Automated Tools for Distance Learning

Abstract

The US Army War College conducted an extensive asynchronous threaded discussion on US national security policy and the future of NATO with 20 sections and approximately 300 participants. The collaborative discussion generated 6 MB of high quality text interchanges that simulated the actual US government interagency process of developing national security policy and strategy. Automatic text understanding tools based on Latent Semantic Analysis (a machine learning software system that understands and represents text in ways similar to humans) were used successfully to assess the quality of groups' and individuals' contributions, and to separate administrative comments from content contributions. This presentation will discuss the results of this exercise, which showed how effective and powerful software agents can be built to moderate online discussions, alert instructors to breakdowns in group communications, monitor content in the discussions, advise and mentor e-learners, and perform embedded assessment of group and individual products.

Introduction

With enrollment in distance learning courses at 2 million students and growing (Allen & Seaman, 2003), there is an increasing need for technology that can support students and instructors in collaborative online learning. Online adult and vocational training, as well as corporate training, are also experiencing tremendous growth (Phillips, 1998). Trends in military training often parallel those in corporate and higher education settings. For instance, as with most workplace and college classrooms, there are increasing demands within the military for learner-centered instruction and technology tools for e-learning, especially tools that foster collaboration (TRADOC, 1999).

Automated knowledge management can assist educators in dealing with this explosion in the quantity of written text. Content assessment and discussion monitoring tools have the potential to improve the quality of distance education for both students and faculty, while reducing the amount of time spent on information management.

Online learning can be just as effective, and in some cases better, than traditional classroom learning. A meta-analysis of 19 empirical studies compared online with face-to-face courses using direct measures of learning outcomes including course grades, exam grades, and participation. This review found that online learning was always as good as or better than equivalent face to face classes (Hiltz et al., 2002). This agrees with faculty perceptions of online learning. In a recent survey of over 255 faculty in New York's state university system, 74% reported that they believed online learning was equivalent or better than other modes of teaching (Shea et al., 2002).

Many of the benefits of online learning are the direct result of the differences between online environments and more traditional classroom environments. Hiltz and Turoff (1993) discovered that the anonymity of online courses leads to increased student participation, especially among minorities and women. Not only do students participate more often, but they also tend to make longer and more thoughtful contributions in an online course. Lochbaum et al. (2002) compared Army officers' discussions of military case scenarios using a threaded discussion group or face to face discussion. The officers in the threaded discussion group demonstrated higher quality work and increased learning

compared to those in the face to face group. This was especially true for the junior officers, whose collaborative responses approached the quality of the senior officers' work. The junior officers' peer-only discussion eliminated poor solutions to the scenarios and reached consensual agreement on the best solutions.

Heckman and Annabi (2003) conducted a content analysis comparing four courses taught as both face-to-face and asynchronous online courses, using the same instructors and questions. They coded the discussions for three kinds of cognitive processes, from low to high: exploration (sharing information), analysis (clarification), and integration (inferencing, judgment, resolution). They found almost twice as many instances of cognition in the face-to-face discussion, but these were mainly at the lowest level: exploration. The online classes had more high level examples of analysis and integration, indicating a deeper level of cognitive processing, even though they used the same instructors and questions.

Instructor Support

While learning is just as good or better in online courses as in traditional courses, the same does not seem to be true for faculty support. Over half of distance learning faculty report spending more time preparing and delivering online courses than equivalent traditional courses (NEA, 2000, Pachnowski & Jurczyk, 2003). Even after teaching an online course 8 times, faculty still reported that online courses take more of their time to teach. Furthermore, most faculty members do not receive reduced course load or additional compensation for the extra time that online courses require. On the other hand, a survey of 255 faculty participating in SUNY's distance education program found that 96% of faculty reported general satisfaction with their experience teaching an online course. They also felt that these benefits spread to the classroom: 85% of these faculty believed that teaching an online course would improve their classroom teaching too (Shea et al., 2002).

Faculty satisfaction most likely varies as a function of the amount and quality of support offered, and the ability to adapt to some of the novel demands of online teaching. DiBiase's (2000) comparison of two courses, one online and one in a traditional classroom, revealed that the online course required more instructor attention and monitoring, although it did not require more overall time. A field experiment comparing online with face-to-face instruction for individuals and groups found that students in the online courses had the best perception of learning, but felt that the discussion had lower quality than students in a face-to-face discussion (Benbunan-Fich & Hiltz, 1999). Closer analysis showed that the online discussions suffered from reduced participation due to login asynchronies, a lack of continuity common in asynchronous online discussions caused by students logging in unpredictably. Additionally, the online students had to spend discussion time planning and coordinating in order to produce their final report, as they had no tools for supporting online collaborative document authoring.

Collaborative Learning Tools

In prior research on threaded discussion groups, we developed an integrated text understanding software agent that automatically suggests the next best note to read (not too hard and not too easy), finds semantically similar notes, and powers searches of a vast electronic library for semantically similar topics (Lochbaum et al., 2002; Psotka et al., 2003). The agent is powered by Latent Semantic Analysis (LSA), a machine learning algorithm that understands and represents text in human-like ways (Landauer et al.,

1998). LSA learns about the relations between words and documents by machine analyzing large collections of text. 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 words or documents. LSA is *not* keyword matching. For example, in LSA similarity space the meaning of the two sentences listed below is extremely close, yet there are no keywords in common between them:

- ***The doctor operates on the patient.***
- ***The physician is in surgery.***

Because it is difficult, if not impossible, for a student to read all of the comments posted in a synchronous learning environment, the text understanding software agent was developed to suggest the next best note to read. This agent is based on the Goldilocks Principle: selecting the next note so that it is not too difficult, nor too easy, but just right for learning (Wolfe et al., 1998; Psotka et al., 2003). We reasoned that the ability of a reader to learn from text depends on the match between the background knowledge of the reader and the difficulty of the text information. LSA was used to measure similarity as a cosine between the students' comments, a good indicator of the students' background knowledge, and other comments in a discussion group.

A low cosine between the two comments would indicate low similarity, and thus reading this text would produce little learning. On the other hand, a high cosine would indicate that the comment was very similar to the student's comment, and thus would be unlikely to contain new information. Ideally, a comment that is similar but not too similar is chosen for the student to read next. Typically, these comments have cosines between 0.5 and 0.6, although this varies by the particular content. These comments tend to be slightly better than the original comment as well. This software agent supports students' learning by choosing appropriate comments, without instructors having to individually select good comments for different students to read.

We have continued to develop LSA's potential for supporting distance learning to include online embedded assessment and discussion moderation tools. These assessment tools can automatically assess the quality of individual and group contributions, while the moderation tools can identify comments that are on-topic or off-topic (such as planning, coordination, social asides and humor, or administrative comments) and automatically alert instructors. By supporting instructors, these tools will make it possible to take advantage of the increasing effectiveness of learner-centered approaches to distance learning and collaborative, online instruction environments (Bonk & Wisher, 2000).

Method

The US Army War College collected a complete electronic copy of an asynchronous distance learning activity that had been offered to several hundred senior officers in January 2003. Using the data from this activity we were able to develop the software agents and test their feasibility in a retrospective study that will prepare for on-line application of the developed toolkit.

The activity was titled "Interagency Process Simulation" (IPS) and simulated the process of interagency development of US foreign and security policy related to the future of NATO. There were 20 separate discussion groups with 12 to 15 participants per group. Participants were US government personnel from around the world—from Kuwait to Kenya to Kansas. Over the course of eight days, students addressed the following issues:

1. Continued US engagement in Europe through NATO
2. Russian membership in NATO
3. European Union security and defense development

Participants were given unique roles, such as Deputy Secretary of State, and assigned to departments and committees, such as the State and Defense Departments and the Policy Coordinating Committee (PCC) and Deputies Committee (DC). The departments and committees were organized according to the structure shown in Figure 1, with members of the PCC also belonging to a department. During the first four 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. By the end of the four days, these students were required to submit a 450-word policy alternatives document to the members of the DC. Over the following four days, the students on the DC 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 six Mb of text with a few hundred comments per group. Comments were often quite long and well thought out, averaging around 150 words, indicating that students were heavily engaged in the simulation. This conclusion was supported by student evaluation of the simulation, which was overwhelmingly positive—194 positive to 16 negative comments.

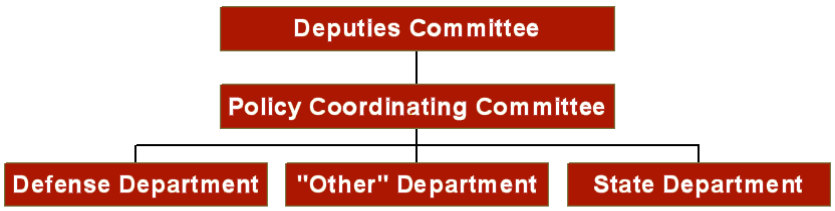

Figure 1. The Interagency Process Simulation Structure.

Results

The success of our automated online learning tools in this retrospective analysis was striking in several areas that support instructors' assessment and knowledge management.

Final Essay Assessment

The 450 word final essays that each group produced provide one way of assessing the performance of the discussion groups and individuals. Four humans (one military

expert, two political science graduate students, and one non subject matter expert) graded the final essays. The inter-grader reliabilities were modest, with the best agreement between two graders reaching 0.43. We then trained the Intelligent Essay Assessor (IEA) on the average of the four grades. The IEA is an automatic essay grader based on LSA that is able to assign grades to text essays that are indistinguishable from a skilled human grader (Landauer et al., 2003). On average, the IEA takes about 1 second to grade an essay, making this an incredibly efficient way of grading text. Grades based on the best IEA model which included two LSA measures and one readability measure, correlated 0.56 with the average human grades. This essay grading model could be used to accurately grade other text on the same topic, including essays for a repeat of this course, leading to the conclusion that the IEA could act as an embedded assessment agent for future Interagency Process Simulations.

However, it would be even more useful to have a flexible assessment agent that could grade essays on any topic – without being trained on human grades. To this end, we used a proprietary, no-training method that compares every essay to every other essay on a series of statistical measures and then produces the best linear ordering of the set of essays. The idea is that a smart human grader, who is not a subject matter expert, should be able to read all the essays and order them from good to bad, and the ordering would be reasonably accurate. In fact, the correlation between the no-training model grades and the average of the four human grades was 0.51, nearly identical to the accuracy of the IEA model based on human grades (0.56). The implication is provocative – that general, embedded assessment agents could be built for use in any content-based course to assess any text produced by individuals or groups. Furthermore, the assessment would be accurate, unobtrusive, automatic and instantaneous.

Monitoring Discussion Comments

As discussed earlier, one of the differences between online courses and face-to-face courses is the increased amount of instructor attention, or monitoring, that online courses require (DiBiase, 2000). An automated software agent that monitors the discussion and alerts instructors to important shifts in topic could reduce the amount of attention required by faculty, and improve the response time when a discussion group gets into trouble. Login lags and reduced participation should be easy to recognize as a sparser discussion with fewer comments. Groups having technological problems or having difficulty coordinating and planning would be harder to identify. It seems reasonable that an increased proportion of off-topic comments or a series of off-topic comments, or changes in students' attitudes, would signal these problems. Our challenge was to create an agent that could identify on- and off-topic comments in a discussion.

We identified all the comments from the 20 discussion groups dealing with the NATO part of the discussion, roughly 1600 in all. A human rated each as either off-topic (0) or on-topic (1). Using the Intelligent Essay Assessor model used to grade the final essays, we scored each comment. Scores ranged from 0.47 to 8.00 and correlated with the binary human scores of the comments 0.72. Separating the two types of comments can be viewed as a signal detection problem—a threshold is needed that makes the greatest number of hits with the fewest number of false alarms. Table 1 shows the best classification based on the threshold selected. Note that only 8% of the comments were misclassified.

Table 1. The classification of comments from the IPS course.

	IEA Off-topic	IEA On-topic	TOTAL
Human Rater Off-topic	300	31	331
Human Rater On-topic	61	1086	1147

The comments were also scored using our proprietary no-training method, and an appropriate threshold selected. The classification based on the no-training method correlated 0.84 with the human scores. Using this method, only 5% of the comments were misclassified, indicating that this model is as accurate as the IEA model. Even without human grades to train on we are able to accurately identify on- and off-topic comments. This suggests that the software agent could look for a small ratio of on-topic to off-topic comments, or a long series of off-topic comments, and notify the instructor that their discussion group may be having problems. Similarly, if the software agent detects unusually high comment scores, indicating a large number of on-topic comments, it could notify the instructor that their group is performing well. An embedded software agent based on this model would be able to identify potential problems in a discussion on any topic, without additional training on human grades, and with only minimal human intervention to pick an appropriate threshold.

Applying Goldilocks to Asynchronous Collaborative Learning

As discussed earlier, we have been able to exploit the Goldilocks effect in previous discussion groups to improve student learning. This effect allows the agent to select comments that are not too easy or too hard, but at just the right level for a student. A comment at the right level should be close to a student's background knowledge, but add something new to what they already know. It is possible to maximize the amount that students learn from a discussion, or even to improve the quality of a discussion, by choosing the right comment for a student to read next.

We analyzed the quality of each contribution compared to the quality of the most similar contribution in any of the groups. Over all 1600 comments in the discussion we found a strong Goldilocks effect. The automated agent selected similar contributions that were significantly higher in overall quality, with a mean difference of 0.44 ($t = 14.32$, $p < 0.001$), but still strongly related in meaning ($r = 0.83$). As we have found in other discussions, the range of similarity cosines from 0.5 to 0.6, appear to have the best characteristics for learning, with similar but somewhat better content. We may be able to take advantage of the Goldilocks effect to help support groups that are struggling to produce good policy papers, by sending them appropriate comments to read. Similarly, groups with few discussion comments, or a long period without any comments, may benefit from reading selected comments from other groups.

Evaluating Overall Group Performance

An ongoing problem for instructors is how to evaluate the relative performance of groups. We summed several measures and arrived at a ranking of the discussion groups. The measures included: average human grades for the policy papers, proportion of content comments, number of comments, and the average Intelligent Essay Assessor grade of all the comments per group. There was, however, insufficient validation data to support the ordering, although the groups identified as lesser performers typically had entry and format problems that detracted from their final product. What measures to use and how to weight them are still open research questions.

Using a multidimensional scaling (MDS) of the cosines between the vectors of all notes is another potentially useful approach for evaluating group performance. In general we have found that the best comments share more common features and concepts with all other comments and so appear near the centroid of the space of all notes. Consistent with this finding, the MDS graph of this space (Figure 2) shows most groups' intermediate and final policy statements lying in a well-defined cluster, and only two groups distinctly outside this cluster. These are two of the groups that other measures agree required improvement. The uniform similarity of most groups, and the divergence of only two groups, provides an argument not only for the success of the course overall, but also for the successful use of this technology for its designed purpose.

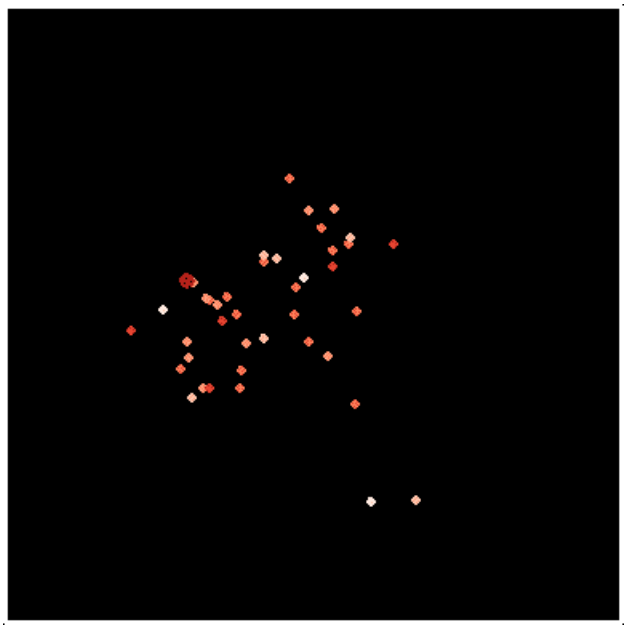

Figure 2. Two groups stand alone in the multidimensional scaling of cosine distances between all groups. They are shown in the bottom right quadrant of Figure 3.

Software Agent Trial

In January of 2004 the Army War College will repeat the Interagency Process Simulation activity. The software agents that we have developed will be used to monitor the discussion and alert instructors to potential problems, to assess the final essays produced by the discussion groups, and to help rank the groups based on various measures of performance, including proportion of content comments, number of comments, and the average IEA grade of all the comments per group. This should provide a critical test of the models the agents are based on, and allow us to create specialized agents that actively monitor many aspects of each group's performance.

Conclusions

Two automated software agents were developed for use in online courses. One agent acts as a contributions assessor, and was used to grade the intermediate and final policy papers produced by each discussion group. One version of the agent used the Intelligent Essay Assessor, a statistical grading model that mimics human graders. Another version used the no-training model, which computes various statistics on the written material and linearly orders the essays from best to worst. Both models were equally accurate, implying that no human grades are necessary for an embedded assessment agent, allowing this agent to function in any content-based course.

The second software agent distinguishes between on-topic and off-topic comments, and could act as an automatic discussion moderator. The separation of the comments was accomplished successfully using either the IEA model or the no-training model. This agent works either by monitoring the number of on-topic and off-topic comments for patterns that indicate a discussion group has gotten off-topic or needs some instructor intervention, or by assessing the type of off-topic comment. When the off-topic comments are disproportionately planning and coordinating comments, the agent can notify the instructor that the group is faltering.

In the context of the burgeoning distance learning community, these automated software agents have the potential for improving the quality and consistency of learning and reducing the workload of instructors. Automated assessment of course papers means that students can receive immediate feedback. Embedded discussion moderation reduces the potential burden on instructors to frequently monitor online courses, and ensures that groups in trouble will receive prompt help.

Endnotes

(1) With two authors from government organizations, we emphasize that the views expressed in this paper are those of the contributors and do not necessarily reflect the official policy or position of the US Army, US Department of Defense, or the US government.

References

Allen, I. E. and Seaman, J. 2002. *Seizing the opportunity: The quality and extent of online education in the United States, 2002 and 2003*. Needham, MA: The Sloan Consortium.

Benbunan-Fich, R. and Hiltz, S. R. 1999. Impacts of asynchronous learning networks on individual and group problem solving: A field experiment. *Group Decision and*

Negotiation. 8.

Bonk, C. J., & Wisher, R. A. 2000. *Applying collaborative and e-learning tools to military distance learning: A research framework.* (Technical Report #1107). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

DiBiase, D. 2000. Is distance teaching more work or less? *American Journal of Distance Education*. 14(3).

Heckman, R. and Annabi, H. 2003. A content analytic comparison of FTF and ALN case-study discussions. In *Proceedings of the 36th Hawaii International Conference on System Sciences*.

Hiltz, S. R. and Turoff, M. 1993. *The network nation*: *Human communication via computer*. Cambridge: The MIT Press.

Hiltz, S. R., Zhang, Y. and Turoff, M. 2002. Studies of effectiveness of learning networks. In *J. Bourne and J.C. Moore (Eds.) Elements of quality online education*. Needham, MA: The Sloan Consortium.

Landauer, T.K., Foltz, P. W., and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Processes*. 25.

Landauer, T. K., Laham, D., and Foltz, P. W. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis and J. C. Burstein (Eds) *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Earlbaum Associates.

Lochbaum, K., Psotka, J., & Streeter, L. 2002. Harnessing the power of peers. *Interservice/Industry Training, Simulation and Education Conference*.

National Education Association. 2000. *A survey of traditional and distance learning higher education members*.

Pachnowski, L. and Jurczyk, J. 2003. Perceptions of faculty on the effect of distance learning technology on faculty preparation time. *Online Journal of Distance Learning Administration*. 6(3).

Phillips, V. 1998. Selecting an online authoring system. *Training*, *35*(4), 53-61.

Psotka, J., Streeter, L., Landauer, T., Lochbaum, K., and Robinson, K. 2003. Augmenting electronic environments for leadership. *NATO Human Factors in Medicine Conference*.

Shea, P., Pelz, W., Frederickson, E. and Pickett, A. 2002. Online teaching as a catalyst for classroom-based instructional transformation. In *J. Bourne and J.C. Moore (Eds.) Elements of quality online education*. Needham, MA: The Sloan Consortium.

TRADOC. 1999. *The Army Distance Learning Plan*. U.S. Army Training and Doctrine Command, Fort Monroe, VA.

Wolfe, M. B., Schreiner, M. E., & Rehder, R. 1998. Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes.* 25 (2-3).