Assessing Argument Diagrams in an Ill-defined Domain

Niels PINKWART[a] , Collin LYNCH[b], Kevin ASHLEY[c] and Vincent ALEVEN[d]
[a] *Clausthal University of Technology, Department of Informatics, Germany*
[b] *University of Pittsburgh, Intelligent Systems Program, Pittsburgh, PA*
[c] *University of Pittsburgh, LRDC, Pittsburgh, PA, USA*
[d] *Carnegie Mellon University, HCI Institute, Pittsburgh, PA, USA*

Abstract. This paper describes a study in which student-created diagrams about arguments in an ill-defined domain were manually graded by two independent human graders. Findings include that the graders overall agreed with each other on their grades, but their agreement was lower than one would expect in well-defined domains, and higher for solutions of extreme quality.

Keywords. Argumentation, diagrams, inter-rater reliability, ill-defined domains

Introduction

Argumentation is a central skill in many aspects of life. As such, learning to argue is central for humans, and teaching argumentation is an important educational goal [1]. Many educational technology systems teach argumentation by having students create or reconstruct arguments in diagrammatic form [2,3,4,5]. Diagrams are often chosen as external representations because they can make the (often complex) structure of arguments salient, and at the same time support automated analysis better than other representations (such as text) do. This is especially important in the AIED field, where educational argumentation systems often attempt to analyze the diagrams in order to give students feedback. Examples of these systems include ArguMed [3] which provides feedback based on structural relations in diagrams, or ARGUNAUT [5], which analyzes arguments using machine learning and text analysis techniques.

Usually, argument diagrams serve primarily representational and procedural purposes in AIED systems: they are a vehicle designed to help the students learn as they create them. They have not played an important role in assessing the student's performance, which is typically measured in a pre/post test design that does not involve diagrams [2,4]. It would be very helpful for AIED researchers if argument diagrams were diagnostic – i.e., if they conveyed information about the learning process or the performance of the student. If, from looking at a diagram that a student produced, a prediction of his learning gains were possible with some accurateness, then this could reduce the need for employing time-consuming post-tests since examining the created diagrams would be sufficient (or would at least provide additional evidence for learning). At the same time, if diagrams were diagnostic and if we could elicit human graders' knowledge about what features they consider when making assessments, this information could be used to inform a system's automated diagram analysis.

This is especially important in dealing with the many forms of argumentation that are ill-defined. It is hard to construct objective tests for assessing argument skills when a hard “correctness” notion for an argument is impossible to define or verify formally, the underlying concepts are open-textured, and the quality of an argument may even be subject to expert disagreement [6]. For such arguments, if a diagram conveyed information about a student’s learning or level of understanding, it would be a positive boon. On the other hand, presumably, diagrams representing forms of arguments that are ill-defined are themselves hard to assess. If no single “ideal” diagram of an argument can be expected, it may even be questionable whether experts agree on their grading of student’s argument diagrams. Only if they do, would it make sense to take a further look at the diagnostic utility of these diagrams for use within AIED systems. The research question addressed in this paper is to what extent experts agree on the grades they assign to student-created diagrams of arguments in an ill-defined domain.

1. Grading procedure

We analyzed material from two prior studies in which law students at the University of Pittsburgh created diagrams about legal arguments using the LARGO software [2]. These diagrams were graphical reconstructions and annotations of textual transcripts of US Supreme Court oral arguments; they were structured according to the argument model described in [7]. As such, the diagrams contained the argument (as analyzed by the student), decomposed into proposed decision rules (“tests”), hypothetical challenges to these tests, and case facts. In their diagrams, the students could relate entities to each other using five different types of relations including “leads to”, “analogized to”, and “modified to”. Although the types of diagram elements the students could use were predefined, the degree of formality of the resulting diagrams is still relatively low (since free-text input was allowed in the diagrams), allowing a variety of good representations of a textual transcript.

57 diagrams covering the petitioners’ side in the legal case *Asahi Metal Industry Co. v. Superior Court*, 480 U.S. 102 (1987), which was the first argument transcript the students annotated in the prior studies, were given to two legal experts who graded these diagrams independently of each other after having contributed to the development of the grading criteria and agreed on them. First, the experts placed all graphs into three equal bins (poor, medium, and good) based solely on a relative ranking according to a “Gestalt grade”. They then partitioned each bin into "worse" and "better" sets depending upon the relative quality of the member diagrams, resulteing in an initial ranking of diagrams on a 6-point scale. Then, having reshuffled the diagrams, they assigned grades for three general criteria (argument *coverage*, *correctness* of the representation, and student’s *comprehension*), assigned detailed grades to every single test and hypothetical in a diagram, and finally gave an overall final grade to each diagram on a 12-point scale based on this in-depth analysis.

2. Results and Discussion

Overall, the two graders agreed with each other in their ranked “Gestalt grades” that they assigned after a cursory first look on the diagrams (Spearman’s $\rho = 0.71$, $p<.001$).

The graders' agreement on the final grades was measured by first normalizing the grades: one grader mistakenly used a 6-point scale instead of the 12-point scale, so we had to transform his grades to the 12-point scale; also, one grader gave consistently lower grades, so we shifted his grades up so that the group means were equal. A weighted Cohen's Kappa analysis with squared weights was then applied to the normalized grades (rounded to the original nominal scale) and revealed an agreement of $\kappa=0.74$ ($p<.001$). Thus, the two expert graders agreed with one-another overall, but their level of agreement was far from perfect. In a well-defined domain, where the distinction between right and wrong is clearer than it is for the type of argument diagrams used here, one would expect a higher level of agreement between experts than what we found in this study. Interestingly, the grader's level of agreement was much higher on the extreme diagrams (good or poor) than on diagrams of middle quality. Excluding the middle diagrams (defined as those where at least one expert assigned an overall grade between 4 and 8) from the analysis revealed a considerably higher level of agreement between the graders ($\kappa=0.83$, $p<.001$) than on the total set. As a comparison, the rater's agreement on the set of middle diagrams was only $\kappa=0.10$ ($p=.2$). We see this as an indication that the ill-defined nature of argumentation comes though more clearly in the *debatable* student solutions, those of medium quality. Here, the experts differed on what was and was not acceptable, while such disagreement was rare for student solutions of very high (or very low) quality.

In summary, we found that experts can agree with each other in assessing student-created diagrams of arguments in an ill-defined domain – but their level of agreement may be lower than in well-defined domains, and higher for diagrams of extremely good or poor quality. We conclude from this that the diagnostic value of argument diagrams is worth further exploration, but should be analyzed with care. A limitation of the analysis presented here is that only the grades on diagrams for one out of three cases the students worked on in the original studies were considered. Our experts are presently grading the remaining cases. Once they complete their work, we will investigate the level of overall agreement on the remaining cases as well as more detailed grading relations. Our aim is to determine the general aspects on which the graders tend to agree (or to disagree), and use this, together with a qualitative analysis on the reasons for disagreement, to determine the factors of domain ill-definedness that impact the assessment of student solutions.

References

[1] J. Andriessen, Arguing to Learn. In: Sawyer, R.K. (Ed.), *The Cambridge Handbook of the Learning Sciences*, p. 443-460, Cambridge University Press, New York, 2006.

[2] N. Pinkwart, C. Lynch, K. Ashley, & V. Aleven, Re-evaluating LARGO in the Classroom: Are Diagrams Better than Text for Teaching Argumentation Skills? In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, p. 90-100, Springer, Berlin, 2008.

[3] B. Verheij, Artificial Argument Assistants for Defeasible Argumentation, *AI* **150** (2003), 291-324.

[4] M. Easterday, V. Aleven, & R. Scheines, R., 'Tis Better to Construct than to Receive? The Effects of Diagram Tools on Causal Reasoning. In *Proceedings of AIED*, p. 93-100, IOS Press, Amsterdam, 2007.

[5] O. Scheuer, & B.M. McLaren, Helping Teachers Handle the Flood of Data in Online Student Discussions. In *Proceedings of ITS*, p. 323-332, Springer, Berlin, 2008.

[6] C. Lynch, K. Ashley, V. Aleven, & N. Pinkwart, Defining Ill-Defined Domains; A literature survey. In *Proceedings of the Workshop on ITS for Ill-Defined Domains at ITS 2006,* p. 1-10. Jhongli, 2006.

[7] K. Ashley, *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press/Bradford Books, Cambridge MA, 1990.