Intelligent Tutoring Technologies for Ill-Defined Problems and Ill-Defined Domains

Proceedings of the 4th International Workshop on Intelligent Tutoring Systems and Ill-Defined domains held at the 10th International Conference on Intelligent Tutoring Systems (ITS 2010) in Pittsburgh, Pennsylvania, U.S.A.

Collin Lynch, Dr. Kevin Ashley, Prof. Tanja Mitrovic,
Dr. Vania Dimitrova, *Dr. Niels Pinkwart*, & *Dr. Vincent Aleven*

(workshop co-chairs)

Preface

Intelligent tutoring systems, and intelligent learning environments support learning in a variety of domains from basic math and physics to legal argument, and hypothesis generation. These latter domains are ill-defined referring to a broad range of cognitively complex skills and requiring solvers to structure or recharacterize them in order to solve problems or address open questions. Ill-defined domains are very challenging and have been relatively unexplored in the intelligent learning community. They usually require novel learning environments which use nondidactic methods, such as Socratic instruction, peer-supported exploration, simulation and/or exploratory learning methods, or informal learning techniques in collaborative settings

Ill-defined domains such as negotiation, intercultural competence, and argument are increasingly important in educational settings. As a result, interest in ill-defined domains has grown in recent years with many researchers seeking to develop systems that support both structured problem solving and open-ended recharacterization. Ill-defined problems and ill-defined domains however pose a number of challenges. These include:

- Defining viable computational models for open-ended exploration coupled intertwined with appropriate meta-cognitive scaffolding;
- Developing systems that may assess and respond to fully novel solutions relying on unanticipated background knowledge;
- Constraining students to productive behavior in otherwise underspecified domains;
- Effective provision of feedback when the problem-solving model is not definitive and the task at hand is ill-defined;
- Structuring of learning experiences in the absence of a clear problem, strategy, and answer;
- Developing user models that accommodate the uncertainty, dynamicity, and multiple perspectives of ill-defined domains; and
- Designing interfaces that can guide learners to productive interactions without artificially constraining their work.

These challenges must be faced in order to develop effective tutoring systems in these attractive, open, and important arenas.

This is the fourth international workshop on Intelligent Tutoring Systems and Ill-Defined domains. Previous workshops have been held at ITS 2006 in Jhongli Taiwan [1], AIED 2007 in Monterey California U.S.A. [2] and at ITS 2008 in Montréal, Canada [3]. As with it's predecessors, this workshop brought together a researchers focusing on domains ranging from strategic planning to scientific and legal argumentation, and employing techniques ranging from textual analysis to agent-based simulations. This diversity supports useful cross-pollination across domains both in the identification of domain-general tutoring techniques and the exploration of shared pedagogical goals.

We thank the reviewers for their efforts, and the authors and participants for their excellent contributions.

Collin Lynch
Kevin Ashley
Tanja Mitrovic
Vania Dimitrova
Niels Pinkwart
Vincent Aleven

References

[1] Aleven, V., Ashley, K.D., Lynch, C., Pinkwart, N. (eds.): Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, vol. 1. ITS Society., Jhongli, Taiwan (2006), at the 8th International Conference on Intelligent Tutoring Systems.

[2] Aleven, V., Ashley, K.D., Lynch, C., Pinkwart, N. (eds.): Proceedings of the Workshop on AIED Applications for Ill-Defined Domains., vol. 1. International Society of AI in Education, Marina Del Rey California (2007), at the 13th International Conference on Artificial Intelligence in Education

[3] Aleven, V., Ashley, K.D., Lynch, C., Pinkwart, N. (eds.): Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains: Assessment and Feedback in Ill-Defined Domains, vol. 1. ITS Society, Montréal, Canada. (2008), held at the 9th International Conference on Intelligent Tutoring Systems.

Table of Contents

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V., Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 5-8)

Borderline Cases of Ill-definedness – and How Different Definitions Deal with Them

Kevin Ashley[1], Collin Lynch[2], Niels Pinkwart[3], and Vincent Aleven[4]

[1]University of Pittsburgh, Learning Research and Development Center,
[2]University of Pittsburgh, Intelligent Systems Program, Pittsburgh, PA, USA
Pittsburgh, PA, USA
[3]Clausthal University of Technology, Department of Informatics, Germany
[4]Carnegie Mellon University, HCI Institute, Pittsburgh, PA, USA
{ashley,lynch}@pitt.edu

Abstract. In prior work, we have proposed definitions of ill-defined problems and domains that focus on the need for framing or recharacterizing the problem in order to generate plausible solutions, where the recharacterizations and solutions are themselves subject to debate. In this paper we elaborate on the framing or recharacterization component of this definition in light of some plausible counterexamples and in contrast with some recent alternative definitions. While the question of defining ill-defined problems and domains is itself ill-defined, we conclude that our formerly proposed definition withstands the counterarguments.

Keywords: ill-defined problems; ill-defined domains, ITSs for ill-defined domains

1 Introduction

In educating students, many problems assigned to them are well-defined. That is, the problem as stated is fairly constrained, the relevant knowledge for solving it is more-or-less apparent and available to the solver, and the answer is unambiguously right or wrong. For instance,

> **Car Example:** A 1000kg car rolls down a 30 degree hill. What is the net force acting on the car?

By contrast, here is an example of an ill-defined problem of a type that instructors, instructional designers, and ITS designers encounter on a daily basis:

> **Best-Pedagogical-Solution-Example:** What is the most pedagogically effective solution of the problem, "A 1000kg car rolls down a 30 degree hill. What is the net force acting on the car?"

In solving this latter problem, various questions necessarily arise that make it ill-defined, questions such as,

- "What does "pedagogically effective" mean?",
- "Effective in what pedagogical context, with what type of students, and for what purpose?", and

- "What are the criteria for assessing a proposed solution's pedagogical effectiveness, and how will solutions be compared?"

Arguments will follow proposing definitions of concepts, criteria for assessing solutions, methods for comparing the solutions in relation to the criteria, etc.

In nearly every domain of human endeavor and in nearly every classroom in higher education, the interesting questions are ill-defined. Certainly, in teaching students ITS design, the goal is learning to solve ill-defined problems like the latter example, not just well-defined problems like the former. Arguably, however, ITS technology is better suited to teaching well-defined problems. The Car Example is typical of questions assigned to high school physics students class. This is the sort of problem that ITS technology has proven it can handle effectively. Students' solutions can readily be compared with expert solutions and used to model the level of students' understanding of the problem and to trigger progressively more explicit hints.

By contrast, in solving an ill-defined problem, the relevant constraints on the problem are not readily apparent from the mere statement of the problem, but the solver must uncover the relevant constraints through a process of exploration. For instance, the solver must explore plausible answers to the questions about what "pedagogically effective" means, and what the relevant context, goals, and criteria are for addressing the problem. Often, the answers are arguments in free-form text justifying one solution over others. Different solvers may frame the ill-defined problem differently according to their knowledge, beliefs, and attitudes, and thus may generate different arguments, many of which may be quite defensible. Since an ITS cannot readily interpret a student's natural language arguments, updating a student model and using it to select guidance, feedback and new problems to tutor students, are especially problematic for a computer tutor teaching ill-defined problem solving.

Nevertheless, in recent years, there has been a growing appreciation in the ITS community of the pedagogical importance of developing methods for addressing ill-defined problems. See, e.g., (Aleven, et al. 2006; 2007; 2008; Pinkwart, et al. 2010). This may reflect the fact that, as students progress through the educational system, the problems at the center of their education become increasingly ill-defined, perhaps because ill-defined problem-solving lies at the core of professional practice in fields not only like law, business, government, and ethics, but even in scientific fields as one moves to the research frontiers of theoretical development and experimental design. In addition, many decisions one makes when *writing* an essay or research article tend to be ill-defined.

Given these considerations, the question of how one should define ill-defined problems and domains, both generally and in the ITS context, becomes increasingly pressing. This question is not likely to be resolved easily or for all time; the question of how to define ill-definedness is itself ill-defined and subject to argument exactly as described above. Indeed, it is useful and enlightening for the community of scholars at this Workshop on Intelligent Tutoring Technologies for Ill-Defined Problems and Ill-Defined Domains to debate this issue and to consider alternative plausible definitions and the arguments for and against them.

To that end, in the remainder of this paper, we present in Section 2 alternative recently proposed definitions of ill-definedness including our own definitions. In Section 3, we consider some plausible counterexamples to our definition and use them in Section 4 to elaborate on the meaning of crucial concepts in our definition as well

as to compare and contrast the various recent definitions. Our goal is to encourage a discussion of the definitional issues at the workshop that will help all of us in this community of interest to consider which definitions work best, generally and in the ITS context, in light of the underlying issues.

2 Definitions of ill-definedness

In this section, we present four alternative recently proposed definitions of ill-definedness (for a history of sources of definitions of ill-defined problems and domains, see Lynch et al. 2010). In prior work, we have proposed the following definitions of ill-defined problems and ill-defined domains (Lynch et al. 2006; 2010):

> [1.1] A *problem* is ill-defined when essential concepts, relations, or solution criteria are un- or under-specified, open-textured, or intractable, requiring a solver to frame or recharacterize it. This recharacterization, and the resulting solution, are subject to debate.

> [1.2] Ill-defined *domains* lack a single strong domain theory uniquely specifying the essential concepts, relationships, and procedures for the domain and providing a means to validate problem solutions or cases. A solver is thus required to structure or recharacterize the domain when working in it. This recharacterization is subject to debate.

A central feature of our definitions is the requirement that a problem solver frame or recharacterize the problem some or all of whose essential concepts, relations or solution criteria are under-specified, open-textured, or intractable. By "framing" or "recharacterizing" the problem, we mean restating or refining aspects of the problem in order to align it with specific domain concepts; redefining existing rules according to the present goals; clarifying the solution criteria; or analogizing the problem to and distinguishing it from prior examples.

Another definition that has received some currency in recent ITS work summarizes a set of criteria proposed by Simon (1978, p. 286).

> [2] An ill-structured problem is defined by Simon (citing Simon 1973) as one that is complex, with indefinite starting points, multiple and arguable solutions, or unclear strategies for finding solutions (citing Fields 2006). (Nkambou et al. 2008)

A further possible strategy is to define ill-defined problems as the complement, in effect, of a definition of well-defined problems as in Le and Menzel (2008).

> [3] [R]equirements … have been proposed as criteria a problem must satisfy in order to be regarded as well-defined: 1) a start state is available; 2) there exist a limited number of relatively easily formalized transformation rules; 3) evaluation functions are specified and 4) the goal state is unambiguous (citing Jonassen, et al. 1999). If one or several of these conditions is violated, the problem is considered

> ill-defined (citing Ormerod 2006). … However, "the boundary between well-defined and ill-defined problems is vague, fluid and not susceptible to formalization" (citing Simon 1973).

A very similar "complementary" definition, also in terms of a state space view, may be found in Mitrovic and Weerasinghe (2009). These authors define as well-defined, tasks or problems having clear start states, goal states, transformations (i.e., problem-solving procedures) that are known to the decision-maker and a correct solution. By negative implication, an ill-defined task or problem has underspecified start states, transformations, or solution criteria. Specifically, the authors emphasize two dimensions for discussing ill-definedness, the definedness of the task and of the domain, which can order the domains/tasks along a continuum, from well- to ill-defined. Their implicit definition is apparent in the following:

> [4] Although the ER model [Entity-Relationship data model] itself is well-defined, the task of developing an ER schema for a particular database (i.e. conceptual database design) itself is ill-defined: the initial state (i.e. the set of requirements) is usually underspecified and ambiguous, there is no algorithm to use to come up with the solution, and finally the goal state is also underspecified, as there is no simple way of evaluating the solution for correctness.

3 Possible counterexamples of the "framing" definition

While the authors of any of the above definitions might well agree that framing or recharacterization of a problem may be useful in dealing with ill-definedness, only definitions [1.1] and [1.2] incorporate it formally into the definitions of ill-defined problems or domains.

One challenge to these definitions is that framing or recharacterization is useful in solving problems generally, including well-defined problems. Consider the following examples (Lynch, et al. 2010):

> **Checkerboard Example:** Given a checkerboard with the two opposing corners removed, is it possible to tile the board (i.e., fully cover it) with dominoes each one covering two adjacent tiles?
>
> **Fraction Example:** Given a proper fraction X/Y, if you add 1 to both the numerator and denominator, will the resulting fraction be larger, smaller, or the same?

Each of these examples illustrates the utility of framing or recharacterizing for problem solving, but the problems are not ill-defined. For instance, the definition of well-defined problems that appears in the complementary definition of ill-definedness in [3] makes that clear. The Checkerboard and Fraction examples each have a clear start state and unambiguously right-or-wrong goal state for which the evaluation criteria are clear and uncontroversial, and there are a few relatively easily formalized transformation rules by which the problem can be framed of recharacterized into a readily solvable form.

The Checkerboard problem is commonly posed in introductory AI classes to illustrate the utility of an appropriate recharacterization of the problem. If one recharacterizes the problem in terms of paired tiles and color matching, it becomes

apparent that removing two opposing corners, each of the same color, necessarily implies that the answer is “no”. The Fraction Example also has one right answer that can be derived in multiple ways, each involving a recharacterization of the problem in terms of logical and algebraic expressions, or of interpretation of appropriate graphical representations of fractions.

Even the Car Example from the introduction illustrates the utility of framing or recharacterizing. It may be solved using linear kinematics or the Work-Energy Theorem, or a combination of these, but recharacterizing in terms of a set of equations is required.

Thus, based on these counterexamples, one might argue that framing or recharacterization is not unique to ill-defined problem-solving and is not a defining characteristic contrary to definitions [1.1] and [1.2].

4 Response to counterexamples

The counterexamples, we submit, are not fatal to proposed definitions [1.1] and [1.2], and the response to them underscores the importance of a proviso of each definition: “This recharacterization, and the resulting solution, are subject to debate.” We argue that framing to bring a problem within a method that will generate a verifiably correct answer is different in kind from that required to bring a problem toward a defensible solution that will not be, indeed, cannot be, verifiably correct.

In solving a well-defined problem, the framing is a search for a solution template whose concepts, relationships, and procedures, when applied, will enable the solution to “click into place.” It involves mapping the problem’s facts into the template, often, a small set of alternative formulae or procedures appropriate to that type of problem, using the procedures associated with the template to generate an answer, and validating that the answer is correct. The templates are “representational tricks” of the trade; knowledge of which templates are appropriate to the problem, how to perform the mappings from the problem to the template, and how to validate the answers are key. While the answer to a well-defined problem may need to be justified with an argument, the criteria for validating an answer usually are accepted by the relevant audience (e.g., physics or math instructors) and are applied in a straightforward way. The generally accepted criteria for validating an answer help constrain the framing.

By contrast, in solving an ill-defined problem, a decision-maker is not framing the problem to apply a template or algorithm; rather he frames the problem to identify solutions and to make arguments for and against the solutions. The framing serves to explore plausible solutions and arguments. It involves not just a search for useful concepts to apply to the problem facts that may lead to a plausible solution. Since the concepts frequently are open-textured (e.g., “pedagogically effective”), one must search both for concepts and ways to define them that lead to a plausible solution. Framing is needed to search even for criteria to assess the proposed solution that are acceptable to the relevant audience, and arguments that justify the solution as a good in terms of the concepts and criteria.

In the Car, Fractions, and Checkerboard Examples, the characterization of the problems for solution is not driven by the need to specify meanings for concepts that

are open-textured, underspecified, or intractable, nor will it change the outcome of the problem.

For the Car Example, linear kinematics and the Work-Energy Theorem are equivalent and each fits into a common theory of classical mechanics. If applied correctly, each produces the same verifiable result. Within the context of classical physics, neither approach is controversial or the subject of reasonable debate. Thus, although the Car Example may require a kind of recharacterization to apply the relevant formulas, it is not ill-defined. This is also the case for the Fraction Example. There are multiple ways to recharacterize the problem for solution, including recharacterizing it in terms of pieces of a pie or glasses of water, but each different representation yields the same solution. The problem does not include any open-textured or unspecified concepts whose alternative plausible meanings need to be explored; the clear, logical solution and the criteria for evaluating it are similarly beyond debate.

Although a different kind of problem, the Checkerboard Example also lacks any open-textured or unspecified concepts. One might recharacterize it as a search problem (with a huge search space), but no recharacterizing of the problem will change the answer nor will the answer be subject to reasonable debate. If the problem were to find the least-creative solution as in (McCarthy, 1999), it would introduce an open-textured concept and raise questions about what "creative" means, in what context, and for what purpose, and how will degrees of creativity be compared, etc. The new problem would be ill-defined; the recharacterization of the problem adding constraints to answer those questions makes it a matter of debate and argument (Buchanan, 2001).

5 Discussion

Apart from challenging our own definitions, the above examples also underscore some differences among the various definitions of ill-defined problems and domains introduced above.

For one thing, according to definitions [3] and [4], the lack of "a limited number of relatively easily formalized transformation rules" or of an "algorithm to use to come up with the solution" renders a problem ill-defined. Obviously, solving the Car and Fraction Examples involves transformation rules and an algorithm; if one knows which rules or algorithm to apply the solution method is straightforward. Is there, however, a set of easily formalized transformation rules or an algorithm for *recognizing* which solution method to apply? If not, would definitions [3] and [4] characterize even these examples as ill-defined?

According to our definition [1.1], however, the problems are well-defined because no essential concepts, relations, or solution criteria are un- or under-specified, open-textured, or intractable, requiring a solver to frame or recharacterize them. Moreover the recharacterizations, and the resulting solutions, are not subject to debate; there is one correct solution regardless of the representation. The problem solver does have to recharacterize these well-defined problems to solve them, but as argued above, that is not the kind of recharacterization that contributes to making the problems ill-defined.

Presumably, on Definition [2] the Car and Fraction Examples are well-defined, too. They are not complex, have definite starting points, only one non-controversial correct solution, and clear strategies for finding solutions.

In addition, (Mitrovic and Weerasinghe 2009) invite taking into account a learner's knowledge or ability (i.e., or power in Simon's sense), but we think this consideration only adds to the difficulties of defining ill-definedness. For instance, the Fraction Example may appear ill-defined to a student unfamiliar with logic or algebra who does not know where to begin. As a result, he may be unable to recharacterize the problem into a solvable form. Or the solver may simply not see the trick that leads to solving the Checkerboard example. That, however, should not make these problems ill-defined. Considerations of learner power are categorically different from truly ill-defined problems where, as recognized by [1.1], because the recharacterizations are subject to debate, no amount of expertise can provide an indisputable answer.

Similarly, a state space view, as in (Mitrovic and Weerasinghe 2009) (and Simon), adds to the problems of defining ill-definedness. If one considers the generalized Checkerboard Example, with all possible board sizes, then the space of possible tilings grows as large as one likes but the problem is still regular and subject to the representational trick that is not subject to debate, and thus remains well-defined.

On the other hand a virtue of the complementary definitions [3] and [4], and of the disjunctive definition [2] is that they accommodate the intuition that problems lie along a spectrum of well- and ill-definedness. On these definitions, a problem becomes more or less ill-defined as more or fewer of the disjunctive criteria are satisfied. In particular, Mitrovic and Weerasinghe (2009) present a two dimensional view representing the ill-definedness of a task or a domain along which problems may vary. While definitions [1.1] and [1.2] do relate the ill-definedness of problems and domains, they do not appear to accommodate the spectra in any obvious way. A problem of the complementary definitions, [3] and [4], however, is that these or'd (i.e., disjunctive) criteria seem too liberally to ascribe ill-definedness to problems (as argued above).

6 Conclusion

A final consideration is the extent to which the definitions take into account the educational context of the problems, and if so, how they do it. Expert instructors in some domains explicitly employ the terminology of ill-definedness. For instance, in domains like law, professional ethics, medicine, business, and public policy decision-making, debate and argumentation about the application of open-textured terms are explicitly incorporated into the pedagogy, and viewing the problem from the (multiple) viewpoints of those affected by the decision is seen as one driver of how to frame the problem. Something like this appears to be the case in design fields, too, where the arguments about whether and how alternative designs satisfy constraints and what those constraints really mean would likely be important in pedagogy. Definitions [1.1] and [1.2], and to a lesser extent [2], employ this language. Interestingly, although definition [4] is applied to a design task, developing an ER

scheme for a database, there is no focus on debating characterizations of the problem. Definition [3] is the most divorced from a particular pedagogical context.

Of course, whether this is a good thing or not is itself, a matter of debate, part of the ill-defined nature of the task of defining ill-definedness. The presentation of alternative current definitions of ill-definedness, the consideration of some possible counterexamples, and their use to underscore the differences among the definitions, it is hoped, will continue the debate on this important issue at the workshop.

References

Aleven, V., Ashley, K., Lynch, C., & Pinkwart, N. (Eds.) (2006). Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the *8th International Conference on Intelligent Tutoring Systems.* Jhongli (Taiwan), National Central University.

Aleven, V., Ashley, K., Lynch, C., & Pinkwart, N. (Eds.) (2007). Proceedings of the Workshop on AIED Applications for Ill-Defined Domains at the *13th International Conference on Artificial Intelligence in Education.* Marina Del Rey California 2007.

Buchanan, B. G. (2001). Creativity at the metalevel: Aaai-2000 presidential address. *AI Magazine*, 22(3):13-28.

Fields, A.M. (2006) Ill-structured problems and the reference consultation: The librarian's role in developing student expertise. Reference services review, Emerald 34(3), 405–420.

Jonassen, D.H., Tessmer, M. and Hannum, W.H (1999). *Task analysis methods for instructional design*, Erlbaum 1999.

Le, N.T. and Menzel, W. (2008) Using Constraint-Based Modelling to Describe the Solution Space of Ill-defined Problems in Logic Programming. Advances in Web Based Learning (ICSL 2007) Lecture Notes in Computer Science. V. 4823. pp. 367-379.

Lynch, C., Ashley, K. D., Aleven, V., and Pinkwart, N. (2006). Defining ill-defined domains; a literature survey. In Aleven et al. (2006), pages 1-10.

Lynch, C., Ashley, K. D., Aleven, V., and Pinkwart, N. (2010). Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*: Special issue on Ill-Defined Domains. (in press).

McCarthy, J. (1999) Creative solutions to problems. In *AISB'99 Symposium on AI and Scientific Creativity*, pages 44–48, Edinburgh, Scotland, April.

Mitrovic, A. and Weerasinghe, A. (2009). Revisiting ill-defiedness and the consequences for ITSs. In Dimitrova, V., Mizoguchi, R., du Boulay, B., and Graesser, A. C., editors, AIED, volume 200 of Frontiers in Artificial Intelligence and Applications, pages 375-382. IOS Press.

Nkambou, R., Nguifo, E.M., and Fournier-Viger, P. (2008) Using Knowledge Discovery Techniques to Support Tutoring in an Ill-Defined Domain. Intelligent Tutoring Systems. Lecture Notes in Computer Science V. 5091. Pp. 395-405.

Ormerod, T. C. (2006) Planning and ill-defined problems, in R. Morris & G. Ward (Eds.): *The Cognitive Psychology of Planning*, London: Psychology Press.

Pinkwart, N., C. Lynch, K. D. Ashley, V. Aleven, eds. (to appear 2010). *International Journal of Artificial Intelligence in Education*, Special Issue on AIED applications for ill-defined domains.

Simon, H. (1973) The structure of ill-structured problems, *Artificial Intelligence*, No. 4, pp. 181-201.

Simon, H. (1978) Information-processing theory of human problem solving. In: Estes, W.K. (ed.) *Handbook of learning and cognitive processes*. Human information, vol. 5.

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V.,
Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 9-16)

The First Report is Always Wrong, and Other Ill-Defined Aspects of the Army Battle Captain Domain

Elizabeth Owen Bratt[1], Eric Domeshek[2], and Paula J. Durlach[3]

[1] Stanford University
[2] Stottler-Henke Associates, Inc.
[3] U. S. Army Research Institute

Abstract. The position of battle captain in United States Army battalions involves responsibility for tracking a large quantity of information, communicating important information to the right people, and making decisions quickly when events deviate from plans. Soldiers assigned as battle captain in maneuver battalions have a wide variation in background and experience, yet rarely receive specific training for the role. A scenario-based intelligent tutoring system (ITS), which adapts to the incoming experience and knowledge of the student, could help fill this training gap; but the domain poses a number of challenges for defining the scope, goals and functionality of the ITS. For example, a skilled battle captain will understand that information may be untrustworthy, as well as clues to that effect. A good battle captain will also be able to adapt to the responsibilities and procedures of the particular tactical operations center (TOC), which may vary depending on the commander, the other personnel in the TOC, the location, and the mission. Defining the best course of action may not be possible (there may be multiple solutions, none of which is objectively the best). The battle captain must balance the demands of higher units for information and lower units for time to act. We discuss these issues in relation to the development of the BC-ITS training system for battle captains and its spoken language interface.

1 Introduction

In this paper, we first describe the battalion battle captain domain, and discuss various situations which involve skill and judgment. Ideally, training for new battle captains should provide opportunities to practice making decisions not only under routine conditions, but also when the routine established procedures do not yield an obvious best course of action. After reviewing the domain, we turn to potential solutions for developing an intelligent tutoring system (ITS), which can incorporate training for both the well- and ill-defined decisions a battle captain will face.

2 Battle Captain Domain

In the Army, a battalion-level battle captain is responsible for overseeing the information flow and tracking ongoing operations in the Tactical Operations Center (TOC) [7], [10]. Other personnel are primarily responsible for monitoring and logging the various information channels, such as the radio nets, e-mail, chat, and battlespace mapping and unit tracking software; but the battle captain is responsible for noticing whenever events have moved beyond the routine execution of the plans and now require actions and decisions. The battle captain, who is not necessarily a captain in rank, defers to the S3, the XO (executive officer), and the battalion commander when they are in the TOC. But these officers are frequently away from the TOC, leaving the battle captain in charge of implementing the planned operations and making decisions to support those operations when needed.

Every TOC is supposed to have a set of "recipes" known as battle drills, which dictate how the battalion staff should behave under particular circumstances (e.g., a patrol suffers a casualty). TOCs also are guided by the commander's critical information requirements (CCIRS). CCIRs are circumstances designated by the battalion commander, which directly affect his decision-making, and of which he wants to be immediately made aware. At first blush, the battle captain domain may seem well-defined, because of the reliance on battle drills, SOPs, CCIRs, and other guidelines. But it turns out that actual procedures vary from one TOC to another, being determined by environment (e.g., urban vs. rural, terrain, infrastructure), threat level, mission type (e.g., offense vs. stability), and "commander's preference." The latter means that every TOC operates somewhat differently, including the SOPs that are followed, the actual information management systems used in the TOC, and the way the commander wants to be briefed [2], [12]. Therefore, training a new battle captain on a particular set of battle drills would be inappropriate. The battle captain will need to adapt behavior to suit both circumstances and superior's expectations. Moreover, real-world events often involve more complexity than laid out in battle drills, and unforeseen conditions may arise. In Army doctrine "exceptional information" results from an unexpected event, such as an unforeseen opportunity for success or an early warning of an unforeseen threat. By its nature, identifying exceptional information relies on the initiative of subordinate commanders and the staff. Thus, an important part of the battle captain's job is turning information into understanding and being proactive with it. The challenge for the development of an ITS is encoding these skills and determining the best way to train and assess them.

3 Ill-Defined Aspects of the Domain

The battle captain domain fits the criteria of an ill-defined domain [6], [1], because the areas where training is seen as most useful are those for which established procedures do not determine the battle captain's actions, and the battle

captain must use intelligence, foresight, and judgment [12]. Two of these areas are described in more detail below.

3.1 When to Act

Experienced battle captains when describing how to react to a new battlefield event often state that *The first report is always wrong* [4], [12]. The reader need only reflect on the Fort Hood shootings of November 2009, and the initial inaccurate reports, to get an understanding of the problem (e.g., initial reports said the shooter had been killed when he was not). The battle captain is responsible for notifying appropriate people, such as his counterpart at brigade, and others who may need to act on the information, such as the QRF (quick reaction force) or Medevac (medical evacuation). But there is a speed-accuracy tradeoff: the need to act promptly vs. the need to have accurate information. Judging when the time to act is right can be difficult. Wrong or incomplete information can mean loss of life; but so can delay. Various strategies can be used: not acting until the second report [4] or requesting roster numbers of wounded personnel so that the casualty count will be more accurate [12]; but these heuristics are not a total solution. Despite knowing that initial information may not be reliable, Army training emphasizes the value in adopting a solution in a timely manner, instead of waiting for the problem to resolve itself some other way [9]. A summary of leadership "lessons learned" includes the point, *Making a decision in combat (even if it is the wrong decision) is better than inaction* [5].

Various communication failures may also present a challenge to the battle captain. Units may fail to acknowledge a radio communication, or to make routine hourly reports. The battle captain will need to determine whether a communication failure is due to a mundane issue (forgetfulness, technical issues) or whether it is the sign of something more serious. If a unit fails to provide mission updates, for example about an unexpected delay, it can have a series of cascading consequences for synchronization of activities. While it is the unit's responsibility to keep the battle captain updated, if the unit fails to report, the battle captain needs to take action. Thus, the battle captain needs to be sensitive to omissions as well as commissions.

3.2 Balancing Conflicting Goals

The battle captain may have to balance competing needs in a pressured situation. If a platoon is engaged in combat, known as a TIC (troops in contact), the platoon leader will be concerned with directing on-going events and will prioritize the immediate actions of his unit above the need to report to the battalion TOC. At the same time, brigade will be pushing the battalion battle captain to supply details of the TIC [12]. The more the battle captain knows about the situation, the more he may be able to commandeer other assets in the aid of the unit in the TIC. The battle captain will need to balance getting accurate information rapidly against distracting the unit in the TIC [4]. Experienced battle captains

say that dealing with these conflicting pressures is one of the most challenging aspects of the job [12].

4 Progress on Development of BC-ITS

4.1 Domain Analysis and Scope

We analyzed the role of the battalion battle captain in order to determine the learning objectives upon which to base the development of adaptive technology-based training for battle captains, BC-ITS. Our analysis included examination of documents [e.g., [4], [7], [10]], observation of live training exercises, and interviews with experienced battle captains and other subject matter experts (SMEs). We narrowed the scope of BC-ITS to cover the battle captain's role in monitoring current operations (this eliminated several battle captain responsibilities including physical set-up of the TOC, administrative duties, and report writing). Our goal is to enable and encourage thinking skills, not just procedural learning; however, we also recognize that competency in basic procedures is a precursor to higher cognitive processes applied in a domain. Therefore, given variation in student background and knowledge, we cannot ignore the need to assess and train both the well- and the ill-defined aspects of the domain. Based on analysis of the domain, we specified a hierarchical set of learning objectives, with five terminal learning objectives: (1) Help the Commander Manage the Force (understand the mission and battalion capabilities), (2) Maintain Situation Awareness (keep track of events and their implications), (3) Information Management (assemble/assess/filter/pull/push), (4) Decision Making and Action (authority, responsibility, timing), and (5) Help Manage the Fight (anticipate, prioritize). Each terminal learning objective has four to seven sub-objectives, and each of these has one to four sub-sub-objectives, for a total of 63 enabling learning objectives.

These learning objectives will serve as the basis for a "situated tutor," by combining the pedagogical approach of traditional ITS (content selection, coaching, feedback, and scaffolding) with a simulation that embeds the student in an ongoing scenario where skills and knowledge will be applied under realistic conditions [8]. While envisioned to integrate instruction and simulation, our initial focus thus far has been on the simulation, including the key events, the student actions to be supported, and how entities should behave in response to events, actions, and resulting world state changes. These sub-elements must provide students the opportunity to demonstrate the mastery (or lack of mastery) of the learning objectives.

We envision delivering different training scenarios to students of different backgrounds such that less knowledgeable students start out with relatively straightforward scenarios requiring application of battle drills and other standard procedures, while more experienced students may start with less well-defined situations requiring judgment, prioritization, and anticipation. SMEs have told us that prior tactical and staff experiences enhance the effectiveness of a battle captain, and so we are hopeful that a fairly short background questionnaire might

allow us to stream different students into different scenario sequences; however, we have yet to identify the specific learning objectives in our set for which we should expect past tactical or staff experience to show mastery. In the absence of this information, creating scenarios (and instruction) appropriate to different backgrounds would just be a best guess. We therefore need to collect data with respect to this question to design scenarios appropriately.

4.2 Natural Language Processing

The battle captain's job is communication-intensive, therefore BC-ITS requires a method of simulating communication-based interactions. Doing this using a menu-based system would be cumbersome, and limit the training to a predetermined set of options. A major research thrust of this project is therefore the development of a natural language interface, supporting spoken and text-based communication. Developing the ability for BC-ITS to process what the student says or types and respond appropriately raises a number of challenges: (1) achieving a high enough level of speech recognition accuracy, (2) interpreting the meaning expressed in the recognized speech, (3) developing appropriate responses for the simulated characters, (4) using the understanding of the student's language to update the student model, (5) supporting interactive dialogs, which can probe student thought processes. The first major step is to be able to understand and respond to short oral or text communications made by the student. For example, if a student requests that a unit report its position, the system must be able to understand what is being asked, examine the underlying simulation for that unit's position (which is dynamic in the scenario), and respond with the correct grid location.

Ultimately, we aim to be able to understand, analyze, and respond to student multi-sentence situation reports (requested by a simulated superior) or shift change briefs. Having the student make these reports would support training by making the student reflect on the scenario experience and organize their thoughts for oral presentation. This can engage the student more deeply while performing a task that is in fact part of their duties [3], [11]. Being able to analyze these briefings will provide insight into the student's situation awareness and attention which can be used as the basis for providing feedback. Moreover, being able to respond with probing questions and conduct interactive dialogs will support an even deeper level of self-reflection, and support contemplation on more ill-defined aspects of the domain. Understanding the factors a student considered in a decision can be more important than which of the competing decisions was chosen, when trying to teach how to think about a problem or situation.

4.3 Implementation Status

At the time of this writing, we have implemented the BC-ITS environment and have preliminary models for natural language processing. We have put into place the infrastructure for the training we would like to provide, but have yet to tackle implementation for the really challenging ill-defined aspects to be trained. The

system has the ability to deliver pre-training and post-training questionnaires, upfront instruction, and training on how to negotiate the simulation interface; but our main focus here is on the student's scenario-based experience. The interface allows the student to interact (via headset and keyboard) with simulated entities, who can be co-located with him in the TOC, or away from the TOC. These latter communications can be conducted via simulated channels such as FM radios, telephones, or text messaging. Because our natural language processing capability is still a work in progress, the system's understanding of the student's language is echoed back to them. In addition, they can open a complete record of all interactions. This allows the student to review what has already been communicated and/or to retrieve specific information that might be needed. Students can also open several documents and figures, which provide them with reference information that would be available in a real TOC. Finally, a situation awareness map allows "blue force tracking;" that is, automated tracking of digitally equipped units in the field.

We have implemented a single scenario, simulating part of a battle captain's shift in the TOC. The scenario combines application of some well-defined knowledge, in that there exist battle drills for each significant event (patrol vehicle breaks down, patrol in a TIC, patrol suffers casualties). Events unfold over time so that some judgment and initiative can be exhibited over and above mere knowledge of the battle drills; however, the extent of uncertainty is not that great, in that the student is not placed in a situation with competing goals. An automated "coach" tracks student responses to unfolding events and provides proactive hints (which can be solicited or unsolicited), or reactive feedback, via text in the coaching window. Our immediate plans include having Soldiers with different levels and types of experience complete this scenario, to give us a better picture of how those prior experiences influence their ability to carry out the battle captain role in this scenario. These findings can then contribute to future efforts to adapt training scenarios to match student capabilities based on prior experience.

5 Future Directions

5.1 Adapting scenarios

Rather than taking each student through a fixed sequence of pre-scripted and progressively more difficult scenarios, our aim is to be able to present each student with a customized set, which will be determined by the accumulating evidence with respect to their capabilities in the battle captain role. We are currently hypothesizing four different possible starting scenarios, assuming that we can develop the ability to categorize students quickly, prior to training, into one of four types, created by crossing the two factors: low vs. high tactical experience and low vs. high staff experience. Of course, we will also need to be able to identify the acquired skills (or lack of skills) associated with each factor in order to design these initial scenarios appropriately. Performance during each

student's first scenario will allow us to refine our initial estimate of student capabilities, update the student model accordingly, and determine which learning objectives to target in the subsequent scenario. We aim to be able to create these subsequent scenarios "on the fly," as opposed to retrieving them from a library. One feasible way to do this would be to prescript fairly complex scenarios, but with particular events designated available for inclusion or deletion, depending on the state of the student model and level of challenge the student is ready for. Of course, the associated coaching would need to be modified accordingly, as well.

5.2 Interactive Tutoring

As previously discussed, we are hoping to be able to process multi-sentence briefings and to be able to address some of the ill-defined aspects of the domain by conducting interactive dialogs with students. These dialogs would encourage student reflection by posing questions and stimulating the student to think about "what if" possibilities. Rather than advising the student on *what to do* these interactions would provide guidance on *how to think* about the situation, and the possible second-order consequences of different courses of action. Prompts for reflection could be interjected during a scenario (e.g., " think about what the platoon leader is coping with right now"), whereas more extensive Socratic-like dialogs could be reserved for post-scenario discussion. Soldiers are accustomed to performing after action reviews (AAR) following training exercises. A good AAR stimulates the trainees themselves to reflect on what went right, what went wrong, and what to change to improve performance. The instructor serves the role of facilitating this self-examination. The challenge is to stimulate the same kind of self reflection with an automated coach that can guide the student down a productive path.

6 Conclusion

The battle captain domain involves many areas in which it is hard to define correct vs. incorrect performance. An ITS for this domain needs to help a student recognize how to turn raw information into an understanding that can guide his/her course of action. Using spoken language briefings to simulated characters can allow the ITS a window into the student's reasoning, and through interactive probing, help guide the student toward recognizing how to weigh the available information appropriately. Even if the first report is wrong, the student still should take action, while being observant of any facts that bear on the situation at hand, and be ready to adapt to the changes that later reports may bring.

References

1. Ashley, K. D., Chi, M., Pinkus, R., and Moore, J. D. 2004. Modeling learning to reason with cases in engineering ethics: A test domain for intelligent assistance. NSF Proposal.

2. Bratt, Elizabeth Owen. In revision, 2010. Intelligent Tutoring for Ill-Defined Domains in Military Simulation-Based Training. International Journal of Artificial Intelligence in Education.
3. Chi, M.T.H. 2000. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), Advances in Instructional Psychology, Hillsdale, NJ: Lawrence Erlbaum Associates. 161-238.
4. Custis, J. 2007. Fundamentals of the Battle Captain. Small Wars Council Doctrine and TTPs. Available at http://council.smallwarsjournal.com/showthread.php?p=28033#poststop
5. Lencioni, Patrick. 2002. The five dysfunctions of a team: a leadership fable. San Francisco: Jossey-Bass.
6. Lynch, C., Ashley, K., Aleven, V., and Pinkwart, N. 2006. Defining Ill-Defined Domains; A literature survey. In V. Aleven, K. Ashley, C. Lynch, and N. Pinkwart (Eds.), Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems (p. 1-10). Jhongli (Taiwan), National Central University.
7. de Oliveira, Marcus F. 1995. WHAT NOW, BATTLE CAPTAIN? The Who, What and How of the Job on Nobody's Books, but Found in Every Unit's TOC. Combat Training Center (CTC) Quarterly Bulletin, 2d Qtr FY 95, Center for Army Lessons Learned (CALL) Available at http://www.globalsecurity.org/military/library/report/call/call_2qfy95_ctcchap1.htm.
8. Schatz, S., Bowers, C. A., and Nicholson, D. 2009. Advanced situated tutors: Design, philosophy, and a review of existing systems. In *Proceedings of the 53rd Annual Conference of the Human Factors and Ergonomics Society.* Santa Monica, CA: Human Factors and Ergonomics Society (HFES).
9. Schirmer, Peter, Crowley, James C., Blacker, Nancy E., Brennan, Richard R., Jr., Leonard, Henry A., Polich, J. Michael, Sollinger, Jerry M.,and VardaLeader, Danielle M. 2008. Development in Army Units: Views from the Field. RAND Corporation Monograph MG648. Available at http://www.rand.org/pubs/monographs/2008/RAND_MG648.pdf.
10. Wampler, Richard L. ; Centric, James ; Salter, Margaret S. 1998. The Brigade Battle Captain - A Prototype Training Product. Army Research Institute Research Product 98-36. Available at http://handle.dtic.mil/100.2/ADA347093.
11. Weerasinghe, A. and Mitrovic, A. 2006. Individualizing Self-Explanation Support for Ill-Defined Tasks in Constraint-based Tutors. Workshop on "Intelligent Tutoring Systems for Ill-Defined Domains", 8th International Conference on Intelligent Tutoring Systems.
12. Interviews with experienced battle captains, conducted by Paula J. Durlach and Elizabeth Owen Bratt in May 2009 at Fort Campbell and Fort Hood, in December 2009 at Fort Carson, and in March 2010 at Fort Drum.

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V.,
Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 17-20)

Is Five Enough? Modeling Learning Progression in Ill-Defined Domains at Tertiary Level

Richard Gluga, Judy Kay, and Tim Lever

University of Sydney, Sydney NSW 2006, Australia

Abstract. Insuring the progressive development of generic and discipline-specific high-level skills in a university degree is hard. This is partly because the skills are defined broadly, in terms of multiple curriculum frameworks which have different granularity and terms. Another key challenge is the definition of *skill level progression*, the very ill-defined notion of *maturity*. This paper describes the vision of CUSP, a system for modeling high-level skills via an overlay student model for long term progression of learning. We argue that the representation of maturity and knowledge should use 5 levels.
Keywords: Curriculum Mapping, Learning Progression, Graduate Attributes, Accreditation Competencies, Learner Model

1 Introduction

A typical university degree is 3-to-5 years long and is bound by multiple sets of curriculum frameworks. These typically include generic, transferable skills defined by the University or Faculty, discipline specific skills defined by vocational or accreditation bodies, and finer-grained learning objectives defined by curriculum bodies. A student completing a 3-to-5 year tertiary degree, with 24 to 60 core and elective subjects, should be able to track his or her progress against all of these skills.

It is widely agreed that generic skills, such *communication* or *teamwork*, are extremely important. These skills are ill-defined in several senses. Firstly, the role of increasing sophistication and higher levels of performance are a key part of their development. As a learner builds a particular skill, their knowledge increases. Additionally, as the learner gains new knowledge, they commonly consolidate their understanding of more basic concepts, developing greater maturity.

Such notions of progression and maturity are very ill-defined. One challenge is that flexible degrees must be designed so that students acquire maturity in generic skills via any of the allowed pathways. Another dimension is very pragmatic: skills are typically taught within the context of disciplines, such as physics, where the lecturer is expert in physics, but not communication skills as a discipline. So, in practice, even though there is a body of knowledge about the generic skill areas, those actually teaching may be unaware of it.

As we worked to model generic skills, we also noted that the notion of maturity is also important but ill-defined in other skill areas. For example, consider

the case of programming, a seemingly very well defined area, with many precise, fine-grained subgoals. Even for this, an important goal for a university degree, and a requirement for accreditation is for a growth in programming maturity. Students must do a sequence of programming subjects, each developing this ill-defined notion of maturity.

We now consider how to design long term learner model frameworks that can capture the progressive learning of high-level skills. A single degree will have skill sets derived from multiple internal and external curriculum frameworks. These skills come in varying granularities. Mapping each skill to each subject activity in which it is taught across a full degree is a very time consuming and intellectually challenging task. Also, modeling learning progression requires a representation of the skill level, which is not always explicitly defined in all curriculum frameworks. The sequencing and delivery of prescribed skills is thus often left to degree designers, who require aids to help with this monumental task.

2 Related Work

There have been numerous attempts to model competencies, generic skills and subjects, for example in standards such as IEEE LOM, IMS LIP, SCORM, HR-XML, IMS-RDCEO and EML (Educational Modeling Language[1] [5]). In parallel, there has been considerable work on ontologies to model learnt skills (for example [1, 7]). We aim to go beyond just mapping competencies/skills to subjects in a single framework, as in Curriculum Central [3] which used 5 levels and UK-SpecIAL [2] which mapped subject multiple-choice questions to UK SPEC Standards for Professional Engineering accreditation attributes.

We draw on work to modeling learning progression and skill maturity in the context of computer programing, such as the use of the 5 maturity levels of the SOLO (Structure of the Observed Learning Outcome) taxonomy or its extension to 6 levels in the context of computer programing [6]. We have found no work that models learning progression of competencies/skills in terms of multiple curriculum frameworks across entire 3-to-5 year degrees with flexible pathways.

3 Approach

We have built CUSP (Course and Unit of Study Portal) which models entire degree programs in terms of core and elective subjects [4]. CUSP models skills in terms of 5 levels, and these are mapped against each subject's learning outcomes and assessments. This enables CUSP to generate report matrices that span entire degrees and show which skills are taught in which subject and at which level. Multiple competency/skill frameworks are supported via a pragmatic approach that allows mapping equivalent attributes from different frameworks to each

[1] Educational modeling Language, `http://www.learningnetworks.org/?q=EML`

other. This has proven to be a workable and scalable solution. However, frameworks of vastly differing granularities are difficult to map against each other. This can lead to translational inconsistencies in the learning-outcome¿primary-skill¿secondary-skill mappings.

Our next phase will be to introduce subject instance level mappings between skills from different frameworks, thus solving the granularity translation issues. We will then move towards creation of individual learner models that can be easily presented in a visualization, such as that shown in Figure 1. Here, a student can see their learner model in terms of skills from one of two different frameworks. Each LM lists the skills vertically downwards, and the learning progression horizontally across. Learning progression is a combination of increasing knowledge at higher levels and strengthened maturity at lower levels. The LM will be supported by evidence collected from assessment task marks. The LM will be navigable such that students can see which skills were developed in which subjects and which assessment tasks.

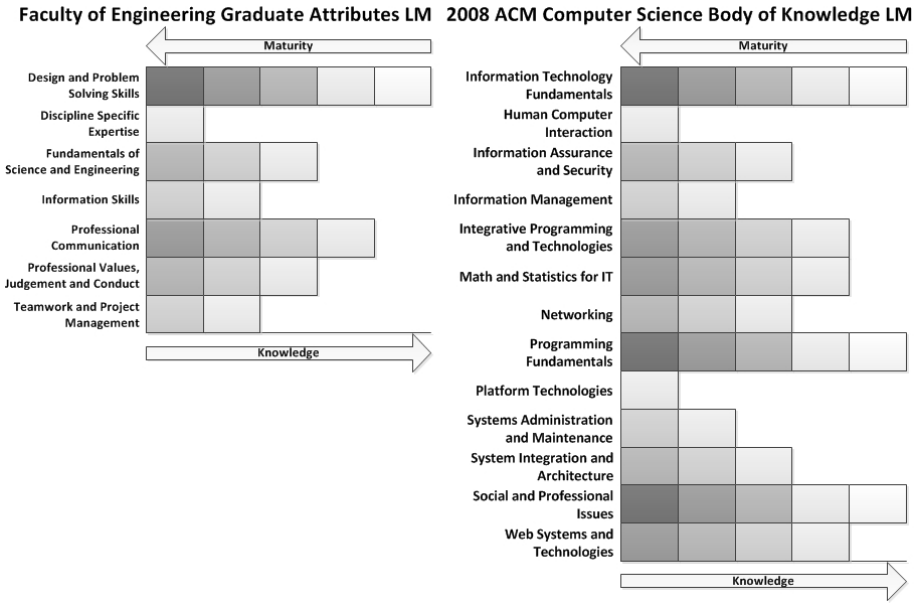


Fig. 1. On the left: learner model representing knowledge and maturity of generic Faculty Graduate Attributes for a fictional student. On the right: the same student's model of knowledge and maturity based on discipline-specific ACM Computer Science high-level topic areas.

4 Conclusions & Future Work

The ill-defined nature of the high-level skills that we need to model has driven us to find a learner model representation that can address the needs of the many stakeholders involved: University level faculty concerned with graduate attributes; Faculty and School level academics concerned with the design of curricula for each degree; academics responsible for the design of their own subject; students who need to select subjects and to understand how the learning objectives and activities in each subject contribute to a big picture development of important long term broad skills.

Our previous experience has indicated that five knowledge levels is manageable for the curriculum design and mapping processes, especially for accreditation. It is enough to show progression over the three to five years of a degree and to capture the broad notion of maturity within an ill-defined skill. The modeling approach and five levels also support a compact representation of the learner's current progress, clearly highlighting areas of relative strength and weakness.

The core contribution of this work is an open learner model design that is intended to be simple and clear enough to support a range of processes. It should assist subject lecturers in making their subjects contribute to the development of broad skills. It should help degree designers ensure their courses satisfy accreditation and curriculum requirements. And finally, it should act as an invaluable tool for students in planning studies, selecting subjects, understanding how each part of a subject contributes to long term learning across the degree, as well as reflecting on their progress in building high-level skills.

References

[1] Assche, F.V.: Linking learning resources to curricula by using competencies. First International Workshop on LO Discovery & Exchange (2007)
[2] Bull, S., Gardner, P.: Highlighting learning across a degree within an independent open learner model. AIED 200, 275–282 (2009)
[3] Calvo, R., Carroll, N., Ellis, R.: Curriculum central: A portal system for the academic enterprise. IJCEELL 17(1), 43–56 (2007)
[4] Gluga, R., Kay, J., Lever, T.: Modeling long term learning of generic skills. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS2010, Proceedings of the Tenth International Conference on Intelligent Tutoring Systems. p. to appear. Springer (2010)
[5] Koper, R.: Modeling units of study from a pedagogical perspective: the pedagogical meta-model behind EML. In: OTEC (2001)
[6] Lister, R., Simon, B., Thompson, E., Whalley, J.L., Prasad, C.: Not seeing the forest for the trees: novice programmers and the solo taxonomy. In: ITICSE 2006. pp. 118–122. ACM, New York, NY, USA (2006)
[7] Paquette, G., Rosca, I., Mihaila, S., Masmoudi, A.: TELOS, a Service-Oriented framework to support learning and knowledge management. E-Learning Networked Environments and Architectures: A Knowledge Processing Perspective p. 434 (2007)

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V., Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 21-28)

Comments of Journalism Mentors on News Stories: Classification and Epistemic Status of Mentor Contributions

Art Graesser[1], Zhiqiang Cai[1], Jonathan Wood[1], David Hatfield[2], Elizabeth Bagley[2], Padraig Nash[2], David Shaffer[2]

[1] Institute for Intelligent Systems, 365 Innovation Drive, University of Memphis, Memphis, TN 38152, {a-graesser, zcai, jwood}@memphis.edu
[2] Educational Sciences, Department of Educational Psychology, 1025 W. Johnson Street, University of Wisconsin, Madison, WI 53706, {ebagley, pnash, dhatfield, dws}@ef-games.com

Abstract. We identified the speech act categories and clusters of discourse comments of journalism mentors who interact with students editing news stories. Two important speech act categories are evaluations and suggestions. Latent semantic analysis and principal components analyses helped us discover clusters of comments involving evaluations and suggestions. The comments of mentors were also significantly aligned with epistemic frame elements that motivate the comments at a deeper level of discourse and pedagogy. Such alignments were validated by logistic regression analyses on a sample of hand-coded judgments of the frame elements. There was some modest transfer from a journalism practicum corpus to a game corpus. These analyses provide an important first step in building a virtual AutoMentor for multiparty epistemic games on ill-defined problems and domains.

Keywords: virtual agents, serious games, epistemic games, discourse, latent semantic analysis

1 Introduction

Our long-term goal is to build an automated virtual mentor, called AutoMentor, that provides guidance to students as they interact in groups in serious games. AutoMentor will vigilantly observe the game states and interactions between players and will periodically offer comments and suggestions to promote learning and productive conversation. Shaffer and his colleagues [1, 2, 3] have established the need for human mentors to promote learning when students interact with multiparty games, such as *Urban Science*, *Science.net*, or the *Land Science* game under current development and testing. These games help students understand the kinds of problems and problem solving that socially valued professions routinely engage in, such as how the development of cities and suburbs are influenced by zoning, roads, parks, housing, and economic investment, or how important developments in science can be

communicated through narrative details and attention to accuracy and source attribution. Student learning is severely limited, however, if there is no mentorship and expertise from professional stakeholders. Our epistemic games group is therefore analyzing the verbal interactions between human mentors and students, with the hope of automating their language and discourse contributions in AutoMentor.

AutoMentor can be viewed as an augmentation of AutoTutor [4, 5, 6], a pedagogical agent that helps students learn by holding a conversation in natural language. The original AutoTutor was developed for one-on-one tutoring in language, but more recent versions have involved trialogs with two agents interacting with one student [7] and interactions among the student, an AutoTutor agent, and an external simulation environment [8]. AutoMentor moves beyond these AutoTutor versions by having a single virtual agent interact with multiple students in groups as they interact with a complex simulation game on urban planning and environmental science.

The quality of such conversational agents depends on their ability to understand and generate discourse. Discourse has multiple levels of analysis that have been identified by researchers in discourse processes [9, 10, 11] and computational linguistics [12]. According to one multilevel discourse framework [11], the levels include the surface code (wording and syntax), the explicit textbase, the referential situation model, genre and rhetorical structure, and pragmatic communication. Discourse becomes more complex as one moves from dialogues to multiparty conversations [9] and from minimal external environments to complex dynamic external environments (such as simulation games). The field of computational linguistics has not advanced to the point of accurately understanding language and appropriately generating language in a broad landscape of discourse worlds [12]. However, a combination of symbolic and statistical architectures have been quite successful in handling conversational interactions in some conversational contexts, including tutoring. Such systems sometimes help student learning and motivation even though the conversation is not perfect [4, 13]. The hope is that these successes will extend to AutoMentor.

The AutoMentor project has to handle many of challenges of ill-defined problems and domains. Understanding the meaning of natural language in student contributions has its own set of issues regarding uncertainty, imprecision, and vagueness. However, the domain knowledge is also open-ended and minimally constrained because there is no perfect well-formed solution to problems in the game space. The dialogue moves of AutoMentor neverthelesss provide feedback and metacognitive guidance.

2 Structure of Games and Mentor Contributions

A few words need to be said about the structure of our games and discourse, although it is beyond the scope of this paper to provide a full specification. The games consist of an ordered sequence of major activities (i.e., game phases) in service of a goal, with a group of players and a mentor in each activity. Each activity also has external media, controls, or products that players can view, manipulate, or create. An activity

in one game may involve students being science journalists with the goal of creating an on-line science news magazine (the external product). An activity in another game might involve a group of players making decisions on how to change resources in a city to reduce pollution. Whatever the activity, the quality of interactions among players should be superior with the intervention of a mentor.

The mentor guides or enters the group discussion through conversational turns and each turn has one or more speech acts. Some speech acts fulfill politeness norms for conversation, such as greetings ("hello"), introductions ("I'm your mentor"), and closing acts ("Let me leave you to work on this by yourselves"). Two important substantive speech acts of the mentor consist of *evaluations* ("The report is fine", "That is a good idea") and *suggestions* ("Make the graph more precise," "Improve the first sentence."). Evaluations consist of judgment about a person, product, or external referent. Suggestions are recommendations, directives, requests, or hints on what actions the students should perform. Table 1 shows some examples of these evaluation and suggestion expressions that came from the Science.net corpus with human mentors. The discourse analysis system needs to segment and classify the contributions of the mentor into speech act categories. One major objective of the present study is to identify the different clusters of evaluations and of suggestions expressed by mentors because these are the two most important speech acts. Once these speech acts are identified, we can develop mechanisms that produce particular evaluations and suggestions in specific game states.

Table 1. Example evaluation and suggestion speech acts by human mentors.

Exemplar	Summary description	Template structure
EVALUATION		
Nice try in the originality of the lead.	The lead is good.	The lead is X.
Readers would want to get to know him better.	Readers want to know more about a person.	Readers want to know more about X.
SUGGESTION		
Strive to get more sizzle into that critical first sentence.	Make the first sentence better.	Make sentence X better.
Ideally, too, there'd be descriptions of action, from watching Bucky at the game.	You need to know what an ideal story is like in order to improve your writing.	Stories should ideally have X, so maybe you should have Y.

There is another important level of discourse course analysis called *epistemic frames* [1, 2, 3]. An epistemic frame is a set of norms, virtues, or criteria that guides the mentor's decisions and actions in the activity. The frame consists of a specific description of a way of talking, listening, writing, reading, acting, interacting, believing, valuing, and feeling (and using various objects, symbols, images, tools, and technologies). The particular frame elements are categorized into skills, knowledge, identity, values, and epistemology (what we call the SKIVE components), as will be

elaborated shortly (see Table 2). The epistemic frame represents the vision of the mentor and hopefully the group of game players after some time while interacting with the mentor. It is indeed a deeper and more abstract ill-defined level of discourse analysis than the particular speech acts. It is essential to have some alignment between the mentor's acts of evaluations or suggestions and the epistemic frame elements. A second major objective of the present study is to examine the alignments between mentor comments and elements of the SKIVE epistemic frame.

3 Corpora of Mentor Contributions in Journalism

We conducted some analyses on two corpora in order to investigate (a) the categories of mentor evaluations and suggestions and (b) the alignment between these comment categories and 18 SKIVE elements of an epistemic frame. A journalism practicum corpus consisted of a journalism professor providing copyedit comments on news stories submitted by student journalists over the course of a semester-long practicum, i.e., line-by-line reactions and suggestions to a news story. The *journalism practicum corpus* consisted of 426 comments. These comments were segregated into 443 evaluation segments and 620 suggestion segments by a graduate research assistant. The comments tended to be copious, detailed, and blunt. A second *game corpus* consisted of copyedit comments from *science.net,* a game designed for middle school students to work as science reporters and publish news stories in an online newspaper. Graduate students outside of the field of journalism were trained to give feedback in the game. These comments were less blunt, but they allegedly retained the salient features of the professional journalists. There were 1620 comments in the game corpus. A sample of comments in these corpora were analyzed on 18 SKIVE elements by two graduate students. They achieved a respectable interjudge reliability score (kappa =.76) when averaging over the 18 SKIVE elements.

Table 2. Elements of Epistemic Frames.

SKIVE element	Description
Skill: investigating	Ability to gather information for a story.
Skill: detail	Need to provide useful information, specific details, and facts in stories.
Knowledge: story	Terms concerning language used by journalists about their stories.
Knowledge: reporting	Terms about reporting, finding, gathering, and analyzing information for stories
Knowledge: Reader	Knowledge about what the reader wants and reader attributes.
Identity: Writer	Feedback on being a writer, including projective identity references that position journalist as a writer.
Value: Informing public	Informing the public on what they want to know about important issues in the community.
Value: Engaging reader	Maintaining the readers' attention by phrases that hook the reader.
Epistemology: Rich details	Guiding principle to tell stories by showing rather than telling and using details to bring the story alive.

Table 1 gives examples of some of the evaluation and suggestion segments. We segregated evaluations and suggestions, crossed with three descriptions: (a) an actual comment verbatim (exemplar from corpus), (b) a succinct summary statement that represents a cluster category of exemplars, and (c) a symbolic specification that could be used as a template to generate comments in AutoMentor when values are bound to elements or parameters within a particular game activity context.

Table 2 presents 9 SKIVE elements in an epistemic frame for an expert in journalism. These were based on ethnographic notes prepared by a graduate student on the comments of the journalism professor. There were 18 SKIVE elements altogether, but the frequencies of some of the elements were too low to analyze. For the 9 elements in Table 2, the proportions of comments that possessed the element in the journalism practicum were .29, .34, .79, .36, .32, 10, .21, 22, and .25, respectively.

4 Analyses of Journalism Comments

Evaluation and Suggestion Categories. This analysis was conducted to discover the clusters of comments within the evaluation and suggestion speech acts. In essence, what were the fundamental clusters of comments expressed by the mentor in the journalism practicum corpus? We computed a similarity matrix between all possible pairs of the 443 evaluation segments using latent semantic analysis (LSA)[14]. In the same fashion we computed a similarity matrix on the 620 suggestion segments.

Each similarity score between segment A and B was computed as the geometric cosine between the two expressions via the dimensions in an LSA space. Our LSA space was based on the Touchstone Applied Sciences Associates (TASA) corpus of 37,651 documents with approximately 11 million words. This is a frequently used corpus to represent what typical high school students have read over their lifetimes. LSA is a useful method of computing similarity because it considers implicit knowledge in addition to the explicit words. LSA is a statistical technique for representing world knowledge, based on a large corpus of documents [14]. A single value decomposition technique is performed on the large document-by-word matrix (from the TASA corpus) that specifies the number occurrences of particular words in particular documents. It reduces the large sparse matrix to approximately 300 dimensions. The conceptual similarity between any two text excerpts is computed as the cosine between the values and weighted dimensions of the two text excerpts.

A principal components (PC) analysis was then conducted on the similarity matrix of the 443 evaluation segments. The top 20 components accounted for 79.2% of the variance in similarity scores. The evaluation segments are clustered according to the sorted loadings on the top 20 components with Varimax rotation. The same analysis was conducted on the similarity matrix of the 620 suggestion segments, with the top 30 components accounting for 85% of the variance, generating 30 clusters of suggestion segments. Table 1 shows exemplars of a few of these clusters.

The PC analysis is useful because we can induce what specific comment categories are relevant in a particular discourse context, in this case the journalism practicum course. This is indeed a useful discovery technique but hardly the end of the story in our analysis of games and mentors. There are at least three fundamental follow-up

questions. First, what states of the game and interactions among students systematically trigger a particular category of evaluation or suggestion? Second, how are these evaluation and suggestion moves (illustrated in Table 1) aligned with the epistemic frame elements illustrated in Table 2? Third, how well can we generalize these categories form the journalism practicum corpus to both a similar corpus and to a different game corpus? It is beyond the scope of this paper to answer the first question, but we did conduct analyses relevant to the second and third questions.

Aligning Evaluation and Suggestion Moves to Epistemic Frame Elements. We conducted some binary logistic regression (BLR) analyses that attempted to predict each epistemic frame element from the principle component loadings of the 20 evaluation and 30 suggestion clusters. We focused on the 9 epistemic frame elements in Table 2 because they had a large enough frequency of occurrences.

Consider first the journalism practicum corpus. We randomly split the observations in half to prepare a training and test set. A BLR analysis was performed on the training set with component scores as predictors and the binary presence of a frame element as the criterion variable. The coefficients from this test set were applied to the test set to generate predictions about presence or absence of a frame element. The predictions from the BLR analysis (binary 0/1 values) were compared with the binary decisions of the human judges. A kappa score served as an index of the accuracy of the predictions; kappa adjusts for base rates and varies from 0 (chance) to 1 (perfect accuracy).

Table 3 shows the kappa scores for the journalism practicum corpus. The left four columns of numbers segregate training versus test sets for suggestion versus evaluation segments. The results support the claim that the component scores indeed can significantly predict the epistemic frame elements. The mean kappa scores were .76, .63, .54, and .50 for training-suggestions, training-evaluations, test-suggestions, and test-evaluations, respectively.

Table 3. Prediction (kappa) of Epistemic Frame Elements from Mentor Comments.

	Journalism Practicum				Game			
	Suggestion		Evaluation		Suggestion		Evaluation	
SKIVE element	Train	Test	Train	Test	Train	Test	Train	Test
Skill: investigating	.79	.57	.68	.53	.83	.44	.83	.56
Skill: detail	.66	.53	.58	.45	.53	.16	.33	.09
Knowledge: story	.55	.40	.58	.43	.59	.49	.57	.54
Knowledge: reporting	.68	.48	.58	.50	.58	.38	.52	.42
Knowledge: Reader	.91	.81	.88	.75	.89	.64	.84	.76
Identity: Writer	1.00	.36	.63	.35	.76	.56	.82	.62
Value: Informing public	.78	.60	.75	.69	.63	.51	.65	.38
Value: Engaging reader	.71	.54	.49	.44	.57	.40	.31	.12
Epistemology: Rich details	.74	.59	.50	.36	.64	.16	.38	.11

We next considered whether our PC categories in the journalism corpus can generalize to the game corpus. For the game corpus, 332 observations were coded on epistemic frame elements by graduate students. We used the coefficients derived from the journalism corpus to compute component scores and predict these coded

elements in the game corpus. Once again, we randomly segregated training and test observations for suggestions and evaluations. In the training BLR we computed a new set of regression coefficients on the game sample; these new regression coefficients were applied to the test sample of the game corpus.

The right 4 columns of Table 3 support the claim that there is some modest transfer of the PC speech act clusters to the game corpus. The mean kappa scores were .67, .58, .42, and .40 for training-suggestions, training-evaluations, test-suggestions, and test-evaluations, respectively. Therefore, we could imagine a methodology in which researchers (a) hand code epistemic frame elements of a modest sample of observations in the new discourse context, (b) compute LSA similarity scores on the new corpus (segregating evaluation and suggestion segments), (c) compute component score coefficients, (d) derive BLR formulas for the observations in a, and (e) compute the predicted elements for other segments in the new corpus.

It should be noted, however, that transfer is quite modest when we do not go through the above process that has humans hand code a sample of observations in the new corpus on epistemic frame elements. For example, we computed a BLR analysis on the 443 evaluation segments in the journalism corpus as a training set and used the same regression coefficients to predict the elements in the game corpus as a test set. The kappa scores for the training and transfer test set were .61 and .23, respectively. It appears that some hand coding in a new game activity is needed for these judgments of epistemic frame elements. However, it is an open question as to how much hand coding is needed. It is also conceivable that the amount of necessary hand coding will decrease substantially as we explore a greater number and diversity of game activities.

5 Discussion

This paper has identified the speech act categories and clusters of comments of journalism mentors who interact with students editing news stories. The basic speech act categories are evaluations and suggestions, in addition to the normal greetings, introductions, and closing moves that speech participants normally perform in multiparty conversation. LSA and principal components analyses helped us discover the different types of evaluations and suggestions. These comments of mentors were also significantly aligned with epistemic frame elements (see Table 2). Such alignments were discovered by logistic regression and validated on a sample of hand-coded judgments of the frame elements. There was some modest transfer from a journalism practicum corpus to a game corpus, but more research is needed to explore how the transfer can improve.

These analyses provide an important first step in building a virtual AutoMentor for multiparty epistemic games. We now have a sketch of AutoMentor's speech act categories, comments, and the epistemic functions that they serve. The next step is to formulate algorithms that generate particular speech acts and discourse moves in a fashion that is sensitive to the game states and interactions among students. The generation algorithms need to be sensitive to top-down goals (i.e., what epistemic frame elements are operating in the current game activity) and bottom-up constraints

(e.g., conflicts between students, progress in the game activity). Moreover, the generation templates in column 3 of Table 1 provide one approach to binding a discourse move to the setting and game parameters at particular points in time. The success of these approaches await the future development and testing of AutoMentor.

Acknowledgments. This research was supported by the National Science Foundation (BCS 0904909, DRK-12-0918409) and the Institute of Education Sciences (R305B070349, R305A080594). The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

1. Shaffer, D.W.: How Computer Games Help Children Learn. Palgrave, New York (2007)
2. Hatfield, D., Shaffer, D.W.: The Epistemology of Journalism 335: Complexity in developing journalistic experience. In: ICLS, Chicago, IL (2010)
3. Bagley, E., Shaffer, D.W.: When People Get in the Way: Promoting Civic Thinking Through Epistemic Game Play. International Journal of Gaming and Computer-Mediated Simulations. 1, 36--52 (2009)
4. Graesser, A.C., Lu, S., Jackson, T.T., Mitchell, H., Ventura, M., Olney, A., Louwerse, M.M.: AutoTutor: A Tutor with Dialogue in Natural Language. Behavior Research Methods, Instruments, and Computers. 36, 180--193 (2004)
5. Graesser, A.C., Jeon, M., Dufty, D.: Agent Technologies Designed to Facilitate Interactive Knowledge Construction. Discourse Processes, 45, 298--322 (2008)
6. D'Mello, S., King, B., Chipman, P., Graesser, A.C.: Towards Spoken Human-Computer Tutorial Dialogues. Human Computer Interaction, In Press
7. Millis, K., Cai, Z., Graesser, A., Halpern, D., Wallace, P.: Learning Scientific Inquiry By Asking Questions in an Educational Game. In: Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2009, pp. 2951--2956. AACE, Chesapeake, VA (2009)
8. Jackson, G.T., Olney, A., Graesser, A.C., Kim, H.T.: AutoTutor 3-D Simulations: Analyzing User's Actions and Learning Trends. In: Proceedings of the 28th Annual Meeting of the Cognitive Science Society, pp. 1557--1562. ErlBaum, Mahway, NJ (2006)
9. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
10. Gee, J.P.: An Introduction to Discourse Analysis: Theory and Method. Routledge, London (1999)
11. Graesser, A.C., McNamara, D.S.: Computational Analyses of Multilevel Discourse Comprehension. Topics in Cognitive Science, In Press
12. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, Upper Saddle River, NJ (2008)
13. VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., Rose, C.P.: When Are Tutorial Dialogues More Effective Than Reading?. Cognitive Science, 31, 3-62 (2007)
14. Landauer, T., McNamara, D.S., Dennis, S., Kintsch, W. (eds.): Handbook of Latent Semantic Analysis. Erlbaum, Mahwah, NJ (2007)

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V.,
Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 29-36)

Towards Intelligent Learning Environments for Scientific Argumentation

Nancy L. Green

University of North Carolina Greensboro
Greensboro, NC 27402 USA
nlgreen@uncg.edu

Abstract. This paper describes the proposed design of an intelligent learning environment for scientific argumentation (ILESA) in the domain of genetic medicine. Informed by a computational model of argumentation, it will support student argument construction and debate. The computational model is based upon our previous research on argument generation in GenIE, a system for increasing transparency of argumentation for genetic counseling clients. Some challenges in design of the ILESA are presented.

Keywords: Intelligent Learning Environments, Argumentation, Science Education

1 Introduction

Argumentation plays an important role in science. Through scientific debate, claims are supported with evidence, subjected to critical appraisal, challenged by alternative explanations and by conflicting data, and defended, modified, or retracted. Thus, it is not surprising that there is significant interest within the field of science education in argumentation (e.g., Bell and Linn 2000; Bricker and Bell 2008; Jiménez-Aleixander et al. 2000; Sandoval and Reiser 2004; Toth et al. 2002; Zohar and Nemet 2002). Argumentation can be used to help students learn science content and to help them better understand the nature of the scientific enterprise, scientific discourse, and scientific knowledge (Bricker and Bell 2008). Recognizing the central role of argumentation in a range of fields, Chinn contends that "learning to argue well should be a central goal of education" (2006, p. 359).

This paper describes the proposed design of an Intelligent Learning Environment for Scientific Argumentation in the domain of Genetic Medicine, ILESA-GenMed. The learning environment could be used as an adjunct to a high school science course or a college biology course designed for non-science majors. The course objectives might include learning to construct arguments and counterarguments about clinical genetics cases in order to increase understanding of scientific thinking, interest in pursuing future studies in science, and critical-thinking skills. Informed by a computational model of argumentation, the learning environment will support student argument construction and debate. A student will be able to construct arguments and counterarguments by selecting "snippets" of text from the learning environment, e.g., a snippet from a fictitious medical record or from a report generated by an imaginary clinical test that the student ordered. The learning environment will provide intelligent assistance in constructing arguments and in evaluating their strengths and weaknesses. In addition, it will support debate with a "virtual" student.

The foundation for the approach to be taken in ILESA-GenMed is the computational model of argumentation developed for the GenIE project (Green 2005; Green 2006; Green 2007; Green 2008; Green 2010a; Green 2010b; Green et al. in preparation). The overall goal of the GenIE project was to make biomedical argumentation more transparent to a lay audience. The research was grounded in the domain/genre of letters written by genetic counselors to their clients. The letters document the reason for a patient's referral to a genetics clinic (e.g. hearing loss), the findings (e.g. a mutation was found in the patient's GJB2 gene and the patient has no syndromic features), the diagnosis (e.g. the hearing loss is due to autosomal recessively inherited mutation of GJB2), and the implications for the biological family members of the patient (e.g. the parents must be carriers of the mutation). Typically the letters contain incomplete arguments in support of the clinic's conclusions. Given findings about a case and a claim such as the patient's diagnosis, GenIE reconstructs the full argumentation for each claim. A qualitative causal domain model of biomedicine (Green 2005) is used to provide warrants for the argumentation. A set of

abstract argumentation schemes representing the patterns of argumentation found in this genre is used by GenIE to compose arguments for a given claim from the findings for a case and the domain model.

The GenIE project's strategy for increasing transparency of these letters was to make the full argumentation for claims in a letter accessible to the intended recipient. However, transparency could be increased further by enabling a client to explore possible counterarguments (and defenses against those counterarguments, and so on). Thus, a prototype interactive interface to GenIE's argument generator was developed (Green 2008). The interface enables a user to request counterarguments to each component of an argument: claim, data, warrant (Toulmin 1958), and to request challenges to those counterarguments, etc. Although GenIE's argument generator was designed to support patient education, it is clear that it can be used to support ILESA-GenMed as well. The rest of this paper discusses the proposed design of ILESA-GenMed in more detail. In the next section, we give a brief description of relevant concepts from argumentation theory. Then argumentation in genetic medicine and the computational model of argumentation used in GenIE are described. In section 4, we discuss critical challenges in adapting the model to support ILESA-GenMed. Lastly, we present a scenario to illustrate use of ILESA-GenMed and to motivate the proposed extensions to GenIE's argument generator.

2 Background in Argumentation Theory

The argumentation theorist who has been most influential in the learning sciences to date, Toulmin (1998) was concerned with modeling argument acceptability in areas such as law and science where, he argued, logical validity is too restrictive a criterion for determining argument acceptability. According to his analysis of the six components of an argument, the *claim* is an assertion to be established by the argument, the *data* is evidence for the claim, the *warrant* is a field-dependent principle linking data to claim, the *backing* provides support for the warrant, the *qualifier* expresses the strength of the conclusion, and the *rebuttal* is a possible challenge to the conclusion. Toulmin's model has been applied to informal analysis of arguments in medical (e.g. Jenicek and Hitchcock 2005) and science education (e.g. Suthers et al. 1995; Jiménez-Aleixandre et al. 2000; Bell and Linn 2000). In addition, Toulmin's ideas have influenced and been refined, extended, and formalized in artificial intelligence (Verheij 2009).

According to the argumentation theorist Walton (2009), there have been different definitions of *argument* and *argumentation* reflecting different perspectives. In one view, an argument is a set of premises and a conclusion; a formal model such as deductive logic determines whether the argument is valid. Another perspective "is called dialogical (or dialectical) in that it looks at two sides of an argument, the pro and the contra" (p. 4). In the dialogical view, the acceptability of an argument is determined by considering the strengths and weaknesses of the competing arguments. An argument can be "attacked" in a number of ways, e.g., by questioning the accuracy or relevance of its premises, providing a counterargument to its conclusion, or asking a "critical question" that raises legitimate doubts about the acceptability of the argument. In addition to deductive validity, an argument can be evaluated by standards of inductive or defeasible reasoning. Formalized in artificial intelligence to handle situations where an agent lacks complete or probabilistic knowledge, defeasible reasoning is used to draw tentative conclusions that may be retracted as new information is acquired (e.g., Birds fly; Tweety is a bird; therefore, assume Tweety flies, at least until you learn that he has a broken wing or ...)

According to Walton et al. (2008), argumentation schemes are abstract descriptions of forms of arguments used in everyday conversation and legal and scientific argumentation, and many schemes reflect defeasible rather than deductive or inductive reasoning. In Walton's formulation, an argumentation scheme consists of a parameterized description of the major premise, minor premise, conclusion, and list of critical questions. The critical questions "represent standard ways of critically probing into an argument to find aspects of it that are open to criticism" (p. 7). The major premise is "best seen as a defeasible generalization, and the argument is defeasible, subject to the asking of critical questions" (p. 7). For example, the Argument from Cause to Effect scheme is described as follows (p. 168): [Major premise] *Generally, if A occurs, then B will (might) occur.* [Minor premise] *In this case, A occurs (might occur).* [Conclusion] *Therefore, in this case, B will (might) occur.* Three critical questions are associated with the scheme: [C1] *How strong is the causal generalization (if it is true at all)?* [C2] *Is the evidence cited (if there is any) strong enough to warrant the generalization as stated?* [C3] *Are there other factors that would or will interfere with or counteract the production of the effect in this case?*

Walton et al. (2008) provide descriptions of a total of 60 argumentation schemes. "The study of argumentation schemes by Walton and others has made a start with the systematic specification of context-dependent, defeasible, concrete standards of argument assessment, as sought for by Toulmin" (Verheij 2009, 233). In artificial intelligence, argumentation schemes have been specified for medical (Fox et al. 2007; Lindgren and Eklund 2005; Rahati and Kabanza 2009) and legal reasoning (Verheij 2003a), and in our project GenIE for modeling biomedical arguments written by genetic counselors.

3 A Computational Model of Argumentation for Genetic Medicine

Genetic medicine is a domain that for the most part cannot be characterized by deductive reasoning from unequivocally true premises. Diagnosis is based upon abductive reasoning from evidence and principles of genetic medicine. Evidence from clinical findings is not necessarily reliable; for example, a test to confirm the presence of a certain mutation may yield a false positive or false negative. Principles of genetic medicine describe a causal, non-deterministic model of inheritance and the effect of genetics on health. Of course, the principles themselves may be questioned. Nevertheless, a corpus study of genetic counseling patient letters identified a variety of argumentation schemes used for communication with a lay audience (Green 2006; Green et al. in preparation).

Each domain model used in GenIE represents an instance of the more abstract, simplified, causal conceptual model of genetic disease that was identified in the corpus of genetic counseling patient letters (Green 2005). The study identified a small set of abstract concepts (*genotype, history, symptom, test result*, etc.) shown to have good inter-rater reliability. In addition, causal/probabilistic relations involving these concepts were specified, e.g., *history* (such as ethnic origin) may increase risk of *genotype* (such as mutation of the CFTR gene), while *genotype* may lead (through *biochemistry*) to *symptoms* and *test results*. The information for each domain model is represented computationally as a qualitative probabilistic network (Druzdzel and Henrion 1993; Wellman 1990). In the network, instances of the abstract concepts are represented as random variables and causal/probabilistic relations are defined in terms of formal qualitative influence and synergy relations. Before generating arguments for a particular patient, the domain model for the related genetic diseases is updated with information about the patient's case (his symptoms, the clinic's diagnosis, etc.). The domain model distinguishes information about a specific patient case (states of variables) from principles of genetic medicine, which are encoded via the influence and synergy relations in the network. This approach to domain modeling enables argumentation schemes to be represented in terms of variables and formal constraints, rather than concepts from genetic medicine.

Argumentation schemes were derived by analysis of arguments in the corpus of genetic counseling letters. For each argument, first, the implicit (Green 2010b) or explicit components -- claim, data, and warrant -- were identified, where the data and claim describe the patient and the warrant is a principle of genetic medicine. Next, claim, data and warrant were mapped to the presumed underlying domain model. The claim and data were analyzed in terms of states of variables, and the warrant was analyzed in terms of the influence and synergy relations. Finally, a formal specification of the argumentation was abstracted. For example, a scheme related to that described in (Walton et al. 2008) as Argument from Cause to Effect (discussed above) is formulated in GenIE as follows (Green et al. in preparation):

Components of scheme	**Example**
Claim: B $\geq$ b	Patient has or will have heart disease
Data: A $\geq$ a	Patient has Familial Hypercholesterolemia (a genetic condition associated with a certain genotype)
Warrant: S*(<A,a>,<B,b>)	Having Familial Hypercholesterolemia increases the risk of heart disease
Applicability Constraint:	Unless
[$\neg\exists C$: $Y(\{C,A\},\neg<B,b>)$: $C \geq c$] &	(1) one is taking preventive medication, and
[$\neg\exists C$: $Y^{+}(\{C,A\},<B,b>)$: $C < c$]	(2) one has no other risk factors (e.g. obesity)

In the above argumentation scheme from GenIE, data and claim represent states of variables *A* and *B*, respectively, and the warrant describes a chain of one or more causal influences in the domain model from *A* to *B*, where *a* is a threshold value such that *A* reaching *a* increases the probability that *B* reaches its

threshold value *b*. The Applicability Constraint of an argumentation scheme in GenIE is a set of critical questions representing other variables that may play a role in the applicability of the associated argumentation scheme. To paraphrase the critical questions for the above example, (1) is there a *C* such that *C* has reached its threshold and is thereby preventing *A* from leading to *B*?, (2) is there a *C* required to enable *A* to lead to *B* but *C* has not reached its threshold? When a domain model contains a variable C satisfying (1) or (2), the scheme would not be used to generate an argument for a patient letter. In an interactive version of GenIE (Green 2008), a user could ask if any of the questions in the applicability constraint hold, thereby defeating the argument. Note that this applicability constraint is a more precise way of formulating critical question C3 of Walton et al.'s (2008) scheme. (Questions C1 and C2 are not included with the above scheme since in the GenIE corpus there were no cases where the healthcare provider voiced doubts over the genetic principles involved.) A variety of other argumentation schemes were specified in GenIE based upon arguments found in the corpus. The schemes reflect causal, diagnostic (e.g. argument from effect to cause, argument by elimination of possible causes) and other argumentation patterns. The GenIE argument generator can construct arguments for or against a given claim by instantiating and composing argumentation schemes.

4 Challenges in Adapting GenIE to Support ILESA

There will be several challenges in adapting GenIE's argumentation generator to support ILESA. First of all, by design GenIE's use of argumentation schemes is domain-independent. As shown in section 3, the schemes and associated critical questions are specified abstractly in terms of formal properties of the domain model and functional role in argumentation. However, to support ILESA-GenMed it is necessary to provide a rich set of domain-specific critical questions. For example, according to Jenicek's textbook on evidenced-based medicine (2003), there are many issues to consider in evaluating the quality and scope of a causal generalization derived from medical research. They include the level of evidence (from expert opinion to randomized controlled clinical trial), the design of the study, etc. In terms of the Cause to Effect scheme used in GenIE (described above), these critical questions could be used to challenge the warrant of the scheme. Following Jenicek's textbook, other critical questions could be added to enable the student to challenge the data of the scheme used in GenIE to argue from Effect to Cause, e.g., *What is the probability of a false positive or false negative?*.

The more serious challenge will be to provide support for evaluating argumentation based upon uncertainty. A problem arises when the expected answer to a critical question is a specification of degree, rather than a simple yes or no. The problem is that the dialectical process for challenging an argument described by Walton et al. does not address how to handle answers to degree questions. To illustrate, consider an Argument from Cause to Effect for the claim that a certain 35-year-old patient will die in the next ten years. The data is the finding that he has the mutation for Huntington's Disease. The warrant is based upon survival data for this genetic disease. Consider the first critical question of the Cause to Effect scheme, *How strong is the causal generalization?*. An appropriate answer might be a quantification of the mortality risk for 35-year-old males with Huntington's Disease. The problem is specifying at what level of risk this argument is defeated.

That problem is related to the problem of evaluating whether a probabilistic claim is justified. Many claims in the corpus are implicitly or explicitly qualified by degree of certainty. For example, suppose that an argument is given for the claim that a patient's risk of developing breast cancer is high (i.e., higher than some baseline risk). The data is that she tested positive for the BRCA1 mutation, indicating that she has inherited one copy of the mutation. The warrant is the so-called "two-hit" model, according to which a cancer develops when the second non-mutated copy of the BRCA1 gene undergoes spontaneous mutation in any cell. However, analysis of this argument in terms of Walton et al.'s Cause to Effect scheme leaves out the qualification of degree of risk from the claim. The claim would be represented as "the patient will or may have breast cancer". Applying the dialectical process proposed by Walton et al. to this argument only allows one to decide whether to accept or reject it, but not to evaluate the degree of certainty expressed in the conclusion.

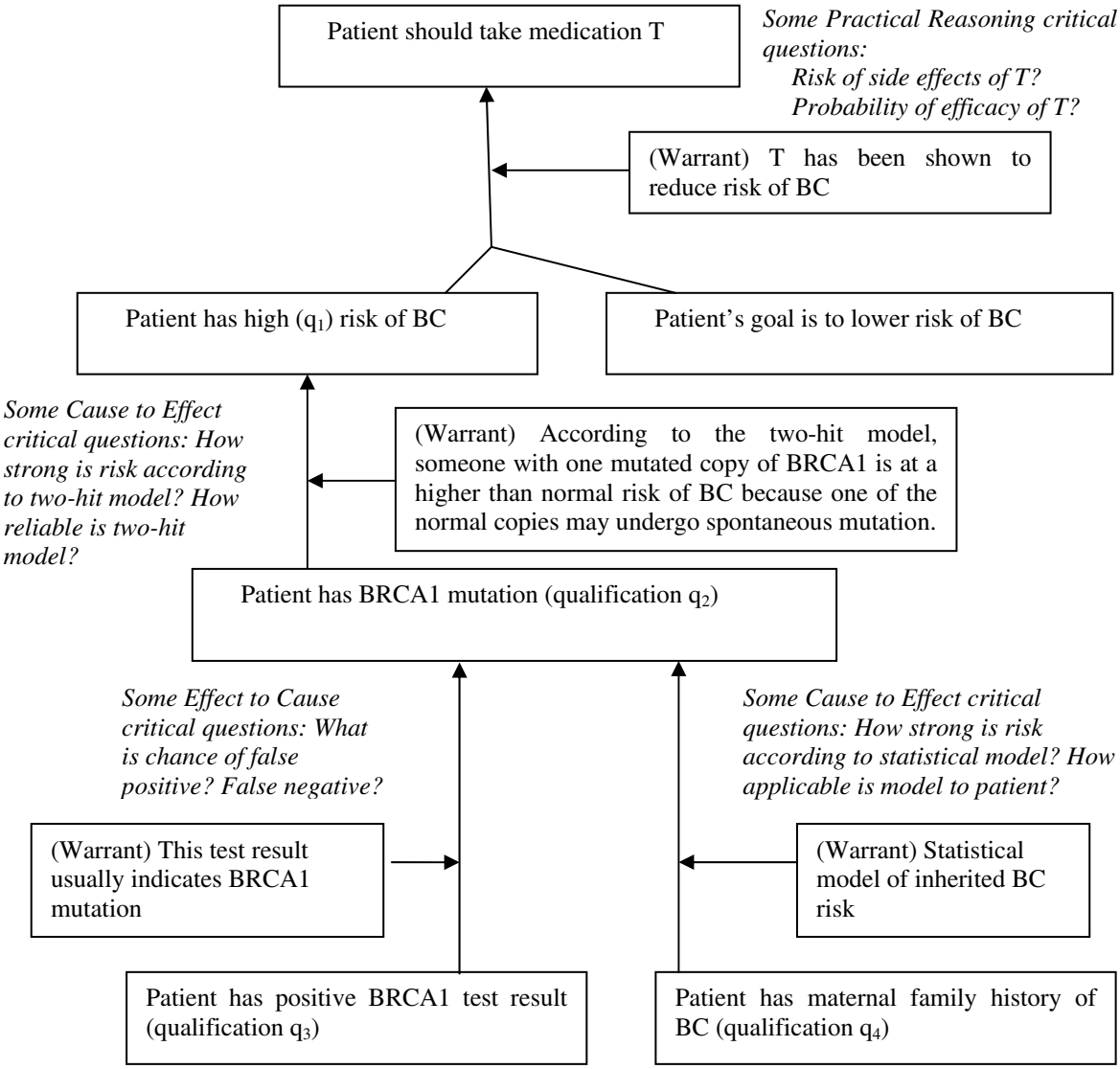


Fig. 1. Diagram of argumentation that patient should take medication T.

A related problem is how to evaluate argumentation composed of subarguments, which compounds the problems noted above (Figure 1). For example, the recommendation that a patient should take a certain cancer prevention medication could be analyzed in terms of an argumentation scheme for Practical Reasoning (Walton et al. 2008, p. 323). The claim is that the patient should take a certain medication T; the warrant is that T has been shown to reduce the risk of breast cancer; the data is that the patient has a higher than normal risk of breast cancer and the patient's goal is to reduce that risk. One of the critical questions of this argument is *what is the probability of the efficacy of T?*. Another is *what is the probability of deleterious side effects of T?*. Now, consider the subargument for the statement that the patient has a higher than normal risk of breast cancer, which could be analyzed in terms of a Cause to Effect scheme. The claim is that the patient has a higher than normal risk; the data is that she has a BRCA1 mutation; the warrant is the two-hit model described above. Critical questions include *How high is the risk that someone with the BRCA1 mutation will have breast cancer?*. Two independent subarguments for the claim that she has a BRCA1 mutation could be given: (1) Her BRCA1 test result was positive, and (2) She has a maternal family history of breast cancer. Critical questions that could be raised about each subargument, respectively, are (1) *What is the chance it was a false positive test result?*, and (2) *How strong is her risk according to the statistical model used, and how applicable is that model to this patient?*. (Note that

determining the proper qualification q_2 requires a procedure for combining evidence, since the strength of the belief that the patient has a BRCA1 mutation may be different based upon subargument (1) or (2). Furthermore, certain types of evidence, such as test results, might have more weight in medicine than other types, such as the output of a statistical model based upon epidemiological data.)

In summary, the dialectical mechanism of debate as described in Walton et al. (2008) is not sufficient to address the challenge of evaluating argumentation in this domain.

5 Envisioning ILESA-GenMed

This section describes how we envision the use of ILESA-GenMed in education. It motivates the future extensions to GenIE, described in the previous section, to support a scientific argumentation learning environment. Finally, it discusses the relevance of this paper to learning in ill-defined domains. To begin, Figure 2 (below) shows the end result of a possible debate between two students, Pro and Con, on the issue of whether a certain patient should take medication T.[1] Suppose that Pro is a real student taking an introductory college biology course for non-majors. Pro has been given the task of arguing in favor of the patient taking medication T. The role of Con could be played by another student in the class or by a virtual student, controlled by ILESA-GenMed. For this scenario, assume that Con is a virtual student. The diagram shown in the figure could have been constructed in several stages as follows.[2]

First, Pro browses the medical record of the (fictitious) patient, noticing the patient's maternal family history of breast cancer (D_5). Using a risk estimation tool (W_5) provided in the learning environment, the student inputs the patient's family history and verifies that the patient has a higher than normal risk of carrying the BRCA1 mutation (C_5). Next, by browsing through multimedia background information on inheritance and breast cancer provided by the ILESA, Pro selects other relevant warrants (W_2, W_3). Manipulating these snippets of text[3], Pro constructs the chained argument (C_2, W_2, D_2 & (C_3, W_3, (C_5, W_5, D_5))) shown near the center of Figure 2 for the claim (C_2) that the patient should take medication T. Other building blocks of this chained argument, such as C_3 and D_2, could be provided by the interface under a list of conjectures, including irrelevant claims and claims that cannot be justified. In addition to providing tools for acquiring evidence (e.g. laboratory tests, risk assessment models) and text snippets, the ILESA will provide a library of argumentation scheme descriptions (Practical Reasoning, Cause to Effect, etc.) with their associated domain-specific critical questions.

Next, Con constructs an unacceptable argument (C_6, W_6, D_6) in an attempt to defeat an element (C_5) of the chained argument constructed by Pro. (In the figure, attacking arguments are linked by arcs ending in diamonds, and the text of Con's contributions is shown in italics.) To defeat this attack, Pro must recognize that although D_6 and W_6 are true, they do not license the claim C_6. After Pro adds this rebuttal (R_8) to the diagram, Con attacks C_5 again, this time with an acceptable argument (C_4, W_4, D_4). Next, Pro formulates a counterargument (C_7, W_7, D_7) based upon the critical question of the argumentation scheme used in Con's argument, i.e., *What is the chance of a false negative?* Lastly, Con constructs a counterargument (C_1, D_1, W_1) to C_2. Although not shown in the figure, note that the debate could continue, e.g., if Con were to propose alternative interventions with less risky side effects.

The current GenIE computational model of argumentation could be used in ILESA-GenMed to support intelligent assistance in argument construction and to control a virtual debater (such as Con in the example). The addition of domain-specific critical questions as described in section 4 would provide more (and better) ways of challenging an argument than currently used in GenIE. The addition of a mechanism for evaluating argumentation as proposed in section 4 is necessary for automated analysis of the strengths and weaknesses of the argumentation, given the uncertainty of many argument components (e.g. C_3) and the use of preferences (e.g. D_1 and D_2) in some argumentation schemes.

In conclusion, the proposed ILESA-GenMed can be characterized as a *discovery support system* for an ill-defined domain according to criteria surveyed in (Lynch et al. 2006). The student's task is to produce argumentation that explains certain observations or that refutes an argument supported by other

[1] Other types of claims that could be debated include, e.g., claims for a certain genetic diagnosis or about the source of inheritance of a patient's genetic condition.

[2] The debate is shown diagrammatically for conciseness. An alternative style of representation could be provided by the user interface, e.g. as a threaded discussion or hypertext document.

[3] The motivation for providing snippets is to avoid the problem of interpreting unrestricted natural language input.

observations. There is no formal theory by which to assess the correctness of a solution, although GenIE's computational model of argumentation can be leveraged to perform automated assessment, to provide intelligent assistance, and to challenge the student by playing the role of a virtual participant in a debate.

Acknowledgments

Project GenIE was supported by the National Science Foundation under CAREER Award No. 0132821.

References

Bell, P., and Linn, M. 2000. Scientific Arguments as Learning Artifacts: Designing for Learning from the Web with KIE. *International Journal of Science Education*, 22(8), 797-817.

Bricker, L. A., and Bell, P. 2008. Conceptualizations of Argumentation from Science Studies and the Learning Sciences and their Implications for the Practices of Science Education. *Science Education*, 92: 473-498.

Chinn, C. A. Learning to Argue. 2006. In O'Donnell, A. M., Hmelo-Silver, C.E., and Erkins, G. (Eds.), *Collaborative Learning, Reasoning, and Technology* (pp. 355-383). Mahwah, NJ: Erlbaum.

Druzdzel, M. J., and Henrion, M. 1993. Efficient Reasoning in Qualitative Probabilistic Networks. In *Proceedings of the 11th National Conference on AI*, 548-553. Washington, DC.

Fox, J., Glasspool, D., Grecu, D., Modgil,S., South, M., and Patkar, V. 2007. Argumentation-based inference and decision-making – a medical perspective. *IEEE Intelligent Systems* 22(6):34-41.

Green, N. 2005. A Bayesian network coding scheme for annotating biomedical information presented to genetic counseling clients. *Journal of Biomedical Informatics* 38, 130-144.

Green, N. 2006. Generation of Biomedical Arguments for Lay Readers. In *Proceedings of International Conference on Natural Language Generation*, 114-121.

Green, N. 2007. A Study of Argumentation in a Causal Probabilistic Humanistic Domain: Genetic Counseling. *International Journal of Intelligent Systems. Special Issue: Computational Models of Natural Argumentation* 22(1): 71-93.

Green, N. 2008. Dialectical Argumentation in Causal Domains. In *Proceedings of 8th International Workshop on Computational Models of Natural Argument*, 31-38.

Green, N. 2010a. Analysis of communication of uncertainty in genetic counseling patient letters for design of a natural language generation system. *Social Semiotics*, 20(1):77-86.

Green, N. 2010b. Representation of Argumentation in Text with Rhetorical Structure Theory. *Argumentation*, 24(2):181-196.

Green, N., R. Dwight, K. Navoraphan, and B. Stadler. In preparation. Natural language generation of transparent arguments for lay audiences.

Jenicek, M. 2003. *Foundations of Evidence-Based Medicine.* New York: Parthenon Publishing.

Jenicek, M., and Hitchcock, D. L. 2004. *Evidence-Based Practice: Logic and Critical Thinking in Medicine*. Chicago, IL: American Medical Association Press.

Jiménez-Aleixander et al. 2000; Jiménez-Aleixandre, M. P., Rodríguez, A. B., and Duschl, R. A. 2000. "Doing the Lesson" or "Doing Science": Argument in High School Genetics. *Science Education* 84: 757-792.

Lindgren, H. and Eklund, P. 2005. Differential diagnosis of dementia in an argumentation framework. *Journal of Intelligent and Fuzzy Systems* 16:1-8.

Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). "Defining Ill-Defined Domains; A literature survey." In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.), Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems,1-10.

Rahati, A., and Kabanza, F. 2009. Persuasive argumentation in a medical diagnosis tutoring system. In *Working Notes of Computational Models of Natural Argument IX*, July 13, Pasadena CA, pp.39-48.

Rahwan, I. and Simari, G. R., eds. 2009. *Argumentation in Artificial Intelligence.* Dordrecht, Springer.

Sandoval and Reiser 2004; Sandoval, W. A., and Reiser, B. J. 2004. Explanation-Driven Inquiry: Integrating Conceptual and Epistemic Scaffolds for Scientific Inquiry. *Science Education* 88: 345-372.

Suthers, D., Weiner, A., Connelly, J., and Paolucci, M. 1995. Belvedere: Engaging Students in Critical Discussion of Science and Public Policy Issues. In *Proceedings of the 7th International Conference on AI in Education*, 266-273.

Toth, E. E., Suthers, D. D., and Lesgold, A. 2002. "Mapping to Know": The Effects of Representational Guidance and Reflective Assessment on Scientific Inquiry. *Science Education* 86: 264-286.

Toulmin, S.E. 1998. *The uses of argument*, Cambridge: Cambridge University Press.

Verheij, B. 2003. Dialectical Argumentation with Argument Schemes: An Approach to Legal Logic. *Artificial Intelligence and Law* 11: 167-195.

Verheij, B. 2009. The Toulmin Argument Model in Artificial Intelligence. In Rahwan and Simari, eds., 219-238.

Walton, D. 2009. Argumentation Theory: A Very Short Introduction. In Rahwan and Simari, eds., 1-22.
Walton, D., C. Reed, and F. Macagno. 2008. *Argumentation Schemes*, Cambridge: Cambridge University Press.
Wellman, M.P. 1990. Fundamental Concepts of Qualitative Probabilistic Networks. *Artificial Intelligence* 44(3): 257-303.
Zohar, A., and Nemet, F. 2002. Fostering Students' Knowledge and Argumentation Skills through Dilemmas in Human Genetics. *Journal of Research in Science Teaching* 39(1): 35-62.

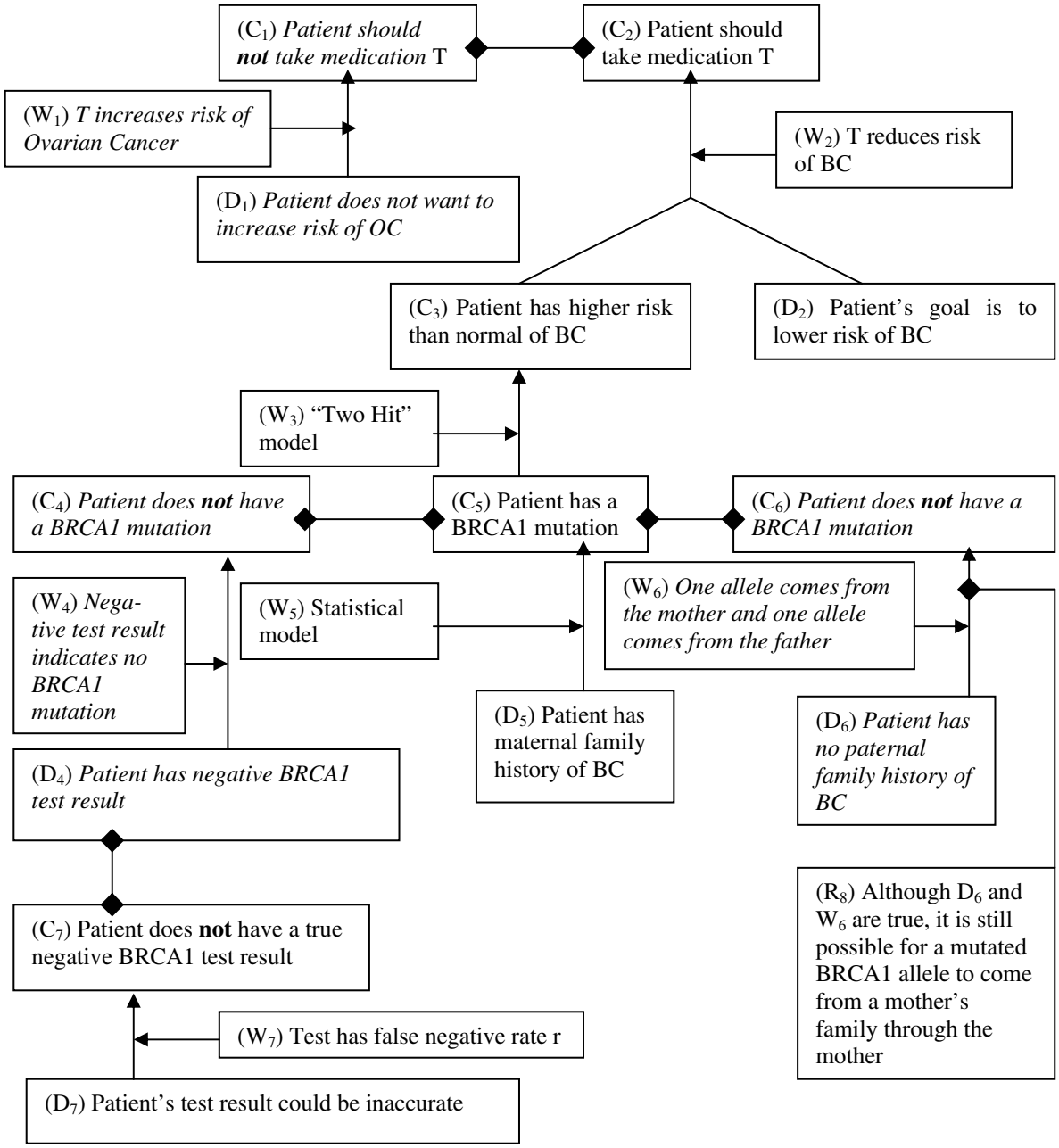


Fig. 2. Diagram of debate constructed in future ILESA-GenMed.

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V., Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 37-44)

The Evolution of Assessment: Learning about Culture from a Serious Game

Matthew J. Hays[1], Amy Ogan[2], H. Chad Lane[1]

[1] Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina del Rey CA 90292, USA

[2] Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA 15213, USA

hays@ict.usc.edu, aeo@andrew.cmu.edu, lane@ict.usc.edu

Abstract. In ill-defined domains, properly assessing learning is, itself, an ill-defined problem. Over the last several years, the domain of interest to us has been teaching Americans about Iraqi business culture via a serious-game-based practice environment. We describe this system and the various measures we used in a series of studies to assess its ability to teach. As subsequent studies identified the limits of each measure, we selected additional measures that would let us better understand what and how people were learning, using Bloom's revised taxonomy as a guide. We relate these and other lessons we learned in the process of refining our solution to this ill-defined problem.

Keywords: learning, technology, assessment, measurement, ill-defined domain, culture, serious game

1 Introduction

As societies and their economic and humanitarian transactions become more globalized, cross-cultural negotiation has emerged as an important ill-defined domain. Culture often dramatically affects people's expectations when they interact with others. These effects can be exacerbated because these expectations are often implicit. That is, the role culture plays becomes salient only when our expectations are violated—and people may not be able to identify cultural differences as the cause of interpersonal difficulty [2].

With several collaborators, we have developed a cultural training system called BiLAT. BiLAT is a serious-game-based learning environment that is designed to teach the preparation for and execution of meetings in a cross-cultural context [12, 13]. The immersive approach [14] and focus on practice [10] were motivated by cognitive psychology and the instructional design literature. Elsewhere, we detail the development and implementation of BiLAT [12, 13]. The present paper provides an overview of BiLAT and an accompanying intelligent tutoring system (ITS), but focuses on our assessments of learning from BiLAT, their evolution, and the lessons we learned along the way.

2 How can BiLAT improve intercultural competence?

Rulebooks, demonstration videos, and lectures are sufficient instructional tools in many learning contexts. When well designed, these mostly passive approaches are effective at conveying facts and examples. However, competence in ill-defined domains is dependent on *contextualized* understanding; learners must determine the circumstances under which particular solutions to problems are appropriate [7, 18]. Without direct experience or live role-play, this task is very nearly impossible [19]. Unfortunately, on-the-job training is rarely an option and live role-play is costly and difficult to scale up. Moreover, even if these were viable alternatives in terms of resources, it would be difficult to ensure consistent pedagogical content and provide appropriate learning scaffolds.

Fig. 1. Meetings with a police officer (left) and a businessman (right) in BiLAT

BiLAT simulates a business meeting in which cultural awareness, adherence to expectations, and relationship building are important. It can be used as a consistent, scalable, lower-cost alternative to role-playing. Figure 1 shows the BiLAT interface, in which learners research and engage in turn-based dialogue with virtual characters by selecting actions from menus.

Success in BiLAT depends on building trust with the virtual characters before discussing potential agreements, a basic principle of Arab business culture [20]. BiLAT therefore emphasizes the timing of actions and their context of use by modeling *meeting phases*, which determine when the actions a user can choose are appropriate. For example, one generally advisable social action is discussion of children; both cultures take pride in the achievements of their young. Talking about your—or your meeting partner's—children is a good idea near the beginning of a meeting, but not while negotiating the terms of an agreement.

Learners with little experience and no external guidance might become confused about when or whether an action is generally advisable. We therefore developed an ITS to help clarify these situations and more broadly support learners through their interactions with the virtual characters. The ITS takes the form of a virtual coach that assists the learner during the meeting. After each turn, the coach decides whether [12] and how [11] to provide feedback about past actions or hints about future actions.

Designing and developing BiLAT and the ITS were extensive, complicated processes. Determining whether the two systems function together as an effective training tool has been an equally intricate process. The next section of this paper details the ways in which we measured how BiLAT and the ITS improved learners' comprehension of and competence with Iraqi business-meeting culture.

3 How can we assess intercultural competence?

The same things that make intercultural interaction difficult to train are those that make its improvement difficult to measure [21]. Is a business meeting successful as long as a mutually beneficial outcome is reached? What if the negotiations came at the cost of the business relationship, making it the last agreement those two parties will ever reach? Perhaps one partner takes from the meeting a negative opinion about all members of the other's culture; is that still a successful meeting? Without hard and fast rules, determining the complete extent of a trainee's ability cannot be accomplished solely by checking multiple-choice responses against a key. Instead, multiple measures are needed to get a complete understanding of trainees' comprehension and competence. Dozens of quantitative studies investigating the effectiveness of non-technological cross-cultural training programs, many including several measures, have been undertaken with exactly this goal [4, 19]. Selecting a subset of these measures appropriate to evaluating learning from BiLAT required several iterations of empirical research. We also used Bloom's revised taxonomy of educational objectives as a framework for our decisions [1, 5]. This taxonomy is a widely accepted hierarchical classification that defines levels of learning, activities that promote learning at each level, and assessments of learning at each level. The rest of this section describes the measures we selected and how we used them to gauge BiLAT's effectiveness as an educational tool.

3.1 Measuring remembering and understanding: a situational judgment test

In Bloom's revised taxonomy, the two most basic levels of learning are remembering and understanding. *Remembering* is the ability to recall or recognize information in the format in which it was learned (i.e., without requiring transfer or application). Students can demonstrate remembering by providing definitions for key terms or labeling components of a system. *Understanding* can be thought of as remembering that has been freed from its original format. Students can demonstrate understanding by summarizing or generating additional examples of a category.

Situational judgment tests (SJTs) are appropriate for measuring remembering and understanding in ill-defined domains [17]. In a common SJT format, learners read several scenarios that describe various problems related to the training domain. Each scenario is accompanied by potential solutions to which learners provide Likert-scale ratings of advisability (i.e., 1 = "very unadvisable"; 10 = "very advisable") [6]. Responses are generated by several subject-matter experts (SMEs), who have substantial familiarity with the training domain. The consensus of the SMEs' answers

is the standard against which trainees' scores are compared [3]. The greater the correlation between a trainee and the SMEs, the greater the trainee's understanding.

Assessments of remembering and understanding must be tailored specifically to the content of instruction. Otherwise, the assessments begin to measure the ability o apply or transfer knowledge, which are at higher levels of Bloom's taxonomy. In the case of assessing learning from BiLAT, we needed measures that explicitly and exclusively addressed Iraqi business culture. A literature search revealed many measures of intercultural competence [23], but none specific to the topics we believed BiLAT taught. We therefore needed to develop one, and worked with several SMEs to create an SJT appropriate to measure learning from BiLAT and the ITS [9].

We used this SJT in several experiments. The participants' first task in each of these experiments was to complete the SJT. We then oriented the participants to the content of BiLAT by showing them a high-production-value video that depicted a live-action American-Iraqi meeting in which the American fails to adhere to the cultural norms of his host [8]. After the video, participants used BiLAT for several hours and then took the SJT again. Thus, we used the SJT in a pretest-posttest design; we defined learning as an increase in the correlation between participants' and SMEs' ratings from pretest to posttest. We found that BiLAT produced substantial overall gains in remembering and understanding [8, 9, 11, 13, 15].

According to Bloom's revised taxonomy, measures of remembering and understanding do not require the interactivity provided by BiLAT or the assistance provided by the ITS. Instead, passive approaches such as watching videos and listening to stories can affect measures of remembering and understanding. In a concurrent experiment, in which the video was shown *prior* to taking the SJT pretest, participants' scores on the pretest were as high as their scores on the posttest in our other experiments [22]. This result suggested that the video was affecting SJT scores. We tested this hypothesis in a subsequent experiment in which we administered the SJT, showed the video, and then again administered the SJT. Even without any use of the BiLAT system, there was a reliable improvement from pretest (M = .474, SE = .032) to posttest (M = .715, SE = .021): $F(1, 17) = 51.225$, $p < .001$, $\eta^2 = .751$. This result—and its magnitude—suggested that the SJT was highly influenced by the video. This result also meant that we could not determine the degree to which gameplay affected SJT scores in our prior studies. However, gameplay caused learning gains on other measures that should not be affected by the video [1]; these measures are discussed below.

3.2 Measuring the ability to apply knowledge: an in-game transfer task

The third level of Bloom's revised taxonomy is applying. *Applying* is the ability to solve a problem similar to those solved during training and modify what has been learned in order to transfer it to another situation. It can be thought of as extending understanding to novel applications.

We were able to use BiLAT itself to measure learners' ability to apply their knowledge. After participants used BiLAT for up to 100 minutes to solve a problem in an Iraqi marketplace, we disabled the ITS and asked participants to solve a new

problem with a different character. Our measure of learning in this transfer task was the probability that participants would select an inappropriate action during their meetings; lower probabilities indicated greater mastery. We chose this measure rather than a pretest-posttest design because interacting with the BiLAT system involves becoming familiar with the interface and how the system models various concepts like trust-building. To the extent that this familiarity affects the likelihood of making errors, pretest scores would have been artificially deflated and would have created the illusion of greater learning gains than were actually generated.

We had three goals in disabling the ITS in this transfer task. First, feedback from the coach could have decreased error probabilities over the course of the task. This effect would have inflated our estimates of learning and would have added noise to the data. Second, the ITS is *designed* to fade over time; it is intended to support practice in a way that helps the learner no longer need support, like training wheels on a child's bicycle. Third, there is no coach in the real world. Assessing learning under similar conditions thus added external validity to our measurements.

In one study, we used this transfer task to compare the pedagogical value of two different coaches. One coach provided action-level, easy-to-follow feedback (e.g., "don't give gifts that contain alcohol"). The other coach provided conceptual feedback, which required learners to more deeply contemplate their potential actions (e.g., "make sure that your gifts are culturally appropriate"). Otherwise, the coaches behaved identically. We found that, while the coaches were active, they were equally helpful; participants made as many errors in the market scenario with the conceptual coach as with the specific coach. However, in the transfer task (without the ITS), a different pattern emerged. Participants who had been assisted by the conceptual coach were reliably less likely to make errors. The deeper thought that the conceptual coach encouraged led the participants to be better able to transfer their understanding to a new character—but did not differentially affect their SJT scores [11]. In other words, both groups of participants had the same *amount* of remembered knowledge, but those who were helped by the conceptual coach were better able to *apply* that knowledge to a new situation.

Unfortunately, there are significant drawbacks to using an in-game measure. Primarily, it invites the criticism that we are "testing to the teach." From that perspective, the new scenario cannot be considered a true transfer task. Indeed, without other measures, one could make the argument that people who use BiLAT may not be learning anything more than how to use BiLAT. To that end, the next section describes yet another measure we used to evaluate the efficacy of BiLAT and the ITS as an instructional system.

3.3 Measuring the ability to analyze: a cultural assimilator

The fourth of the six levels in Bloom's revised taxonomy is analyzing. *Analyzing* is the ability to deconstruct and examine instructional materials. It results in the student understanding why some solution can be applied to a particular set of problems. This understanding allows the student to infer the *causes* of problems and what makes particular solutions appropriate.

By definition, analyzing extends beyond the learner's experiences and reaches throughout the training domain. Thus, measuring learners' analytical skills does not require assessments to be tailored precisely to the content of the training system (vs. remembering or understanding). As a result, we were able to use an existing measure rather than creating our own. The measure we chose was the cultural assimilator (CA) created by Cushner and Brislin [7]. The implementations of this and many other CAs appear similar to the SJT, in that each item consists of a scenario that occurs in a target culture. However, whereas the SJT asks learners to rate various solutions to the problem described in the scenario, the CA asks learners to select the best *explanation* of the problem. On each item, selecting a culturally sophisticated explanation yields a score of two points; explanations reflecting some insight are worth one point; and inappropriate selections are worth zero points. Selecting a two-point explanation requires deep understanding for two reasons. First, the situations in the CA are not those encountered in gameplay, and so the gameplay experience must be analyzed in order to extract the needed information. Second, some one-point explanations are "attractive lures," meaning that people with a less sophisticated understanding of cross-cultural interaction will be likely to select them instead.

According to Bloom, measures of analytical skill are affected by interactive learning tools but not passive instructional tools like stories and videos [5]. Thus, the CA should have been affected by gameplay but not by the orientation video. We have found evidence in recent studies that gameplay improves CA scores—and is especially helpful for learners who started out in the bottom half of scores on the pretest [16, 22]. On the other hand, we included the CA in the video-only study described above (immediately following the SJT, pre- and post-gameplay). Unlike with the SJT, we found that the video caused negligible change in CA scores from pretest (M = 9.375, SE = .460) to posttest (M = 9.333, SE = .462): $F < 1$ *ns*. In summary, BiLAT has been shown to improve scores on both the SJT and the CA, whereas the video only affects SJT scores. Together, these results suggest that BiLAT and the ITS combine to form an effective teaching system. These effects manifest at least at the level of analysis—and probably at yet higher levels in Bloom's revised taxonomy [5].

4 What did *we* learn through this process?

The development of BiLAT took years. It began with an intensive study of cross-cultural negotiation. This effort resulted in storyboards and board-game prototypes, which were developed and refined into a simulation prototype. This prototype was refined through systematic review by SMEs. We used a similar process to develop the training support provided by the ITS [12].

Likewise, the assessment of learning from BiLAT has undergone iterative development. Initially, we used only the SJT and found results consistent with our hypotheses; BiLAT appeared to be an effective pedagogical tool. As we conducted further experiments tied more strongly to learning theory, it became clear that the video could be affecting the SJT results. We directly tested this idea and found that at

least some of the improvements in SJT scores in our earlier studies were likely driven by the video. In subsequent studies, we introduced additional measures that evaluate deeper levels at which one might learn from gameplay and from the ITS. These measures highlighted the result that BiLAT does not simply provide training for *remembering* and *understanding,* but furthermore supports *applying* and *analyzing* in an intercultural domain. Above all, our experience emphasized the need to analyze learning in an ill-defined domain more completely and at a deeper level. As will often be the case in ill-defined domains, understanding student learning is almost certainly going to require employing multiple—and more refined—measures. In our experience, Bloom's revised taxonomy was an informative guide for this exploratory process.

As our work continues, we will further approach our problem from multiple perspectives. Mendenhall has created several dimensions of cultural assessment from a review of many different instruments [19]. Some of these dimensions include measures of learners' satisfaction with the instructional tool, which is absent from Bloom's hierarchy but is increasingly important as learners more frequently are required to manage their own learning. Even as we diversify and improve our assessments, we must also strive to be realistic about their limitations. Cross-cultural interaction will always be imperfectly measured by sets of questions or rubrics for behavioral change, regardless of their refinement. Researchers operating in such ill-defined domains must reconcile the need for better assessment with the reality of the difficulties inherent in such an endeavor, and continue to draw conclusions from their data with the appropriate amount of caution.

References

[1] Anderson, L.W., & Krathwohl, D.R. (eds.): A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Outcomes: Complete Edition. New York: Longman (2001)

[2] Bennett, M.J.: Towards Ethnorelativism: A Developmental Model of Intercultural Sensitivity. In Paige, R.M. (ed.) *Education for the Intercultural Experience*, pp. 21–71, Yarmouth, ME: Intercultural Press (1993)

[3] Bergman, M.E., Drasgow, F., Donovan, M.A., Henning, J.B., Juraska, S.E.: Scoring Situational Judgment Tests: Once You Get the Data, Your Troubles Begin. *International Journal of Selection and Assessment, 14*, pp. 223-235 (2006)

[4] Black, J.S., Mendenhall, M.: Cross-Cultural Training Effectiveness: A Review and a Theoretical Framework for Future Research. *The Academy of Management Review, 15,* pp. 113-136 (1990)

[5] Bloom B.S., Krathwohl, D.R.: *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: Longman (1956)

[6] Chan, D., Schmitt, N.: Situational Judgment and Job Performance. *Human Performance, 15*, pp. 233-254 (2002)

[7] Cushner, K., Brislin, R.W.: *Intercultural Interactions: A Practical Guide* (2nd ed.). Thousand Oaks, CA: Sage Publications (1996)

[8] Durlach, P.J.: Issues in Deployment of Serious Games. *Proceedings of the 31st Interservice/Industry Training, Simulation, and Education Conference* (2009)

[9] Durlach, P.J., Wansbury, T.G., Wilkinson, J.G.: Cultural Awareness and Negotiation Skills Training: Evaluation of a Prototype Semi-Immersive System. *Proceedings of the 26th Army Science Conference* (2008)

[10] Ericsson, K.A., Krampe, R.T., Tesch-Romer, C.: The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review, 100*, pp. 363-406 (1993)

[11] Hays, M.J., Lane, H.C., Auerbach, D., Core, M.G., Gomboc, D., Rosenberg, M.: Feedback specificity and the learning of intercultural communication skills. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pp. 391-398 (2009)

[12] Hill, R.W., Belanich, J., Lane, H.C., Core, M., Dixon, M., Forbell, E., Kim, J., Hart, J.: Pedagogically Structured Game-Based Training: Development of the ELECT BiLAT Simulation. *Poster presented at the 25th Army Science Conference* (2006)

[13] Kim, J.M., Hill, R.W., Durlach, P.J., Lane, H.C., Forbell, E., Core, M., Marsella, S., Pynadath, D. V., Hart, J.: BiLAT: A Game-Based Environment for Practicing Negotiation in a Cultural Context. *International Journal of Artificial Intelligence in Education* (in press)

[14] Lane, H.C.: Metacognition and the Development of Intercultural Competence. *Proceedings of the Workshop on Metacognition and Self-Regulated Learning in Intelligent Tutoring Systems at the 13th International Conference on Artificial Intelligence in Education*, pp. 23-32 (2007)

[15] Lane, H.C., Hays, M.J., Auerbach, D., Core, M., Gomboc, D., Forbell, E., & Rosenberg, M.: Coaching Intercultural Communication in a Serious Game. *Proceedings of the 16th International Conference on Computers in Education*, pp. 35-42 (2008)

[16] Lane, H.C., Hays, M.J., Auerbach, D., Rosenberg, M.: Investigating the relationship between presence and learning in a serious game (under review at ITS2010)

[17] Legree, P.J., Psotka, J.: Refining Situational Judgment Test Methods. In *Proceedings of the 25th Army Science Conference* (2006)

[18] Lynch, C.F., Ashley, K., Aleven, V., Pinkwart, N.: Defining "Ill-Defined" Domains: A Literature Survey. In Aleven, V., Ashley, K., Lynch, C., Pinkwart, N. (eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, pp. 1-10 (2006)

[19] Mendenhall, M.E., Stahl, G.K., Ehnert, I., Oddou, G., Osland, J.S., Kuhlmann, T.M.: Evaluation Studies of Cross-Cultural Training Programs: A Review of the Literature from 1988-2000. In Landis, D., Bennett, J.M., Bennett, M.J. (eds.) *Handbook of Intercultural Training* (3rd ed.), pp. 129-144, Thousand Oaks, CA: Sage Publications (2004)

[20] Nydall, M.K.: *Understanding Arabs: A Guide for Modern Times* (4th ed.). Boston: Intercultural Press (2006)

[21] Ogan, A., Aleven, V., Jones, C.: Culture in the Classroom: Challenges for Assessment in Ill-Defined Domains. In Aleven, V., Ashley, K., Lynch, C., Pinkwart, N. (eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, pp. 92-100 (2006)

[22] Ogan, A., Kim, J., Aleven, V., Jones, C.: Explicit Social Goals and Learning in a Game for Cross-Cultural Negotiation. In *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education* (2009)

[23] Paige, R.M.: Instrumentation in Intercultural Training. In Landis, D., Bennett, J.M., Bennett, M.J. (eds.) *Handbook of Intercultural Training* (3rd ed.), pp. 85-128, Thousand Oaks, CA: Sage Publications (2004)

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V., Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 45-52)

What is the Real Problem? Using Corpus Data to Tailor a Community Environment for Dissertation Writing

Lydia Lau, Royce Neagle, Sirisha Bajanki, Vania Dimitrova, and Roger Boyle
School of Computing, University of Leeds, UK

{ L.M.S.Lau, R.J.Neagle, S.Bajanki, V.G.Dimitrova, R.D.Boyle }@leeds.ac.uk

Abstract. Training in soft skills is becoming paramount in today's educational and societal climate, and receives increasing attention in the area of intelligent learning environments for ill-defined domains. We present a study that analyses written feedback given to undergraduate students by tutors at a key stage of dissertation preparation. This allows us to identify key problems students are facing and to understand how these problems are articulated and addressed by tutors. The results of the study are applied to tailor an existing social semantic web environment (AWESOME Dissertation) to address the needs of a particular community for dissertation writing in Computing.

Keywords: dissertation writing, semantic wikis, social computing, scaffolding.

1 Introduction

Dissertation writing, which is a major challenge faced by most students in higher education, is an example of soft skill training as the process requires the learners to explore, interpret, communicate, and manage their own work and progress during a sustained period of time. It fits into the 'IDIT quadrant' (for ill-defined domain and ill-defined task) of the classification scheme proposed by Mitrovic et al [10]. A fundamental step in developing such intelligent learning environments for ill-defined domains is to articulate *what problems learners are facing and how to shape the learning environment to effectively address these problems.*

Although intelligent technological solutions for writing development have been built, they focus mainly on discrete aspects of the dissertation process, for example argumentation or research methods [1,2]. An earlier attempt has been made in developing the AWESOME Dissertation Environment (ADE) which exploits social computing to provide holistic support throughout the dissertation process [3]. The intention was to make ADE a generic platform for any students writing a dissertation.

The issue of generic versus subject specific issues quickly emerged in the pilot testing of the ADE and, subsequently, instances for different disciplines were developed for further trials [4]. Experiences from these trials led us firstly to question the adequacy of high level (generic) views of the dissertation writing issues for supporting students who face individual and discipline-specific problems; and

secondly to drive for an environment which can be 'tailored' and 'evolved' with usage in the community where understanding and interpretation of domain-specific vocabulary and concepts can be shared.

This paper presents a domain-specific study on the use of ADE for Computing. On one hand, we analysed how some dissertation writing problems were handled in the current practice of tutors giving written feedback to final year students at a key stage of dissertation preparation. In parallel, we tailored an initial AWESOME-Computing instance by integrating examples of previous dissertations; and seeding the environment with content that corresponds to some typical problems and tutor feedback. We developed example scenarios of students and tutors interacting with the AWESOME Computing environment to simulate the process of social scaffolding which enables further 'tailoring' as the ADE evolves with use.

2 The AWESOME project

The platform for the study is a novel community environment 'AWESOME Dissertation Environment (ADE)' which uses semantic wikis to implement the pedagogical approach of 'social scaffolding'. It uses MediaWiki[1] and its extension Semantic Media Wiki [9] which provides a user-friendly interface to create and query semantic content. The ADE was developed within a UK research project called AWESOME (Academic Writing Empowered by Social Online Mediated Environments) which involved the universities of Leeds, Coventry and Bangor[2]. The environment was instantiated and trialed in several domains: Education, Fashion and Design, Philosophy and Religious Studies, and an Academic Writing Centre.

The ADE architecture consists of a core ontology which supports semantic mark-ups for

- a scaffold in the form of main stages of dissertation writing process; for example: getting an overview of dissertation, choosing a topic, adopting an appropriate research methodology, literature `review`, writing up, and project management;
- some common issues associated to each of these stages; for example: for choosing a topic, students need to consider whether the topic "has a research question" or "is appropriate for the discipline";
- personal contributions in terms of related top tips, or examples of good writings.

As part of the tailoring process, a separate emerging ontology is built during content creation for additional community driven scaffolds. Features are also provided for users to link content between community and personal spaces. Readers are referred to [4] for a more detailed description of the ADE architecture.

Following both the encouraging feedback from the trial instantiations and the challenges faced in deploying the environment in practice in earlier studies, we conducted a systematic approach in understanding the disparity between generic and

[1] http://www.mediawiki.org

[2] See the AWESOME web site http://awesome.leeds.ac.uk/ for more information.

domain specific vocabulary when adapting the ADE to dissertation writing in Computing.

3 Dissertation Writing Problems Faced by Computing Students

The final year project (or dissertation) is a hallmark of most Computing and Informatics programmes worldwide [5]. Common to many other programmes is the difficulty experienced by Computing students in recording their work: the write-up represents a challenge they have often not encountered earlier in their studies and usually represents the primary (or sole) artifact that is used for assessment.

In the auhtors' School of Computing, it is established practice for every dissertation student to prepare a mid-project report under the guidance of the tutor. It is typically 10 pages long, containing background research, progress to date and initial bibliography. This report will be commented on by another academic member of staff (i.e. 'assessor') to provide early written feedback to the student, with no marks given. This collection of assessor feedback forms is potentially a good resource for us to identify common problems and the feedback given to the students as a comparison with the scaffold/core ontology to be provided by the tailored ADE.

A systematic analysis was conducted on 63 authentic feedback forms from the academic year 2008/09 with the aim of identifying common early problem indicators flagged up by the assessors to the students, the suggestions they made and the language used.

3.1 Procedure

1) Content analysis on twenty mid-project report feedback forms was conducted independently by four staff members (as coders), three of whom had considerable experience in assessing and supervising dissertations. Initially, each coder chose his/her own way to annotate the categories of the problem and the associated issues, with the broad understanding of the need to relate the annotation to the final dissertation marking scheme as the student/tutor may use it to judge the impact of the feedback on the work. The marking scheme considered the criteria: 'Understanding the problem', 'Produce a solution', 'Evaluate the solution', 'Write up' and 'Reflection'[3] .
2) For each issue extracted from the assessors' feedback, the following were recorded: a) the problem as cited, b) the solution as cited, c) annotation to capture the general category that represented the problem and solution (e.g. write-up), and d) annotation to capture the issue category associated to the general category (e.g. scientific style, referencing).
3) A joint review of the annotations used for the first twenty feedback forms was conducted by the four coders with the aim of arriving at a taxonomy of annotated issues.

[3] http://www.comp.leeds.ac.uk/tsinfo/projects/assessment-criteria.shtml

4) The updated list of annotations (or taxonomy) was then used by the coders to analyse the remaining feedback forms as in step (2) above.

3.2 Results and Analysis

A total of 250 issues were identified from the 63 feedback forms provided by 25 tutors. Some feedback was positive encouragement reinforcing what had been done correctly; but most were constructive feedback for further improvements. Following is an example of feedback which was classified as "evaluation" for generic category and "criteria" for issue category:

- issue cited: "Your evaluation criteria need work";
- problem cited associated to this issue: "what were put in the report were subjective .. and unconvincing";
- solution cited associated to the issue: "focus groups won't help unless the users have a real task they are trying to achieve".

Table 1 summarises the taxonomy emerged from the analysis and the frequency of each being raised as an issue. It is clear that 'write-up' was most problematic as it was highlighted 86 times and with a wide range of issues being commented on. 'Methodology' came second by being mentioned 40 times.

For a specific issue category, such as 'scientific style', a range of comments can be found. For examples: "*no evidence of three prototypes claimed to be produced ... no pointer to the corpus generated*", "i*t is important to clearly identify what you have created/developed yourself, and what you have 'inherited' from others"*, and "*there should be more x-referencing*". Quite often, these comments represent a range of 'take-for-granted' common knowledge by academics experienced in scientific writing, but the concept of which has not been grasped properly by the students concerned.

Another example of common feedback for which students are often at a loss for remedial action is: "*there is a lack of critical analysis on the literature read*". Tutors found themselves needing to stress this issue repeatedly, despite formal timetabled classes which were run to discuss 'literature review' annually.

3.3 Implications

Our study revealed that the problem for Computing students in recording their work persists: the write-up represents a challenge they have often not encountered earlier in their studies and the dissertation usually represents the primary (or sole) artifact that is used for assessment for this piece of independent study. Previous studies suggested that common dissertation problems are due to students' unfamiliarity with the dissertation as a genre and inability to effectively engage with the processes associated with dissertation writing, and delay between information delivery and the time when students actually face the complexities of dissertation processes [6,7].

Table 1. Taxonomy for Dissertation Writing Issues for Computing Students

Generic category	Sub-total	Issue category	Frequency
evaluation	29	criteria	14
		depth	10
		missing	5
literature review	35	criticality	11
		depth	22
		web dependence	2
methodology	40	aims	6
		justification	10
		methods	16
		missing	3
		requirements	5
project definition	12	complexity unclear	7
		requirements specification	5
project management	24	milestones	6
		schedule unclear or delayed	18
topic selection	24	novelty	3
		problem clarity	19
		suitability	2
write-up	86	acronyms	4
		code	1
		explanation unclear	7
		formatting (diagrams/maths)	10
		presentation	1
		referencing	13
		scientific style	32
		specific content	4
		structure	2
		Use of english	12

Perhaps, the solution lies not only in the prevention of problems but also in the provision of support when issues arise. A deeper understanding of how feedback is being acted on by the students and tutors is needed. The current practice in Computing is for a tutor to discuss the feedback face to face with his/her student. Experience shows that most students need assistance from their tutors to interpret the feedback which often followed by further assimilation in order to understand how to address the issues. This process could take between a week and some months, with the exact path of inquiry taken by an individual student unpredictable. A combination of learning-by-example, and social and individual learning processes is expected for an individual student to proceed with the rest of the dissertation journey. Pedagogically, it would be beneficial to provide some kind of structure or ‘scaffold’

[8] for channeling and focusing these activities and a 'public platform' for the sharing of experience.

4 AWESOME-Computing

We suggest that "learning by example" is of great value, but more than that, open discussion by peers and staff of favourable and unfavourable examples *as the issues arise* would be of most help. Thus, a semantic wiki framework in which earlier dissertations exhibiting fragments of annotated good practice, which allowed the student to add her own annotations or questions may be productive.

4.1 Tuning and Seeding

The first step to tailor the generic ADE for Computing was by adding links to previous dissertations. Although these dissertations were available online on a website, there was no facility to comment on specific good practice or examples. AWESOME-computing enabled this by allowing students or tutors to add comments to a dissertation in a wiki fashion (see Fig. 1). Additional semantic markups enabled the example to be pulled into other appropriate wiki pages on specific issues.

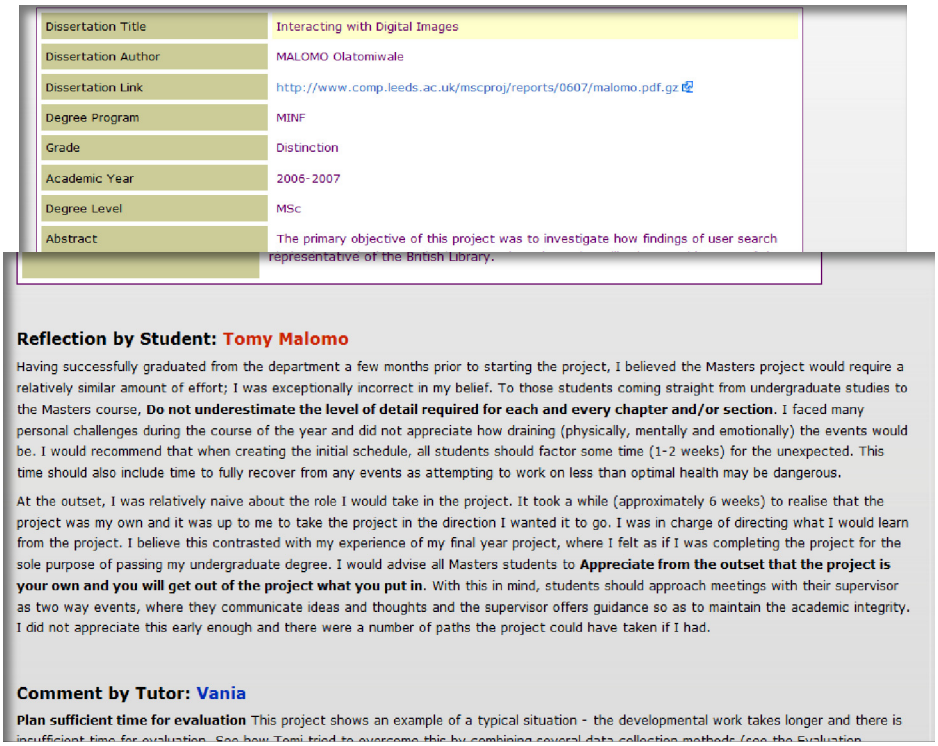


Fig. 1. Comments and tips linked to a previous dissertation for writeup

Secondly, we scaffolded and seeded the environment with content for some anticipated problems and tutor feedback (see Fig. 2 for an example).

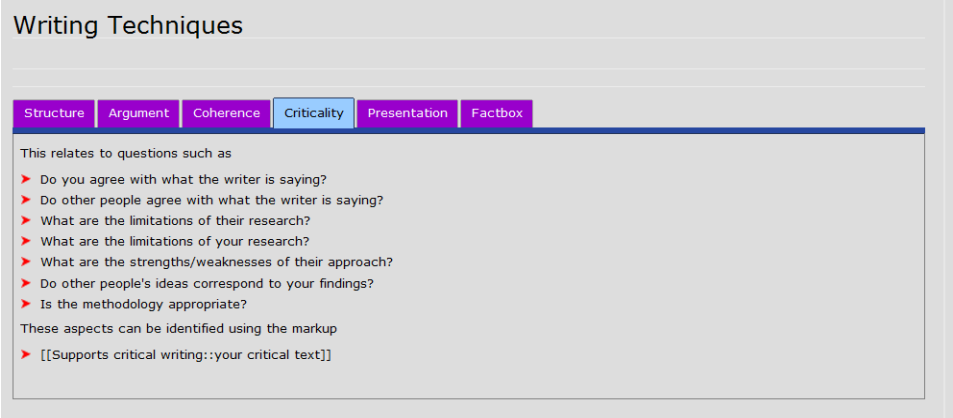


Fig. 2. Having problems relating to 'critical writing'?

4.2 Scenarios for Further Seeding of Content

To provide some useful initial content, a number of scenarios were developed based on real experiences by some tutors. These scenarios were then 'walked through' to populate the ADE with authentic content. Fig. 3 showed an example of the end result of this seeding process, which appears on the home page.

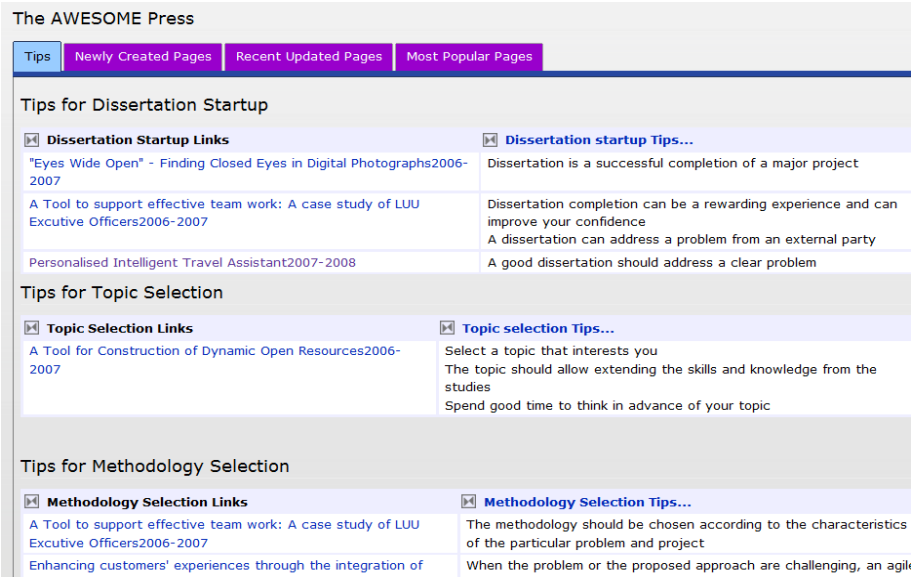


Fig. 3. AWESOME Press on the ADE home page

5 Conclusion

This paper presented an empirical study on a cohort of Computing students in their dissertation writing by analysing all feedback forms on their mid-project reports. The top three problems were: (i) not writing in the expected scientific style, (ii) lack of depth in literature review and (iii) lacking problem clarity. Although classes were held every year to prepare students in tackling these issues, experience showed that many students still struggled to fully understand their relevance when taught.

AWESOME-computing, based on semantic wiki technology, was proposed as a solution to provide complementary support for students to get further assistance in a social context. We believe in the pedagogical approaches of 'scaffolding' and 'learning by example'. We learned from previous trials that a well-seeded environment is vital for the success of the system launch. Hence, specific scenarios were design to seed a scaffolded environment with real examples and feedback.

References

1. McLaren, B.M., Scheuer, O., De Laat, M., Hever, R., De Groot, R., Rose, C.P. : Using Machine Learning Techniques to Analyse and Support Mediation of Student E-discussions. In:Proceedings of AIED07, pp.331--338, IOS Press (2007)
2. O'Rourke, S.T., Calvo, R.A.: Analysing Semantic Flow in Academic Writing. In: Frontiers in Artificial Intelligence and Applications; Vol. 200; Proceeding of the 2009 conference on Artificial Intelligence in Education, pp.173--180, IOS Press (2009)
3. Dimitrova, V., Lau, L., Le Bek, A.: Sharing of Community Practice through Semantics: A Case Study in Academic Writing. In: Sixth Int'l Workshop on Ontologies and Semantic Web for E-Learning, in conjunction with Int. Conf. on Intelligent Tutoring Systems - ITS'08, pp.30--39 (2008)
4. Bajanki, S., Kaufhold, K., Le Bek, A., Dimitrova, V., Lau, L., O'Rourke, R., Walker, A.: Use of Semantics to Build an Academic Writing Community Environment. In: Proceedings of AIED09, pp.357—364, IOS Press (2009)
5. Fincher, S., Petrie, M., Clark, M.: Computer Science Project Work: Principles and Pragmatics, Springer (2001)
6. Ganobcsik-Williams, L. (ed.): Teaching Academic Writing in UK Higher Education, Palgrave (2006)
7. Lea, M., Stierer, B. (eds.) : Student Writing in Higher Education: New Contexts, OU Press (2000)
8. Pea, R.D.: The Social and Technological Dimensions of Scaffolding and Related Theoretical Concepts for Learning, Education, and Human Activity. In: The Journal of the Learning Sciences, 13(3), pp.423--451 (2004)
9. Krötzsch, M., Vrandečić, D., Völkel, M., Haller, H., Studer, R. : Semantic Wikipedia. Journal of Web Semantics, 5, pp. 251-261 (2007)
10. Mitrovic, A., and Weerasinghe, A. : Revisiting ill-definedness and the consequences for ITSs. In: Proceedings of AIED09, pp. 375–382, (2009)

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V.,
Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 53-60)

Layered Learner Modelling in ill-defined domains: conceptual model and architecture in MiGen

Manolis Mavrikis, Sergio Gutierrez-Santos, Darren Pearce-Lazard,
Alexandra Poulovassilis, and George Magoulas*

London Knowledge Lab, 23-29 Emerald Str, London, UK.

Abstract. The design of learner modelling components for Exploratory Learning Environments (ELEs) presents a significant challenge, particularly when pertaining to ill-defined tasks and knowledge domain. We argue that representing a learner's knowledge just in relation to concepts is not adequate in such cases. We focus particularly on microworlds and present the conceptual model and architecture of the learner model of MiGen system that aims to support 11–14-year-old students develop the complex cognitive skill of mathematical generalisation.

1 Introduction

The work presented in this paper focuses on modelling learners as they are undertaking ill-defined tasks within exploratory learning environments (ELEs), and microworlds (MWs) in particular. A recent review of ill-defined domains [1] refers to such environments as 'model building' systems and identifies them as belonging to a particular genre of discovery learning whereby learners are provided with model-building tools and are encouraged to 'test their own intuitions about a domain' [1]. Although most ELEs provide non-adaptive feedback designed to scaffold students' learning, as with other constructivist approaches learning can be hampered in the absence of explicit support (c.f. [2]).

Our overarching objective is to enable the provision of adaptive feedback to students and information to teachers that will assist them in their efforts to integrate MWs into the classroom. However, the nature of the interaction in MWs, and their underlying pedagogical orientation, introduce difficulties in modelling the epistemological development of the learner, thus placing an additional hurdle for the provision of support in any form, e.g. explicit feedback, assistance for self-regulation through open learner modelling, assistance for teachers, etc.

The onus thus falls on the learner modelling component of an intelligent system, which is required to provide a substrate that describes, stores and manages short-term and long-term information about learners. However, following the usual approach in the field, whereby a learner's knowledge is represented only in relation to domain concepts, is not straightforward in this context. Similar to other intelligent learning environments that support learning in ill-defined

* Supported by ESRC/TLRP Grant RES-139-25-0381

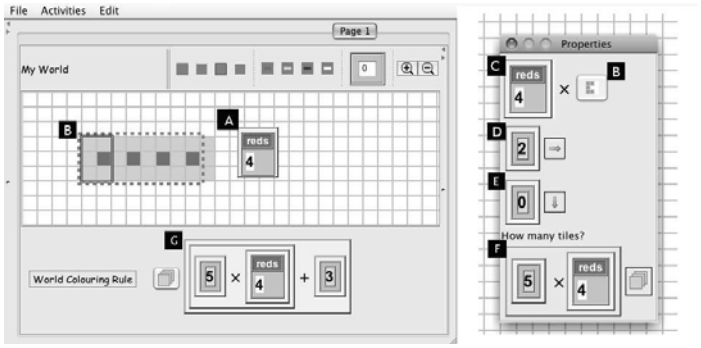


Fig. 1: Constructing a pattern in the eXpresser and describing it with a rule. Letters highlight the main features: (A) An 'unlocked' number that acts like a variable is given the name 'reds' and signifies the number of red (dark grey) tiles in the pattern. (B) Building block to be repeated to make a pattern. (C) Number of repetitions (in this case, the value of the variable 'reds'). (D,E) Number of grid squares to translate B to the right and down after each repetition. (F) Units of colour required to paint the pattern. (G) General expression that gives the total number of units of colour required to paint the whole pattern.

domains, MWs are designed to provide an empirical basis and opportunities for learners to develop complex cognitive skills rather than just to learn declarative knowledge of particular concepts. Moreover, even in cases where the knowledge domain underlying a MW is well-defined, the tasks that students are usually asked to undertake are open-ended in nature, have multiple approaches to a valid solution, and encourage students to explore the environment and follow a variety of strategies, not all of which can be sequenced or pre-defined (c.f. [3] for a detailed discussion on the need to distinguish between domain and task when discussing ill-definedness).

The MiGen project is creating a system to help 11–14 year olds to develop an appreciation of mathematical generalisation, which is considered one of the major routes to algebra in the UK curriculum. The system comprises a MW and several intelligent components that analyse the actions of students and provide support on different tasks. Fig. 1(A) shows the MiGen MW, called eXpresser. The MW encourages students to construct patterns and to find general expressions (i.e. rules) underpinning such patterns. Students use building blocks that they construct from unit tiles in order to make their patterns. To represent the generalities that they perceive, they can use numbers which they can 'unlock' to become variables. Locked and unlocked numbers can be used in expressions. The eXpresser gives a lot of freedom to students, who may construct their patterns in a multitude of different ways. For a detailed description of the eXpresser, see [4].

It is important to emphasise that interaction with a MW, such as eXpresser, does not necessarily provide direct evidence for assessing students' understanding of concepts. Behind the surface activities that students undertake, lies a different

objective. The learning challenge addressed in eXpresser is not the creation of the patterns nor the building of the algebraic expression (both of which are well-defined), but rather the development of generalisation skills (e.g. finding common structures, identification of general unknowns) by means of those tasks, and the development of mathematical 'ways of thinking' (WOTs) [5], including abstracting a general rule from a set of specific examples, finding a common structure from several samples of a series or set, using variables to represent universal unknowns, and developing heuristics for verification of hypotheses and falsification of false conjectures, among others.

In this respect, it is necessary to make a distinction between the overall subject domain for which the MW is designed, and what is referred to as the 'epistemological domain of validity' of the microworld [6] i.e. the knowledge domain as it has been transformed by the affordances and interface of the environment. For example, the notion of a variable in the case of eXpresser is linked with the view of a variable as a 'generalised number' in the eXpresser's affordances and is operationalised as an 'unlocked number' (see Fig 1(A)). If the objective behind the modelling process were just to develop a model of the 'user', then this distinction between the subject domain and the MW domain would not be so important. However, our goal is to model learners and their learning progress, outside of the boundaries of the MW. Moreover, teachers need a correspondence between learners' interactions with the MW and the subject domain, as they are often required to identify and work towards specific learning objectives. Taking both requirements into account, it is imperative to represent explicitly the relationships between the subject domain and the MW domain.

In the next section, we present the conceptual model and the architecture of the learner model we use in MiGen, that satisfy these requirements. Section 3 presents our conclusions and future work.

2 Learner Modelling in MiGen

2.1 Conceptual Model

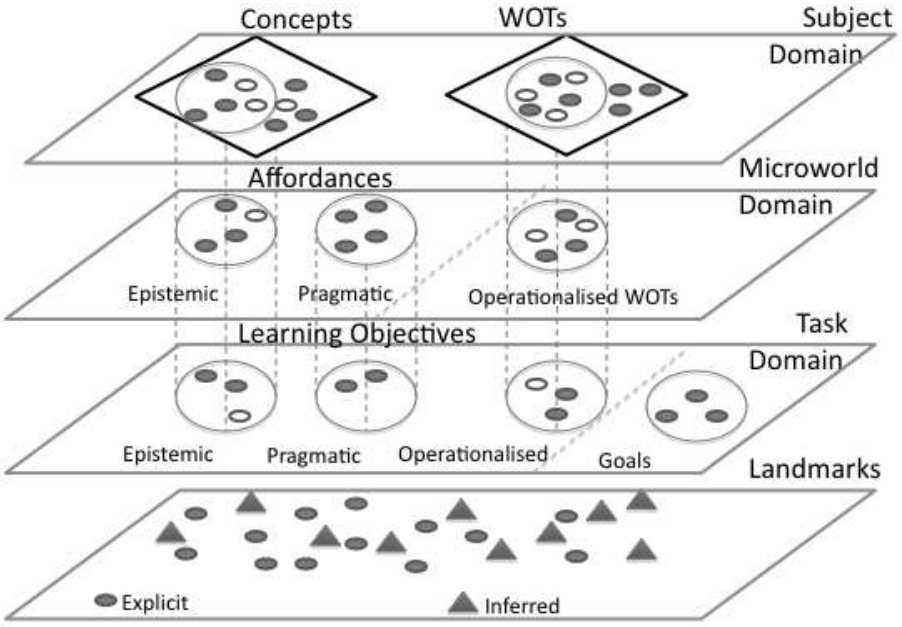
We address the requirements mentioned in the Introduction first by advocating that apart from the usual approach of modelling valid and invalid concepts in particular contexts, the learner model should also include epistemic and 'unproductive' ways of thinking. This approach is in line with [7] which argues for extending the scope of learner models with aspects outside the subject domain [1].

Second, in order to take into account the transformation of the domain, we consider a 'layer' of knowledge that involves microworld-specific concepts and that operationalises both the concepts of the subject domain and the ways of thinking (WOTs). Subject domain concepts are operationalised in the MW through the actions available using the objects and tools of the MW, i.e. its affordances. Because of their direct relationship to knowledge, we refer to these as

[1] We recognise that the learner model should include affective factors and learner beliefs about the domain. Currently, such information is encapsulated under 'ways of thinking', as a fully-fledged affect modelling is beyond the scope of this research.

'epistemic' affordances. In order to distinguish those actions in the MW that are independent of any epistemic basis (e.g. 'knowledge of creating a building block' see Fig. 1(B)), we refer to these as 'pragmatic' affordances. In addition, this layer includes what we refer to as 'operationalised' ways of thinking. The two top layers of Fig. 2 present schematically the relationship between subject domain and the MW domain. This conceptualisation has the additional advantage of enabling us to take into account that learners' previous knowledge about the subject domain (or their intuitions) play an important role and can influence the way they perceive and interact with the MW. By representing both the subject domain and its operationalisation through the MW, it is possible to use such information (if available) to guide the adaptation process.

Finally, we recognise the critical role of specific tasks that are designed to contextualise students', otherwise unbounded, interaction with the MW. This provides specific goals that the learner is required to achieve during a task.



.

Fig. 2: Conceptual model for learner modelling in microworlds. Concepts in the top layer (including 'invalid' conceptions for a particular context — blank ovals) are operationalised to epistemic affordances in the microworld layer. The same applies to productive and unproductive ways of thinking (WOTs). The microworld layer also includes pragmatic affordances corresponding to actions independent of any epistemic basis. It is projected to the task layer comprising concrete goals and learning objectives. Lastly, landmarks indicate the completion of goals and attainment of learning objectives.

Such goals include tangible objectives such as 'find a general expression to colour a [certain] pattern' — see Fig 1(G). Learning objectives are assigned to each task, e.g. an introductory task could have the simple objective to 'explore how a [certain] tool behaves' while a more complex task could have the objective to 'appreciate the power of unlocked numbers'. As the task domain layer in Fig. 2 shows, there are three types of learning objectives: pragmatic learning objectives correspond to pragmatic affordances of the MW and are independent of the subject domain e.g. 'knows how to drag numbers on the canvas'; epistemic affordances (e.g. 'unlocking a number') are mapped to epistemic learning objectives; and WOTs to operationalised WOTs (e.g. 'validates the generality of construction by animation'). In MiGen, learning objectives, tasks and goals are co-designed by the research team with teachers (see also [8]); though we intend that in the future it will be possible for teachers (or trained learning designers) to define their own.

Ensuring that tasks have tangible goals and are associated with learning objectives, enables a much more tractable diagnostic aim: measuring beliefs in relation to learning objectives and not abstractly in relation to a student's state of mind. This is achieved through the automatic inference of *landmarks* as students interact with the MW (bottom layer of Fig. 2). In particular, Explicit landmarks occur when specific actions are undertaken by the student, e.g. 'clicking the animate button to validate their construction', while Inferred landmarks are derived from occurrences of combinations of actions, e.g. 'the student has started to construct generally', or 'the student is exploring in a systematic way'.

2.2 Learner Modelling Architecture

We now formalise the main entities of the learner model architecture and the relationships between them. An earlier paper [9] described the conceptual and architectural design of the overall MiGen system. Here we extend that work by focusing on details of the learner modelling aspects.

Figure 3 shows the major entities comprising the MiGen Learner Model, as well as some associated entities. For each eXpresser task undertaken by a Student, information on their ongoing progress through the task is maintained within a TaskShortTermModel. This information is derived from the occurrence of Landmarks as the student undertakes the task. The TaskShortTermModel is used to derive a longer-term model of the student's strategies and outcomes in relation to a task — the TaskLongTermModel. This in turn is used to derive a model of the students' attainment in relation to learning objectives pertaining to the whole MW — the MicroworldLongTermModel. Finally, this is used to derive a model of attainment of learning objectives related to the domain of mathematical generalisation — the DomainLongTermModel. Thus, overall, a student's learner model consists of their TaskShortTermModels, TaskLongTermModels, MicroworldLongTermModel and DomainLongTermModel. For implementing these derivation processes, we are employing a hybrid of rule-based and case-based reasoning (see [10]) in order to infer the occurrence of Inferred landmarks and to update the learner's TaskShortTermModel. Once a student

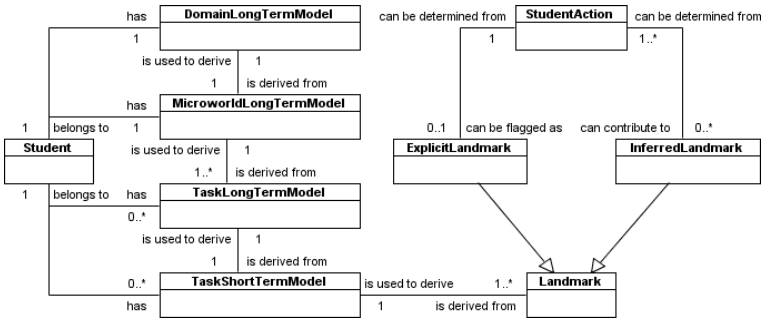


Fig. 3: MiGen learner model. At each end of an edge linking two entities is an indication of the cardinality of that end of the relationship and a verb phrase, e.g. a StudentAction can contribute to zero or more InferredLandmarks. A single-headed arrow indicates a sub-class relationship between two entities.

completes a task, the higher layers of the learner model are updated by additional rule-based components which successively infer updates to each higher layer based on its current values and the values of the layer below it.

The student's DomainLongTermModel is consistent with the subject domain model of MiGen, which includes concepts such as 'constants', 'variables', 'constructions' and 'expressions' and the corresponding learning objectives mapped from the U.K. National Maths Curriculum. e.g. 'visualise and draw on grids of different types where a shape will be after a translation', 'understand and use the rules of arithmetic in the context of positive integers', 'explore number relationships and propose a general statement involving numbers'. Similarly, the MicroworldLongTermModel is consistent with the second layer of Figure 2, and the TaskShortTermModels and TaskLongTermModels with the third layer. In practice, a teacher may initialise some attributes in a student's DomainLongTermModel, or may make explicit modifications to them over time.

Figure 4 shows the major types of Learning Objectives in MiGen and the relationships between them, as well as some associated entities. In brief, DomainLearningObjectives are separated into epistemic objectives (shown as ConceptualLearningObjective in the diagram) and objectives related to mathematical ways of thinking, e.g. 'appreciation of the use of variables'. Each TaskLearningObjective may be associated with a number of TaskShortTermModels and TaskLongTermModels. Likewise, each DomainLearningObjective may be associated with a number of students' DomainLongTermModels. Landmarks provide evidence for TaskLearningObjectives, as well as for LearnerInconsistencies — these are context-specific stumbling blocks, e.g. 'using more variables than needed'. Each TaskLearningObjective corresponds to a MicroworldLearningObjective; though there may be additional instances of the latter that have no counterpart TaskLearningObjective. There is a looser many-to-many correspondence between sets of MicroworldLearningObjectives and DomainLearningObjectives.

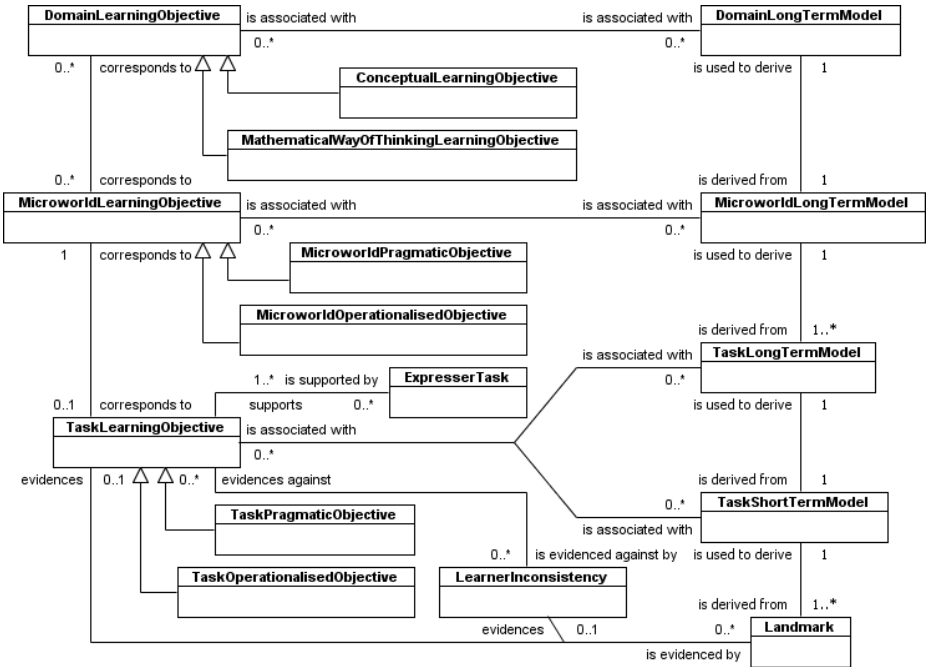


Fig. 4: Learning Objectives

3 Conclusions

This paper has argued that learner modelling in MWs introduces the need to extend the standard approach of representing the knowledge domain as concepts, with additional information representing learners' epistemic ways of thinking. The conceptual model we have presented here takes into account the transformative nature that MWs (and other ELEs) have on the nature of knowledge and on the domain they represent. Our layered approach simplifies the learner modelling problem by contextualising the domain first to its operationalisation in the MW, and subsequently to tasks with goals and particular learning objectives.

Several MWs have been reported in the educational technology literature. However, in most cases integration into the clasroom has been hindered because of the extensive requirement on teachers for helping students both with pragmatic and epistemic aspects. The ill-defined nature of the tasks that students undertake in MWs invites learner modelling methods such as the ones used in ill-defined tasks and domains (see [1, 3]). Independently of the precise techniques used to infer information and update the learner model, representing and maintaining the required knowledge is an important prerequisite, particularly when we need to expose such knowledge to stakeholders (e.g. teachers or students).

Our approach relies on explicit definition of the goals and the learning objectives underlying students' interactions, as well as their mapping to the MW and Domain layer. As a proof of concept, in MiGen, these mappings are captured as rules elicited from domain experts and teachers experienced with the MW. However, the conceptualisation is independent of the updating and inference techniques employed (for a review of related techniques see [11]).

This paper has also presented the learner model architecture of the MiGen system. For more details on the process by which the learner model is updated as students undertake tasks in the eXpresser see [11], where we also provide a preliminary investigation of how this approach can be applied to other ELEs with ill-defined tasks and domains. For the future, as more learners interact with the eXpresser, and teachers are exposed to our conceptual model in operation, we plan to investigate different inference techniques and design tools that enable the iterative refinement of the entities represented in the learner model.

References

1. Lynch, C.F., Ashley, K.D., Aleven, V., Pinkwart, N.: Defining ill-defined domains;a literature survey. In: Intelligent Tutoring Systems (ITS 2006): Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. (2006)
2. Mayer, R.E.: Should there be a three-strikes rule against pure discovery learning? - the case for guided methods of instruction. American Psychologist (2004) 14–19
3. Mitrovic, A., Weerasinghe, A.: Revisiting ill-definedness and the consequences for ITSs. In: Proceeding of the 2009 conference on Artificial Intelligence in Education, Amsterdam, The Netherlands, The Netherlands, IOS Press (2009) 375–382
4. Noss, R., Hoyles, C., Mavrikis, M., Geraniou, E., Gutierrez-Santos, S., Pearce, D.: Broadening the sense of 'dynamic': a microworld to support students' mathematical generalisation. Int. Journal on Mathematics Education **41**(4) (2009) 493–503
5. Papert, S.: Teaching children to be mathematicians vs. teaching about mathematics. Artificial Intelligence Memo Number 249 (July 1971)
6. Balacheff, N., Sutherland, R.: Epistemological domain of validity of microworlds: the case of logo and cabri-géomètre. In: Proceedings of the IFIP TC3/WG3.3 Working Conference on Lessons from Learning, Amsterdam, The Netherlands, North-Holland Publishing Co. (1994) 137–150
7. Bull, S., Brna, P., Pain, H.: Extending the scope of the student model. User Modeling and User-Adapted Interaction **5**(1) (March 1995) 45–65
8. Mavrikis, M., Gutierrez-Santos, S.: Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. Computers & Education **54**(3) (2010) 641–651
9. Pearce, D., Poulovassilis, A.: The conceptual and architectural design of a system supporting exploratory learning of mathematics generalisation. In Cress, U., Dimitrova, V., Specht, M., eds.: Learning in the Synergy of Multiple Disciplines. Volume 5794. Springer Berlin Heidelberg, Berlin, Heidelberg (2009) 22–36
10. Gutierrez-Santos, S., Cocea, M., Magoulas, G.: A case-based reasoning approach to provide adaptive feedback in microworlds. In: Intelligent Tutoring Systems, ITS'2010. (to appear)
11. Mavrikis, M., Gutierrez-Santos, S., Pearce-Lazard, D., Poulovassilis, A., Magoulas, G.: Learner modelling in microworlds: conceptual model and architecture in MiGen. Technical Report BBKCS-10-04, Birkbeck College, University of London (2010) Available at `http:/www.dcs.bbk.ac.uk//research/techreps/2010/`.

Proc. IllDef2010 Vol. 4: Lynch, C., Ashley, K. D., Mitrovic, T., Dimitrova, V., Pinkwart, N., & Aleven, V. (Eds). Pittsburgh, Pennsylvania, June. (pp 61-68)

Using a Quantitative Model of Participation in a Community of Practice to Direct Automated Mentoring in an Ill-Defined Domain

David Williamson Shaffer[1] and Arthur Graesser[2]

[1] University of Wisconsin—Madison, Department of Educational Psychology
1025 West Johnson Street, Madison, WI 53711, USA. dws@education.wisc.edu

[2] Institute for Intelligent Systems, 365 Innovation Drive, University of Memphis, Memphis, TN 38152, a-graesser@memphis.edu

Abstract. We describe a system for producing automated professional mentoring using a quantitative model of enculturation that is applied to ill-defined problems and domains. The system under development is an automated mentoring technology, called AutoMentor, that builds on previous research on on AutoTutor, a computer tutor that helps students learn about science and technology topics by holding a conversation in natural language with the learner. We do this by exploring a specific hypothesis about mentoring in ill-defined domains: using sociocultural model as the basis of an automated tutoring system can provide a computational model of participation in a community of practice and will produce effective professional feedback from non-player-characters in a learning game.

Keywords: Epistemic frames, epistemic games, epistemic network analysis (ENA), automated mentoring, AutoTutor, AutoMentor, tutoring systems, learning games, educational games, mentoring.

1 Background

In this paper we address a critical research question about intelligent tutoring systems: Can a quantitative model of participation in a community of practice automate professional mentoring in a learning game involving ill-defined problems? This project on AutoMentor attempts to many of challenges of ill-defined problems and domains. Understanding the meaning of natural language in student contributions has its own set of issues regarding uncertainty, imprecision, and vagueness. However, the domain knowledge is also open-ended and minimally constrained because there is no perfect well-defined solution to problems in the game space. The dialogue moves of AutoMentor neverthelesss provide feedback and metacognitive guidance.

The automated mentoring technology is being developed and tested within the context of a specific discipline (the study of ecology and the development of systems thinking more broadly) and within the context of a specific computer game for middle

school students. In this paper we describe the general principles and techniques that provide professional mentoring within the context of learning games in ill-structured domains. Unfortunately, space constraints make it impossible to provide a concrete example of the game environment.

1.1 Learning Theory: The Epistemic Frame Hypothesis

Learning to solve ill-defined problems typically comes from being part of a community of practice [1]: a group of people who share similar ways of solving problems. Learning does not end with the mastery of pertinent skills and knowledge; it must also include developing a sense of what kinds of judgments are compatible with the values and practices of a field. Within a ill-defined domain, there are particular ways of justifying decisions and developing solutions [2].

The epistemic frame hypothesis suggests that any community of practice has a culture [3] and that culture has a grammar, namely a structure composed of:

1. Skills: the things that people within the community do;
2. Knowledge: the understandings that people in the community share;
3. Values: the beliefs that members of the community hold;
4. Identity: the way that members of the community see themselves; and
5. Epistemology: the warrants that justify actions or claims as legitimate within the community.

This collection of skills, knowledge, values, identity, and epistemology forms the epistemic frame of the community [4].

This theory has been developed and tested in the context of epistemic games: games where players engage in simulations of training in professions such as engineering and urban planning to develop the epistemic frame of thinking [5, 6]. These games are developed by studying professional practica and by creating a game storyboard that describes the key activities of the practicum: the actions that professionals-in-training take in the practicum and the occasions for reflection between professionals-in-training and their mentors [4, 7].

The storyboard for a game is then expanded into a frameboard that describes, for each activity: (a) the activity of the players in the game (such as chat, decisions of resource allocation, or negotiations); (b) the activity of the mentors, including key dialogue moves or talking points; (c) the expected work product, output, or action of the players (such as critiquing a solution or writing a report); (d) the criteria for evaluating the work product or output; (e) the expected elements of the epistemic frame of the profession; and (f) the sources of evidence that will be used to determine whether the elements of the epistemic frame are used in the particular activity.

The specific elements and actions of the game are built from this frameboard. This includes simulations and other professional tools, non-player character (NPC) responses, requests for information and feedback from the mentor, and instructions for game mentors who interact with players through instant messaging (IM) and e-mail. The game engine that controls an epistemic game automatically records player interactions with game mentors via IM and e-mail, which can be later analyzed for both the forms of mentoring and content they contain.

1.2 Ill-defined Domain: Ecological Thinking

The game we are using for this project, *Urban Science*, is a computer-based game in which late elementary, middle, and high school students learn ecological thinking by role-playing as members of an urban planning firm dealing with land use issues in ecologically sensitive areas. In the game, players interact with NPCs in the form of stakeholders in the community and other planners in the firm. These computer-generated characters represent different interest groups with competing agendas, as well as supervisors and other members of the firm who provide professional resources, information about ecological issues, and advice about the planning process.

Urban planning is a domain of practice traditionally taught at the postsecondary level, but it is a potentially fruitful context for the development of science understanding in middle school students. Work in urban planning addresses elements of the National Science Education Standards [8] that call for understanding systems, order, and organization; evolution and equilibrium; and form and function in natural systems. The Geographical information system (GIS) tools that planners use make these complex processesmore accessible to middle school students. [9, 10]

We have constructed a set of land use models that integrate geographic features and sets of secondary attributes into interactive visual models of complex systems. In our previous studies of Urban Science, we developed five such models for Madison, Wisconsin. Land use models use a GIS system to generate feedback from virtual stakeholders in the community in response to players' land use decisions. The feedback is not based on an ideal solution but it does integrate a land use impact model that quantifies the impact of land use decisions on key environmental, economic, and social indicators, such as pollution, tax revenue, and acreage of wildlife habitat. The models represent ecological and ecosocial relationships in a computational form that allows players in Urban Science to explore, propose, and defend solutions to complex ecological and economic issues.

As an example, one land use model explores development on the north side of Madison, Wisconsin, adjacent to a large wetland area known as Cherokee Marsh. The project raises a number of significant economic and ecological issues related to wetland ecology and conservation. While working in the Cherokee Marsh land use model, players of Urban Science have to investigate, analyze, understand, and communicate about a number of scientific issues, including local species, their life cycle, and their habitat; the role of wetlands in the local ecological system; and specific pollutants, their sources, and their impacts.

A key component of the game is that players interact with professional mentors, namely undergraduate and graduate students playing the role of more senior urban planners in the firm. These mentors help players in the game take action as urban planners to deal with ecological issues in the land use problems they are solving. But more important, they help players reflect on their actions in the game, a form of natural language that is ill-defined and sometimes has no established answer or solution. Previous research on Urban Science has shown that (a) the game is effective in developing ecological understanding for students, and (b) the time players spend reflecting with mentors is a key part of that process [5].

Automating professional STEM feedback in the game would be an important component for scaling up such interventions. Therefore our current project is exploring the use of AutoMentor as a virtual agent serving as a professional mentor.

1.3 Existing Intelligent Automated Tutoring Research

Our project builds on previous research on intelligent tutoring systems. Intelligent tutoring systems track the knowledge states of learners in fine detail (called user modeling) and adaptively respond with activities that incorporate computational models in artificial intelligence and cognitive science, such as production systems, case-based reasoning, Bayes networks, theorem proving, and constraint satisfaction algorithms. We base our project on previous research on AutoTutor, a computer tutor that helps students learn about science and technology topics by holding a conversation in natural language with the learner. Previous research on AutoTutor has focused on the learning of STEM topics, such as physics, computer literacy, biology, and scientific methods, and has shown that AutoTutor improves learning by nearly one letter grade compared with reading a textbook an equivalent amount of time, or compared with a pretest of students' abilities in a STEM domain [11, 12].

AutoTutor helps students compose answers to deep-reasoning questions and solutions to problems by expressing a variety of dialogue moves, such as: feedback (positive, neutral, negative), pumps for more information ("Tell me more"), hints, prompts to fill in missing words, summaries, corrections of student misconceptions, answers to student questions, and requests for students to perform actions in interactive simulation environments.

The system architecture of AutoTutor has five key components:

1. A state table that maintains and updates the states of the student-system dialogue and the tasks in the learning environment.
2. A student model that records the student's knowledge, progress on covering expected material, emotional states, and other learner characteristics.
3. A curriculum script of pedagogical tasks (problems to solve or difficult questions to answer), the expected correct answers for each task, alternative tutor dialogue moves, and other content that is task-specific.
4. A set of computational linguistic modules that include lexicons, syntactic parsers, speech act classifiers, shallow semantic analyzers, and latent semantic analysis (LSA) spaces [13, 14] for analyzing the meaning of what the student expresses verbally.
5. A dialogue planner that formulates the dialogue moves of the next conversational turn of AutoTutor in a fashion that is sensitive to the state table, the student model, and interpretation of student input via the computational linguistics modules.

The conversations managed by AutoTutor are imperfect, but smooth enough for students to work with minimal difficulties. Dialogue is sufficiently tuned so that a bystander who observes tutorial dialogue in print cannot tell whether a particular turn was generated by AutoTutor or by an expert human tutor [15].

1.4 Existing Research on Models of Participation in Communities of Practice

Our project uses Epistemic Network Analysis (ENA), a methodology to assess students' ability to think and act like professionals through epistemic game play [16]. ENA is a measurement and modeling technique that is designed to be intellectually responsive to current theories of learning that emphasize the contextual and connected aspects of complex problem solving.

ENA is based on two key concepts: (a) that thinking in an ill-defined domain can be characterized by the application of an epistemic frame composed of the linkages between skills, knowledge, identity, values, and epistemology and (b) that the development of thinking in an ill-defined can be quantified, analyzed, and visualized with a dynamic network model of the developing epistemic frame. In this sense, ENA provides a computational model of a player's (or a mentor's) participation in the culture of a profession—the extent to which a player has adopted the ways of knowing, being, talking, and acting that characterize a particular community of practice.

ENA adapts the tools of social network analysis to a different domain. Social network analysis provides a robust set of analytical tools for representing networks of relationships, including complex and dynamic relationships of the kind that characterize epistemic frames [17, 18]. If we take the epistemic frame hypothesis as a basis of a student model for assessment, then we can quantify the developing epistemic frame for each participant at time T by summing, for each pair of frame elements (i,j) the number of times they are coded as occurring at each point of time in the game t<T. That is, we can construct a cumulative adjacency matrix ${}_{T}A^{p}_{i,j}$ which shows the strength of association between each pair of frame elements (i,j) for a given player in the data set. Once an epistemic frame is represented as a series of cumulative adjacency matrices, we can quantify the hypothesized epistemic frame using concepts from social network analysis, such as density and centrality [16]. We can use ENA to track the state of the epistemic frame of players in an epistemic game, and also characterize events, actions, and interactions in an epistemic game in terms of their effect on the players' developing epistemic frames. That is, we can track how specific features and events in a learning environment (or combinations of events, the current state of a learner's network, and interactions with peers and mentors) lead to significant changes in understanding of the ill-defined domain. ENA is thus a model of the extent to which an individual has the ways of thinking, talking, and acting that are characteristic of a particular community of practice.

2 Hypothesis and Method

We bring together these lines of research in the following hypothesis: If we use a sociocultural rather than traditional cognitive model as the basis of an automated tutoring system, we can build a computational model of participation in a community of practice in addition to task analysis models to produce professional feedback from automated mentors in a learning game.

Our program of research is to link (a) epistemic games, which prepare players to participate in STEM communities of practice through professional mentoring, with (b) AutoTutor, which provides automated feedback on learners' actions. We will do so by creating expert feedback based on (c) ENA, which is a model of an individual's ability to participate in a community of practice.

We accomplish this using a Wizard of Oz methodology [19], in which we collect data about player/mentor interactions over multiple instances of game play. We use this database of player questions and actions, as well as the accompanying expert mentor responses, to develop and validate a system for automatically coding interactions for elements of the epistemic frame of urban planning: the skills, knowledge, identity, values, and epistemology (abbreviated as SKIVE) of the planning profession. We then use the coded database to generate automated responses to player actions in the game, and test whether players' learning with automated mentoring are comparable to outcomes with live mentors. As a result, we will determine if, how, for whom, and when automated professional mentoring (based on a computational model of participation in a community of practice) affects learning in an ill-defined domain.

In the future we will produce and test a version of the game Urban Science that includes the AutoMentor module for producing automated mentor feedback through e-mail and IM during game play. That is, AutoMentor will be a conversational agent, but not a "talking head" or animated agent. This version will be a 10-hour game designed to be played in schools or as part of out-of-school enrichment programming for middle school students. It will be a Web-based game in which players become interns at the office of a fictitious urban and regional planning firm that develops land use plans for local and national sites.

We will also produce a set of algorithms and code to implement AutoMentor within the context of epistemic games, and learning games more generally. One important companion module is AutoFramer, a set of algorithms and code for automated coding of epistemic frame elements in game actions and interactions that will be necessary to implement the AutoMentor system. AutoFramer may also be used separately to assess players' participation in a community of practice during game play.

While this is fundamentally a project to produce dialogic material, another contribution of the project is to produce appropriate quantification of the textual material. More specifically there are three forms of quantified data derived from qualitative verbal reponses: (1) results from pretests and posttests that involve verbal content, (2) human-produced textual material from players and human mentors , and (3) textual material produced by AutoMentor. As such, our fundamental data analytical techniques are (a) the statistical methods used to quantify textual material in general, and (b) comparative statistical analyses of the relationships between these three data types. There will also be exploratory investigations of the applicability of innovative measurement models to extend, enhance, or tune the foundational ENA data.

3 Preliminary Results and Conclusions

We do not have complete results to report on any phase of the project. However, at this point of the project we have developed the following:

1. We have developed a version of Urban Science suitable for data collection and are currently pilot testing it.
2. We have isolated a subset of interactive episodes where students and mentors are working in a focused manner. Limiting the range of game episodes for initial analysis (e.g., focusing on some checkpoints or common experiences that are important moments in game play) is allowing us to build more focused models of interactions, which will later be validated in the larger data set.
3. We have hand coded a subset of the observations on SKIVE elements and are refining the codes already in use for Urban Science by characterizing key aspects of situations and mentor interaction that are salient for statistical/measurement modeling. This includes codes for the kind and level of support that mentors provide and specific forms of urban planning SKIVE elements.
4. We have used latent semantic analysis (LSA) and clustering techniques to extract a set of prototype exemplars of each coding category of mentor contributions. We are currently implementing a snowballing methodology in which we identify a pool of initial prototype responses, and then use LSA to identify close matches to new discourse responses. We are adding additional prototypes and refining the matching parameters against hand coded schemes.
5. We are validating the initial AutoFramer algorithms on a second, larger subset of mentor/player interactions. We are currently able to achieve a Cohen's kappa of .70 or higher (for the presence of a category, or the selection of a category) between human and automated coding for some categories.

We conceive of this project as a form of design research, where initial hypotheses about game design, assessment, and automated mentoring technologies are revised by subsequent experiments in each area. An advantage of design research in this context is the relatively continuous stream of implementation activity that we have planned. Our primary objective of this work is to use these game contexts as an occasion to develop and validate an automated assessment and mentoring system based on a model of participation in a community of practice.

Acknowledgements

This work was funded in part by the Macarthur Foundation and the National Science Foundation through grants REC-0347000, DUE-091934, DRL-0918409, and DRL-0946372. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

1. Lave, J. and E. Wenger, Situated learning: Legitimate peripheral participation. 1991, Cambridge, MA: Cambridge University Press.
2. Shaffer, D.W., Epistemic Frames for Epistemic Games. Computers and Education, 2006. 46(3): p. 223.
3. Rohde, M. and D.W. Shaffer, Us, ourselves, and we: Thoughts about social (self-) categorization. Association for Computing Machinery (ACM) SigGROUP Bulletin, 2004. 24(3): p. 19-24.
4. Shaffer, D.W., Pedagogical praxis: The professions as models for post-industrial education. Teachers College Record, 2004. 106(7): p. 1401-1421.
5. Bagley, E.A.S. and D.W. Shaffer, When people get in the way: Promoting civic thinking through epistemic gameplay. International Journal of Gaming and Computer-Mediated Simulations, 2009. 1(1): p. 36-52.
6. Svarovsky, G.N. and D.W. Shaffer, SodaConstructing knowledge through exploratoids. Journal of Research in Science Teaching, 2007. 44(1): p. 133-153.
7. Schon, D.A., The reflective practitioner: How professionals think in action. 1983, New York: Basic Books. x, 374.
8. National Research Council, National Science Education Standards. 1995, Washington, DC: National Academy Press.
9. Baxter, R. and H. Broda, Using GIS and GPS technology as an instructional tool. The Clearing House, 2002. 76(1): p. 49-52.
10. Chawla, L., Insight, creativity and thoughts on the environment: Integrating children and youth into human settlement development. Environment & Urbanization, 2002. 14(2): p. 11-22.
11. Graesser, A.C., M. Jeon, and D. Duffy, Agent technologies designed to facilitate interactive knowledge construction. Discourse Processes, 2008. 45: p. 298-322.
12. VanLehn, K., et al., When are tutorial dialogues more effective than reading? . Cognitive Science, 2007. 31: p. 3-62.
13. Graesser, A.C., et al., eds. AutoTutor: A cognitive system that simulates a tutor that facilitiates learning through mixed-initiaive dialogue. Cognitive systems: Human cognitive models in systems design, ed. C. Forsythe, M.L. Bernard, and T.E. Goldsmith. 2005, Erlbaum: Mahweh, NJ.
14. Landauer, T., et al., eds. Handbook of Latent Semantic Analysis. 2007, Erlbaum: Mahwah, NJ.
15. Person, N.K. and A.C. Graesser, Human or computer?: AutoTutor in a bystander Turing test, in Intelligent Tutoring Systems 2002, S.A. Cerri, G. Gouarderes, and F. Paraguacu, Editors. 2002, Springer: Berlin.
16. Shaffer, D.W., et al., Epistemic Network Analysis: A prototype for 21st Century assessment of learning. The International Journal of Learning and Media, in submission.
17. Brandes, U. and T. Erlebach, Network Analysis: Methodological Foundations. 2005, Berlin, Heidelberg: Springer-Verlag. 16-61.
18. Wasserman, S. and K. Faust, Social Networks Analysis: Methods and Applications. 1994, Cambridge: Cambridge University Press.
19. Kelley, J.F., An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. ACM Transactions of Office Information Systems, 1994. 2(1): p. 26-41.