Can Diagrams Predict Essay Grades?

Collin F. Lynch[1], Kevin D. Ashley[1], and Min Chi[2]

[1] ISP, LRDC, & School of Law University of Pittsburgh, Pittsburgh, Pennsylvania.
collinl@cs.pitt.edu; ashley@pitt.edu,
http://www.cs.pitt.edu/∼collinl/ http://www.lrdc.pitt.edu/Ashley/
[2] North Carolina State University, Raleigh North Carolina.
mchi@ncsu.edu http://www.csc.ncsu.edu/people/mchi

Abstract. Diagrammatic models of argument have grown in prominence in recent years. While they have been applied in a number of tutoring contexts, it has not yet been shown that student-produced diagrams can be used to effectively grade students or predict their future performance. We show that manually-assigned diagram grades and automatic structural features of argument diagrams can be used to predict students' future essay grades, thus supporting the use of argument diagrams for instruction. We also show that the automatic features are competitive with expert human grading despite the fact that semantic content was ignored in automatic processing.

Keywords: Argument Diagrams, Essay Grading, Argumentation, Educational Datamining, Writing, Automatic Grading

1 Introduction

Argumentation is an essential skill, particularly in scientific domains where students must articulate and defend clear, testable, hypotheses and frame or recharacterize research problems in order to solve them. Argumentation is difficult for novices who often fail to comprehend arguments or formulate coherent new ones. Students' argumentation skills are often masked by their speaking and writing abilities, or lack thereof, which can limit the effectiveness of expert assessments and peer review. Despite this, argumentation is not always taught explicitly, even in domains such as law where its importance is widely acknowledged. Argumentation is also a challenging domain for AI as real-world arguments are open-ended, typically presented orally or as written text, rely on domain-specific conventions, and are often largely implicit. Thus argumentation presents unique and important challenges for Intelligent Tutoring Systems (ITSs).

Diagrammatic models of argument have been growing in prominence in recent years as theoretical models, practical tools, and educational interventions. The models make argument schema explicit, reifying the essential components and the structured relationships between them as a graph. This reification both makes the structure *salient* and imposes productive *constraints* on novices [**?**]. This unfamiliar structure, however, can be unfamiliar and challenging to master,

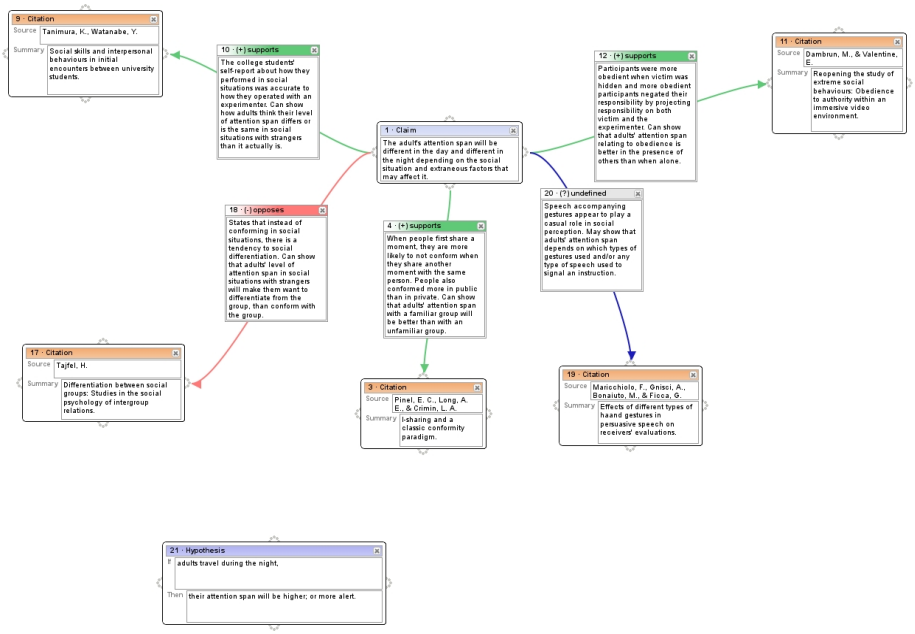

Fig. 1. A segment of a student-produced LASAD diagram representing an introductory argument. It contains a central *claim* node surrounded by *citation* nodes. The isolated node is a *hypothesis* that has not been integrated into the argument.

thus imposing additional cognitive load which can, in turn, inhibit performance [**?**]. Equally importantly, argument diagrams are amenable to computer processing. Making the structure of the argument explicit enables programmatic assessment and feedback. Argument diagrams have been used in a variety of domains including science [**?**], law [**?**], and philosophy [**?**]. A sample argument diagram of the type used in this study is shown in Figure **??**.

While argument diagrams have shown some success in tutoring contexts their overall performance has been mixed (see [**?**]) and important open questions remain. In particular, it has not yet been shown that student-produced argument diagrams are *empirically-valid*. That is, we have not yet shown that the diagrams can be graded and that the features of those diagrams can be used to predict subsequent performance on natural argumentation tasks such as essay writing. Some prior studies (e.g. [**?**]) have included qualitative analyses of existing diagrams but that has not been connected to subsequent student performance. In more recent work we have shown that some *a-priori* features of student diagrams (e.g. incorrect arcs) can be used to predict students' argument comprehension [**?**]. That work, however, focused solely on note-taking diagrams where students were annotating a shared example and did not consider their ability to make novel arguments. In subsequent work we showed that general features of student-produced arguments (e.g. size, length of summative text) could be used to predict subsequent assignment grades. Those grades, however, reflected crite-

ria such as students' presentation and the depth of their background research as well as argument quality. Nor did the study involve grading the diagrams themselves. Thus while argument diagrams have been used in ITSs, they have been promoted chiefly as pragmatic or *effective* interventions that improve student performance, not *diagnostic* ones. Much like a cricket player cross-training with a soccer game, the practice is helpful but doesn't show off your bowling.

This question of diagnosticity is important, however, both for theoretical and practical reasons. If one of the primary benefits of argument diagramming is the reification of argument structures then the diagram should reflect natural practice. If, however they are not diagnostic, then explicit scaffolding is not a useful explanation. Similarly, if the diagrams are not diagnostic then it will be difficult to convince often skeptical domain experts to use them in place of traditional representations. Moreover, if the diagram structure is not diagnostic it is not clear that the skills of argument diagramming are actually *transferable* to more traditional domains. Our goal in the present study is to address these questions by testing whether or not student-produced argument diagrams can be used to predict subsequent essay grades. We will test the following hypotheses:

H_a Manual diagram grades can be used to predict subsequent essay grades.
H_b Automatic diagram features can be used to predict subsequent essay grades.
H_c Feature-based predictions can be competitive with manual grade predictions.

2 Methods

We tested these hypotheses by means of a grading and machine learning study conducted with an exploratory dataset. The data consisted of a set of paired diagrams and essays collected from a course on psychological research methods (RM) held at the University of Pittsburgh in 2011. The diagrams were produced using LASAD and were graded using a set of parallel grading rubrics. We also defined a set of *a-priori* diagram rules that flagged pedagogically-relevant features. We then applied greedy linear regression to induce a set of predictive models connecting diagram features and grades to the essay grades.

LASAD is an online diagramming toolkit that supports complex diagram ontologies including node and arc types, subfields, and optional text links [**?**]. The ontology used here has 8 types: (nodes) *hypothesis*, *claim*, *citation*, and *current-study*; (arcs) *supporting*, *opposing*, *undefined*, and *comparison*. All contained flexible text fields for semantic information such as explanations of the relationships or citation information. A sample diagram is shown in Figure **??**. While LASAD has an optional help system (see [**?**]) it was not used here.

RM is a threshold course that covers ethics, study design, and analysis. It is subdivided into 9 lab sections. Students in each section are required to complete 2 empirical research projects. Each section collaborates on the general study design and data collection. Students author their research reports independently or in teams of 2-3. The reports follow a clear pattern. The students are instructed to present their overall argument in the *introduction* section stating their general research question, hypothesis, claims, and citing relevant work. The subsequent

sections are expected to support this basic structure. In non-study years the students are given lectures on hypothesis formation and selection of relevant citations but are not always given explicit instruction in argument formation. That is done implicitly through readings and discussion.

The study was integrated into the first writing assignment. Students were given an introductory lecture on argumentation, argument diagrams, and LASAD. They were then tasked with reading 1-3 published research papers and diagramming the arguments found using LASAD. They then used LASAD to diagram their own argument before writing their essays. Diagramming began in class and continued as a homework assignment with students submitting the final diagram and essay for grading. Further details may be found in [**?**].

The diagrams and essays were graded by an independent grader using a pair of parallel grading rubrics, one for diagrams and the other for essays. The grader had served as a TA in the course in 2012 where LASAD was used again. The rubrics each contained 14 questions, 11 of which focused on specific features of the arguments such as the use of citations and the quality of the hypothesis. The rest focused on the *gestalt* features of coherence, persuasiveness, and overall quality. 13 were graded on a scale of -2 to 2 in $\frac{1}{2}$ point increments. *G/E.14 (Arg-Quality)* was graded on a scale of -5 to 5 in $\frac{1}{2}$ increments given its broader scope. These scores were normalized to the range of 0 to 1 for analysis.

We tested the inter-grader reliability of the rubrics in a separate study [**?**]. In that study we found that suitably-trained graders can achieve statistically-significant or marginally-significant agreement on all of the diagram grades and most of the essay grades. In the present study we focused on the 5 features for which both criteria had statistically-significant agreement. 4 of these were specific criteria: (*E.01 (RQ-Quality)*) the quality of the research question; (*E.04 (Hyp-Testable)*) whether or not the hypothesis is *testable*; (*E.07 (Cite-Reasons)*) whether or not the author explains the relevance of the cited works; and (*E.10 (Hyp-Open)*) whether or not the author defends the novelty of the research hypothesis. The remaining question, *E.14 (Arg-Quality)*, addressed *gestalt* quality.

In other diagram-based systems such as LARGO [**?**], students are provided with automated advice driven by *a-priori* rules that detect violations of an ideal argument model or assignment-specific constraints. In this study we defined a set of 77 diagram features that we use for basic evaluation. 34 of these features were simple general features of the type examined in [**?**] such as the order and size of the diagram. The remaining 43 features were complex features that detect important components of the argument, such as pairs of counterarguments, and violations of argument constraints, such as claims without supporting citations.

We developed five predictive models for each essay question: $M_{baseline}$ is a static model that guesses the most common grade. M_{direct} is a simple linear model of the form $E_i = \alpha_i + \beta_i G_i + \epsilon$ that predicts each essay grade from the corresponding graph grade. M_{grade}, $M_{feature}$, and $M_{combined}$ are linear models that predict the essay grade based upon a subset of the diagram grades, diagram features, or both. These were induced via a two-pass process that first eliminates multicollinear features and then iteratively constructs predictive models based

upon the Root Mean Squared Error (RMSE). RMSE is an empirical measure of model error calculated under cross-validation. RMSE gives the absolute value of the expected error of each prediction. The candidate models were selected using a greedy hill-climbing approach. They were trained using least-squares regression with RMSE scores calculated using 10-fold cross-validation with balanced random assignment. The final RMSEs below were calculated via leave-one-out cross-validation. For more details on the algorithm see [**?**].

3 Results

We collected and graded 105 unique diagram-essay pairs. 74 were authored by a team, 31 by individuals. The model performance is shown in Table **??**. On every question $M_{combined}$ outperformed $M_{feature}$ which outperformed M_{grade}. M_{grade} met or beat M_{direct} which beat $M_{baseline}$. On question *E.10*, for example, the baseline RMSE was 0.463, or 1.8 points out of 5. M_{direct} and M_{grade} beat $M_{baseline}$ by 0.12, while $M_{combined}$ beat it by 0.152 or more than $\frac{1}{2}$ a point out of 5. On question *E.14* $M_{combined} < (M_{grade} \approx M_{feature}) < M_{direct} < M_{baseline}$ with $M_{combined}$ beating the baseline by 0.043 or almost $\frac{1}{2}$ a point out of a range of 11. Therefore both the expert grades (M_{direct} & M_{grade}) and diagram features ($M_{feature}$) were better predictors of students' subsequent essay grades than the baseline model $M_{baseline}$ while the combined models ($M_{combined}$) beat the others on every question.

4 Analysis and Conclusions

Proponents of argument diagrams, including ourselves, have long argued that they can be used for both *effective* and *diagnostic* tutorial interventions. Our goal in this study was to determine whether or not student-produced argument diagrams can be used to predict subsequent essay grades. In this work we showed: that manual diagram grades (M_{direct} & M_{grade}) were better predictors of the essay grades than the baseline model ($M_{baseline}$) thus validating hypothesis H_a; that models based upon diagram features ($M_{feature}$) also beat $M_{baseline}$ thus validating H_b; and that the grade and feature-based models were competitive

Table 1. RMSE scores for the five predictive models for the essay grades. The scores were calculated using leave-one-out cross-validation.

Question	$M_{baseline}$	M_{direct}	M_{grade}	$M_{feature}$	$M_{combined}$
E.01 (RQ-Quality)	0.344	0.311	0.311	0.29	0.284
E.04 (Hyp-Testable)	0.237	0.232	0.232	0.212	0.202
E.07 (Cite-Reasons)	0.27	0.248	0.245	0.243	0.223
E.10 (Hyp-Open)	0.463	0.339	0.334	0.316	0.311
E.14 (Arg-Quality)	0.245	0.214	0.206	0.207	0.202

($M_{feature} \leq M_{grade}$) thus validating H_c. This is surprising given that the human grader was able to evaluate the semantic content of the diagram fields while the automatic models did not. Therefore argument diagrams can be used for diagnostic educational interventions and this form of empirical modeling can be applied fruitfully even where natural language understanding is unavailable.

Interestingly, while M_{grade} and $M_{feature}$ were competitive, $M_{combined}$ dominated on every problem. Therefore either the semantic content was not used by the grader, contrary to instructions, or it conveyed different information than the diagram structure but conferred no substantive advantage. We plan to address this in future work and to test both the generality of these models and their use in ITSs to support individuals, peers, and instructors. In LARGO, for example, help is provided upon request and students are free to ignore it. Given these results, we plan to test whether help in argumentation should be compulsory for lower-performing students and then faded over time. We also plan to test whether diagnostic models such as these can be used to improve peer review and expert instruction by helping to rank students by skill level, to match appropriate mentors, and to flag students in need of expert guidance.

Acknowledgments

This work was Supported by National Science Foundation Award #1122504, "DIP: Teaching Writing and Argumentation with AI-Supported Diagramming and Peer Review," Kevin D. Ashley PI, Chris Schunn & Diane Litman, co-PIs.

References

1. Chryssafidou, E., Sharples, M.: Computer-supported planning of essay argument structure. In: Proc. of the 5th International Conference of Argumentation (2002)
2. Harrell, M., Wetzel, D.: Improving first-year writing using argument diagramming. In: Knauff, M., Sebanz, N., Pauen, M., Wachsmuth, I. (eds.) Proc. of the 35^{th} Annual Conf. of the Cognitive Science Society. pp. 2488–2493
3. Loll, F., Pinkwart, N.: Lasad: Flexible representations for computer-based collaborative argumentation. Int. J. Hum.-Comput. Stud. 71(1), 91–109 (2013)
4. Lynch, C.F.: The Diagnosticity of Argument Diagrams Univ. of Pittsburgh (2014)
5. Lynch, C.F., Ashley, K.D., Falakmassir, M.H.: Comparing argument diagrams. In: Schäfer, B. (ed.) JURIX 2012: The 25^{th} Annual Conference, University of Amsterdam, The Netherlands, 17-19 December. vol. 250, pp. 81–90. IOS Press (2012)
6. Lynch, C.F., Ashley, K.D., Pinkwart, N., Aleven, V.: Argument graph classification with genetic programming and c4.5. In: de Baker, R.S.J., Barnes, T., Beck, J.E. (eds.) EDM. pp. 137–146. www.educationaldatamining.org (2008)
7. Pinkwart, N., Ashley, K.D., Lynch, C.F., Aleven, V.: Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. International Journal of Artificial Intelligence in Education 19(4), 401–424 (2009)
8. Scheuer, O., Niebuhr, S., Dragon, T., McLaren, B.M., Pinkwart, N.: Adaptive support for graphical argumentation - the LASAD approach. IEEE Learning Technology Newsletter 14(1), p. 8 - 11 (2012)

9. Scheuer, O., Loll, F., Pinkwart, N., McLaren, B.: Computer-supported argumentation: A review of the state of the art. International Journal of Computer-Supported Collaborative Learning 5, 43–102 (2010)
10. Shum, S.J.B., MacLean, A., Bellotti, V.M.E., Hammond, N.V.: Graphical argumentation and design cognition. HCI 12(3), 267 300 (1997)
11. Suthers, D.D.: Empirical studies of the value of conceptually explicit notations in collaborative learning. In: Okada, A., Buckingham Shum, S., Sherborne, T. (eds.) Knowledge Cartography, pp. 1–23. Springer Verlag (2008)