Intelligent Argument Grading System for Student-produced Argument Diagrams

Linting Xue
North Carolina State University
Raleigh, North Carolina, USA
lxue3@ncsu.edu

ABSTRACT

Current automated essay grading systems are typically focused on the semantic and syntax analysis of written arguments via Natural Language Processing techniques. Few systems focus on the automatic assessment of argument *structure*. In this work, we propose to build an Intelligent Argument Grading System to automatically assess and provide feedback on the structure of arguments of student-produced argument diagrams, which are graphical representations for real-word argumentation. The proposed system contains two stages. In the first, it automatically induces empirically-valid graph rules for expert-graded argument diagrams. An assessment model is trained from the dataset of manually-graded argument diagrams with the feature of induced graph rules. In the second stage, the assessment model automatically grades and provides feedback by identifying both good features and structural flaws in students' work. The significance of this work will be that the proposed system can save high cost of labor by automatically inducing empirically-valid rules, grading, and providing feedback on the structure of arguments for students. We anticipate that the automatic feedback can help students revise their structural plans accordingly before they start to write essays, which will in turn lead them to produce more high-quality arguments.

Keywords

Argument Diagrams, Structure of Arguments, Automated Grading System, Automatic Feedback

1. INTRODUCTION

Argumentation is an essential skill in scientific domains including physics, engineering, and computer science, where students must articulate and justify testable hypotheses through argumentative reasoning. As a consequence, automated essay grading systems have become particularly useful tools for argument assessment (e.g. [1, 3, 9]). Prior research has shown that automated assessment systems can be used to assess student-produced arguments correctly and cost-effectively. Current automated grading systems rely on either surface-level analysis of linguistic features within a bock of text (as in [3]) or deeper Natural Language Processing (NLP) that utilizes machine learning techniques (as in [9, 1]). These systems are typically designed to evaluate on the basis of readability (e.g. the number of prepositions and relative pronouns or the complexity of the sentence structure), shallow semantic analysis (e.g. lexical semantics or the relationships analysis among named entities), and syntax analysis (e.g. grammatical analysis). Ultimately, these systems return the scores or feedback on the content and the qualities of the students' writing based on a predictive model that is trained by the dataset stored in the system.

However, very few active systems are focused on automatic analysis of the rhetorical structure of arguments to address structural flaws. Argument structure refers to the organization of the key components of argumentation (e.g. hypotheses, citations, or claims), which can reveal how the students justify their research hypotheses by using relevant evidence to support or oppose conclusory statements. In real-life teaching, the students are encouraged to structure their argumentative essays before they start writing by formulating a research hypothesis based on the research question, listing relevant evidence and factual information, and identifying the logical relationships between them. Evaluating the draft structure of these arguments and identifying flaws can help students to revise their plans and to produce high-quality arguments in the future. It is possible for human experts to grade draft arguments. However that process is costly and time-consuming.

In this work, we propose to build an Intelligent Argument Grading System that can automatically grade and provide feedback on the structure of students' arguments. The system will be based upon LASAD [4], an online tool for argument diagramming and collaboration. The input to the system will be a valid argument diagram, the output is the grade and feedback pointing out the outstanding substructures and structural flaws in the student's work.

2. BACKGROUND

2.1 Argument Diagrams

Argument diagrams are visual representations of real-world argumentation that reify the essential components of arguments such as *hypotheses* statements, *claims*, and *citations* as nodes and the *supporting*, *opposing*, and *clarification* relationships as arcs [6]. These complex nodes and arcs can

include text fields describing the node and arc types or free-text assertions, links to external resources and other data. Argument diagrams have been used in a variety of domains, including science [10], law[8] and philosophy [2] to help students learn written argumentation. Prior researchers have shown that argument diagrams can be used to scaffold students' understanding of existing arguments [2] and can help to support scientific reasoning [10].

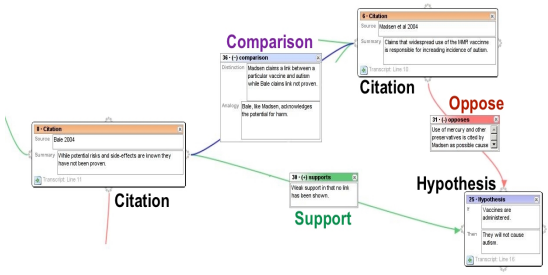

Figure 1: A student-produced Argument Diagram.

A sample student-produced diagram is shown in Figure 1. The diagram includes a *hypothesis* node at the bottom right, which contains two text fields, one for a conditional or *if* field, and the other for a consequent or *then* field. Two *citations* are connected to the *hypothesis* node via *supporting* and *opposing* arcs colored green and red, respectively. They are also connected via a *comparing* arc. Each citation contains two fields: one for the citation information and the other for a summary of the work; each arc has a single text field explaining what purpose the relationship serves.

3. PRELIMINARY RESULTS

In Lynch's study of diagnosticity of argument diagrams [5], a set of 104 paired diagrams and essays were collected at the University of Pittsburgh in a course on Psychological Research Methods. The diagrams and essays were independently graded by an experienced TA according to a parallel grading rubric. They showed that hand-authored graph rules were *empirically-valid* and were correlated with the diagram and essay grades; and thus that they could be used as the basis of predictive models for automatic grading.

Our prior work has also shown that Evolutionary Computation (EC) can be used to automatically induce empirically-valid graph rules for student-produced argument diagrams, and that the induced graph rules can be used as features for automatic grading [11, 12]. It is possible to harvest a set of diverse rules that were filtered via post-hoc Chi-Squared analysis [7]. This includes both good rules that are positively correlated with the diagram and essay grades and bad rules which are negatively correlated with the former representing positive structural features and the latter indicating flaws in the argument.

Figure 2 shows an example of a positive graph rule (P-G) and a negative graph rule (N-G) induced in our prior work. P-G shows a graph structure where the students identified at least two related citations ($c0$ & $c1$) that can be synthesized to support a single claim ($k0$) and where they included both a separate hypothesis (h) and an additional claim ($k1$).

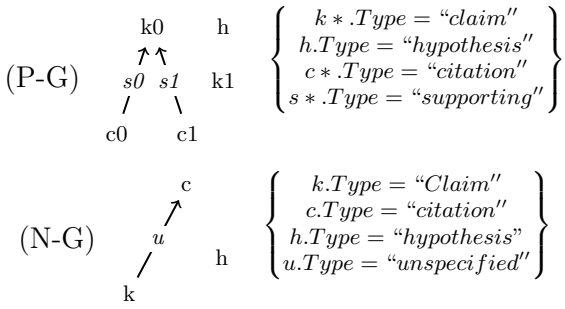

Figure 2: Examples of positive and negative graph rule.

It shows one of the structures that students have been encouraged to incorporate into their arguments as it shows an ability to synthesize citations to form a complex claim.

N-G is a negative rule that contains a single claim node (k) which is connected to a citation node (c) via an undefined arc (u), and a separate hypothesis node (h) which may or may not be connected to the rest structure. This rule is a clear violation of the semantic guidance that students were given. In our experiment, the students were instructed to use unspecified arcs for definitions or clarifications. Some students instead used them only when they were unsure about the strength of their evidence or did not understand the citation.

4. PROPOSED SYSTEM

In this work, we propose to build an Intelligent Argument Grading System (iARG) for student-produced argument diagrams. Our goal is to automatically grade the structure of arguments for students and provide feedback that reflects the good features and structural flaws in students' work. The proposed system includes two stages, which are shown in Figure 3.

The top part of Figure 3 illustrates the first stage, **Automatic Rule Induction**, in which the system automatically induces empirically-valid graph rules for expert-graded argument diagrams. The system will contain a database of argument diagrams and expert-assigned grades, along with a database of graph rules induced by the EC algorithm with a χ-Squared filter as described in [11, 7]. After the system produces a set of individual rules, the induced rules are evaluated by domain experts to determine whether or not they are semantically valid. Only valid rules will be incorporated into the database. Note that the induced rules contain both positive and negative examples. At the end of the process, we will use supervised learning methods to train an assessment model based upon the feature of induced rules and other graph feature (e.g. the degree of diagram nodes, the complexity of diagrams, and the attribute of the hub nodes in diagrams).

In the second stage of **Automatic Grading and Feedback**, the trained model will automatically grade and provide feedback on students' submissions by identifying both good features and structural flaws of the arguments. After

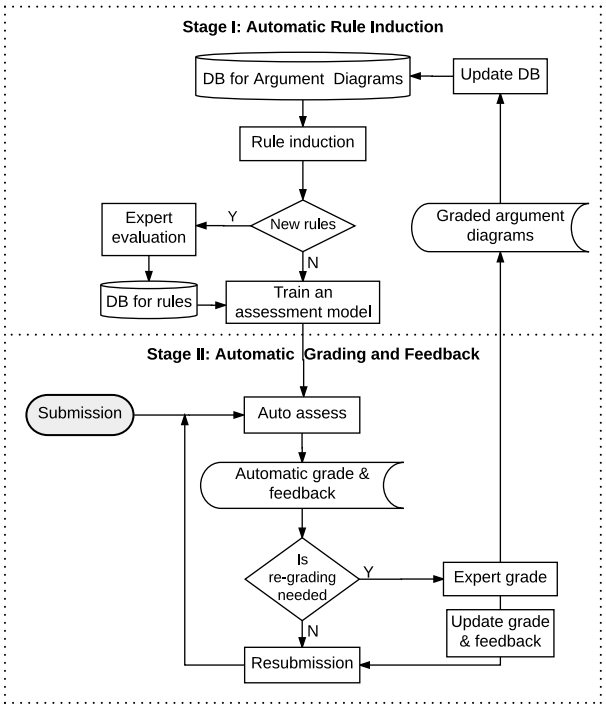

Figure 3: Flowchart for the proposed iARG

this, we will have experts re-evaluate the automatic grades and give feedback periodically, and if necessary, to re-grade the submission. We include this step because the students' submissions may include novel structures that are not included in the current rule database. In this case, the assessment model may treat these novel structures as outliers and provide uncorrected feedback. If the submissions are re-graded by experts, they will be updated to the database for argument diagrams. The rule database and assessment model will also be updated for future use.

5. FUTURE WORK & OPEN QUESTIONS

In the future work, we plan to achieve the following:

1. In Fall 2017, we plan to work with domain experts to determine whether the induced graph rules are semantically valid; whether they can be used for automatic grading; and whether they include all of the good features and structural flaws in students' work. This gives rise to our first research question: *how can we improve the performance of the graph rule induction algorithm by inducing more empirically-valid graph rules?*

2. In Spring 2018, we will leverage different supervised learning methods to train an assessment model from our current dataset of expert-graded argument diagrams with the feature of valid graph rules and other graph features. We will evaluate the assessment model on a new set of student-produced argument diagrams. Our second research question is that *what other graph features can we use to build the assessment model?*

3. In Fall 2018, we plan to implement the proposed system based upon LASAD by building databases for the argument diagrams and for the graph rules, and integrating the assessment model into the system.

4. In 2019, we will test the performance of our system in an augmentative writing class at NCSU. We will focus on accessing the automatic grades and feedback from the student's perspective and determine whether they find the automatic feedback to be useful. Thus we will not have experts to examine the automatic feedback in the second stage. Based upon the students' feedback, we will consider whether to have experts to regrade the new submission and to update the database and assessment model.

6. REFERENCES

[1] J. Burstein, C. Leacock, and R. Swartz. Automated evaluation of essays and short answers. 2001.

[2] M. Harrell and D. Wetzel. Improving first-year writing using argument diagramming. In *The 35th CogSci*, pages 2488–2493, 2013.

[3] M. A. Hearst. The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, 15(5):22–37, 2000.

[4] F. Loll and N. Pinkwart. Lasad: Flexible representations for computer-based collaborative argumentation. *International Journal of Human-Computer Studies*, 71:91–109, Januart 2013.

[5] C. F. Lynch and K. D. Ashley. Empirically valid rules for ill-defined domains. In J. Stamper and Z. Pardos, editors, *Proceedings of The* 7^{th} *International Conference on EDM*. IEDMS, 2014.

[6] C. F. Lynch, K. D. Ashley, and M. Chi. Can diagrams predict essay grades? In S. Trausan-Matu, K. E. Boyer, M. E. Crosby, and K. Panourgia, editors, *ITS*, Lecture Notes, pages 260–265. Springer, 2014.

[7] C. F. Lynch, L. Xue, and M. Chi. Evolving augmented graph grammars for argument analysis. GECCO, 2016.

[8] N. Pinkwart, K. D. Ashley, C. F. Lynch, and V. Aleven. Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. *IJAIED*, 19(4):401 – 424, 2009.

[9] L. M. Rudner and T. Liang. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2), 2002.

[10] D. D. Suthers. Empirical studies of the value of conceptually explicit notations in collaborative learning. In A. Okada, S. Buckingham Shum, and T. Sherborne, editors, *Knowledge Cartography*, pages 1–23. Springer Verlag, 2008.

[11] L. Xue, C. Lynch, and M. Chi. Unnatural feature engineering: Evolving augmented graph grammars for argument diagrams. In *Internatinal Educational Data Mining*, pages 255–262. IEDMS, 2016.

[12] L. Xue, C. F. Lynch, and M. Chi. Mining innovative augmented graph grammars for argument diagrams through novelty selection. EDM, 2017.