J. Comput. Educ.
DOI 10.1007/s40692-015-0040-3

Mining learner–system interaction data: implications for modeling learner behaviors and improving overlay models

Tenzin Doleck[1] · Ram B. Basnet[2] · Eric G. Poitras[3] · Susanne P. Lajoie[1]

Received: 4 April 2015 / Revised: 10 July 2015 / Accepted: 14 July 2015

Abstract A growing body of empirical evidence suggests that the adaptive capabilities of computer-based learning environments can be improved through the use of educational data mining techniques. Log-file trace data provides a wealth of information about learner behaviors that can be captured, monitored, and mined for the purposes of discovering new knowledge and detecting patterns of interest. This study aims to leverage these analytical techniques to mine learner behaviors in relation to both diagnostic reasoning processes and outcomes in BioWorld, a computer-based learning environment that support learners to practice solving medical problems and receive formative feedback. In doing so, hidden Markov models are used to model behavioral indicators of proficiency during problem solving, while an ensemble of text classification algorithms are applied to written case summaries that learners' write as an outcome of solving a case in BioWorld. The application of these algorithms characterize learner behaviors at different phases of problem solving which provides corroborating evidence in support of where revisions can be made to provide design guidelines of the system. We conclude by discussing the instructional design and pedagogical implications for the

✉ Tenzin Doleck
tenzin.doleck@mail.mcgill.ca

Ram B. Basnet
rbasnet@coloradomesa.edu

Eric G. Poitras
eric.poitras@utah.edu

Susanne P. Lajoie
susanne.lajoie@mcgill.ca

[1] McGill University, 3700 McTavish St., Montreal, QC H3A 1Y2, Canada

[2] Colorado Mesa University, Grand Junction, CO, USA

[3] University of Utah, Salt Lake City, UT, USA

Published online: 05 August 2015

novice–expert overlay system in BioWorld, and how the findings inform the delivery of feedback to learners by highlighting similarities and differences between the novice and expert trajectory toward solving problems.

Keywords Intelligent tutoring systems · Medical education · Educational data mining · Learning analytics · Hidden Markov models · Overlay models

Introduction

The utility of computer-supported education for learning cannot be understated; such learning experiences enrich the learner in varied ways. Progress in instructional technology research and development has brought various affordances and achievements in computer-based learning environments (CBLEs). CBLEs present important learning opportunities (VanLehn 2011), highlighted by the body of work replete with examples of CBLEs that have been developed, deployed, and assessed in an effort to aid learners in their scholastic activities. CBLEs have been used for a wide variety of purposes and in a gamut of contexts, including science (e.g., Lajoie 2009), math (e.g., Matsuda et al. 2013), computer science (e.g., Mitrovic 2003), and history (e.g., Poitras et al. 2012). The literature is rife with studies documenting the positive outcomes of the use of CBLEs (e.g., Beal et al. 2007; Biswas et al. 2010; Dodds and Fletcher 2004; Graesser et al. 2005; Matsuda et al. 2013; VanLehn et al. 2005). Research has suggested that there are beneficial effects of incorporating adaptation in such learning systems (Anderson and Gluck 2001; Durlach and Ray 2011).

Computer-based learning environments like intelligent tutoring systems (ITSs) are designed to provide adaptive instruction through the careful tracking of learner performance that determines the level of assistance needed at specific points in time. Some examples of adaptive systems include: Autotutor (Graesser et al. 2004), Andes Physics Tutor (VanLehn et al. 2005), and Cognitive Tutor in Algebra (Koedinger and Corbett 2006). ITSs provide learners with a one-on-one tutoring experience, which as Bloom (1984) found, can result in learning outcomes two standard deviations above regular classroom instruction. Over the years, research on the design, implementation, and evaluation of such learning systems has been accruing, yet challenges remain. One perennial challenge for learning technology researchers is how to adapt learning systems and improve response mechanisms. Modeling learner behaviors in such systems is crucial to better comprehend the learning trajectory which is needed to improve the learning system. Durlach and Ray (2011) note that generally the input (data) used to determine the level of adaptation is based on the "student ability to answer questions or solve problems" (p. 21). Thus, to enhance the adaptation and response mechanisms in learning systems, and ultimately learning, user–system interaction data need to be considered in light of its importance to improving instructional systems.

The unbridled growth in digital details that are created by people and systems is a widely told story. Similarly, a wealth of educational data is being amassed by researchers designing computer-based learning environments. An important aspect

of learning in an ITS is the monitoring of learners' progress by capturing the traces they leave as they interact with the learning system. Technological advances have afforded the ability to capture students' learning on a more granular level than ever before. One of the most common means of capturing learner–system interactions in computer-based learning environments is via log files (Romero and Ventura 2010). The use of log files can be especially helpful in ascertaining learning without being intrusive, and also providing a flexible means of capturing data at different levels of granularity. Leveraging the learning data could be used to create more effective educational experiences. Recent advances in educational data mining and learning analytics have opened up new avenues for conducting educational research. Specifically, these techniques have created new opportunities to investigate learners' learning trajectories, and thus, increases the ability to facilitate learning. Concomitantly, there has been an explosion of interest in mining the vast and rich repository of data in CBLEs as is shown by the use of educational data mining (EDM), a fast-growing field based on the optimal use of educational data, to facilitate and augment educational research. Data mining has been applied with much success on educational data (Baker and Yacef 2009; Romero and Ventura 2010). EDM research has focused on mining and modeling educational data toward comprehending various dimensions of learning such as off-task behavior (Baker 2007), learner emotions (Craig et al. 2008), gaming the system (Baker et al. 2013), detecting programing strategies (Blikstein 2011), and learner engagement (Cocea and Weibelzahl 2009). Durlach and Ray (2011) highlight that "one way to tune parameters of adaptation is through analysis of past student performance data using data mining techniques" (p. 24). Thus, EDM techniques and tools can be useful analysis tools that can lead to improvements in the types of student modeling afforded by ITSs.

Enhancing learner experience is a constant goal for instructional technology researchers, and the ways to achieve this goal are plentiful. One of the predominant features of intelligent tutoring systems is the provision of feedback to learners. Studies have suggested the benefits of feedback toward better learning outcomes (Hattie and Timperley 2007; van der Kleij et al. 2012) and further, guided feedback through appropriate and timely feedback is also posited to be important for learning (Anderson et al. 1995; Merrill 2002). Similarly, individualized feedback is considered an important component of ITSs (Park and lee 2004). Employing various data mining techniques on usage data can be used to improve the learning system, specifically toward the design of feedback mechanisms. In this paper, we present and discuss our efforts to leverage data mining techniques in the context of BioWorld.

BioWorld (Lajoie 2009; Lajoie et al. 2015) is an ITS developed to scaffold novice physicians in developing clinical reasoning skills as they diagnose virtual patient cases. Appropriate scaffolding and adaptation is based on an accurate learner model, which is an essential step toward identifying learner behavior, ascertaining the use of the learning material, and improving the learning system. Analysis of the sequence of actions employed by learners in problem solving can provide insights into the use of the learning material, and facilitate a deeper understanding of the learning process. For the proposed study, we employ a data mining approach, which

automatically builds a learner model via the behavior patterns observed that illustrate specific learning trajectories. Hidden Markov models (HMMs) provide a simple and effective way for modeling sequence data (Rabiner 1989; Rabiner and Juang 1986), and have received a growing recognition for its applicability in varied problems. In the present study, we employ HMMs on the sequence data extracted from the log files generated by BioWorld, to explore insights that can be harvested by such a modeling technique.

Scaffolding is provided in the form of expert feedback about the clinical reasoning processes taken to solve a specific case. BioWorld dynamically assesses novices' reasoning trajectory against expert paths. The novice–expert overlay model is used to provide requisite feedback to learners (Naismith and Lajoie 2010; Doleck et al. 2014a). An important facet of clinical reasoning is the ability to write case summaries. However, the current overlay model does not include these summaries. Motivated by the desire to improve the Novice–Expert overlay model in BioWorld, we explore the feasibility and efficacy of text mining in ascertaining the differences between case summaries written by experts and novices in BioWorld.

The primary goals and contributions of this paper are to illustrate the possibilities and efficacy of two popular mining techniques, namely, HMMs and text mining on learner–system usage data from an ITS called BioWorld. This paper is organized as follows: (1) the first section presents a brief overview of the learning environment used in the study, namely, BioWorld; (3) the second section describes the methodology; (4) the third section presents the procedure, analysis, and results of HMM analysis across the three clinical cases; (5) the fourth section outlines the experimental setup and findings of using text mining for the novice–expert classification task; and (6) the final section discusses the findings, highlights limitations, and offers future directions of the present study.

Learning environment: BioWorld

BioWorld (Fig. 1) is an ITS that simulates clinical reasoning and is designed to support medical students in practicing diagnostic reasoning skills with virtual patients while receiving feedback (Lajoie 2009; Lajoie et al. 2015). The system was created using a cognitive apprenticeship framework (Collins 2006) where learners practice realistic tasks and are scaffolded in the context of their learning with expert models. BioWorld was designed to stimulate individual metacognitive awareness by providing tools to support students in both monitoring and controlling their learning (Lajoie et al. 2013). Students use these tools dynamically to document their hypotheses, their confidence level in these hypotheses, and evidence that supports their hypotheses. Bioworld consists of four main learning spaces (The Patient Problem, Chart, Library, and Consult). Using these tools, novice physicians can diagnose virtual patient cases by identifying relevant symptoms, ordering lab tests, and reasoning about the nature of the underlying disease (Lajoie 2009). Each clinical case begins with the patient problem where learners review the patient history and gather evidence by highlighting relevant symptoms and other relevant information (which is sent to the evidence palette). Based on the collected evidence, students propose preliminary hypotheses (with the help of the Hypothesis Manager

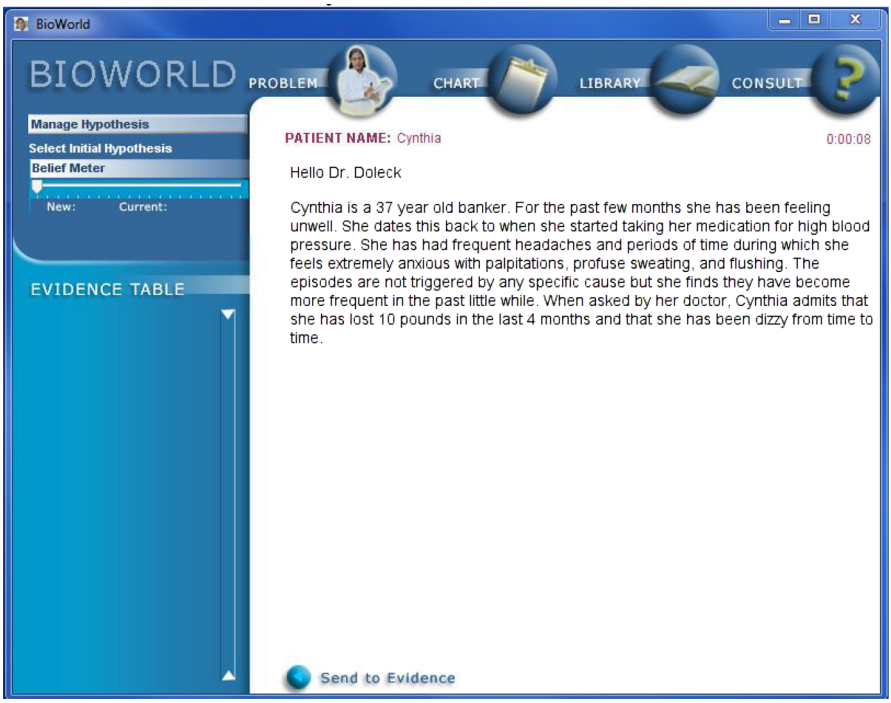

Fig. 1 A screenshot of the BioWorld interface

Tool) and report their corresponding confidence (as a percentage) in each hypothesis (via the Belief meter). In order to confirm or disconfirm a particular hypothesis, students can order a range of laboratory tests (the results of which are saved in the evidence palette). Help-seeking facilities are available via the Library (glossary of medical terminology and diagnostic testing procedures, as well as, the typical symptoms and transmission routes of a specific disease) and Consult tools (context-specific hints delivered in increasing order of specificity). This string of activities helps students arrive at a diagnosis. Upon submitting their final diagnosis, students are tasked with justifying their selection by sorting and prioritizing all the evidence used to arrive at the final diagnosis. Students receive individualized feedback on their solution based on an aggregated expert solution via the systems' dynamic assessment. The learning cycle ends with the student writing a final case summary of the patient's case; this summary is meant to capture the steps taken to solve the case, and how each step contributed toward the end diagnosis. An example of a line of diagnostic reasoning where Pheochromocytoma is the main hypothesis for what is wrong with Cynthia is illustrated in Fig. 2. The case description is as follows:

> Cynthia is a 37-year-old banker. For the past few months she has been feeling unwell. She dates this back to when she started taking her medication for high blood pressure. She has had frequent headaches and periods of time during which she feels extremely anxious with palpitations, profuse sweating, and

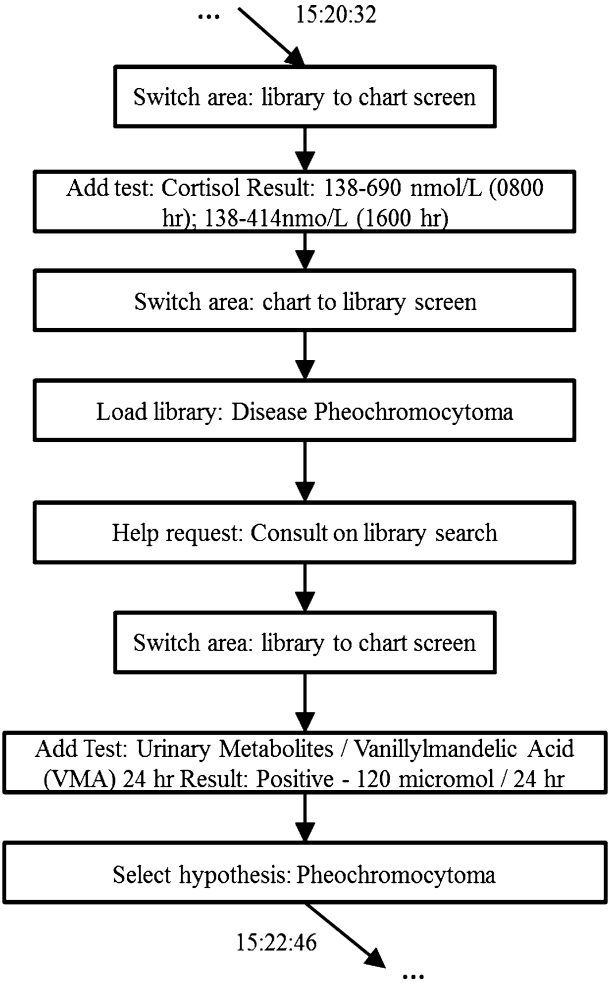

Fig. 2 A line of diagnostic reasoning where Pheochromocytoma is the main hypothesis

flushing. The episodes are not triggered by any specific cause but she finds they have become more frequent in the past little while. When asked by her doctor, Cynthia admits that she has lost 10 pounds in the last 4 months and that she has been dizzy from time to time.

Methods

Participant profile

Participation in the study was solicited through advertisements and newsletter at a North American research university. Thirty volunteer undergraduate students

participated in the study, and were compensated $20 at the completion of a 2-h study session. The convenience sample comprised 19 women (63 %) and 11 men (37 %), with an average age of 23 ($SD = 2.60$). All 30 participants were registered in the same classes, where 28 were medical students and 2 were dental students. The data for this study were collected as part of a larger project that investigated the antecedent factors that led to attention allocation toward feedback in the BioWorld environment (Naismith 2013). All participants consented to the use of the data collected for research purposes.

Experimental procedure

Participants completed both a demographic questionnaire and the achievement goal questionnaire. This was followed by a training session, where participants completed a training case, which allowed them to learn how to navigate and use the BioWorld system. Participants were also provided instructions on how to think aloud while solving cases in BioWorld. Following the training case, the actual experimental study began, where participants solved each of the three cases in BioWorld on an individual basis for 2 h. The three endocrinology cases were: Amy, Cynthia, and Susan Taylor. The correct diagnosis for each was diabetes mellitus (type 1), pheochromocytoma, and hyperthyroidism, respectively. The order of the cases was counterbalanced to mitigate practice effects. Upon completion of each case, participants completed a retrospective outcome achievement emotions questionnaire. The learners' processes and think alouds were recorded. As a measure of case difficulties, we used the results of an earlier study as a baseline; Gauthier et al. (2008) ascertained the difficulty levels of the various patient cases based on accuracy alone; the anticipated success rates (represented as percent accuracies) for the three cases, ordered from easiest to the most difficult, were Amy (94 %), Susan Taylor (78 %), and Cynthia (33 %).

Measures

In computer-based learning environments, one of the most common means of capturing learner–system interactions is via log files. As learners use BioWorld, the system captures learner actions, unequivocally attached to a learner, in log files (Fig. 3). Three types of performance metrics are captured: diagnostic efficacy (e.g., percentage of matches with experts), efficiency (e.g., number of tests ordered), and affect (e.g., confidence). The usage data include the attempt identifier (participant and case ID), a timestamp, the BioWorld space (e.g., chart), the specific action taken (e.g., add test), and details in relation to the action (e.g., (TSH) Result: 0.2 mU/L). An example of a line of diagnostic reasoning, as logged by the BioWorld system, is illustrated in Fig. 2.

Examining learner behavior

Although the topic of clinical reasoning has previously been examined extensively, the bulk of the studies focus on diagnosis correctness. However, understanding how

id	area	time	evidence	action
25117	history	12/10/10 14:56	splash - history screen switch	switch area
25118	history	12/10/10 14:58	1when she started taking her medication for	add evidence
25119	history	12/10/10 14:58	0frequent headaches	add evidence
25120	history	12/10/10 14:59	1extremely anxious	add evidence
25121	history	12/10/10 14:59	1palpitations	add evidence
25122	history	12/10/10 14:59	1profuse sweating	add evidence
25123	history	12/10/10 14:59	1flushing	add evidence
25124	history	12/10/10 14:59	0not triggered by any specific cause	add evidence
25125	history	12/10/10 14:59	0lost 10 pounds in the last 4 months	add evidence
25126	history	12/10/10 14:59	0dizzy	add evidence
25127	history	12/10/10 15:01	Panic Attack	add hypothesis
25128	history	12/10/10 15:01	Diabetes Mellitus (type II)	add hypothesis
25129	history	12/10/10 15:02	Pheochromocytoma	add hypothesis
25130	history	12/10/10 15:03	Hyperthyroid (Grave's disease)	add hypothesis
25131	history	12/10/10 15:04	Diabetes Mellitus (type II)	select hypothesis
25132	history	12/10/10 15:04	49%2- Diabetes Mellitus (type II)	change hypothesis conviction

Fig. 3 Snapshot of log file generated by BioWorld

learners arrived at the diagnosis can provide a more detailed examination of the diagnostic reasoning processes that lead to patient case solutions than diagnosis accuracy alone. The learning trajectory, i.e., actions taken by learners in solving a problem, holds considerable information that can be useful for modeling and improving learning systems. Biswas et al. (2010) recommend that shifting focus from the frequency and relevance of learner activities to considering the internal states and related learning strategies, can be useful in illuminating additional information for examining learning in learning systems. Recent research in computer-supported education has leveraged data mining techniques for examining varied problems (Baker and Yacef 2009; Romero and Ventura 2010). One method that has proven useful in investigating a range of problems is the HMMs (Rabiner 1989). HMMs have found widespread use in modeling sequential data in a range of contexts, yet the application of HMMs to usage data from computer-based learning environments is relatively new. Recent examples of application of HMMs in educational contexts include studies by: (a) Beal et al. (2007) who used HMM to ascertain the level of engagement in a tutoring system for high school mathematics to predict the subsequent actions in the tutoring environment; and (b) Jeong et al. (2008) who employed HMM in predicting transitions between learner activities in a teachable agent environment and exhibited the efficacy of HMMs in ascertaining learners' pattern of activities. Leveraging HMM analysis may facilitate considerable progression and improvements in learner modeling by providing an alternate understanding of learner activities that will lead to better adaptive environments.

As mentioned earlier, the actions that learners take to solve a problem, hold considerable information. As learners use BioWorld, the system captures user actions in log files. An example of actions taken in solving a real case, where Pheochromocytoma is the main hypothesis, is illustrated in Fig. 2. The set of actions

captured by the BioWorld system include: AE = 'add evidence'; AH = 'add hypothesis'; CHC = 'change hypothesis conviction'; LE = 'link evidence'; SEH = 'select hypothesis'; AT = 'add test'; SH = 'submit hypothesis'; P = 'prioritize'; EM = 'expert match'; FP = 'final priority'; LL = 'load library'; C = 'categorize'; RP = 'reprioritize'; SU = 'submit summary'; UL = 'unlink evidence'; DH = 'delete hypothesis'; SLC = 'select library category'; ASH = 'abort submit hypothesis'; SL = 'search library'; U = 'unprioritize'; RC = 'recategorize'.

Hidden Markov models

A hidden Markov model (HMM) is a double stochastic process with an underlying stochastic process that is unseen, but can only be seen through another set of visible stochastic processes that generate the observation sequence (Rabiner and Juang 1986). HMMs are formal foundations for making probabilistic models of linear sequence 'labeling' problems (Rabiner 1989; Rabiner and Juang 1986). A formal definition and additional details on HMMs can be found in Rabiner and Juang (1986). The applicability of HMMs is offered by a growing body of literature and has been applied to a range of applications including protein sequence modeling, profile searches, speech recognition, multiple sequence alignment, and regulatory site identification. A state diagram is used to visually represent the HMM model (Fig. 4). The rectangles (labeled 1 and 2) represent the possible states of a process and the arrows represent transitions between states. The label on each arrow represents the probability of the transition. At each step of the process, the model may generate an emission depending on which state it is in and then make a transition to another state. An important characteristic of HMM is that the next state depends only on the current state and not on the history of transition that leads to the current state. Since we are only given the observed sequence, this underlying state path is hidden—these are the residue labels that are needed to be inferred. The state path is essentially a hidden Markov chain. In the following subsections, we showcase the methodology employed for modeling problem solving in BioWorld using HMM.

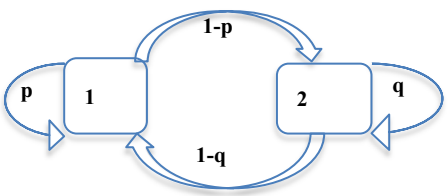

Fig. 4 HMM example

Experimental setup

One major affordance of intelligent tutoring systems is that learner interactions are captured, and can be analyzed to provide adaptive instruction. Such fine-grained user–system data can be beneficial in learner modeling and analytics. When solving a case in BioWorld, learners' actions and interactions with several tools, such as the library and chart during problem solving are logged by the system. The learner actions are time stamped and categorized according to both superordinate (e.g., Action: add lab test) and subordinate categories (e.g., Evidence: Thyroid Stimulating Hormone (TSH) Result: 0.2 mU/L).

From the log files, a line of diagnostic reasoning (i.e., sequence data) can be extracted. We extracted the sequence data for each of the endocrinology cases (Amy, Cynthia, and Susan Taylor) from the BioWorld logs. To this end, a parser (Doleck et al. 2014b) was used to extract the sequence data for the three cases from the log files. The HMM that captures learners' aggregated behavioral patterns was then generated using this sequence data. In our work, we elected to use the HMM generation tool (Biswas et al. 2010), which is used for modeling generative sequences, to calculate and generate the HMMs. The tool generates visualizations that illustrate the states, the action emission probabilities for each state, and the transitions between states. In this paper, we present the visualizations illustrating the state connections and the action emission probabilities for the three cases.

HMM results

We provide the results of applying HMM on the sequence data for three cases: Amy, Cynthia, and Susan Taylor. The learned HMM structure, using the HMM tool, for the three cases are given in Figs. 5, 6, and 7, respectively. The HMM structure is defined by a set of states, the transition probabilities among these states, and the emission probabilities for each state. The arrows in the generated HMM output illustrates the transition probabilities between and within states, and the percentage attached to each arrow indicates the transition probabilities. The analysis of the generated HMM involves investigating the interactions between and within states; these interactions elucidate how learners interact with the learning system. Once the possible transitions are plotted, researchers have the task of dissecting and interpreting the output. The qualitative aspect of model interpretation involves the naming/labeling of the various states based on the emission probabilities. The emission probabilities are the probability density functions that characterize each state. The goal of this exercise is to demonstrate the viability and utility of HMM for modeling learners' behaviors; therefore, it is out of scope of this research to delve into each behavioral pattern for the three cases. Thus, for the sake of simplicity of analysis, the state transition with the highest probability is discussed for each of the three cases.

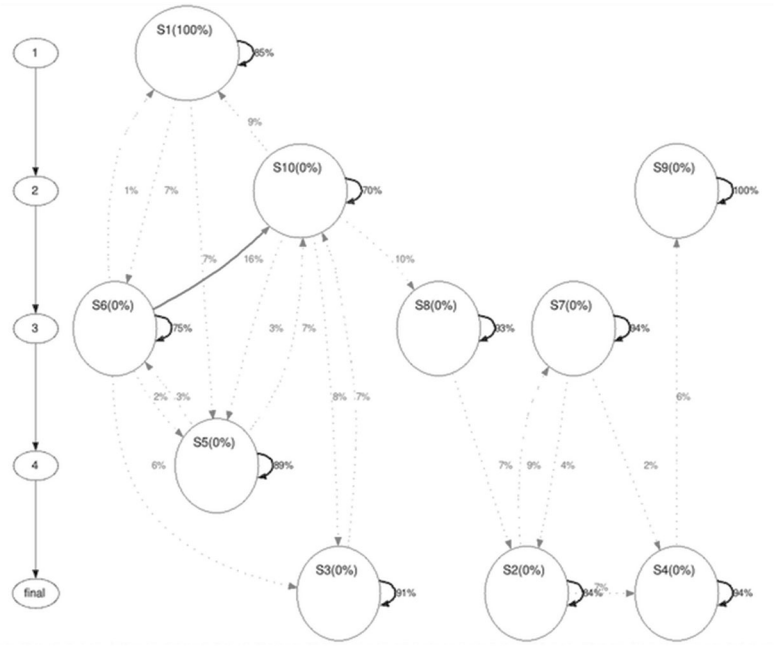

Fig. 5 Amy—HMM output

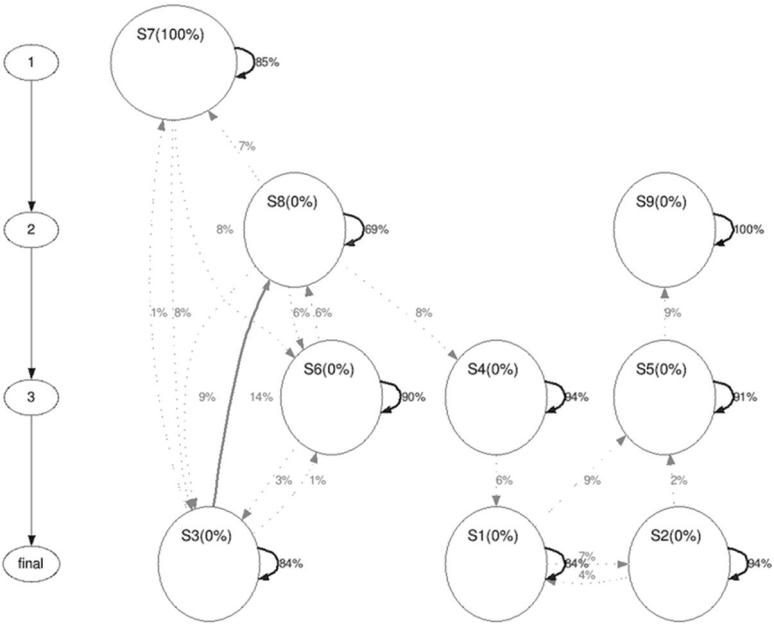

Fig. 6 Cynthia—HMM output

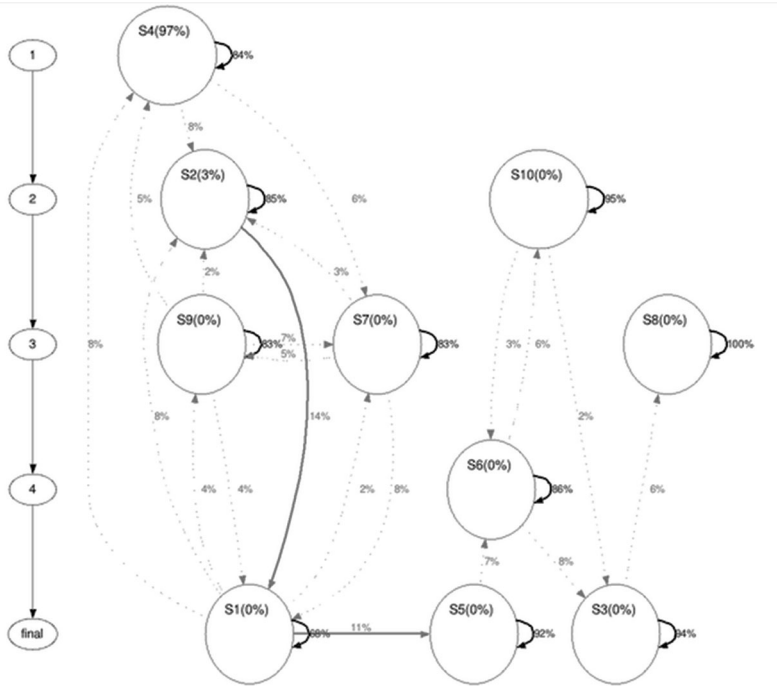

Fig. 7 Susan Taylor—HMM output

Amy

The generated structure for the Amy case has ten hidden states. The transition from S6 to S10 (16 %) had the highest probability. To understand the transition from S6 to S10, we consider the emission probabilities in each state (Fig. 8). For S6, the majority emission probabilities included: Add Hypothesis (0.73) and Change Hypothesis Conviction (0.16). For S10, the majority emission probabilities included: Change Hypothesis Conviction (0.61) and Submit Hypothesis (0.29). Thus, considering the descriptions of the distinct states undergirded by the emission probabilities help interpret the states and reveal the transition from S6 (diagnosis formulation) to S10 (Evaluation), which represents learners' transition from diagnosis formulation to evaluation.

Cynthia

The generated structure for the Cynthia case has nine hidden states. The transition from S3 to S8 (14 %) had the highest probability. To understand the transition from S3 to S8, we consider the emission probabilities in each state (Fig. 9). For S3, the majority emission probabilities included: Add Hypothesis (0.22) and Link Evidence (0.75). For S8, the majority emission probabilities included: Change Hypothesis Conviction (0.60) and Submit Hypothesis (0.29). Thus, the transition from S3 (evidence-driven diagnosis formulation) to S8 (Evaluation) represents learners' transition from evidence-driven diagnosis formulation to evaluation.

Susan Taylor

The generated structure for the Susan Taylor case has ten hidden states. The transition from S2 to S1 (14 %) had the highest probability. To understand the transition from S2 to S1, we consider the emission probabilities in each state (Fig. 10). For S2, the majority emission probabilities included: Add Hypothesis (0.23), Change Hypothesis Conviction (0.08), and Link Evidence (0.67). For S1, the majority emission probabilities included: Change Hypothesis Conviction (0.55) and Submit Hypothesis (0.38). Thus, the transition from S2 (evidence-driven diagnosis formulation) to S1 (evaluation) represents learners' transition from evidence-driven diagnosis formulation to evaluation.

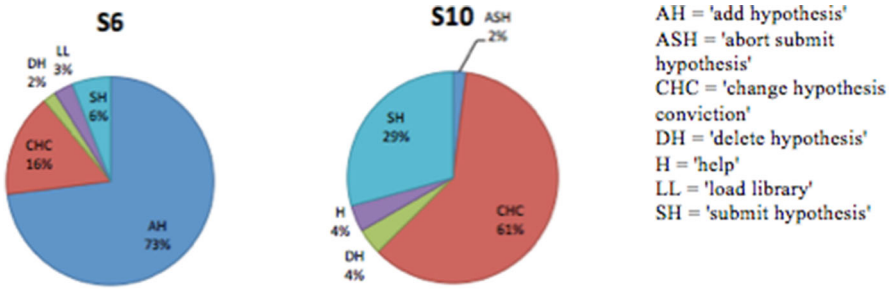

Fig. 8 Amy—state emission probabilities

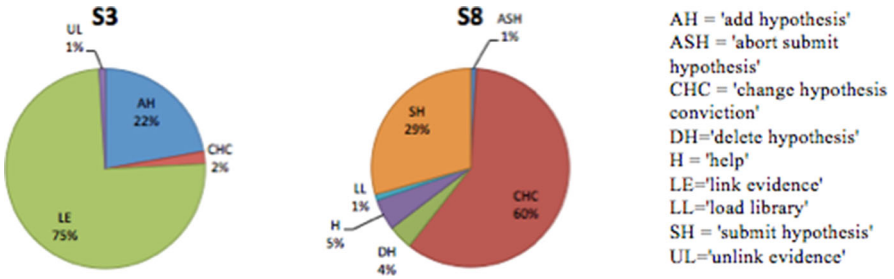

Fig. 9 Cynthia—state emission probabilities

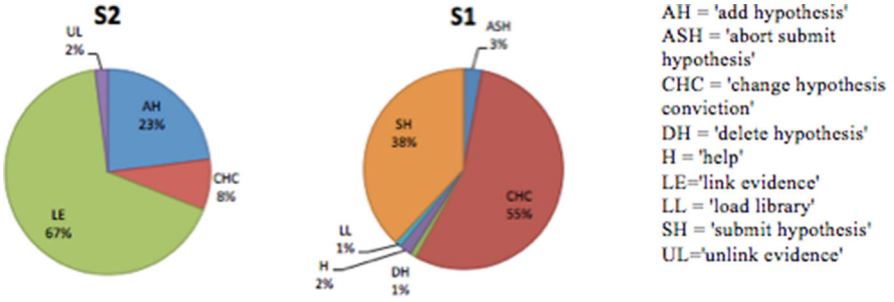

Fig. 10 Susan Taylor—state emission probabilities

Table 1 Comparison of the HMM derived for the three cases

	Amy	Cynthia	Susan Taylor
State transition with the highest probability	S6–S10	S3–S8	S2–S1
Transition description	Diagnosis formulation to evaluation	Evidence-driven diagnosis formulation to evaluation	Evidence-driven diagnosis formulation to evaluation

Modeling relations between learners and their actions can be used to harvest interesting and meaningful structural characteristics of the use of the learning material. The HMM analysis helped trace and illuminate the different decision learners made in solving a case in BioWorld. Several interesting observations can be derived from the comparison of the HMM derived for the three cases. Specifically, the examination of the HMM's across different cases that vary in levels of difficulty suggests important differences in the problem-solving trajectories. In particular, in the easier case, the Amy case, learners engaged in a pattern of formulating a hypothesis, and then evaluating their final diagnosis. On the other hand, in the more complex cases, Cynthia and Susan Taylor, the evidence that is linked to a particular hypothesis is also predictive of a shift from diagnosis formulation to evaluation. This suggests a proclivity for linking evidence in the more difficult cases. Comparison of the HMM derived for the three cases is highlighted in Table 1.

Discussion: modeling behaviors via HMM

Biswas et al. (2014) note that theory-driven metrics and context-driven hypotheses have in the past been the vehicles for assessing learning behaviors in learning systems; furthermore, they highlight the shift toward adoption of data mining techniques for examining learner behaviors. Much work has been done to collect and model user–system usage data, so that the information they contain can be most effectively used. Various mining methods can be applied to educational data to improve learning experiences and outcomes by increasing comprehension of learner behaviors. A deep and appropriate examination of learner actions in learning systems can lead to insights on behavioral patterns in learning and problem solving. The HMM analysis on the clinical reasoning data from BioWorld represents an effort toward revealing rich information about learner behaviors that can be valuable from both an instructional design and technology perspective. The findings from the HMM analysis indicate that behavioral patterns are indeed identifiable, where the HMM model did capture a pattern in the lines of diagnostic reasoning in the context of BioWorld. In particular, for the Amy case, learners engaged in a pattern of formulating a hypothesis, and then evaluating their final diagnosis. For the Cynthia and Susan Taylor cases, the evidence that is linked to a particular hypothesis is also predictive of a shift from diagnosis formulation to evaluation. As highlighted earlier, the three cases have varying difficulty levels; the Cynthia and Susan Taylor cases are the more difficult cases. Linking evidence is more prominent in the two

aforementioned cases, suggesting the proclivity of evidence linking in the more complex cases. The HMM results provide valuable insights into behavioral differences across the three cases under consideration. This provides some evidence of the case-specific (van der Vleuten and Swanson 1990; Fitzgerald et al. 1994; Doleck et al. 2015) nature of clinical reasoning. The assessment of these reoccurring behaviors stands to elucidate the important factors to task performance. Learner modeling is an important research topic in computer-supported education; the improvements and increased availability of mining methods have opened the area of learner modeling to vast possibilities, enabling increased knowledge discovery from educational data. Overall, HMMs can be used effectively in meeting the growing interest in the analysis of educational usage data to detect behavior patterns in learning environments. Future directions involve looking at other state transitions for each of the cases to understand other differences in behavior patterns.

Similar to the popularity of HMMs, text mining has also been widely applied in various disciplines as improving learning systems like ITSs continues to be of interest to learning technologists. Mining data logged by such systems can play a key role in system improvements. Text mining has been widely applied in extracting knowledge from various forms of text data. In the following sections, we illustrate our approach to improving the novice–expert overlay model in BioWorld by employing text mining in ascertaining the differences between case summaries written by experts and novices in BioWorld.

Improving the novice–expert overlay model

The ill-structured nature of problem solving involved in diagnosing diseases raises several challenges in modeling learning. Studies have documented the existence of multiple routes toward attaining the correct solution (Lajoie 2003, 2009; Gauthier and Lajoie 2014). Because of the knotty nature of such problems, these learning or problem-solving routes should be made explicit in order to represent common misconceptions in progressing along the trajectory toward competency. An overlay model "is a novice–expert difference model representing missing conceptions, often implemented as either an expert model annotated for missing items or an expert model with weights assigned to each element in the expert knowledge base" (Shute and Zapata-Rivera 2008, p. 284). A novice–expert overlay model (Shute and Zapata-Rivera 2012) has been applied in a variety of applications (e.g., Zapata-Rivera and Greer 2000). Similarly, BioWorld employs a novice–expert overlay system (Naismith and Lajoie 2010) to assess clinical reasoning during learning. The system dynamically assesses student performance against expert solution paths; the system compares similarities and differences in learners' solution path against an expert solution path, allowing learners to self-reflect on their problem-solving approach (Lajoie and Poitra 2014). For instance, after novices submit their final hypothesis, learners are able to compare their solution with an expert's solution (Fig. 11). Along with a comparison of the novice–expert on diagnosis and evidence, the system also provides a detailed explanation of an expert's reasoning. However, the current version of this user model omits an important aspect of the novices' clinical reasoning, i.e., the final written case summary. The written case summaries

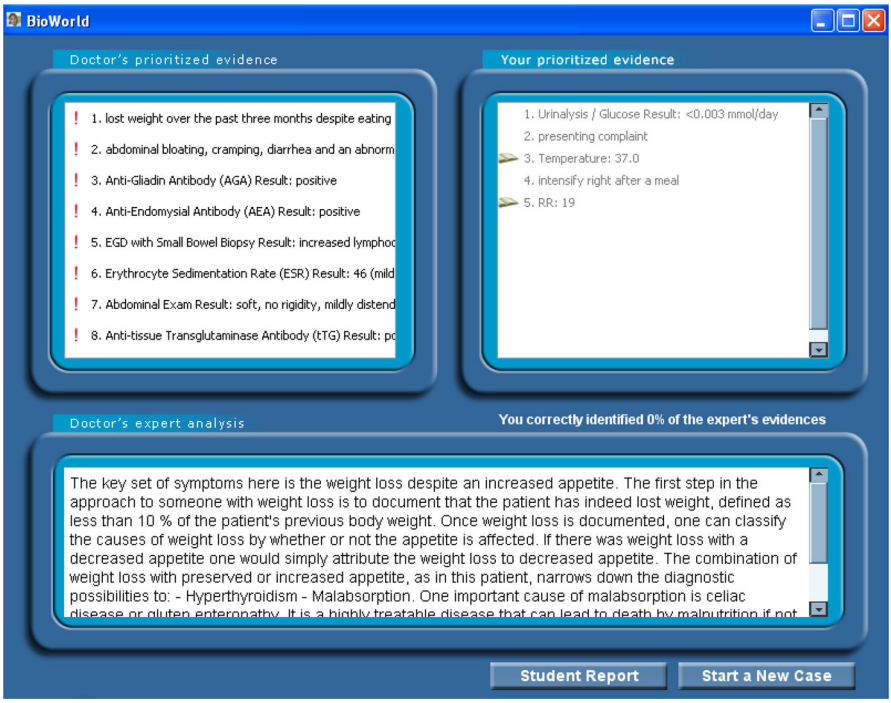

Fig. 11 Novice–expert evidence comparison

contain highlights apropos the symptoms, vital signs, and lab tests that were germane to diagnosing the patient case. In order to address this gap (case summaries divorced from the overlay model) and to augment the current novice–expert overlay model, this paper represents the first steps in exploring the efficacy of text categorization algorithms in differentiating between novice and expert case summaries written in BioWorld to augment the current novice–expert overlay model in BioWorld. The following subsections explicate the text-mining approach that set the scope of this work.

Text classification for novice–expert overlay model

Text mining has become an important means of knowledge-based discovery from varied data sources and types, and has assumed a central method with multifarious potential applications in varied fields ranging from commerce to education. Text classification is employed in a gamut of domains for varied purposes (Sebastiani 2002), such as news filtering and organization, document organization and retrieval, opinion mining, email classification, and spam filtering (Aggarwal and Zhai 2012). Text classification algorithms are also commonly used in intelligent tutoring systems for assessment purposes (McNamara 2007). When investigating the potential for augmenting the novice–expert overlay model in BioWorld, one natural idea was to consider the use of various text-mining classifiers for this task. A survey of the literature yields a gamut of classifiers for solving text classification problems;

for our study we limited our experiments to a set of commonly used text-mining algorithms (Aggarwal and Zhai 2012; Kibriya et al. 2004; Platt 1998). The goal of text classification is the classification of text into a number of predefined categories; this is achieved by first transforming the text (which are typically strings of characters) into a representation suitable for the learning algorithm and the classification task (Joachims 1998). Essentially, a general text classification problem involves assigning a new unseen text to one of the given classes. For example, a binary classification task for email messages would involve assigning a new unlabeled email as either spam or non-spam. The general approach toward solving such problems is to train classifiers utilizing labeled texts. Text-mining approaches can be incorporated into learning systems like BioWorld in order to improve such learning systems. We investigate the efficacy of text mining in the case summary classification problem. Specifically, we test an ensemble of machine learning algorithms for the novice–expert problem.

Data set: case summaries

In BioWorld, after learners receive individualized feedback on their solution based on an aggregated expert solution, the learners' final task in the diagnostic process involves writing a final case summary of the patient's case. The case summaries written by novices and experts serve as the data for our experiments. The case summaries were extracted from the log files generated by BioWorld. The final data for the novice–expert classification problem included a total of 74 case summaries, with 60 summaries written by novices and 14 summaries written by experts. In order to evaluate our method, we used the 74 case summaries for our experiments.

A sample of a case summary written by a student is presented below:

> Patient has elevated T3, T4; low TSH, and elevated thyroid stimulating antiglobulin. This is very suggestive of hyperthyroidism due to an autoimmune process. Listed symptoms (anxiety, weight loss, elevated HR, BP, tremor, sweating) all support this diagnosis.

A sample of a case summary written by an expert is presented below:

> 37-year-old female, presenting after starting high blood pressure pill with episodes of palpitations, flushing, and sweating. On exam, hypertensive, relative tachycardia. Labs revealed: normal TSH, T4, T3, glucose. Elevated free urinary cathecolamines but normal total. CT abdo normal.

Initial exploration

For the text classification task, we explored the use of an efficient optimization package and decided to utilize the popular Library for Support Vector Machines (LibSVM) for model selection (Chang and Lin 2011). LibSVM provides a parameter selection tool using cross-validation via a grid search. For a predictive task like classification, both the model and parameter estimation method selection are important for achieving high levels of performance. The task of learning

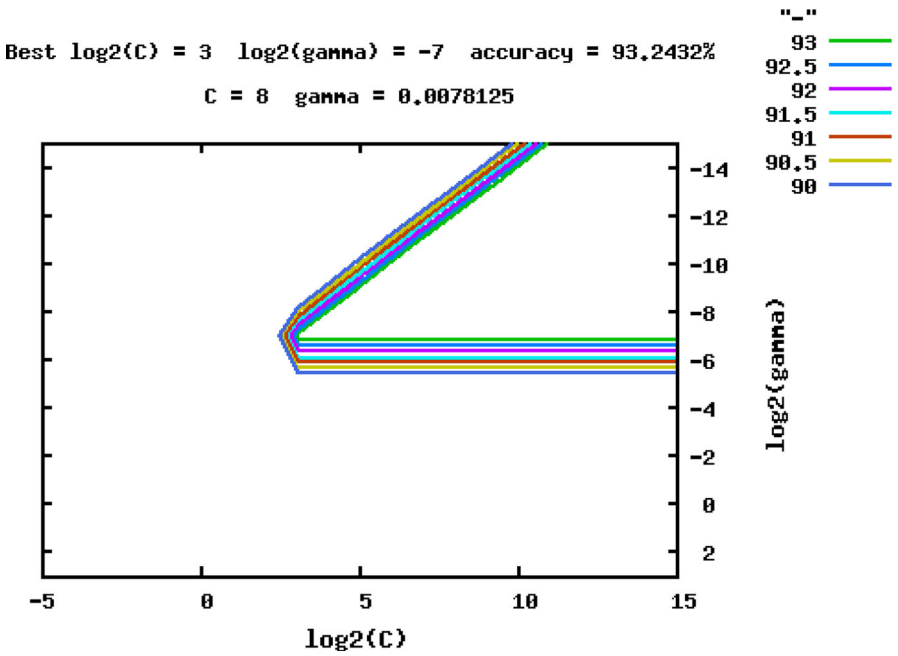

Fig. 12 LibSVM model file

involves choosing the model of optimal complexity and estimating the parameters from the data (Chapelle and Vapnik 2000). The case summaries were extracted from the log files generated by the BioWorld system. The performance (model selection) of the method is evaluated through the contour graph shown in the model file (Fig. 12). The accuracy rate achieved was 93.2432 % ($C = 8$; $\gamma = 0.00078124$).

Receiver operating characteristic (ROC) curves are generated by considering the rate at which true positives (vertical axis) accumulate versus the rate at which false positives accumulate (horizontal axis). Figure 13 illustrates the ROC curve of the LibSVM model. Generally, a data point in the upper left corner of the graph represents optimal high performance. AUC represents a measure of accuracy (area under the ROC curve). An area of 1 represents a perfect test, while an area of .5 represents a worthless test. The ROC curve generated for our model had an AUC of 0.9917.

Ensemble of text-mining algorithms

Our initial experiment with LibSVM demonstrated the feasibility of achieving highly accurate prediction rate in the Novice–Expert case summary classification task. In data mining, classification performance evaluation is an essential and key way to ascertain the optimal learning algorithm from an ensemble of algorithms. We wanted to extend this work by comparing an ensemble of algorithms. By

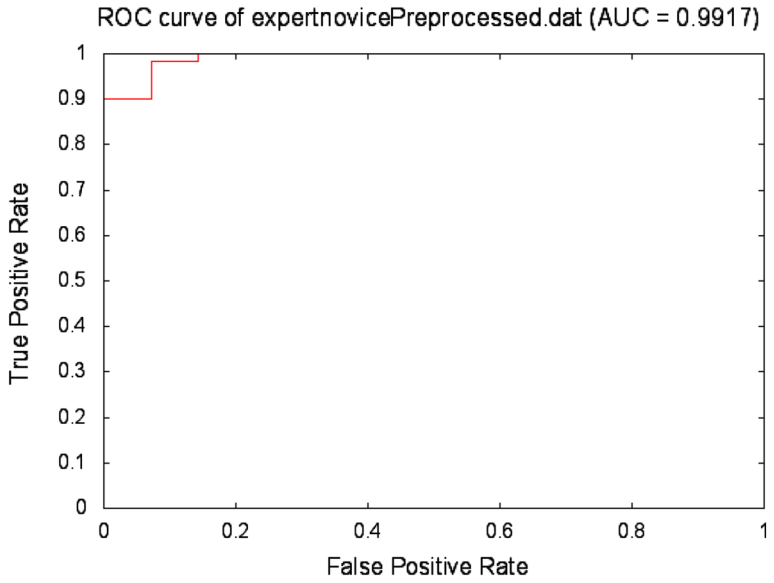

Fig. 13 ROC curve: LibSVM accuracy

empirically comparing a number of classifiers, we can ascertain the classifier that yields the best performance.

Experimental setup

The preprocessing filter *StringToWordVector* provided by WEKA (Hall et al. 2009) was used to extract feature vector from the summary texts. The default values were used except for the parameter that converted texts into lowercase. Stemmer is commonly used to reduce the feature size in text classification problems where the feature size tends to be in the order of tens of thousands. Because our dataset is fairly small in terms of both the size of each summary, as well as the total number of summaries written by expert and novices, the stemmer was not utilized. Stopwords were not removed—they were kept as a part of feature vector. Using the aforementioned preprocessing technique, word tokens were extracted as features. The resulting *arff* file format was then utilized for the experiments.

Results with all the features

We employed a number of classifiers available in WEKA to learn to classify the novice–expert dataset. The classifiers were applied on the generated dataset using 10-fold cross-validation. In k-fold cross-validation, generally, the dataset is segmented into k segments. The classifier is then tested for each of these segments and learned over other $k - 1$ segments. The initial results with all the features are presented in Table 2.

Table 2 Results with all the features

	SMO	Naïve bayes	Naïve bayes multinomial	J48	Random forest
Accuracy (correctly classified)	92.1053 %	86.8421 %	88.1579 %	78.9474 %	85.5263 %

Table 3 Results with feature selection

	SMO	Naïve bayes	Naïve bayes multinomial	J48	Random forest
Accuracy (correctly classified)	84.2105 %	82.8947 %	84.2105 %	80.2632 %	80.2632 %

Results with feature selection

Further experiments were conducted with feature selection to ascertain whether the classifier results could be improved. Feature selection techniques are commonly applied in data mining problems to extract the most discriminative features to reduce the feature space as well as to improve classifier performance. The Correlation-based Feature Subset Selection (*CfsSubsetEval*) technique provided by WEKA (we used the default values for the parameters) was utilized; this technique evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them (Hall 1999). The *GreedyStepwise* (Hall et al. 2009) forward search technique in WEKA was employed through the space of attribute subsets to search for the most discriminative feature set and rank them based on their individual predictive ability. A new dataset was generated using these features, and a classifier was applied to this dataset. Having applied the aforementioned feature selection technique to our dataset, the classification results are presented in Table 3.

Results with cost-sensitive classifiers

The different classifiers showed considerable promise in the novice–expert classification task. However, it should be noted that the data set under consideration is unbalanced; unbalanced datasets tend to have a bias toward the majority class and can lead to high accuracy rates even though the classifier may not be particularly good. Frank and Bouckaert (2006) note that it is common to have text classification problems that are unbalanced. Several approaches have been proposed to counter the problem of unbalanced datasets. Cost-sensitive classifiers are common approach used in scenarios when the text categorization problem is unbalanced. Thus to mitigate the unbalanced class problem with our dataset, we utilized WEKA's *CostSensitiveClassifiers* (Hall et al. 2009). We then ran the same classifiers and the results are presented in Table 4.

Table 4 Results with cost-sensitive classifiers

	SMO	Naïve bayes	Naïve bayes multinomial	J48	Random forest
Accuracy (correctly classified)	92.1053 %	86.8421 %	84.2105 %	78.9474 %	80.2632 %

Discussion: improving novice–expert overlay model

The initial results obtained with LibSVM (the accuracy rate of the prediction was 93.2432 %) showed promise. We then tested an ensemble of classifiers. Using the full features in the dataset, the various applied classifiers were able to achieve high prediction rates. SMO performed the best among the classifiers (92.1053 %). We then experimented with feature selection techniques to ascertain if the performance of the various classifiers could be improved. Interestingly, the results with feature selection did not result in improved performance, except for J48 (minor improvement from 78.9474 to 80.2632 %). The results obtained using feature selection were worse than results obtained using feature selection for the rest of the classifiers. We noted that the dataset used in our experiments was unbalanced. Thus, to mitigate that limitation, we decided to use cost-sensitive classifiers. Using the *CostSensitiveClassifiers*, there was no deterioration in the performance. Overall, the experimental results revealed the feasibility and efficacy of using text mining to achieve highly accurate predictions for the Novice–Expert case summaries classification task. In doing so, the linguistic features that differentiate case summaries written by novices and experts may inform revisions to the novice–expert overlay model, allowing the system to use the unstructured data obtained from case summaries to better tailor the content shown in the feedback panel. As such, the novice–expert overlay model may be further improved by examining the descriptors that characterize these learner behaviors and how they differ across lines of reasoning that are found to be correct as opposed to those that are incorrect. On the basis of our theoretical framework, learners are expected to benefit from instruction on how to adaptively monitor and control their own reasoning processes to avoid common pitfalls in their interpretation of the evidentiary data collected from the patient. The revisions made to the novice–expert overlay system allow the system to individualize instruction to the specific needs of different learners.

Conclusion

The present work examined the feasibility and efficacy of employing HMMs and text mining to usage data from a computer-based learning environment called BioWorld. On the one hand, the formulation and revision of diagnoses on the basis of evidence collected from the virtual patient are the most probable underlying sequence of behaviors that characterize problem solving in BioWorld. We are currently investigating this finding by exploring sequential patterns of behaviors

through the use of subgroup discovery method (Poitras et al. 2015). On the other hand, we have shown that the linguistic features that characterize case summaries written by learners are indicative of proficiency differences in synthesizing the evidence and diagnostic processes after solving problems in BioWorld.

Two techniques were used to determine both process and product differences of how medical students solve medical cases using BioWorld. In addition to diagnostic accuracy, we were interested in the processes learners use to diagnose a patient case. What steps do they take to reason about the disease? HMM was used to identify state probabilities and transitions within and between 3 cases. HMM was used to determine the types of learner patterns that existed for solving cases of varying difficulty level. Another important advance in computational analyses is the use of text-mining techniques to look at patterns in data, which we used to look at expert/novice differences in how physicians wrote patient case summaries at the completion of a case. These summaries are used in the real world as a hand-off method to inform the next physician about the patient case.

The goal of the first part of our paper was to employ the use of HMMs for examining the learner behaviors across three different virtual patient cases. The findings demonstrate that the HMM model was effective at capturing patterns in the lines of diagnostic reasoning that mediate performance in solving problems in the context of BioWorld. More specifically, the results exhibited that learners' behaviors could be modeled and analyzed effectively with HMMs. Similar to previous examples of successful application of HMMs in solving various problems, the results from our explorations of learner behaviors exhibited the utility of HMMs in discover, detection, explication, and comparison of learner behaviors. However, there are some limitations for this approach. Firstly, there is some subjectivity in model explication; the derived states of the model are assigned labels/meaning by the researchers using qualitative analysis. One way to mitigate this limitation is to have domain experts involved in this stage of the analysis. Secondly, large data sets are usually required to produce generalizable models; our data set is fairly small and thus, the next logical step is to collect more observations and retest the model. Our research demonstrates that using HMMs in this particular context can illustrate how novices' problem-solving actions differ from experts. This contribution is part of a larger ongoing series of studies to better understand learners' behaviors in BioWorld; a forthcoming contribution proposes the use of process mining for modeling learner behaviors (Doleck et al., in prep).

We have shown empirically that the ensemble of text-mining algorithms is especially useful in generating high predictive accuracy for the classification task. As such, the proposed case summary classifier allows the system to differentiate between case summaries written by novices and experts, which suggests that these case summaries are not only recognizably different by a machine, but also that their constituent content is different in some important manner. The novice–expert case summary classification task represents an important step in broadening the nature of the information that is processed by the current version of the novice–expert overlay model embedded in the BioWorld. This model is currently limited to the structured data that is collected as learners interact with tools embedded in the interface; however, the unstructured texts that constitute the case summaries may provide a

wealth of information that is unaccountable by such tools, for instance, the patient management plan or a justification for a differential diagnosis. One of the most important limitations of the proposed revisions to the algorithms that underlie the novice–expert overlay system is its scalability. It is yet to be determined whether the model performs as well for novel examples of case summaries, or if the parameters are applicable for case summaries written in relation to other cases that exhibit variations in symptoms and vital signs indicative of the same disease. Although the extensiveness of the training dataset is an important factor in addressing this challenge, we maintain that the current method benefits the learner model embedded in the system by targeting complementary channels of data about the learner. The current version of the overlay model is severely limited in its inferential capabilities by relying solely on evidence items, including highlighted symptoms and vital signs as well as lab test procedures that warrant the main hypothesis for the case under investigation. In order to build more scalable models, we suggest that establishing common standards for logging case summary data across the field is necessary in order to attain sufficiently large labeled datasets. We call for researchers to establish such standards for intelligent tutoring systems in the medical domain to enable further progress in this area.

The findings of this study suggest that sequential and linguistic features extracted from the log-file database of intelligent tutoring systems are a meaningful source of information to differentiate the mechanisms that underlies superior performance in the medical domain. One promising line of research involves fine-grained examinations of case summaries in order to better understand how novices write case summaries differently than experts. A sentence- or proposition-level analysis may lead the novice–expert model to make detailed recommendations regarding the type of different feedback that most benefit a particular learner. Alternatively, the features of the case that are most commonly reported by both novices and experts alike should be considered as self-evident, and not worthy of extensive consideration beyond the feedback that is currently delivered by the system, and which has been shown to be effective in terms of promoting learning. In future studies, we will focus our efforts on applying the text classifiers that were found to be successful in the current study to a training dataset that has been labeled at a fine-grained level with the aim of evaluating the impacts on classification accuracy.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 77–128). New York: Springer.

Anderson, J., Corbett, A., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*(2), 167–207.

Anderson, J., & Gluck, K. (2001). What role do cognitive architectures play in intelligent tutoring systems? In D. Klahr & S. Carver (Eds.), *Cognition & Instruction: 25 years of progress* (pp. 227–262). Mahwah, NJ: Lawrence Erlbaum.

Baker, R. S. J. D. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of ACM CHI 2007: computer-human interaction*, 1059–1068.

Baker, R. S. J. D., Corbett, A. T., Roll, I., Koedinger, K. R., Aleven, V., Cocea, M., et al. (2013). Modeling and studying gaming the system with educational data mining. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 97–116). New York: Springer.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Beal, C. R., Mitra, S., & Cohen, P. (2007a). Modeling learning patterns of students with a tutoring system using hidden Markov models. In R. Lucking, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education* (pp. 238–245). Amsterdam: IOS Press.

Beal, C. R., Walles, R., Arroyo, I., & Woolf, B. P. (2007b). On-line tutoring for math achievement testing: A controlled evaluation. *Journal of Interactive Online Learning, 6*(1), 43–55.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology-Enhanced Learning, 5*(2), 123–152.

Biswas, G., Kinnebrew, J. S., & Segedy, J. R. (2014). Using a Cognitive/Metacognitive Task Model to analyze Students Learning Behaviors. In *Proceedings of the 16*th *International conference on human-computer interaction*. Crete, Greece.

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. *Proceedings of the 1st international conference on learning analytics and knowledge (lak '11)* (pp. 110–116). New York,: ACM.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16. doi:10.3102/0013189x013006004.

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27.

Chapelle, O., & Vapnik, V. (2000). *Model selection for support vector machines. Advances in neural information processing systems* (Vol. 12). Cambridge, MA: MIT Press.

Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Model and User-Adapted Interaction, 19*(4), 341–385.

Collins, A. (2006). Cognitive apprenticeship. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 47–60). NY: Cambridge University Press.

Craig, S., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with Autotutor: Applying the facial action coding system to cognitive–affective states during learning. *Cognition and Emotion, 22*(5), 777–788.

Dodds, P., & Fletcher, J. D. (2004). Opportunities for new "smart" learning environments enabled by next-generation web capabilities. *Journal of Educational Multimedia and Hypermedia, 13*(4), 391–404.

Doleck, T., Jarrell, A., Chaouachi, M., Poitras, E., & Lajoie, S. (in prep). A tale of three cases: Examining accuracy, efficiency, and process differences in diagnosing virtual patient cases.

Doleck, T., Basnet, R. B., Poitras, E., & Lajoie, S. (2014). BioWorldParser: A suite of parsers for leveraging educational data mining techniques. In *Proceedings of 2nd IEEE International Conference on MOOCs, Innovation & Technology in Education* (MITE), (pp. 32–35), India: IEEE. doi: 10.1109/MITE.2014.7020236

Doleck, T., Basnet, R. B., Poitras, E., & Lajoie, S. (2014). Augmenting the novice-expert overlay model in an intelligent tutoring system: Using confidence-weighted linear classifiers. In *Proceedings of IEEE International Conference on Computational Intelligence & Computing Research (IEEE ICCIC)*, (pp. 87–90), India: IEEE

Doleck, T., Jarrell, A., Poitras, E., & Lajoie, S. (2015). Towards investigating performance differences in clinical reasoning in a technology rich learning environment. In Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M. F. (Eds.), *Artificial intelligence in education* (pp. 567–570). Lugano: Springer International Publishing. doi: 10.1007/978-3-319-19773-9_63

Durlach, P. J. & Ray, J. M. (2011). *Designing adaptive instructional environments: Insights from empirical evidence*. Army Research Institute Technical Report 1297. U.S. Army Research Institute, Arlington, VA.

Fitzgerald, J., Wolf, F., Davis, W., Barclay, M., Bozynski, M., Chamberlain, K., et al. (1994). A preliminary study of the impact of case specificity on computer-based assessment of medical student clinical performance. *Evaluation and the Health Professions, 17*(3), 307–321. doi:10.1177/016327879401700304.

Frank, E., & Bouckaert, R. R. (2006). Naive Bayes for text classification with unbalanced classes. In *Proceedings of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 503-510. Berlin: Springer.

Gauthier, G., & Lajoie, P. S. (2014). Do expert clinical teachers have a shared understanding of what constitutes competent case-based reasoning? *Instructional Science, 42*(4), 579–594.

Gauthier, G., Lajoie, P. S., Naismith, L., & Wiseman, J. (2008). Using expert decision maps to promote reflection and self-assessment in medical case-based instruction. In *Proceedings of Workshop on the Assessment and Feedback in Ill-Defined Domains at ITS*, Montréal, Canada.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education, 48*, 612–618.

Graesser, A., Lu, S., Jackson, G., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. (2004). Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 180–192.

Hall, M. A. (1999). *Correlation-based feature subset selection for machine learning*. (doctoral dissertation). Department of Computer Science, University of Waikato, New Zealand.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*(1), 10–18.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., & Schwartz, D. (2008). Using hidden Markov models to characterize student behaviors in learning-by-teaching environments. In *Proceedings of Intelligent Tutoring Systems: Vol. 5091. Lecture Notes in Computer Science* (pp. 614–625). Montreal: Springer.

Joachims, T. (1998). Text Categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial naïve bayes for text categorization revisited. In G. I. Webb & X. Yu (Eds.), *Advances in artificial intelligence* (pp. 488–499). Berlin, Heidelberg: Springer.

Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The cambridge handbook of the learning sciences* (pp. 61–78). Cambridge: Cambridge University Press.

Lajoie, S. P. (2003). Transitions and trajectories for studies of expertise. *Educational Researcher, 32*(8), 21–25.

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from avionics and medicine. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 61–83). Cambridge: Cambridge University Press.

Lajoie, S. P., Naismith, L., Hong, Y. J., Poitras, E., Cruz-Panesso, I., Ranellucci, J., et al. (2013). Technology rich tools to support self-regulated learning and performance in medicine. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 229–242). New York: Springer.

Lajoie, S. P., Poitras, E. G., Doleck, T., & Jarrell, A. (2015). Modeling metacognitive activities in medical problem-solving with bioworld. In A. Peña-Ayala (Ed.), *Metacognition: Fundaments, applications, and trends* (pp. 323–343). New York: Springer Series: Intelligent Systems Reference Library.

Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., William, W. C., Stylianides, G. J., & Koedinger, K. R. (2013). Cognitive anatomy of tutor learning: Lessons learned with SimStudent. *Journal of Educational Psychology, 105*(4), 1152–1163.

McNamara, D. S. (2007). IIS: A marriage of computational linguistics, psychology, and educational technologies. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the twentieth international florida artificial intelligence research society conference* (pp. 15–20). Menlo Park, California: The AAAI Press.

Merrill, M. (2002). First principles of instruction. *Educational Technology Research and Development, 50*(3), 43–59. doi:10.1007/bf02505024.

Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education, 13*(2), 173–197.

Naismith, L. (2013). *Examining motivational and emotional influences on medical students' attention to feedback in a technology-rich environment for learning clinical reasoning (Unpublished doctoral dissertation)*. Montreal: McGill University.

Naismith, L., & Lajoie, S. P. (2010). Using expert models to provide feedback on clinical reasoning skills. In *proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 242–244).
Park, O. C., & Lee, J. (2004). Adaptive instructional systems. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (2nd ed., pp. 651–684). Mahwah, NJ: Lawrence Erlbaum.
Platt, J. C. (1998). A fast algorithm for training support vector machines. *Technical Report MSR-TR-98-14*.
Poitras, E. G., Lajoie, S. P., Doleck, T., & Jarrell, A. (in press). Subgroup discovery with user interaction data: An empirically guided approach to improving intelligent tutoring systems. *Educational Technology & Society*.
Poitras, E., Lajoie, S., & Hong, Y.-J. (2012). The design of technology-rich learning environments as metacognitive tools in history education. *Instructional Science, 40*(6), 1033–1061.
Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286. doi:10.1109/5.18626.
Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine, 3*(1), 4–16.
Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.
Shute, V. J., & Zapata-Rivera, D. (2008). Adaptive technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 277–294). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.
Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach & A. Lesgold (Eds.), *Adaptive technologies for training and education*. New York, NY: Cambridge University Press.
van der Kleij, F., Eggen, T., Timmers, C., & Veldkamp, B. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*(1), 263–272.
van der Vleuten, C., & Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58–76. doi:10.1080/10401339009539432.
Vanlehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.
Vanlehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education, 15*, 147–204.
Zapata-Rivera, D., & Greer, J. (2000). Inspecting and visualizing distributed Bayesian student models. In *Proceedings of Intelligent Tutoring Systems* (pp. 544–553).

Tenzin Doleck is a doctoral student in the Learning Sciences program at McGill University, Montréal, Canada, and is currently a member of the Advanced Technologies for Learning in Authentic Settings (ATLAS) Lab. His research interests include learning design & technology, learning analytics, pedagogical agents, intelligent tutoring systems, and STEM education.

Ram B. Basnet is an Assistant Professor in the Department of Computer Science at Colorado Mesa University. He received his Bachelor of Science in Computer Science from Colorado Mesa University and then went on to earn his Master of Science and PhD, both in Computer Science, from New Mexico Tech. His research interests include information assurance, phishing detection, machine learning, and data mining.

Eric G. Poitras is an Assistant Professor for Instructional Design and Educational Technology in the Department of Educational Psychology at the University of Utah. He graduated from McGill University, where he earned a graduate degree in the Learning Sciences and worked as a postdoctoral researcher at the Learning Environments Across Disciplines research partnership. His research aim to improve the adaptive capabilities of instructional systems and technologies designed as cognitive and metacognitive tools as a means to foster self-regulated learning. In particular, the capabilities of intelligent tutoring systems and augmented reality applications to capture and analyze learner behaviors in order to deliver

the most suitable instructional content in domain areas such as medical diagnostic reasoning, historical thinking, and teacher professional development.

Susanne P. Lajoie received her Doctorate from Stanford University in 1986. She is a Canadian Research Chair Tier 1 in Advanced Technologies for Learning in Authentic Settings in the Department of Educational and Counselling Psychology at McGill University. She is a Fellow of the American Psychological Association, appointed for her outstanding contributions to the field of Psychology as well as an Inaugural Fellow of the American Educational Research Association. Dr. Lajoie is a recipient of the McGill Carrie Derick Award for graduate supervision and teaching. Dr. Lajoie has engaged in a wide array of innovative research and scholarly activities where she designs technology-rich learning environments for educational and professional practices. She uses a cognitive approach to identify learning trajectories that help novice learners become more skilled in the areas of science, statistics, and medicine. She has designed effective computer-based learning environments in these domains based on her research findings. She had been invited to present her research worldwide including Australia, France, Germany, Hong Kong, Korea, Singapore, Spain, Sweden, Taiwan, Mexico, the UK, and the Ukraine. She has numerous publications including 2 volumes on Computers as Cognitive tools published by Erlbaum. These volumes have highlighted the necessity for theory-driven design of technologies for education and training.