Empirically Valid Rules for Ill-Defined Domains

Collin F. Lynch
Center for Educational Informatics
North Carolina State University
Raleigh, North Carolina, U.S.A.
collinf@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
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-
pared counterarguments. The rule shows a two citation augmented graph grammar rule that detects uncom-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-

plex features that were designed to identify pedagogically-
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

Proceedings of the 7th International Conference on Educational Data Mining 240