Empirically Valid Rules for Ill-Defined Domains

Collin F. Lynch
Center for Educational Informatics
North Carolina State University
Raleigh, North Carolina, U.S.A.
collinl@cs.pitt.edu

Kevin D. Ashley
Learning Research & Development Center
University of Pittsburgh
Pittsburgh, Pennsylvania, U.S.A.
ashley@pitt.edu

ABSTRACT

Ill-defined domains such as writing and design pose challenges for automatic assessment and feedback. There is little agreement about the standards for assessing student work nor are there clear domain principles that can be used for automatic feedback and guidance. While researchers have shown some success with automatic guidance through *a-priori* rules and *weak-theory structuring* these methods are not guaranteed widespread acceptance nor is it clear that the lessons will transfer out of the tutoring context into real-world practice. In this paper we report on data mining work designed to *empirically validate a-priori* rules with an exploratory dataset in the domain of argument diagramming and scientific writing. We show that it is possible to identify diagram rules that correlate with student performance but that direct correlations can often run counter to expert assumptions and thus require deeper analysis.

Keywords

Empirical Validity, Argument Diagramming, Ill-Defined Domains, Writing, Assessment, Intelligent Tutoring Systems

1. INTRODUCTION

Ill-defined domains such as writing and design pose key challenges for automatic assessment and feedback. Solvers of ill-defined problems must reify implicit or open-textured concepts or solution criteria to make problems solvable and then justify those decisions [8, 9]. Consequently, ill-defined problems lack widely-accepted domain theories or principles that can be used to provide automatic assessment and feedback. Moreover it is not always clear that automatic advice can generalize to a wider domain or transfer out of the study context into the real world. Our present goal is to identify *empirically valid* rules that both correlate with subsequent performance on the real-world tasks that are the target of instruction and can be used for guidance and assessment.

Prior researchers have advanced a number of techniques for guidance in ill-defined domains such as peer review and microworlds [9]. Researchers have also developed successful systems which guide students via optional rules or constraints [12, 10], a method known as *weak-theory scaffolding* [9]. This type of scaffolding can include use of constraints to bound otherwise open student solutions [15], or use of structured graphical representations combined with on-demand feedback as in Belvedere [14] and LARGO [12, 11].

LARGO, for example is a graph-based tutoring system for legal argumentation. Students use the system to read and annotate oral argument transcripts from the U.S. Supreme Court. As students read the transcript they identify crucial passages in the text containing legal *tests*, *hypothetical cases*, or *logical relationships* and represent them as elements in a graph with textual summaries. They are guided in this analysis via *a-priori* graph rules that detect violations of the argument model. While systems of this type have shown success, particularly with lower-performing students, no broad systematic attempt has yet been made to demonstrate the empirical validity of these graphical structures or rules. Validity is essential, especially in ill-defined domains, where the utility of the models have been assumed but where we cannot always be sure that a given violation of the model is a student error and not a judicial prerogative. Demonstrating empirical validity of the argument models would support their use both pragmatically, by helping to persuade skeptical domain experts that they are effective, and functionally, by providing us with an empirical confidence measure that can be used to evaluate or weight their implementation.

We have previously evaluated the individual predictiveness of the rules used in LARGO and found that while some could be used to classify students by performance few of rules were strongly predictive [7]. This assessment, however, is qualified by the fact that the rules were used to give advice to the students as they worked. Thus the students flagged by the rules in the analysis either received the advice and ignored it or did not ask. Moreover the performance measures used were comprehension tests and not the production of novel arguments. Some prior researchers (e.g. [2, 1]) have discussed the relationship between student-produced argument diagrams and written essays. Those analyses, however, are purely qualitative. In more recent work we examined the relationship between basic features of student argument diagrams, such as order and size, and found that they could be used to predict students' overall grades [6]. The features

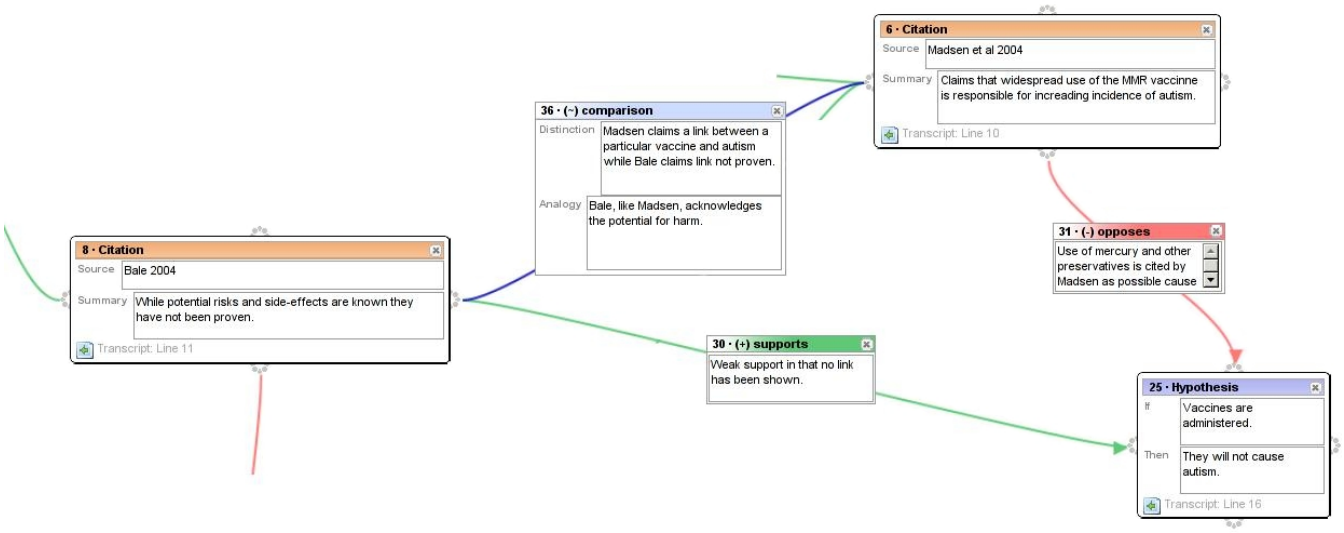

Figure 1: A segment of a student-produced LASAD diagram showing a hypothesis node (lower right) and two conflicting citation nodes with a comparison arc between them.

chosen, however, do not always lend themselves to robust feedback and the grades chosen incorporate a number of criteria beyond performance at argument.

In the present work we focus solely on an *exploratory dataset* where advice was not given and *planning diagrams* in which students plan novel arguments using a domain-specific argument model and the LASAD diagramming toolkit [3]. Students in this study were *not* provided with advice nor were they annotating an existing text. LASAD supports advice through an optional JESS-based system called the AFEngine [13, 3] which we are using in present studies. As part of this work we have shown that it is possible to reliably grade student-produced argument diagrams and essays, and that the expert-assigned diagram grades can be used to predict essay performance [4]. We have also shown that it is possible to make automatic predictions of student essay grades via regression models [5]. In the present paper we focus on individual rule evaluation.

2. METHODS

Data for this study was collected in a course on Psychological Research Methods in Fall 2011 at the University of Pittsburgh (see: [4]). Students in this course are taught study design, analysis, and ethics. The course is divided into lab sections. As part of the course students are required to conduct two empirical research projects including hypothesis formation, data collection, analysis, and writeup. Each lab jointly identifies a research topic and collaborates in data collection. The remaining aspects of planning, analysis, and writeup, are completed independently.

For the purposes of the study we augmented the traditional assignment with a graphical planning step. Once the students had completed the study design and data collection they were now required to plan their arguments graphically using the LASAD diagramming toolkit [3]. LASAD is an online tool for argument diagramming that allows for customized ontologies, peer collaboration, and annotation. The students were given a customized ontology with specialized nodes representing *hypothesis statements*, *citations*, and *claims*, and arcs representing *supporting*, *opposing*, and *undefined* relationships as well as *comparisons* between items.

Part of a representative student diagram is shown in Figure 1. The diagram shows a single hypothesis node (#25) at the lower right-hand corner. This node is supported by a citation node located on the left-hand side of the diagram (#8) and opposed by citation node (#6) at the top. These two citations are, in turn, connected to one-another by a comparison arc that states both analogies or similarities between the nodes and distinctions or differences. This structure forms a *paired counterargument with comparison*. Students were instructed to use it to express conflicting citations and to explain the source of the disagreement.

The diagrams and essays collected in the course were graded using a parallel rubric focused on the clarity, quality, persuasiveness, and other aspects of the argument. Grading was carried out by an experienced TA and reliability was tested in a separate inter-grader agreement study (see [4]). In that study we found that 5 of the 14 criteria were reliable and we focus on them below. The criteria chosen focus on: the quality of the research question (*RQ-Quality*); whether or not the hypothesis can be tested (*Hyp-Testable*); whether the author explained why the cited works relate to their argument (*Cite-Reasons*); and whether or not the hypothesis was open or untested (*Hyp-Open*). The final one measured the overall quality of the argument presented (*Arg-Quality*).

As part of this study we identified a set of 77 unique graph features for analysis. 34 of these were *simple features* such as the order and size of the diagram, the number of nodes and arcs of each type, and the amount of text in each node. Some of these were previously evaluated with legal arguments and found to be informative [6]. We also identified 43 com-

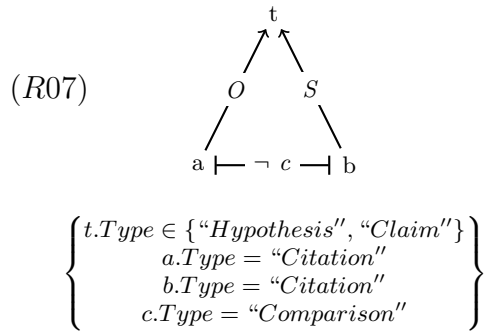

Figure 2: *R07: Uncompared Opposition* A simple augmented graph grammar rule that detects uncompared counterarguments. The rule shows a two citation nodes (*a*, & *b*) that have opposing relationships with a shared hypothesis or claim node (*t*) and do not have a comparison arc (*c*) drawn between them. The arcs *S* and *O* represent recursive supporting and opposing paths.

plex features that were designed to identify pedagogically-relevant subgraphs such as the paired counterarguments discussed above, unfounded hypotheses, and incorrect applications of individual arcs. These complex features were encoded as Augmented Graph Grammars and were evaluated using the AGG Engine [4]. Augmented Graph Grammars are a rule formalism that supports complex field constraints such as text criteria and multiple sub-fields as well as comparisons between nodes. The features were identified by domain experts based upon examination of previously-collected diagram and essay data and *a-priori* assumptions about the structure of good and poor data.

One such rule is shown in Figure 2. This rule detects *R07: Uncompared Opposition*. This occurs when two citation nodes, *a* & *b*, disagree about a shared hypothesis or claim node *t* with one opposing it and the other supporting and the user has *not* drawn a comparison arc between them to explain the disagreement. Students were trained to represent opposing citations with distinct nodes and then to explain that disagreement via a comparison arc. This rule tracks violations of that guidance and would not match the subgraph shown in Figure 1 which has a comparison arc.

For each diagram we collected a frequency count for the individual features and performed a series of pairwise comparisons mapping the observed frequencies to the paired essay grades. The comparisons were made using three candidate distributions: raw count, logarithmic, and binary. For analysis purposes the essay grades were normalized to a range of $(0-1)$. The statistical comparisons were made using Spearman's ρ, a nonparametric measure of correlation in the range $(-1 \leq \rho \leq 1)$ with -1 indicating a strong monotonically decreasing relationship and 1 a strong increasing one. ρ was chosen as it is robust in the face of nonlinear relationships.

3. RESULTS

We collected a total of 132 original diagrams and 125 essay introduction drafts from the course. After dealing with dropouts and incomplete assignments we obtained 105 unique diagram-essay pairs 31 of which were authored by individuals with the remaining 74 were authored by a team of 2-3. Of the features tested we found that eight of the simple features had statistically- or marginally-significant correlations between one or more of the grades and that all of the grades were significantly correlated with at least one simple feature. The weakest such correlation was between the *Order* or the number of nodes in the diagram ($\ln|G_n|$) and the grade *Cite-Reasons* (log: $\rho = 0.162, p < 0.098$). This is consistent with our expectations given the work in [6]. The strongest such correlation was between the presence of a hypothesis node (*Elt Hypothesis*) and the testability of the hypothesis (*Hyp-Testable*) (bin: $\rho = 0.383, p < 0.001$).

We found that 19 of the complex features were correlated with at least one of the grades and again all the questions were related to at least one feature. Here the weakest correlation was between the absence of a hypothesis node and *Cite-Reasons* (bin: $\rho = -0.166, p < 0.09$). The strongest correlation was between the amount of *uncompared opposition* (See Fig 2), that is the number of opposing citations without a comparison arc, and the grade for the openness of the hypothesis (*Hyp-Open*) (log: $\rho = 0.396, p < 0.001$).

Of these correlations some, such as the correlation between the presence of a hypothesis and the testability mentioned above, validate our *a-priori* assumptions. Hypothesis nodes are central to the entire argument and the empirical results validate their importance. We also found that the number of *paired counterarguments*, conflicting supporting and opposing nodes of the type shown in Figure 1, was positively correlated with the openness of the *Hyp-Open* ($(raw)\rho = 0.323, p < 0.001$). This was consistent with the instructions given to the students about how disagreements were to be presented and thus these results are promising. Moreover, we found that the presence of hypothesis nodes with no connection to a citation (*undefined ungrounded hypothesis* is negatively correlated with both *Cite-Reasons* (log: $\rho = -0.226, p < 0.02$), and *Arg-Quality* (log: $\rho = -0.219; p < 0.025$). This too reflects the need to ground the discussion in the appropriate literature.

While these results were positive a number of other significant correlations did not validate our assumptions. One notable example was the positive correlation between the uncompared opposition and *Hyp-Open* discussed above. We also found that the presence of unopposed hypotheses positively correlated with *Hyp-Testable* (log: $\rho = 0.196, p < 0.045$), and that the number of unfounded claims, claim nodes not connected to a citation, was positively correlated with *Hyp-Testable* (log: $\rho = 0.225, p < 0.021$).

4. ANALYSIS & CONCLUSIONS

Our goal in this research was to test the individual empirical validity of our *a-priori* diagram rules and to demonstrate the utility of empirical validation for ill-defined domains. To that end we collected a set of planning diagrams in a Research Methods course paired with graded argumentative essays. Unlike prior studies these diagrams were collected in an exploratory system where no automated advice was given to the students, and the argumentative essays were both novel, and graded independently with a focus on spe-

cific features of the arguments and their gestalt quality. In general, we found that some but not all of the features were significantly correlated with subsequent grades. Those correlations, however, were not always consistent with the *a-priori* assumptions that motivated their construction.

These counter-intuitive results are difficult to explain and highlight the central challenge of data-driven rule validation. The *Paired Counterarguments*, for example, are a positive diagram structure. Students were instructed to use them to indicate disagreement and, by extension, the openness of the hypothesis and research question. The rule defining them, however, is less precise than the rule defining uncompared opposition shown in Figure 2. Paired counterarguments omit any test for the comparison arc *c*. Thus all subgraphs detected by the latter rule will also be detected by the former. Given that the students were explicitly instructed to explain any opposing citations we expected that the latter rule would be strongly negative while the former would have a weak correlation at best. The fact that this was not the case suggests that either the students violated the instructions consistently or that the data is otherwise skewed, or that the rules are insufficiently precise to capture our *a-priori* assumptions.

We plan to address these limitations in future work by conducting a more detailed analysis of the existing data and by testing *conditional correlations*. In the case of uncompared opposition, for example, the author must have paired counterarguments in order to have the option of drawing a comparison arc. Thus it may be more informative to evaluate the impact of the uncompared opposition on graphs where paired counterarguments are found. This form of conditioning may address the generality of the rules but may require a larger dataset for us to draw robust conclusions. We also plan to test this approach on other related datasets that are presently being collected and to examine the alignment between the diagrams and essays. While the two elements were produced and graded separately, we anticipate that a more detailed tagging process should identify direct mappings between the diagram components and the essay structures. These mappings, if found, should enable us to perform a more direct evaluation of the role that individual structural elements play in the subsequent essay quality.

Acknowledgments

NSF Award 1122504 "DIP: Teaching Writing and Argumentation with AI-Supported Diagramming and Peer Review" Kevin D. Ashley PI, Chris Schunn & Diane Litman co-PIs.

5. REFERENCES

[1] Chad S. Carr. Using computer supported argument visualization to teach legal argumentation. pages 75–96. Springer-Verlag, London, UK, 2003.

[2] Evi Chryssafidou and Mike Sharples. Computer-supported planning of essay argument structure. In *Proceedings of the 5th International Conference of Argumentation*, June 2002.

[3] Frank Loll and Niels Pinkwart. Lasad: Flexible representations for computer-based collaborative argumentation. *Int. J. Hum.-Comput. Stud.*, 71(1):91–109, 2013.

[4] Collin F. Lynch. The Diagnosticity of Argument Diagrams, 2014. (defended January 30th 2014).

[5] Collin F. Lynch, Kevin D. Ashley, and Min Chi. Can diagrams predict essays? In Stefan Trausen-Matu and Kristy Boyer, editors, *Intelligent Tutoring Systems, 12th International Conference, ITS 2014, Honolulu, Hawai'i, USA*, Lecture Notes in Computer Science. Springer, 2014. (In Press).

[6] Collin F. Lynch, Kevin D. Ashley, and Mohammad H. Falakmassir. Comparing argument diagrams. In Burkhard Schäfer, editor, *Legal Knowledge and Information Systems - JURIX 2012: The Twenty-Fifth Annual Conference, University of Amsterdam, The Netherlands, 17-19 December 2012*, volume 250 of *Frontiers in Artificial Intelligence and Applications*, pages 81–90. IOS Press, 2012.

[7] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Argument graph classification with genetic programming and c4.5. In Ryan Shaun Joazeiro de Baker, Tiffany Barnes, and Joseph E. Beck, editors, *EDM*, pages 137–146. www.educationaldatamining.org, 2008.

[8] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, 19(3):253–266, 2009.

[9] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Ill-defined domains and adaptive tutoring technologies. In Paula J. Durlach and Alan M. Lesgold, editors, *Adaptive Technologies for Training and Education.*, chapter 9, pages 179–203. Cambridge, UK: Cambridge University Press., 2012.

[10] Antonija Mitrovic and Amali Weerasinghe. Revisiting the definition of ill-definedness and the consequences for itss. In *Proceedings of AIED2009*, 2009.

[11] Niels Pinkwart, Kevin D. Ashley, Collin F. Lynch, and Vincent Aleven. Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. *International Journal of Artificial Intelligence in Education*, 19(4):401–424, 2009.

[12] Niels Pinkwart, Collin F. Lynch, Kevin D. Ashley, and Vincent Aleven. Re-evaluating largo in the classroom: Are diagrams better than text for teaching argumentation skills? In Beverly Park Woolf, Esma Aïmeur, Roger Nkambou, and Susanne P. Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 90–100. Springer, 2008.

[13] O. Scheuer, S. Niebuhr, T. Dragon, B. M. McLaren, and N. Pinkwart. Adaptive support for graphical argumentation - the lasad approach. IEEE Learning Technology Newsletter 14(1), p. 8 - 11, 2012.

[14] Daniel D. Suthers. Representational guidance for collaborative inquiry. In *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments*, page 27âĂŞ46. 2003.

[15] Amali Weerasinghe, Antonija Mitrovic, and Brent Martin. Towards individualized dialogue support for ill-defined domains. *International Journal of Artificial Intelligence in Education*, 19(4):357–379, 2009.