Toward Assessing Law Students' Argument Diagrams

Collin Lynch
Intelligent Systems Program
Univ of Pittsburgh
Pittsburgh, Pennsylvania, USA
collinl@cs.pitt.edu

Kevin Ashley
LRDC & School of Law
Univ of Pittsburgh
Pittsburgh, Pennsylvania, USA
ashley@pitt.edu

Niels Pinkwart
Department of Informatics,
Clausthal Univ of Technology
Clausthal, Germany
np@tu-clausthal.de

Vincent Aleven
HCII, SCS,,
Carnegie Mellon Univ
Pittsburgh, Pennsylvania, USA
aleven@cs.cmu.edu

ABSTRACT

The development of graphical argument models is an active and growing area of research in Artificial Intelligence and Law. The aim is to develop models which may be readily used by legal professionals and novices to produce and parse arguments. If this goal is to be realized it is important to develop models that human reasoners can manipulate and assess consistently. We report on an ongoing study of graph agreement in the context of the LARGO system.

Keywords

Argument models, argument diagrams, intelligent tutoring systems, hypothetical legal reasoning

1. INTRODUCTION

Graphical argument models are a strong and growing area of research, especially in the area of legal argument [11, 5]. These models provide a framework for computationally modeling complex legal decision-making by instantiating argument schema and domain-specific critical questions about an argument's potential weaknesses [10]. The diagrams typically represent arguments as an accumulation of moves, or the results of moves in the argument scheme.

Diagrams reify argument structure making inferences, dependencies, and logical structures explicit in a form comprehensible both to human and machine reasoners. Thus they have the potential to facilitate argument comprehension particularly for novice arguers. Diagrams have been employed in legal education [1, 9, 2], critical thinking [13], causal reasoning [4], and natural science [12].

It is necessary to show that graphical arguments can be constructed and assessed reliably if they are to realize this potential, an empirical question. While many existing models are subject to strong formalisms, no set of rules can easily account for all possible variations. This is especially acute when dealing with arguments created by students or other non-experts who may misunderstand or disagree with the requirements of diagrammatic rules. In this paper we describe ongoing work on argument assessment. Our goal is to address the reliability of our argument model and to identify areas of future work.

So far little work has been done on the assessment of problem-solving and argumentation skills using diagrams. In [7] the authors manually compared Toulmin-style diagrams constructed by students in one of two conditions: free argument generation and argument reconstruction. Our research differs from their work in a variety of respects. First and foremost we are focusing on expert human graders employing agreed-upon criteria rather than a formal scoring algorithm as Lund did. This is similar to the route taken by McClure, Sonak and Suen [8]. However their focus was on the assessment of concept maps rather than functional arguments or procedural annotations. Most of the prior work employs Toulmin-style diagrams as opposed to process-model diagrams of the type used in LARGO. Additionally, our work focuses on argument reconstructions rather than novel argument generation as in [13].

2. STUDY

LARGO is an Intelligent Tutoring System for legal argumentation with tests and hypotheticals [1, 9]. In brief, students use the system to annotate oral arguments taken from the U.S. Supreme Court using a graphical model. While this task is more constrained than novel argument construction robustness is still important and results here have implications for general argument structures. In the present study, we have collected a total of 57 student diagrams produced by first and final-year law students. Each represents the petitioner's argument for the case of Asahi Metal Industry Co. v. Superior Court of California, 480 U.S. 102 (1987). This is the first of three cases that students' annotated within the system. We have collected and begun grading the diagrams for the remaining cases but will focus on Asahi here.

In the present work we are focused on the problem of inter-grader agreement – are grades assigned by different legal experts consistent? We engaged a pair of senior faculty from the University of Pittsburgh's School of Law to grade the diagrams. Both were trained on the system using the

ICAIL-2009 Barcelona, Spain

same training materials as the students. They were then provided with a sample of 6 graphs drawn from a different case and a set of draft grading criteria. Each marked up the cases independently before meeting to compare 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. The graders were then provided with the 57 Asahi diagrams. The graphs were provided in anonymized form with each grader being given a different ordering to avoid bias. They were also provided with a transcript from which the graph was drawn and a copy of their own Asahi diagram for reference.

Each grader began by partitioning the graphs to three equally-sized bins of "poor", "medium" and "good" graphs. They then subdivided each bin equally into "better" and "worse" sets. This resulted in an initial grading of the graphs on a six-point scale based upon a *gestalt* comparison. They then randomly shuffled the graphs and assigned detailed grades to the graph features, (tests and hypotheticals) as well as the graph's *coverage* of the argument, its' *correctness*, and the students' *comprehension* of the argument and the model. Finally the graders assigned a 12 point *overall* score reflecting their complete and detailed judgment of the graph.

3. PRELIMINARY RESULTS

Due to space limitations we will focus solely on the *gestalt* and *overall* grades. A comparison of the gestalt grades using Spearman's ρ shows that the graders agreed on the graph rankings ($\rho = 0.71$, $p < .001$). Additionally, this gestalt grade was highly correlated with the final grade ($\rho = 0.73$ for Grader A and 0.83 for Grader B, $p < 0.001$). This suggests that, while the detailed grading compelled the graders to spend more time on individual analysis, it did not consistently change their opinion of the graphs. Rather the detailed analysis confirmed their prior assessments.

A comparison of their *overall* grades revealed a more variable pattern. Unlike grader A, grader B assigned grades on a 6 point scale rather than the recommended 12 point scale. We were thus required to normalize both grades. Additionally grader B consistently ranked graphs higher than grader A. In order to account for this we applied a mean-score correction to the overall grades. This correction was not performed for the final *gestalt* grades which guaranteed equal means by design. After having applied this continuity correction we identified a substantial agreement between the two groups using a weighted Cohen's Kappa ($\kappa = 0.73$, $p < .001$; squared weights) [3, 6].

4. CONCLUSIONS

Argument diagrams have their strong champions and we count ourselves among their number. Our purpose in this analysis is to address the robustness of these models and to provide empirical support for their use as educational and communicative tools. Our preliminary analysis supports the contention that argument diagrams can serve as such tools. The high level of agreement between the experts' gestalt and final grades indicates that the graders can make gestalt judgments consistently. Moreover, the high levels of overall agreement supports our contention that graphs can be assessed consistently by different analysts. We anticipate comparing these scores to our full analysis in future work.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] K. Ashley, C. Lynch, N. Pinkwart, and V. Aleven. A process model of legal argument with hypotheticals. In *Legal Knowledge and Information Systems, Proc. Jurix 2008:* 21^{st} *Annual Conf.*, pages 1–10, 2008.

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

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

[4] M. Easterday, V. Aleven, and R. Scheines. 'tis better to construct than to receive? the effects of diagramming tools on causal reasoning. In R. Luckin, K. Koedinger, and J. Greer, editors, *Proc. of the* 13^{th} *International Conference on AI in Education*, pages 93–100. Amsterdam, IOS Press., 2007.

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

[6] J. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics.*, 33:159–174, 1977.

[7] K. Lund, G. Molinari, A. Sjourn, and M. Baker. 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).

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

[9] N. Pinkwart, V. Aleven, K. Ashley, and C. Lynch. Evaluating legal argument instruction with graphical representations using largo. In *Proc. AIED2007. Marina Del Rey, CA.*, July 2007.

[10] H. Prakken, C. Reed, and D. Walton. Dialogues about the burden of proof. In *Proc.10 Intl Conf. AI and Law.* ACM Press., 2005.

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

[12] D. D. Suthers and C. D. Hundhausen. Learning by constructing collaborative representations: An empirical comparison of three alternatives. In P. Dillenbourg, A. Eurelings, and K. Hakkarainen, editors, *European Perspectives on Computer-Supported Collaborative Learning, Proc. of the* 1^{st} *European Conference on CSCL.*, pages 577–584. Maastricht, the Netherlands., 2001.

[13] T. van Gelder. The rationale for rational. *Law, Probability and Risk: Special Issue on Graphic and Visual Representations of Evidence and Inference in Legal Settings.*, 6(1-4):23–42, 2007.