Computational Argument as a Diagnostic Tool: The role of reliability.

Collin Lynch
Intelligent Systems Program
Univ. of Pittsburgh
Pittsburgh, Pennsylvania,
USA

Kevin D. Ashley
ISP & School of Law
Univ. of Pittsburgh
Pittsburgh, Pennsylvania,
USA

Niels Pinkwart
Computer Science Institute
Clausthal Univ. of Technology
Clausthal, Lower Saxony,
Germany

Vincent Aleven
HCII
Carnegie Mellon Univ.
Pittsburgh, Pennsylvania,
USA

Abstract

Formal and computational models of argument are ideally suited for education in ill-defined domains such as law, public policy, and science. Open-ended arguments play a central role in these areas but students of the domains may not have been taught an explicit model of argument. Computational models of argument may be ideally suited to act as argument tutors guiding students in the formation of arguments and argument analysis according to an explicit model. In order to achieve this it is important to establish that the models can be understood and evaluated reliably, an empirical question. In this paper we report ongoing work on the diagnostic utility of argument diagrams produced in the LARGO tutoring system.

Introduction

Argumentation plays a central role in many domains such as the law, natural sciences and philosophy. Practitioners of these disciplines communicate through the use of open-ended arguments and novices approach the domain through the analysis of these expert arguments. Instruction in argumentation is therefore a primary educational goal. Argumentation instruction is typically Socratic in nature. Students examine both good and poor examples of argument, and engage in the the production of arguments in classroom situations. While this process depends upon a degree of scaffolding, guidance is typically implicit. Students are often given carefully designed examples to study but may not guided in their markup. Nor is the *process* of argument reified or made explicit to the students during their interaction. Rather the students propose arguments about a topic in the domain and encounter and respond to counterarguments in return. It is left to them to reify the process.

Diagrammatic and computational models of argument such as those described in (Ashley 2006; Gordon, Prakken, and Walton 2007; Gordon 2007; Reed and Rowe 2004) are ideally suited to fill this gap. These models are designed to provide a robust framework for argumentation that: a) reifies the essential structural and functional components of an argument thus enabling a computational reasoner to work effectively; and b) are comprehensible and natural for human arguers permitting robust communication. By reifying the essential argument structures a computational model can scaffold novice arguers focusing their attention on *relevant* interrelationships and argument moves. Similarly, retaining a robust and natural argument structure helps to ensure that the user-generated diagrams can be diagnostic, that is, used to diagnose student errors or misconceptions and to provide relevant feedback.

If this goal is to be realized it is important to establish that the existing models can be understood and evaluated reliably, an empirical question. While many of the existing models (e.g., Toulmin diagrams) are subject to strong formalisms, no set of formal rules can easily account for all possible variations. Nor can one expect a sufficiently complex ruleset to be followed exactly by all practitioners. This is especially acute when dealing with arguments created by students or other non-experts where accurate assessment is crucial to providing robust diagnoses and feedback.

In this paper we report on a series of studies that we have conducted with LARGO, an Intelligent Tutoring System (ITS) for legal argumentation with cases and hypotheticals (Pinkwart et al. 2007). Students use the LARGO system to annotate or reconstruct oral arguments made before the U.S. Supreme Court using a graphical argument model. In brief the diagrams reify the process of arguing with cases and hypotheticals as a series of moves and countermoves in which advocates propose legal tests or rules that are then challenged by means of hypothetical cases. This form of reasoning is particularly common in higher level courts, particularly those in common-law domains, whose decisions set precedents for deciding future cases (Ashley et al. 2008), but it is also observed in civil law legal reasoning (MacCormick and Summers 1997).

Students using the system are presented with an oral argument transcript. They then summarize the essential tests, hypotheticals, and factual assertions in the argument using the graphical markup language. This summary process results in a diagrammatic representation of the argument which may be examined for diagnoses or used as a summary representation of the underlying debate. As students use the system, they are guided by the system. This feedback chiefly takes the form of self-explanation prompts where students are guided to reconsider their work and to explicate why a given choice was made. Additional guidance is provided in the form of collaborative feedback where students compare

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

their work to others'.

In (Pinkwart et al. 2008) we presented empirical evidence that features of argument diagrams made with LARGO are correlated with two independent measures related to argumentation ability: standardized test scores that assess ability to evaluate reasoning and arguments and students' number of years in law school. In this paper, we report on a grader study in which two legal instructors jointly defined a set of evaluation criteria for LARGO graphs, and then independently scored the students' diagrams in a double-blind manner. Our analysis here will focus on the inter-grader and intra-grader reliability of the scoring. Reliability in interpreting and evaluating argument diagrams, whether of the LARGO variety or others, is a necessary precondition for the use of argument models as a basis for communication and instruction.

Related Work

In related work, researchers have developed instructional programs to teach problem-solving, argumentation, and reasoning skills through the use of diagrammatic argument models. Carr developed an instructional program in which law students created Toulmin-style argument diagrams as a medium for constructing and communicating their own arguments (Carr 2003). Similar systems have been employed in teaching philosophical reasoning (Easterday, Aleven, and Scheines 2007; Harrell 2007) and critical thinking (van Gelder 2007). A survey of four systems can be found in (Van den Braak et al. 2006).

One system, Belvedere (Suthers and Hundhausen 2001; Paolucci, Suthers, and Weiner 1996) was deployed in natural science and, like LARGO, offered the students advice based upon an analysis of their diagrams. Less systematic work has been done on the assessment of problem-solving or argumentation skills using the argument diagrams as evidence of students' understanding. In (Twardy 2004) the author describes some common student errors but focuses on error description, not grading.

In the most relevant related work (Lund et al. 2007), the researchers manually compared Toulmin-style argument diagrams dealing with public health issues (e.g., the desirability of introducing genetically-modified organisms into the human food chain). Students in the study either conducted a debate by constructing novel argument diagrams using the Toulmin model, or they used the model to reconstruct the argument process of a prior debate. Of special interest is the ADAM method of analysis the authors employed to assess the quality of the students' diagrams in terms of: 1) a diagram's form (i.e., branches extending linearly from a claim versus sub-branches); 2) the number of arguments and relations, number of opinions expressed both "pro" and "con"; 3) the breadth of topics broached; 4) the variety and degree of elaboration of the arguments (i.e., single word versus use of propositions); and 5) the correctness of argumentative relations (i.e., if a link correctly expresses a phrase supporting or attacking a claim or does something else such as using an incorrect direction or incorrect sign, a non-argumentative relation, or an unspecified relation.)

The current research differs from the above work in a variety of respects. First and foremost we are focusing on expert human graders employing agreed-upon criteria rather than a specific scoring algorithm as Lund did. This is similar to the route taken in (McClure, Sonak, and Suen 1999). However their focus was on the assessment of concept maps rather than functional arguments or process models. Secondly, most of the above work employs Toulmin-style datum-claim diagrams; LARGO's diagrams correspond to a process model of arguing with hypotheticals. Finally LARGO students do not use the system to construct novel arguments as in (Carr 2003) but to reconstruct experts' arguments as done by (van Gelder 2007)

LARGO Diagrams

Figure 1 illustrates a LARGO student diagram drawn from our present study. The diagram in question represents a snippet of the oral arguments in the case of Asahi Metal Industry Co. v. Superior Court of California, 480 U.S. 102 (1987). This is a classic personal jurisdiction case commonly employed in first year legal process courses. The issue turns on whether Asahi Metal Industry Co., a Japanese company, may be called into the Superior Court of California to answer a civil suit based upon a faulty tire of which they manufactured one component. All of the students involved in our study had previously examined the decision in *Asahi* but none had read the oral argument transcript.

As noted above, the focus of LARGO diagrams is on the process of argument. In making arguments before the U.S. Supreme Court an advocate will routinely pose a *test* or legal rule that, if adopted, achieves his desired outcome. The justices, or under some circumstances the opposing counsel, will respond by posing *hypothetical cases* or what-if scenarios that put pressure on some aspect of the test. The advocate may then respond by analogizing or distinguishing the posed hypothetical from the facts of the case at hand and modifying his test as necessary. For more details of the process model see (Ashley et al. 2008).

In the LARGO diagrams this process is illustrated through the use of a small palette of node types (*test*, *hypothetical*, and *fact*) as well as a larger palette of directional relationships including *modified-to* and *distinguished-from*. The test and hypothetical nodes represent utterances of the designated type in the underlying dialogue. Each may be linked to the portion of the dialogue where the specific item is proposed. Students may then enter a structured summary of the item in question into the box itself. Arcs represent a relationship between two elements. Some relations are *explicit* such as when an advocate or justice specifically enumerates the facts of a case in order to draw an analogy with the present hypothetical. Others are more implicit such as when an advocate reframes his test in order to include or exclude a prior hypothetical. U.S. Supreme Court oral arguments are conducted by expert arguers who share a great deal of background knowledge under strict time pressure. As a consequence the arguments contain a great deal of implicit information as their focus is more often on the process of stating new information rather than reifying

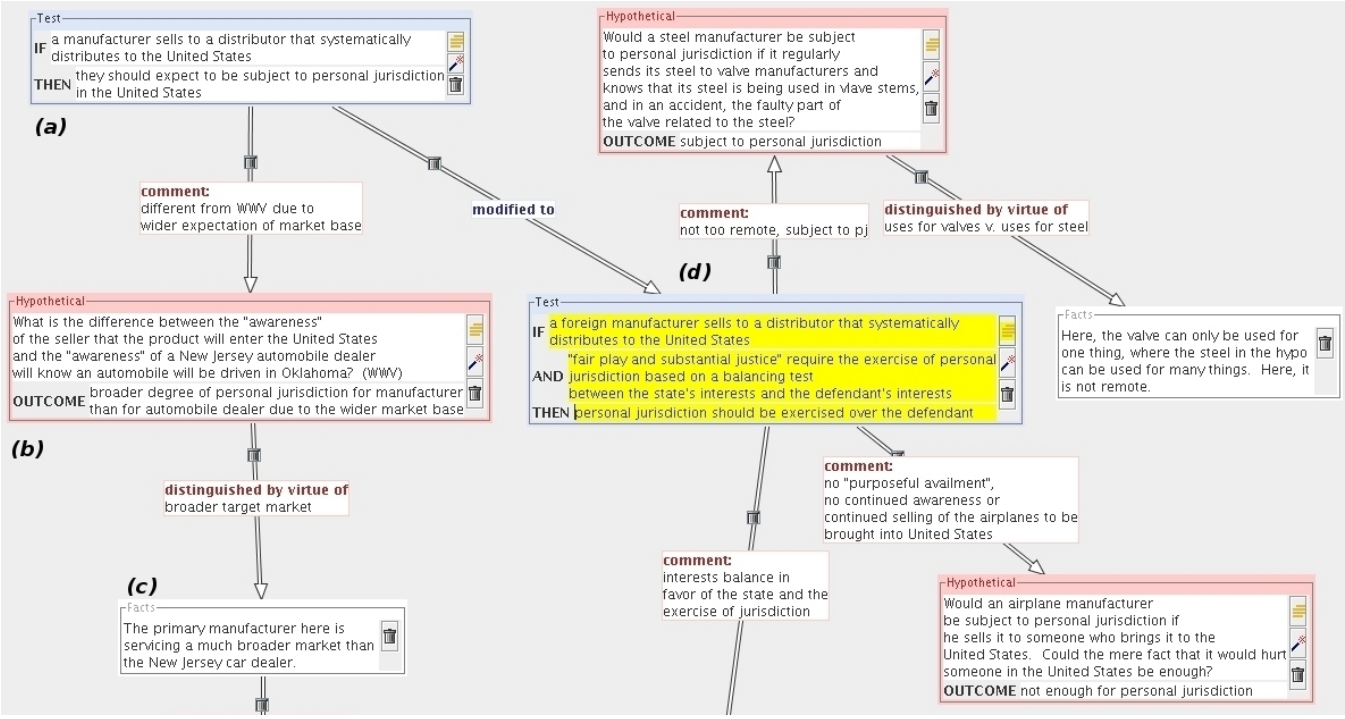

Figure 1: Sample Student Graph.

shared knowledge.

In Figure 1 for example the advocate begins by posing a test (a) shown in the upper left hand corner. In the summary box they have entered the following text:

> *"IF a manufacturer sells to a distributor that systematically distributes to the United States THEN they should expect to be subject to personal jurisdiction in the United States."*

This follows the legal format of relating legally relevant facts to legal consequences relevant to the case at hand. A hypothetical case (b) is put forth in response to the test with the relationship between them being noted by a general "comment" arc. This hypothetical is then, in turn, distinguished from the facts of the case at hand as shown by the fact box (c) and the "distinguished-by" arc. Test (a) is then modified to test (d) and the argument continues. In this way the graph reifies essential components of the arguments as well as the relationships between them. The resulting graphs are typically acyclic but rarely follow a linear pattern as novel lines of inquiry are introduced in an argument only to be abandoned, taken up again, and then synthesized with other strains of discussion in new hybrid tests.

While the graph in Figure 1 was produced by an expert subject it diverges in some ways from our argument model. For example the relationship between test (a) and hypothetical (b) should have been noted with a *leads-to* arc rather than the more general *comment* arc that was employed. Yet despite this and other divergences the student is following both the letter and spirit of the model quite closely.

In this present work we are focused on the problem of inter-grader agreement – are grades that different legal experts manually assign to diagrams consistent? Assessment of inter-grader agreement is of prime importance in the testing and educational assessment literature. In order for an argument model to be diagnostic of student behavior it must be the case that faculty members can reliably assess both the strengths and weaknesses of student arguments.

Expert Grading Experiment

We have conducted a series of studies at the University of Pittsburgh's School of Law engaging both first and third-year students in the use of LARGO. In the United States, law is a graduate degree taking three years to complete. First year law students are typically recent graduates of a four-year baccalaureate program while third-year students are due to graduate and receive legal accreditation. In each study a set of students was tasked with annotating a series of oral arguments using LARGO. Overall learning was measured by pre- and post-test scores. Analyses of these gains as well as a comparison between LARGO and text-based tools and discussion of automatic diagnoses may be found in (Pinkwart et al. 2007; Lynch et al. 2008).

In this paper we have drawn LARGO diagrams from three study populations: a 2006 set of paid *volunteer first-year* students who annotated two cases using the system, Burger King Corp. v. Rudzewicz, 471 U.S. 462 (1985), and Burnham v. Superior Court of California, 495 U.S. 604 (1990); a

2007 set of *non-volunteer first-year* students that annotated three cases *Asahi*, *Burnham* and *Burger King* as part of their first-year legal process course; and a 2008 set of *expert third-year* students who annotated the same three cases as paid volunteers. The resulting groups yielded a total of 57 graphs for *Asahi*, and 71 each for *Burnham* and *Burger King*.

We engaged a pair of senior law school faculty from the University of Pittsburgh's School of Law to grade the graphs. Prior to the grading both faculty members trained on the system using the same series of cases as the students. Their graphs were made available to them during grading to act as a reference. Additionally one faculty graph was inserted into the other graders' stack in order to obtain additional cross-grader comparisons.

The faculty were then provided with a sample of 6 graphs drawn from a different case and a set of draft grading criteria. They marked up the cases independently of one another and then met to discuss the results and refine the grading criteria. This process was designed to ensure that the grading criteria were "legally sensible" and to avoid any spurious sources of error.

All graphs were provided to the faculty in anonymized form with each graph being designated by a randomly assigned ID that did not identify the student or study group. While each faculty member saw the same ID for the same graph they were presented the graphs in a randomly shuffled order to ensure that they did not grade them in the same order. In addition to the graphs themselves the graders were also provided with a copy of the argument transcript for which the graphs were constructed. Annotations on the graphs indicated to what segment of the transcript, if any, a particular node was linked. This facilitated lookup when assessing the individual Test and Hypothetical nodes.

Each grader began by partitioning the graphs into one of three equally-sized bins of "poor", "medium" and "good" graphs. They then further divided each bin equally into "better" and "worse" graphs. This binning resulted in a six-point grading scale for the graphs with roughly equal set sizes based upon an initial "gestalt" comparison. The purpose of this initial binning was to avoid the "reassessment" phenomenon whereby graders alter their criteria as they work through a set of materials. Having assigned the "gestalt" grade they then reshuffled the graphs and began assigning detailed grades and individual item grades.

For the detailed grades the faculty were given three general categories: *coverage*, how much of the essential argument does the graph include; *correctness*, how well does the graph represent the underlying argument; and *comprehension*, how well does the student understand the argument model. For each category the graders were given a set of criteria such as: "How well does the graph cover all of the essential hypotheticals in the argument?" For each criterion they assigned a six point score ranging from 0 ("poorly" or "not at all") to 5 ("completely").

They were then given a set of detailed questions for each individual test and hypothetical in the diagram (e.g., "Is this test correctly related to the relevant hypotheticals?"). These were graded on the same six point scale. Once that process was completed they then assigned an overall grade to each graph on a 12 point scale reflecting their now more complete judgment of the graph quality. Due to confusion in instructions one grader (grader B) assigned overall grades on a 6 point scale for *Asahi* and *Burger King* but not *Burnham*.

Case Name	ρ	*p-value*
Asahi	0.71	$p < 0.001$
Burger King	0.73	$p < 0.001$
Burnham	0.7	$p < 0.001$

Table 1: Inter-grader ranking agreement.

	A		B	
Case Name	ρ	*p-value*	ρ	*p-value*
Asahi	0.73	$p < 0.001$	0.83	$p < 0.001$
Burger King	0.85	$p < 0.001$	0.85	$p < 0.001$
Burnham	0.88	$p < 0.001$	0.87	$p < 0.001$

Table 2: Intra-grader ranking to score agreement.

Results & Discussion

Agreement metrics used in the literature range from Cohen's Kappa (Cohen 1960), to linear correlations, to generalizability theory as used by (McClure, Sonak, and Suen 1999), to ranked coefficients such as Spearman's Rank Sum Correlation or ρ (Spearman 1904). Kappa, while common in the literature, is designed for taxonomic tasks such as species classification where no ordinal relationship exists between the candidate classifications. As a consequence it is overly sensitive to minor grade variations. In this paper we made use of linear grade correlations to assess the inter-grader agreement on overall grades and use Spearman's ρ to report agreement with the gestalt grades.

A comparison of the gestalt grades using Spearman's ρ shown in Table 1 shows that the graders agreed substantially on the graph rankings for all three cases. While the agreement is not perfect this is a high degree of correlation and shows that the graders' initial assessments of the graphs tally. Additionally, we compared the authors' gestalt grades to their final grades for all three cases. The results are shown in Table 2. Again there is a high level of agreement, higher than between graders, indicating that the subsequent detailed grading tended to confirm the grader's initial assessments, not alter them.

	Slope		Intercept	
Case Name	*est.*	*p-value*	*est.*	*p-value*
Asahi	0.32	$p < 0.001$	0.92	$p < 0.001$
Burger King	0.30	$p < 0.001$	0.88	$p < 0.001$
Burnham	0.57	$p < 0.001$	2.85	$p < 0.001$

Table 3: Inter-grader overall score agreement.

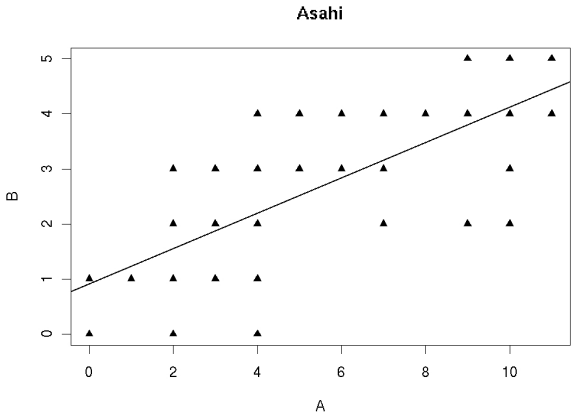

Figure 2: *Asahi* Grade Correlations.

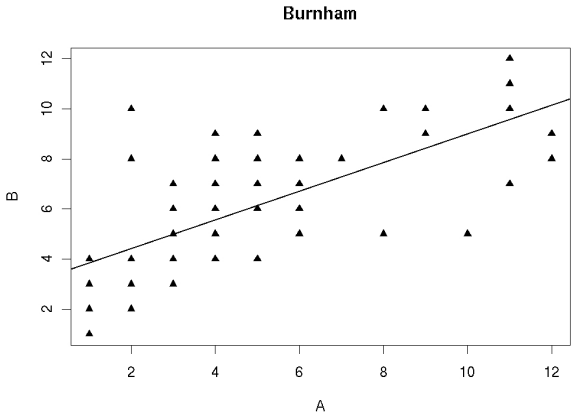

Figure 4: *Burnham* Grade Correlations.

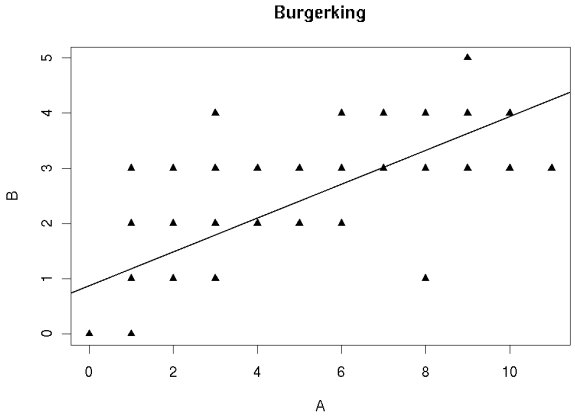

Figure 3: *Burger King* Grade Correlations.

For the purposes of the overall grading we computed an unweighted linear correlation between the two graders. The estimates and p-values for these models are shown in Table 3. Plots of the estimated values with fitted linear models are shown in Figures 2, 3, & 4. As noted above, one grader, B graded both the *Asahi* and *Burger King* cases on a six point scale. This, however, does not impair the correlational analysis. As with the gestalt grades the strength of these correlations indicates a high level of agreement between the graders. Despite this high level of overall agreement, however, there exist notable outliers. One *Burger King* graph, for example, received a score of 8/12 from grader A and a score of 1/6 from grader B. Similarly one *Burnham* graph received a score of 1/12 from grader A while receiving a score of 9/12 from grader B.

Conclusions & Future Work

During law school, skills in argumentation are rarely taught with explicit argument models. It is generally assumed that students understand the basic processes of argumentation prior to entering law school or will learn them as they go. Arguments are the language of law school classrooms and the main vehicle by which more complex domain knowledge is communicated both within class and in professional life. A student who fails to understand legal argument early in his or her education will be seriously impaired. This is particularly problematic as many misunderstandings are latent, based upon assumptions that neither the student nor professors explicitly identify. This means that initial misconceptions may grow increasingly embedded and thus harder to identify and correct. The adoption of computational argument models, argument diagrams, or other tools for reifying arguments within the classroom and legal practice might aid strongly in this endeavor.

As we stated above, computational models of argument are increasing in use both as communicative tools and educational supports. These models have their strong champions and we count ourselves among their number. Our purpose in this analysis was to focus on the argument diagrams' diagnostic utility thus addressing a basic prerequisite for their use in educational domains, inter-rater reliability.

Our analysis leads us to conclude that that LARGO diagrams can be used to diagnose overall student comprehension. Our analysis found strong agreement both in gestalt rankings and overall grades. Moreover the high intra-grader agreements suggest that the relatively short training process we engaged in is sufficient to make the grading process a natural one. While specific disagreements exist, this is not unusual. Expert disagreement is a feature of ill-defined domains such as law. This disagreement is endemic to all open-ended argument models and, like student misconceptions, is often based upon implicit conceptual differences. As open ended communication tools, the diagrams do not, in their present forms, eliminate all ambiguity, nor should we expect them to do so.

We are continuing our analysis of the argument diagrams.

As we described above, prior to assigning their overall grade, both graders assigned a number of detailed criteria covering features such as *Completeness*, *Correctness*, and *Comprehension*. They also assigned detailed grades to each test and hypothetical case. It remains to be seen whether this level of overall agreement extends to the more fine-grained grading. Our preliminary indications are that it does for some but not all of the criteria.

Coupled with this detailed analysis we are focusing our attention on the outliers. While the faculty established good general overall agreement, the outliers present interesting questions. Do they differ because of a disagreement about the meaning of a particular grading criterion, a disagreement about the argument model, or about the argument itself? After analyzing the outliers in greater detail, we will take them back to our graders in order to identify the sources of these disagreements. Having shown good overall agreement, we have now established a benchmark set of student grades. We plan to use this benchmark both to test the predictiveness of individual graph grades, and to assess the quality of LARGO's automated feedback. While this feedback is useful it remains to be seen whether or not it correlates with the facultys' assessments. Finally, we are now planning extensions to the LARGO system based upon our findings here. We plan to conduct additional studies with LARGO in the Fall of 2009 and are presently implementing changes to the system to facilitate both the students' and graders' work.

Acknowledgments

NSF Grant IIS-0412830, Hypothesis Formation and Testing in an Interpretive Domain, supported this work.

References

Ashley, K.; Lynch, C.; Pinkwart, N.; and Aleven, V. 2008. A process model of legal argument with hypotheticals. In *Legal Knowledge and Information Systems, Proc. Jurix 2008.*, 1–10.

Ashley, K. D. 2006. Hypothesis formation and testing in legal argument. invited paper. In *Inst. de Investig. Jurdicas 2d Intl. Meet. on AI and Law, UNAM, Mexico City.*

Carr, C. 2003. Using computer supported argument visualization to teach legal argumentation. In *In Visualizing Argumentation*. London, Springer. 75–96.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Easterday, M.; Aleven, V.; and Scheines, R. 2007. 'tis better to construct than to receive? the effects of diagram tools on causal reasoning. In Luckin, R.; Koedinger, K.; and Greer, J., eds., *Proc of AIED 2007*, 93–100. IOS Press.

Gordon, T.; Prakken, H.; and Walton, D. 2007. The carneades model of argument and burden of proof. *Artificial Intelligence* 171:875–896.

Gordon, T. 2007. Visualizing carneades argument graphs. *Law, Probability and Risk* 6(109).

Harrell, M. 2007. Using argument diagramming software to teach critical thinking skills. In *Proc. of the 5th International Conf. on Education and Information Systems, Technologies and Applications.*

Lund, K.; Molinari, G.; Sjourn, A.; and Baker, M. 2007. How do argumentation diagrams compare when student pairs use them as a means for debate or as a tool for representing debate? *Computer-Supported Collaborative Learning* 2(273).

Lynch, C.; Pinkwart, N.; Ashley, K.; and Aleven, V. 2008. What do argument diagrams tell us about students' aptitude or experience? a statistical analysis in an ill-defined domain. In Aleven, V.; Ashley, K.; Lynch, C.; and Pinkwart, N., eds., *Proc. of the Workshop on ITS for Ill-Defined Domains at the ITS 2008*, 56–67.

MacCormick, D. N., and Summers, R., eds. 1997. *Interpreting Precedents: a Comparative Study.* Ashgate/Dartmouth.

McClure, J.; Sonak, B.; and Suen, H. K. 1999. Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching* 36(4):475–492.

Paolucci, M.; Suthers, D.; and Weiner, A. 1996. Automated advice-giving strategies for scientific inquiry. In *Proc. ITS-1996*.

Pinkwart, N.; Aleven, V.; Ashley, K.; and Lynch, C. 2007. Evaluating legal argument instruction with graphical representations using largo. In *Proc. AIED2007.*

Pinkwart, N.; Lynch, C.; Ashley, K.; and Aleven, V. 2008. Re-evaluating largo in the classroom: Are diagrams better than text for teaching argumentation skills? In Woolf, B.; Aïmer, E.; Nkambou, R.; and Lajoie, S., eds., *Proc. of ITS-2008 Montral, Canada.*, 90–100.

Reed, C., and Rowe, G. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools* 13(4):961–980.

Spearman, C. 1904. The proof and measurement of association between two things. *Amer. J. Psychol.* (15):72101.

Suthers, D. D., and Hundhausen, C. D. 2001. Learning by constructing collaborative representations: An empirical comparison of three alternatives. In Dillenbourg, P.; Eurelings, A.; and Hakkarainen, K., eds., *Proc. of the* 1^{st} *European Conf. on CSCL.*, 577–584. Maastricht, the Netherlands.

Twardy, C. 2004. Argument maps improve critical thinking. *Teaching Philosophy* 27(2):95–116.

Van den Braak, S.; Van Oostendorp, H.; Prakken, H.; and Vreeswijk, G. 2006. A critical review of argument visualization tools. In *ECAI-06 Workshop on Computational Models of Natural Argument.*

van Gelder, T. 2007. The rationale for rational. *Law, Probability and Risk* 6(1-4):23–42.