# Ontology-Based Argument Mining and Automatic Essay Scoring

**Nathan Ong, Diane Litman, and Alexandra Brusilovsky**
Department of Computer Science, University of Pittsburgh
Pittsburgh, PA 15260 USA
nro5,dlitman,apb27@pitt.edu

## Abstract

Essays are frequently used as a medium for teaching and evaluating argumentation skills. Recently, there has been interest in diagrammatic outlining as a replacement to the written outline that often precedes essay writing. This paper presents a preliminary approach for automatically identifying diagram ontology elements in essays, and demonstrates its positive correlation with expert scores of essay quality.

## 1 Introduction

Educators tend to favor students providing a minimal-writing structure, or an outline, before writing a paper. This allows teachers to give early feedback to students to reduce the amount of structural editing that might be needed later on. However, there is evidence to suggest that standard text-based outlines do not necessarily improve writing quality (Torrance et al., 2000). Recently, there has been growing interest in graphical outline representations, especially for argumentative essays in various domains (Scheuer et al., 2009; Scheuer et al., 2010; Peldszus and Stede, 2013; Reed and Rowe, 2004; Reed et al., 2007). Not only do they provide a different outlining format, but they also allow students to concretely visualize their argumentation structure. Our work is part of the ArgumentPeer project (Falakmassir et al., 2013), which combines computer-supported argument diagramming and peer-review with the goal of improving students' writing skills.

In this paper, we follow the lead of others in discourse parsing for essay scoring (Burstein et al., 2001), and we preliminarily attempt to answer two questions: Q1) Can an argument mining system be developed to automatically recognize the argument ontology used during diagramming, when processing a student's later written essay? Q2) If so, is the number of ontological elements that can be recognized in a student's essay correlated with the essay's argumentation quality? Potentially, answering these questions in the affirmative would allow us to assist students with their writing by allowing computer tutors to label sentences with the ontology, determine which elements are missing, and suggest adding these missing elements to improve essay quality.

## 2 Corpus

Our corpus for argument mining consists of 52 essays written in two University of Pittsburgh undergraduate psychology courses. In both courses, students were asked to write an argumentative essay supporting two separate hypotheses that they created based on data they were given. The average essay contains 5.2 paragraphs, 28.6 sentences, and 592.1 words.

Before writing the essay, students were first required to generate an argument diagram justifying their hypotheses using the LASAD argumentation system[1]. LASAD argument diagrams consist of nodes and arcs from an instructor-defined ontology, as shown in Figure 1. Next, students were required to turn their diagrams into written argumentative essays. Automatically tagging these essays according to the 4 node types (**Current Study, Hypothesis, Claim, Citation**) and 2 arc types (**Supports, Opposes**) common to both courses is the argument mining goal of this paper. The tagged essay corresponding to Figure 1 is shown in Table 1.[2] While the diagram is required to be completed by students, this work does not utilize the student diagrams.

[1] http://lasad.dfki.de

[2] Both diagrams and papers were distributed to other students in the class for peer review. While the diagrams were not required to be revised, students needed to revise their essays to address peer feedback. To maximize diagram and essay similarity, here we work with only the first drafts.

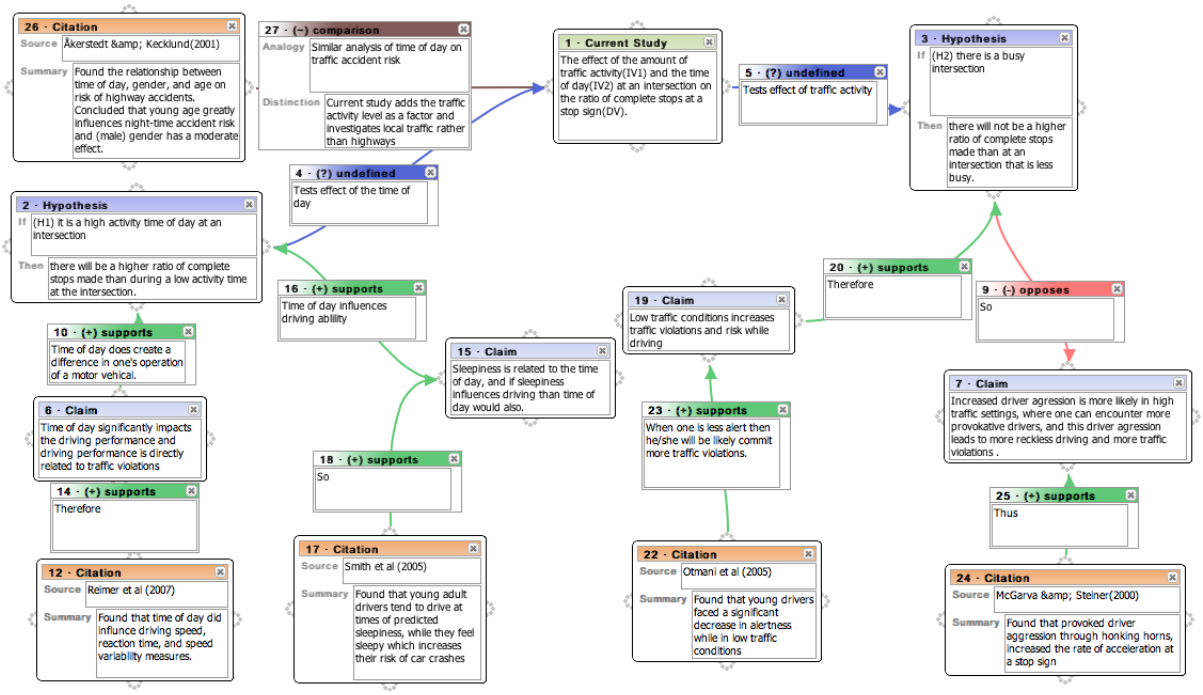


Figure 1: An argument diagram from a research methods course.

After the courses, expert graders were asked to score all essays on a 5-point Likert scale (with 1 being the lowest and 5 being the highest) without the diagrams, using a rubric with multiple criteria. For the essay as a whole, graders not only checked for correct grammar usage, but also for flow and organization. In addition, essays were graded based on the logic behind their argumentation of their hypotheses, as well as addressing claims that both supported and opposed their hypotheses. While not an explicit category, many of the criteria required students to present multiple citations backing their hypotheses. The average expert score for the 52 essays is 3.03, and the median is 3, with the scores distributed as shown in column four of Table 2.

## 3 Methodology

**Essay Discourse Processing.** Firstly, raw essays are parsed for discourse connectives. Explicit discourse connectives are then tagged with their sense (i.e. *Expansion, Contingency, Comparison*, or *Temporal*) using the Discourse Connectives Tagger[3], as shown in Table 1.

**Mining the Argument Ontology.** We developed a rule-based algorithm to label each sentence in an essay with at most one label from our target argument ontology. Our rules were developed using our intuition and informal examination of 9 essays from the corpus of 52. The algorithm consists of the following ordered[4] rules:

**Rule 1:** If the sentence begins with a *Comparison* discourse connective, or if the sentence contains any string prefixes from {conflict, oppose} and a four-digit number (intended as a year for a citation), then tag with **Opposes**.

**Rule 2:** If the sentence begins with a *Contingency* connective and does not contain a four-digit number, then tag with **Supports**.

**Rule 3:** If the sentence contains a four-digit number, then tag with **Citation**.

**Rule 4:** If the sentence contains string prefixes from {suggest, evidence, shows, Essentially, indicate} (case-sensitive), then tag with **Claim**.

**Rule 5:** If the sentence is in the first, second, or last paragraph, and contains string prefixes from {hypothes, predict}, or if the sentence contains the word "should" and contains no *Contingency* connectives, and does not contain a four-digit number and does not contain string prefixes from {conflict, oppose}, then tag with **Hypothesis**.

**Rule 6:** If the previous sentence was tagged with Hypothesis, and this sentence begins with an *Expansion* connective and does not contain a four-

[3] http://www.cis.upenn.edu/~epitler/discourse.html

[4] When multiple rules apply, the tag of the earliest is used.

| # | Essay Sentence | Label | Rule |
|---|---|---|---|
| 1 | The ultimate goal of this study is to investigate the relationship between stop-sign violations and traffic activity. | Current Study | 7 |
| 2 | To do this we analyzed two different variables on traffic activity: time of day and location. | None | 8 |
| ... | ... | ... | ... |
| 6 | Stop-signs indicate that the driver must come to a complete stop before the sign and check for oncoming and opposing traffic before[-*Temporal*] proceeding on. | Claim | 4 |
| 7 | For a stop to be considered complete the car must completely stop moving. | None | 8 |
| ... | ... | ... | ... |
| 16 | The first hypothesis was: If[-*Contingency*] it is a high activity time of day at an intersection then[-Contingency], there will be a higher ratio of complete stops made than during a low activity time at the intersection. | Hypothesis | 5 |
| 17 | The second hypothesis was: If[-*Contingency*] there is a busy intersection then[-*Contingency*], there will be a higher ratio of complete stops made than at an intersection that is less busy. | Hypothesis | 5 |
| 18 | So[-*Contingency*] essentially, it was expected that when[-*Temporal*] there was a higher traffic activity level, either due to location or time of day, there were to be less stop-sign violations. | Supports | 2 |
| 19 | There have been many studies which indicate that people do drive differently at different times of day and[-*Expansion*] that it does have an impact on driving risk. | Claim | 4 |
| 20 | Reimer et al (2007) found that time of day did influence driving speed, reaction time, and speed variability measures. | Citation | 3 |
| ... | ... | ... | ... |
| 24 | However[-*Comparison*], McGarva & Steiner (2000) oppose the second hypothesis because[-*Contingency*] they found that provoked driver aggression through honking horns, increased the rate of acceleration at a stop sign. | Opposes | 1 |
| ... | ... | ... | ... |

Table 1: Essay sentences, their mined ontological labels, and rules used to determine the labels, for the essay associated with Figure 1. Inferred discourse connective senses are *italicized* in square brackets.

digit number, then tag with **Hypothesis**.

**Rule 7:** If the sentence is in the first or last paragraph and contains at least one word from {study, research} and does not contain the words {past, previous, prior} (first letter case-insensitive) and does not contain string prefixes from {hypothes, predict} and does not contain a four-digit number, then tag with **Current Study**.

**Rule 8:** Do not assign a tag to the sentence.

Some sample output can be found on Table 1. Note that sentence 24 could have been tagged as **Citation** using Rule 3, but because it fits the criteria for Rule 1, it is tagged as **Opposes**.

**Ontology-Based Essay Scoring.** We also developed a rule-based algorithm to score each essay in the corpus. These rules were developed using our intuition in conjunction with the examination of the expert grading rubric. These rules take a labeled essay from the argument mining algorithm and outputs a score in the continuous range [0,5] using the following procedure:[5]

**1:** Assign one point to essays that have at least one sentence tagged with **Current Study** (**CS**).

**2:** Assign one point to essays that have at least one sentence tagged with **Hypothesis** (**H**).

**3:** Assign one point to essays that have at least one sentence tagged with **Opposes** (**O**).

**4:** Assign points based on the sum of the number of sentences tagged with **Claim** (**Cl**) and the number of sentences tagged with **Supports** (**S**), all divided by the number of paragraphs (**#¶**). If this

[5]Score 0 occurs when no labels are assigned to the essay.

value exceeds 1, assign only one point.

**5:** Assign points based on the number of sentences tagged with **Citation** (**Ci**) divided by the number of paragraphs (**#¶**). If this value exceeds 1, assign only one point.

**6:** Sum all of the previously computed points.

For the three paragraph essay excerpted in Table 1 (assigned expert score 3), there were three sentences tagged with **Current Study**, three with **Hypothesis,** one with **Opposes**, one with **Supports**, two with **Claim** and three with **Citation**. The score is computed as follows:

$$1_{CS} + 1_H + 1_O + \frac{2_{Cl} + 1_S}{3_{\#\P}} + \frac{3_{Ci}}{3_{\#\P}} = 5$$

## 4 Results

Since our essays do not have gold-standard ontology labels yet, we cannot intrinsically evaluate the argument mining algorithm. We instead performed an extrinsic evaluation via our use of the mined argument labels for essay scoring.

The average automatic score for the corpus is 3.42 and the median is 3.5, while the corresponding expert values are 3.03 and 3, respectively. A paired t-test of the means has a significance of $p < 0.01$, suggesting that our algorithm over-scores the essays. We also ran a one-sample t-test on each expert score value to see if the automatic scores were similar to the expert scores. We hypothesized that within each expert score category predicted accurately, we should not see a significant difference ($p \geq 0.05$). Table 2 shows that while the automatic score is not significantly different for expert score 4, the scores are significantly different for scores 2 and 3.

We also examined the Spearman's rank correlation between the computed and expert scores.[6] We see that the Spearman's rank correlation shows significance of $p < 0.0001$ with a rho value of 0.997. Together these metrics suggest that our automated scores are currently useful for ranking but not for rating.

## 5 Conclusion and Future Work

We have presented simple rule-based algorithms for argumentation mining in student essays and essay scoring using argument mining. Based on preliminary extrinsic evaluation, our pattern-based recognition of a basic argumentation ontology seems to provide some insight into essay scores across two courses. While the automatic scores did not necessarily reflect the expert scores, the ranking correlation demonstrated that more argumentative elements were related to higher scores. Even with the limitations of this study (e.g. no intrinsic evaluation, a small essay corpus, a limited argument ontology, a scoring algorithm using only ontology features, application of discourse connector for a different genre), our results suggest the promise of using argument mining to trigger feedback in a writing tutoring system.

| expert score | avg. auto score | t | n | p |
|---|---|---|---|---|
| 1 | 4.33 | – | 1 | – |
| 2 | 3.23 | 3.21 | 8 | 0.013 |
| 3 | 3.30 | 2.10 | 31 | 0.044 |
| 4 | 3.80 | -1.00 | 12 | 0.337 |

Table 2: One-sample t-test results for scores.

To develop a more linguistically sophisticated and accurate argument mining algorithm, our future plans include exploiting discourse information beyond connectives, e.g., by parsing our essays in terms of PDTB (Lin et al., 2011) or RST relations (Feng and Hirst, 2012). We also plan to look at the helpfulness of argumentation schemes (Feng and Hirst, 2011), and other linguistic and essay features for automatic evaluation (Crossley and McNamara, 2010). In addition, our essays are being annotated with diagram ontology labels, which will enable us to use machine learning to conduct intrinsic argument mining evaluations and to learn the weights for each rule or determine new rules. Finally, we plan to explore using the diagrams to bootstrap the essay annotation process. While some sentences in an essay can easily be mapped to the corresponding diagram (e.g. sentence 1 in Table 1 to node 1 in Figure 1), the complication is that essays tend to be more fleshed-out than diagrams, and at least in our corpus, also contain argument changes motivated by diagram peer-review. While sentence 6 in Table 1 is correctly tagged as a **Claim**, this content is not in Figure 1.

## Acknowledgments

This work is supported by NSF Award 1122504. We thank Huy Nguyen, Wenting Xiong, and Michael Lipschultz.

[6] A Pearson correlation did not give significant results.

## References


[Burstein et al.2001] Jill Burstein, Karen Kukich, Susanne Wolff, Ji Lu, and Martin Chodorow. 2001. *Enriching automated essay scoring using discourse marking*. ERIC Clearinghouse.

[Crossley and McNamara2010] Scott A Crossley and Danielle S McNamara. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In *Proceedings of the 32nd annual conference of the Cognitive Science Society*, pages 984–989. Austin, TX: Cognitive ScienceSociety.

[Falakmassir et al.2013] Mohammad Falakmassir, Kevin Ashley, and Christian Schunn. 2013. Using argument diagramming to improve peer grading of writing assignments. In *Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education*, Memphis, TN.

[Feng and Hirst2011] Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

[Feng and Hirst2012] Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 60–68. Association for Computational Linguistics.

[Lin et al.2011] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997–1006. Association for Computational Linguistics.

[Peldszus and Stede2013] Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

[Reed and Rowe2004] Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

[Reed et al.2007] Chris Reed, Douglas Walton, and Fabrizio Macagno. 2007. Argument diagramming in logic, law and artificial intelligence. *The Knowledge Engineering Review*, 22(01):87–109.

[Scheuer et al.2009] Oliver Scheuer, Bruce M. McLaren, Frank Loll, and Niels Pinkwart. 2009. An analysis and feedback infrastructure for argumentation learning systems. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, pages 629–631, Amsterdam, The Netherlands, The Netherlands. IOS Press.

[Scheuer et al.2010] Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.

[Torrance et al.2000] Mark Torrance, Glyn V. Thomas, and Elizabeth J. Robinson. 2000. Individual differences in undergraduate essay-writing strategies: A longitudinal study. *Higher Education*, 39(2):181–200.