a little bit unsure on what happens next. Do we sell Kennedy on trying to revise his language to be more direct and if so who does? I am reaching here. For now though I will look and see what the IPS Gazette has to say for today.

Classifications from both comment scoring methods were very accurate. The classification based on the unsupervised learning method is well suited for use in courses where human scored essays are not available to train IEA. The unsupervised learning method will be used as the basis for automatically coding comments and monitoring the 2004 run of the IPS activity.

C. Results of Monitoring the 2004 IPS Activity

During the activity we received via email all messages that were posted to the forum. The emails were processed each day at 3:00 pm, in order to return feedback by 5:00 pm. A total of 1829 comments were received. The course author, Charles Krupnick, invited us to provide a daily feedback report to all the instructors, customized to their discussion groups, detailing the students' performance, and indicating whether any groups needed instructor intervention. In essence, this feedback could act as a stand-in for the normal instructor monitoring.

Feedback was returned beginning with the third day of the PCC discussion and continuing through the first week. Feedback was returned for every day of the DC discussion during the second week except the first day. The first two days of the first week, and the first day of the second week did not yield much discussion, as students were busy reading instructions and course materials, and learning to navigate the forum. The daily feedback file sent to instructors included a brief summary at the beginning, where important information was listed including any groups that had nonparticipating key members (responsible for submitting policy papers) or had a high percentage of administrative comments, followed by a list of participation by student. Following is an example of the feedback returned to student discussion group 1 on the fourth day of discussion:

IPS feedback for comments posted between 5pm EST January 21 and 5pm EST January 22.

Notes: The following discussion groups had no posted comments from the Asst Sec of State for European and Eurasian Affairs for the last 2 days: SDG 06.

SDG 06

Student Name	Number of Contributions	
Student 1	0	Asst Sec of State for European and Eurasian Affairs
Student 2	0	
Student 3	1	
Student 4	1	
Student 5	1	
Student 6	1	
Student 7	1	

Student 8	1
Student 9	3
Student 10	3

Total Contributions: 12
0% Administrative (off-topic)
100% Content (NATO)

The comments were scored using both the automated essay grader and the unsupervised learning method based on grades from the previous IPS administration (2003). The cut-offs to discriminate between administrative and content comments were based on cut-offs from the previous IPS (2003) as well. The course author, Charles Krupnick believed that the feedback was useful and contained pertinent information that could reduce the amount of time instructors spend monitoring the discussions. In future runs of the IPS activity, fewer instructors might be able to manage the course, while maintaining the same high level of discussion among the students.

THE AUTOMATED MODERATOR: RESEARCH WITH THE AIR FORCE ACADEMY

Scenario based instruction has become a popular way of instilling practical or tacit knowledge in students. However, scenarios can fail as a learning mechanism if the subject matter does not grab the students' attention, if there is not enough to say about the topic, or if there is not sufficient diversity of opinion among the participants. Even if a scenario is representative of problems encountered in the field, it may thus fail to generate discussion. For these reasons we constructed a scenario based on actual events, geared to military cadets, that engaged tactical and strategic military thinking, and for which there had been considerable debate among experts when the event occurred.

A. Developing an engaging scenario

Our scenario was based on the 2002 Moscow theater incident, in which a Chechen terrorist group took over a theater and held approximately 700 theater-goers hostage for three days. The real event was resolved when Russian forces released a sedative gas into the theater and stormed the building. As a twist on the actual incident, the scenario was set in the Philippines, some of the hostages in the theater were Americans, and the hostage-takers claimed that the incident was a direct reaction to the American military re-entry into the Philippines. The discussants played the role of an ADVON (advanced echelon) team commander, and were asked to develop reasonable courses of action. In addition to the scenario, background news articles were provided as part of the exercise. The news articles were adapted from real news stories that were published about the Moscow theater incident.

B. Collecting expert responses to the scenario

Colonels and Lieutenant Colonels from the National Defense University contributed their knowledge and expertise to the terrorist scenario in several pilot studies. Three Colonels and one Lieutenant Colonel in the U.S. Air Force were recruited to review the scenario for accuracy and generate expert comments. They were given the terrorist scenario and background news articles, and asked to discuss the scenario in a face-to-face group. A moderator prompted the officers to discuss various aspects of the scenario that cadets would later be asked to discuss in Knowledge Post. These experts included information that was missing from the scenario and background information, possible courses of action, and personal experiences that contributed to their decision-making. The officers' comments were recorded and transcribed. A total of 104 comments were generated.

One U.S. Army Colonel and two Lieutenant Colonels contributed to an online Knowledge Post discussion that involved Air Force and Army cadets as part of an in-class exercise. An additional 16 comments were generated by the Colonels. An example interchange is shown below:

Colonel 1: "All Commanders in similar situations should immediately review their initial mission statements and guidance. In this case Rules of Engagement and political priorities are important since you represent your country and its interests. Careful consideration of the consequences of your actions should include immediate and close coordination with the country team - including the embassy staff."

Colonel 2 replied: "Agree with this advice. In addition to being cautious, absolutely review the ROE's and any mission statements, OPORD's etc provided to me as CDR of the advance team, and be sure we adhere to these orders."

C. Selecting the Right Comment to Interject

To be effective, an automated moderator should interject relevant comments during the discussion. Additionally, the automated moderator will only have a chance of improving the quality of the discussion if the participants read the comments that it posts. If many of the comments seem to be off-topic or irrelevant, the participants will learn to ignore the automated moderator's contributions.

To make comments relevant, we used LSA to select the comment that was most semantically similar to the ongoing discussion at any given point in time. LSA was used to compare the text of the discussion to the database of officer comments. The officer comment that was most similar to the discussion up to that point (had the highest cosine) was selected automatically and posted to the discussion in real-time. Because the automated moderator added comments from senior officers, the hope was that the comments contained important additional information for the participants to consider. As the discussion unfolded and new information was introduced by the participants, the agent continued to select and post new and relevant officer comments.

Based on pilot testing with Army ROTC students from the University of Colorado, we determined that the automated agent should interject a comment for every seven student comments. When comments were interjected more frequently, the agent's comments tended to be more frequent than any one student's comments. When the agent's comments were interjected less frequently, the agent failed to contribute more than a few meaningful pieces of information during the discussion. Selecting a rate of one agent comment for every seven student comments allowed the students to explore the bulk of the scenario on their own, but ensured that enough comments were added by the automated moderator that students were exposed to several important pieces of information they had failed to generate on their own. In an additional pilot study at Fort Hood, we gave transcripts of the scenario discussions from other officers. We asked the Fort Hood Majors and Lieutenant Colonels to indicate when a comment should be added and what the comment should be. There was no general agreement on when to interject comments. The officers did not seem to feel that there was an optimal time to correct the cadets. However there was high agreement that comments needed to be added in order to instruct the cadets on more appropriate responses to the scenario. Thus, the ratio of 1 to 7 used here may neither be optimal or appropriate for all circumstances.

D. Evaluating the Automated Moderator

We found this terrorist scenario to be successful in eliciting rich discussions within Knowledge Post because of its realism. Responses to this scenario also seem to discriminate well between officers at different ranks. Junior officers often focus on rescuing the hostages with direct military action, while more senior officers are generally concerned with the safety of the ADVON team they are commanding, rules of engagement, informing the appropriate chain of command, and possible diplomatic courses of action. Improvements in the junior officers' responses based on expert interjections should be measurable.

In order to determine how effective the automated moderator was in increasing the quality of the discussion, we conducted an experiment at the U.S. Air Force Academy in the fall of 2004.

Method. One hundred and twenty-six U.S. Air Force Academy cadets volunteered to participate in discussions of the terrorist scenario using Knowledge Post. The cadets participated in groups of approximately ten, for a total of twelve separate groups. The cadets were provided with the background news articles as well as the scenario. The cadets were asked to discuss various aspects of the scenario culminating in a decision about the best course of action. The discussion was conducted in three sections. During the first part, the cadets were asked to discuss the following questions:

1. What problem(s) needs to be solved in the scenario? What is the main problem that must be solved? Why do you consider this to be the main problem?

2. What information did you feel was most relevant and why?

3. Based on your collective wisdom, what additional information is still needed?

During the second phase, the cadets were asked to focus on making connections between the information presented during the discussion, by addressing these questions:

1. Additional information was needed to supplement the news articles. Did you attempt to find this information, if so what did you find? Is the additional information still relevant?

2. How does the information you identified as most relevant relate to the leadership topics discussed in class? At a minimum, discuss the CPP, Philippines, and U.S. in terms of the following leadership topics: i) Values and Ethics, ii) Personality.

3. After considering other discussants comments, evaluate the information considered most relevant in terms of the connections between the relevant information and/or with the leadership topics. Without actually discussing response options, what two or three pieces of relevant information when combined provide insight on an appropriate course of action? What are the associations that support combining the information? Why is this combination of information important?

In the third phase the cadets were asked to discuss these questions:

1. Discuss a personal experience that is relevant to the problem discussed, focusing on the collective information you combined during the previous discussion.

2. After considering other discussants' personal experiences, what course of action would you recommend to solve the problem? Why do you consider this course of action to be most appropriate?

3. What course of action would you recommend to end the hostage crisis? Why do you consider this course of action to be most appropriate?

The cadets were assigned to one of three conditions: 1. Knowledge Post with the automated expert, 2. Knowledge Post without the automated expert, and 3. a face-to-face classroom discussion of the scenario with a human moderator. All cadets discussed the same scenario and received the same instructions. They were all given one hour to complete the discussion with approximately twenty minutes spent on each of the three sections. To allow direct comparisons between Condition 1 and Condition 3, the human moderator in Condition 3 read aloud a selection of the expert comments that were interjected into Knowledge Post and did not add any of their own comments or moderation behaviors. The human moderators selected the comment from the database that they felt best matched the face-to-face discussion. We felt this was analogous to the way in which LSA was used to automatically select the best comment form the database. Figure 6 is a screenshot of Knowledge Post showing a discussion with expert comments added by the automated expert. The expert comments are presented in a table above the cadet discussion. This screenshot is taken from a completed discussion, and shows the second of the three parts of the discussion.

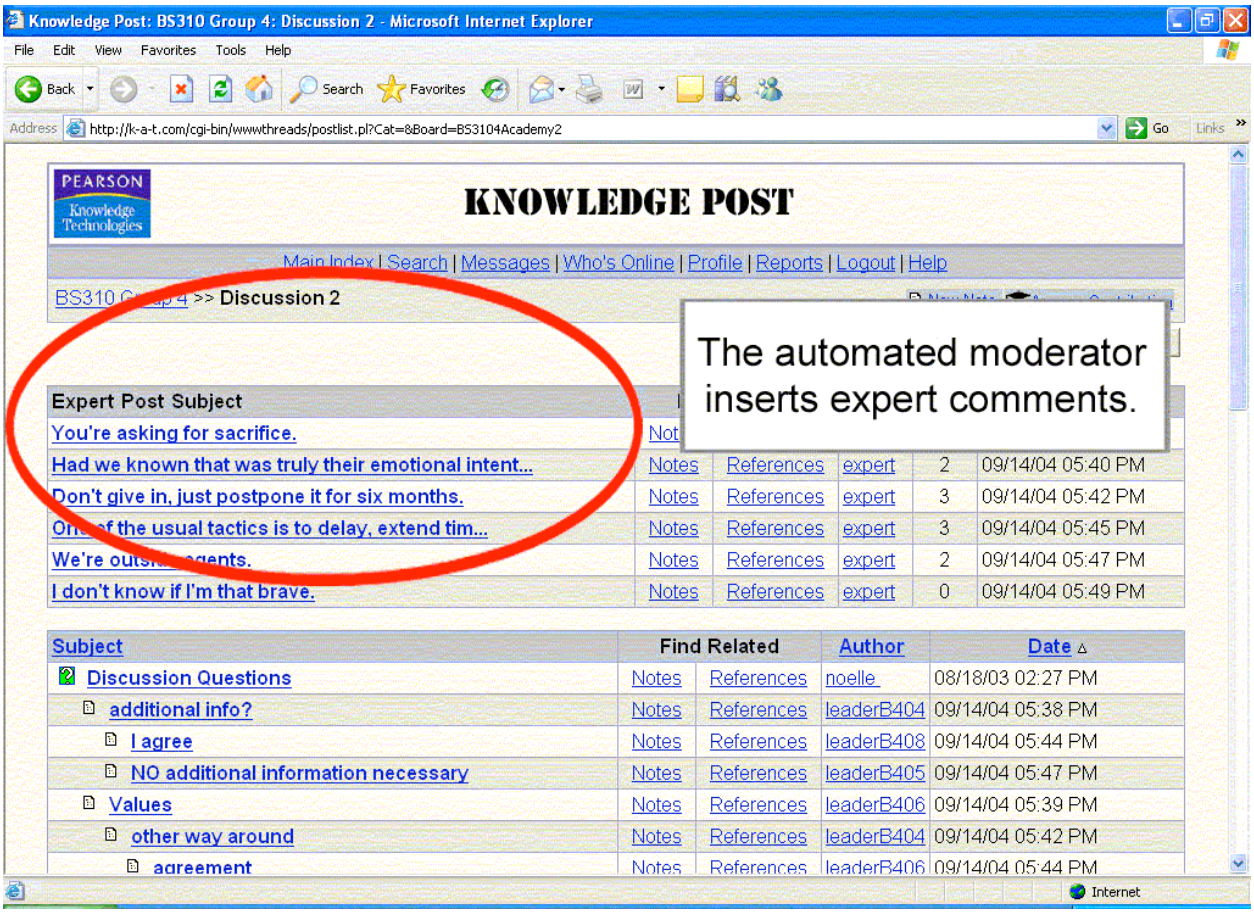

Figure 6. A screenshot of Knowledge Post showing the automated moderator's comments at the top.

Results. Based on previous research [4], [13], we expected Knowledge Post groups to perform better than the face-to-face cadets. We also expected that the cadets who had the benefit of the automated moderator would have the highest quality comments because of the additional knowledge and perspective they were exposed to through the senior officer comments. We did not expect these comments to have the same impact in the face-to-face group, mainly because face-to-face discussions offer fewer opportunities for participants to perform deeper levels of cognitive processing [6]. Further, fewer participants are likely to participate actively in a face-to-face discussion compared to a well-run online discussion [8].

We assessed the effectiveness of the automated moderator using three separate metrics: the quality of the cadets' comments, the number of moderator comments read, and the cadets' ratings of the moderator comments. We looked at the quality of the cadets' comments by comparing them with the senior officers' discussions. These discussions represented the highest quality of discussion around the terrorism scenario and included discussions among groups of Lieutenant Colonels and Majors from Fort Hood and Fort Sill, as well as NDU. Using LSA, we were able to compare the text generated by the cadets to the text of the entire senior officer expert discussion (not just the comments used by the automated moderator) and assess the similarity between the cadet comments and the expert comments. Analyses focused on the third portion of the discussion. The third section was chosen because the cadets discussed optimal courses of action and the resolution of the scenario in this section after having been exposed to a large number of expert interjections over the course of the exercise. Similarity was measured as the cosine between the cadets' comments and the expert discussion. Cosines vary between 1 and -1, with 1

indicating perfect similarity between the texts being compared, 0 indicating that there is no similarity, and -1 indicating perfect dissimilarity.

Quality of Discussion Comments. A one-way analysis of covariance was conducted to examine the differences between the three conditions (Knowledge Post with moderator, Knowledge Post without moderator and face-to-face). Discussion group was included as a covariate because the quality of cadet comments may be dependent upon the discussion group. Further, contrast codes were used to test the a priori hypotheses that the cadets who received moderator comments in Knowledge Post would have higher quality comments than those in the other conditions, and that cadets in the face-to-face condition would have lower quality comments than the cadets in either Knowledge Post condition. There was a significant main effect of discussion condition, over and above any effect of discussion group, $F(2, 112) = 9.7$, $MSE = .08$, $p < .001$. The contrast coded comparisons revealed that cadets who participated in a Knowledge Post discussion with the automated moderator had higher quality comments ($M = .64$) than did cadets who participated in a Knowledge Post discussion without the moderator ($M = .59$), mean difference = .065, std. error = .021, $p = .002$. Also as predicted, the cadets that participated using Knowledge Post (either with or without the moderator) had higher quality comments ($M = .62$) than those that participated in face-to-face discussions ($M = .54$), mean difference = .129, std. error = .033, $p < .001$. These results are shown in Figure 7.

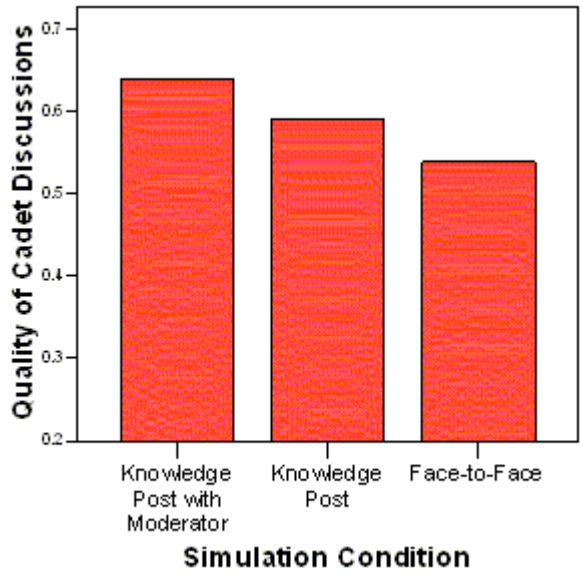

Figure 7. Quality of discussion comments in the three discussion conditions.

The covariate, discussion group, was significant as well, $F(1, 112) = 4.3$, $MSE = .04$, $p = .041$, suggesting that the separate discussion groups varied in quality. This is of interest, especially in light of research suggesting that dynamics within a group can mitigate learning, regardless of the quality of individual contributions [22]. This would be an interesting issue to follow-up in a future experiment, designed to examine the impact of group processes on collaborative learning.

These results clearly indicate that the automated moderator improves the quality of cadet comments through the online discussions. Further, the results suggest that the cadets incorporated the new information gleaned from the senior officer comments into their discussion – even though no senior officer was present. Knowledge Post, or other online discussions, may be an ideal environment for this kind of knowledge transfer, as the cadets in the face-to-face condition did not get the same benefits from the senior officer comments they were exposed to as the cadets in the automated moderator condition did.

Number of Moderator Comments Read. The cadets who discussed the scenario face-to-face had a live

moderator who read aloud some of the comments collected from the senior officers during these face-to-face discussions (drawn from the same comments that were added by the automated moderator in Knowledge Post). On average, the human moderator added fifteen comments over the course of the hour's discussion. The face-to-face discussions were recorded and transcribed for later analysis. The cadets who used Knowledge Post with the automated moderator received an average of 30 expert comments over the three parts of the discussion. We assessed how many times cadets read the expert comments posted by the automated moderator in Knowledge Post. We reasoned that if the cadets thought the comments were useful they would read them more often, and continue to read them throughout the course of the discussion.

We looked at how many times the cadets who received comments from the automated moderator actually read those comments during the hour long discussion. Because the students had only one hour for the discussion, they were limited in the number of comments they were able to read. On average each individual cadet read 19% of the comments, six of the 32 added by the automated moderator. However, the group of cadets as a whole read 63% of the expert comments (20 out of the 32), suggesting that much of the information contained in the expert comments may have been shared during the discussion. As a comparison, on average students read 42% of the comments written by other students over the course of the discussion. The students also continued to read the moderator's comments periodically throughout the hour long discussion, reading two comments per section, on average. This is a good indication that the cadets found the moderator's comments relevant throughout the discussion, although perhaps not as relevant as they found each others' comments.

Cadet Ratings. Following the discussion, the cadets were asked for feedback on the automated moderator including their impressions of the automated moderator's comments. They rated the usefulness, relevance and how much they liked the moderator's comments on paper using a 7 point Likert scale. The Likert scale had a midpoint labeled "somewhat", and end points labeled "not at all" and "very much" for each of the attributes they were asked to rate.

We also looked at the comments collected from the cadets following the discussion. When asked whether they thought the moderator's comments were relevant, the average rating was 4.9 out of 7. Similarly, the ratings of comment usefulness averaged 4.5 out of 7. When asked how much they liked the moderator, cadets responded with an average rating of 4.4. All ratings fell between "somewhat" and "very," providing further evidence that the cadets found the moderator's comments relevant and useful and that they generally liked the moderator's presence in the discussion.

DISCUSSION

A. Making Knowledge Post enhancement available to a more general audience

Knowledge Post differs from a vanilla online discussion group in many useful ways, including semantic searching and automated monitoring and moderation. These enhancements are currently available only within Knowledge Post, and integrating with all of the online discussion group vendors would be prohibitive as the vendor space is large and the APIs required for integration do not yet exist. Instead, we envision an architecture as shown in Figure 8 in which the LSA search facility and automated moderator functionality operate as commercially available internet-based services that can be integrated into any forum software package or group discussion software (e,g., Wiki's, IRC chat, etc.) via an exchange of XML messages.

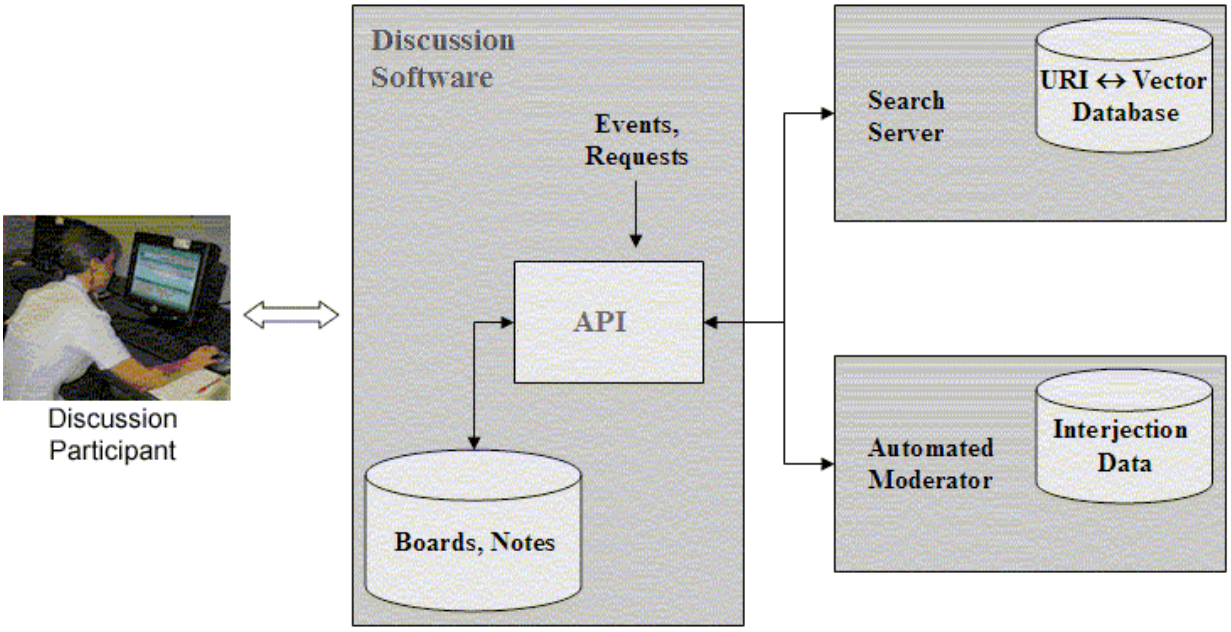

Figure 8. Knowledge Post Service Architecture.

The architecture provides two cooperating information services accessible via APIs we would provide and publish. The Search Service maintains a database of domains (corresponding to a board in a discussion forum, a stream of IRC chat or a Wiki topic area) containing elements (corresponding to posts in a forum, statements in a chat stream, or pages in a Wiki). Client applications label domains and elements using a unique identifier, such as a URL or URI, and send them to the search service for indexing. An LSA vector is created for each element received by the search service and indexed against the corresponding unique identifier. In addition, the search service computes a summary sentence for each element, which is returned to the discussion board software to be displayed whenever a concise summary of the message is needed (for example, as a mouse tooltip). The element text is discarded by the search service, and only the LSA vector is stored.

While this architecture allows the Knowledge Post features to be more accessible to other online discussion group software, there would be some development work required to adapt our tools for use with other software. There are also additional requirements for applying the automated moderator in other courses, namely a selection of "expert" comments appropriate for interjecting into a discussion. However, comments written by the instructor during a previous course on the same topic could be collected and used. Another option would be for the instructor to set up a discussion group and invite a few colleagues to join in a discussion of a scenario problem that the students in the course will be discussing. As long as the comments collected represent a greater understanding of the domain, the students should benefit from being exposed to the comments during their own discussion.

B. Conclusions

We have developed a distance learning tool set that includes automatic embedded assessment, monitoring and moderation, as well as unique reports. With the current boom in distant learning courses, tools are needed to help faculty better manage the extra effort involved in teaching distant learning courses. Through a series of studies with both the U.S. Army War College and the U.S. Air Force Academy we have demonstrated a) the effectiveness of the automated monitor for reducing faculty workload and improving responsiveness to students, b) the accuracy of our embedded assessment, even for never before seen content, and c) the learning benefits of participating in an online discussion with an automated moderator.

The primary contributions from the studies with the Army War College are methods to do automatic embedded assessment of groups and individuals in collaborative learning environments, as well as the development of an automated discussion group monitor. Two methods for embedded assessment were outlined: one that created a statistical grading model mimicking human grading and the other an unsupervised learning method that required no human grading, ordering the samples from best to worst. The methods performed nearly equivalently, implying that the unsupervised learning method could be used with new courses and operate as a valid embedded assessment tool.

The monitor that distinguishes content from administrative comments has great generality and could be applied to automatic discussion alerts for instructors. The algorithm runs concurrently with ongoing discussion groups, analyzing the time series pattern of administrative and content comments. Instructors could be alerted to potentially troublesome patterns—e.g., groups in which the proportion of administrative comments is quite high, indicating either coordination problems or off-topic discussions. In addition, unusually high content scores by an individual or group is also worthy of instructor attention, perhaps to prompt words of praise to the group or to direct other groups to the noteworthy interchanges.

Our work with the Air Force Academy yielded an impressive automated moderator that demonstrated the ability to mimic the presence of senior military officers in a collaborative learning environment. This moderator was able to improve the collective knowledge of a group of cadets discussing a military scenario over the course of a single hour.

The automated moderator is an effective method of disseminating knowledge from senior officers to cadets or junior officers. The collection of expert comments is not automated at this point. However, Knowledge Post itself is an easy mechanism for collecting expert comments, and does not require that a group of experts be co-located in space or time, as the experts may participate in an asynchronous discussion. Contributions from the experts are then loaded directly into a Knowledge Post database, and can be interjected into subsequent student discussions. We found that LSA was able to select comments that were relevant and contained enough additional knowledge to improve the quality of cadets' discussions around a military scenario. The cadets chose to read these comments over the course of the discussion, suggesting that they saw the value in the comments. Feedback provided by the cadets following the discussions supports this conclusion as well. The cadets on average reported finding the moderator's comments relevant and useful.

The moderator selects comments that humans find useful and it does so quickly and automatically. The moderator uses a set of senior officer comments that take very little time to collect, and in turn passes this information on to hundreds of junior officers with minimal resources. Using the same method, the moderator is capable of choosing appropriate reference material from the electronic library, and automatically adding this information to the discussion. The important implication of the automated moderator we developed is that it provides an efficient manner of disseminating large amounts of institutional knowledge contained in key, senior people to the rest of the institution.

Because the techniques we describe are based on LSA, they are relatively content-independent, and can be applied in a variety of subject matter areas. However, additional refinements would no doubt be required to maximize the benefits of this tool set for other courses. We have outlined the steps needed to integrate our distance learning tools with existing discussion environments, which would make these innovations available to the discussion group world at large. The tools we have developed offer a substantial improvement over traditional distance learning environments, providing enhanced learning while reducing instructors' labor intensive moderation and monitoring.

C. Future Directions

There are several studies that we feel would further the results reported here. It would be very interesting to assess the direct impact of using our automated tools on instructor workload by having instructors record the amount of time spent performing various tasks in courses while using traditional online discussions or online discussions with our automated tools. We would expect to find some reduction in workload when our tools were used, but would these reductions be across the board? Or would the tools

only reduce the time spent on certain tasks? Similarly, instructor ratings of student discussions and comments would provide additional evidence that our tools can increase student knowledge.

In the current study we compared the automated moderator to a human moderator, but the human moderator was limited in the number of comments they could add to the discussion. One benefit of a human moderator is that they can think on their feet and adapt to changing discussion topics. A direct comparison between the automated moderator and a skilled human moderator would provide the ultimate test of the automated moderator's effectiveness.

Both the Army War College and the Air Force Academy have student populations that are typically well above average. Using these tools in a wider range of educational institutions may require some adaptation of our technology to be as effective in monitoring and moderating more average students. A demonstration that the tools can be used across a spectrum of different courses and student bodies would greatly increase the generalizability and applicability of our results.

REFERENCES

1. **Allen, I. E. and Seaman, J.** Sizing the opportunity: The quality and extent of online education in the United States, 2002 and 2003. Needham, MA: The Sloan Consortium, 2003.
2. **Setzer, J. C. and Lewis, L.** Distance education courses for public elementary and secondary school students: 2002-03 (NCES 2005-010). U.S. Department of Education. Washington, DC: National Center for Education Statistics (2005).
3. **U.S. Department of Education**, Office of Postsecondary Education, Office of Policy, Planning and Innovation, Third Report to Congress on the Distance Education Demonstration Program, February 2005, Washington, DC, 20006.
4. **Hiltz, S. R., Zhang, Y, and Turoff, M.** Studies of effectiveness of learning networks. In J. Bourne and J. C. Moore (Eds.) *Element of quality online education* (pp. 15-41). Needham, MA: The Sloan Consortium, 2002.
5. **Koschmann, T.** Dewey's contribution to the foundations of CSCL research. In Proceedings of the CSCL (2002).
6. **Heckman, R. and Annabi, H.** A content analytic comparison of FTF and ALN case-study discussions. (pp. 1-10). In Proceedings of the 36th Hawaii International Conference on System Sciences (2003).
7. **Benbunan-Fich, R., Hiltz, S. R., and Turoff, M**. A comparative content analysis of face-to-face vs. asynchronous group decision making. *Decision Support Systems, 34* (4), 457-469 (2003).
8. **Benbunan-Fich, R. and Hiltz, S. R.** Impacts of asynchronous learning networks on individual and group problem solving: A field Experiment. *Group Decision and Negotiation, 8*, 409-426 (1999).
9. **Hislop, G. W. and Ellis, H. J. C.** A study of faculty effort in online teaching. *Internet and Higher Education*, 7, 15-31 (2004).
10. **Pachnowski, L. M. and Jurcyzk, J. P.** Perceptions of faculty on the effect of distance learning technology on faculty preparation time. *Online Journal of Distance Learning Administration*, 6 (3) (2003).
11. **Thompson, M. M.** Faculty self-study research project: Examining the online workload. *Journal of Asynchronous Learning, 8,* 84 – 88 (2004).
12. **Goodman, B. A., Linton, F. N., Gaimari, R. D., Hitzeman, J. M., Ross, H. J. and Zarrella, G.** Using dialogue features to predict trouble during collaborative learning. *User Modeling and User Adapted Interaction, 15*, 85-134 (2005).
13. **Lochbaum, K., Psotka, J., and Streeter, L.** Harnessing the power of peers. Interservice/Industry Training, Simulation and Education Conference, Orlando, FL, (December, 2002).
14. **Hedlund, J., Horvath, J. A., Forsythe, G. B., Snook, S., Bullis, R.C. and Williams, W. M.** Tacit knowledge in military leadership: Evidence of construct validity (Technical Report 1018). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences (1998).
15. **Landauer, T.K., Foltz, P. W., and Laham, D.** An introduction to latent semantic analysis. Discourse Processes, 25, 259-284 (1998).
16. **Landauer, T. K., and Dumais, S.** A solution to Plato's problem: The Latent Semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240 (1997).
17. **Landauer, T. K., Laham, D., and Foltz, P. W.** Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis & J. C. Burstein (Eds) *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Earlbaum Associates, Publishers, 2003.
18. **Donmez, P., Rose, C., Stegmann, K., Weinberger, A. and Fischer, F.** Supporting CSCL with automatic corpus analysis technology. *Proceedings of Computer Supported Collaborative Learning*, 1-10 (2005).
19. **Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R.** Indexing

by Latent Semantic Analysis. Journal of the American Society of Information Science, 41 (6), 391–407 (1990).
20. **Dumais, S.** LSA and information retrieval: Getting back to basics. To appear in: T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch (Eds.), *LSA: A Road to Meaning.* Lawrence Earlbaum, in press.
21. **Krupnick, C.** U.S. Army War College's interagency process simulation. Interactive Technologies Conference, Arlington, VA, (August, 2004).
22. **Barron, B.** When smart groups fail. *The Journal of the Learning Sciences, 12*, 307-359 (2003).

ACKNOWLEDGEMENTS

This research was supported by a Small Business Innovation Research (SBIR) grant from the U.S. Army Research Institute.