# Rubric Reliability and Annotation of Content and Argument in Source-Based Argument Essays

**Yanjun Gao**[*] and **Alex Driban**[*] and **Brennan Xavier McManus**[**] and **Elena Musi**[†]
**Patricia M. Davies**[‡] and **Smaranda Muresan**[**] and **Rebecca J. Passonneau**[*]
[*]Department of Computer Science and Engineering
Penn State University
{yug125, akd524, rjp49}@cse.psu.edu
[**]Department of Computer Science, Columbia University
{bm2530, smara}@columbia.edu
[†]Department of Communication and Media, University of Liverpool
Elena.Musi@liverpool.ac.uk
[‡]College of Sciences and Human Studies, Prince Mohammad Bin Fahd University
pdavies@pmu.edu.sa

## Abstract

We present a unique dataset of student source-based argument essays to facilitate research on the relations between content, argumentation skills, and assessment. Two classroom writing assignments were given to college students in a STEM major, accompanied by a carefully designed rubric. The paper presents a reliability study of the rubric, showing it to be highly reliable, and initial annotation on content and argumentation annotation of the essays.

## 1 Introduction

Researchers in education have long recommended the use of rubrics to assess student writing and to inform instruction, especially regarding feedback to students (Graham et al., 2016; Jonsson and Svingby, 2007). Writing is important not only as a means to demonstrate knowledge, but also to acquire understanding of subject matter, including in STEM (Sampson et al., 2013; Klein and Rose, 2010; Gunel et al., 2007; Norris and Phillips, 2003). Argumentative writing plays a key role in such learning (Hand et al., 2015). It is difficult, however, for instructors in subject areas to provide writing instruction alongside the disciplinary content (Gillespie et al., 2014; Graham et al., 2014; Kiuhara et al., 2009). We are investigating the use of rubrics to support instruction in argumentation writing, with two goals in mind. Our first goal is to investigate effective instruction of argument writing skills, including the design and application of rubrics. Our second goal is to investigate how natural language processing techniques can facilitate instructors' use of rubrics.

The study described here is a collaboration among three computer science faculty: one specializing in educational technology, and two in natural language processing (NLP), who apply NLP to educational data. To investigate how a rubric can support instruction in argument writing, we designed a sequence of two argument essay assignments and rubrics. The collaborator in educational technology gave the assignments to college freshman enrolled in her academic skills class in their first semester. Both assignments asked students to do a critical analysis of source material, and write an argumentative essay in response to a prompt by stating a claim, providing arguments in support of their claim, as well as counterarguments, before reaching a conclusion. The instruction, and therefore the rubrics, emphasized students' ability to understand source material (content), to write a coherent essay (coherence), and to construct an argument (argumentation).

The assignments asked students to summarize the source material before writing the argument. To support a fine-grained analysis of the students' essays and provide data for evaluating NLP techniques, the students' essays are manually annotated for content and argument. The following sections present the assignments and rubrics, the essay data set, the reliability study, and the annotation methods for content and argumentation. We present initial findings on the comparison of grades assigned in the class to those assigned by reliable raters, and on the relation of the annotation to the reliable grades. We discuss questions that can be investigated about student learning, and about the interdependence of students' ability to articulate content and construct an argument.

## 2 Related Work

Previous work has looked at automated methods to support rubric-based writing (Passonneau

*Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 507–518
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

et al., 2018). Rubric-based writing assessment has recently been brought to researchers' attention, particularly on designing automated assessment methods. Gerard et al. (2016); Gerard and Linn (2016) have demonstrated that automated assessment using rubrics successfully identifies the students at risk of failing, and facilitates effective guidance and meaningful classroom interactions. Agejev and Šnajder (2017) uses ROUGE and BLEU in assessing summary writing from college L2 students. Santamaría Lancho et al. (2018) show that using G-Rubric, an LSA-based tool applying rubric assessment, helps the instructors grade the short text answers to open-ended questions, and proves to be reliable, with a Pearson correlation between human graders and G-Rubric of 0.72.

Recent work investigates fine-grained writing assessment, especially on content quality evaluation by diving into linguistic phenomena and structures, combined with various NLP techniques. Klebanov et al. (2014) investigated the correlations between essays scores with a content importance model. Another line of research has studied the role of argumentative features in predicting the overall essay quality (Ong et al., 2014; Song et al., 2014; Klebanov et al., 2016; Ghosh et al., 2016; Persing and Ng, 2015). For example, Klebanov et al. (2016) and Ghosh et al. (2016) examine the relations between argumentation structure features and the holistic essay quality (low, medium and high) applied to TOEFL essays. In this paper, we use the argumentative features introduced by Ghosh et al. (2016), but correlate them with the rubric related to the quality of the argument on a scale of 0-5.

## 3 Assignments and Rubrics

Two argument essays were assigned in fall 2018 to computer science freshman in a university in the United Kingdom. In the first of these (Essay 1, assigned early in the semester), students were asked to choose one of three articles on a current technology topic, with the number of students per article capped at one third of the class comprising 141 students. These students are enrolled in a variety of degree programs, ranging from information technology to computer engineering, offered by a department of mathematics and computer science. They form a heterogeneous group, both in educational background and age, since many are admitted through the university's mission to provide learning opportunities for the whole community.

The first part of the assignments required that students **summarize** the readings in one hundred and fifty to two hundred and fifty words. The second part elicited a three to five hundred-word **argument essay** addressing a given question. The list of topics and associated questions is shown below. Students were allowed to use external sources to back up their arguments, but had to reference these.

1. Autonomous Vehicles: will these change how we travel today?
2. Cybercrime: will education and investment provide the solution?
3. Cryptocurrencies: are they the currencies of the future?

For the second assignment (Essay 2), all students were provided with the same three journal articles relating to uses of AI in education. They were asked to **summarize**, in one hundred and fifty to two hundred and fifty words, key issues relating to the use of AI in teaching and learning as stated in the articles. Then, they had to write a three to five hundred word **argument essay** addressing the question: *Should artificial intelligence be used in teaching and learning?* Both essays were assessed using a rubric designed by the three collaborators, based on existing rubrics: SRI's Source-Based Argument Scoring Attributes (AWC) (Gallagher et al., 2015) and Ferretti's well known argument rubric (Ferretti et al., 2000). The four dimensions and their weights (in parentheses) are shown below. Each dimension or subdimension was rated on a 6-point scale ([0 to 5]; see Appendix A which gives the rubric for Essay 1.)

1. Content (3/7) - quality, coverage, coherence;
2. Argument (2/7) - claims, support, counterargument;
3. Conventions (1/7) - lexis and grammar;
4. Referencing (1/7) - sources and citations.

For Essay 2, some of the details of the Content-quality and the Referencing dimensions of the rubric were revised in recognition of the fact that in the second essay, students were not allowed to use external sources.

Students received three hours of preparatory instruction prior to the essays being assigned. The first two hours focused on how to write argument essays - engaging with the prompt, formulating a claim, developing arguments and counterarguments, and concluding the essay - as prescribed by Simon Black (2018) in his text, *Crack the Essay: Secrets of Argumentative Writing Revealed.*

The third hour of instruction, given later in the course, provided students with feedback, exemplified by student submissions for the first essay. During this lecture, they were shown good and poor examples of essay writing through an application of the rubric to several anonymized examples. These were later made available for their reference. Using the rubric to provide formative feedback may have resulted in better performance by students on Essay 2. Many of the students performed better on the second assignment.

Of the 141 enrolled students, 123 completed Essay 1, 101 completed Essay 2, and 97 completed both. Essay 2 was due 4 weeks following the submission of Essay 1, which made it possible for students to receive feedback on Essay 1 before submitting Essay 2. The framework used in designing the instruction is the cyclical process suggested by Jonassen (2008). Grading of Essay 1 was done by three of the five tutors teaching the course; each tutor graded all the essays for one of the three topics. The two remaining tutors split the Essay 2 submissions between them. To ensure consistency between the graders, the instructor moderated the grading by randomly selecting one-tenth of each set to regrade.

Using rubrics in higher education is well documented (Reddy and Andrade, 2010). Although mainly used for defining and grading assignments, rubrics can also be incorporated into instruction. Here, the feedback provided following the scoring of Essay 1 using the rubric was part of a developmental process, which culminated in Essay 2. For many of these students, it was their first attempt at writing an argument essay. A large proportion of them reported that the rubric helped them to understand the assignment better, and that they used it as a guide. In a survey examining how students used the rubric for Essay 1, 84 students responded, and 34% believed they achieved a good mark because they used the rubric. Only 11.3% felt the rubric made them lose marks.

Over 65% of students who submitted both essays received the same or a better grade on the second essay. Most who received the same grade on both essays ranked in the 95th percentile on both. The students had very positive things to say about what they learned from the assignment. These included: how to read critically; recognizing and questioning an author's argument; how to structure and write an argument; how to support a claim with evidence; and how to analyze complex issues. All of these competencies underpin critical thinking and problem solving, which are the fundamental skills taught to STEM majors.

## 4 Essay Dataset

The composition of the dataset supports simultaneous investigation of summary content analysis and argumentation mining: the former reflects the skills of reading comprehension and summarization, and the latter includes logical reasoning, argumentation, and writing skills. While summary and argument serve distinct roles, the combination into a single writing assignment allows us to assess the interdependence between reading comprehension and argument writing.

Below, we present descriptive statistics of the dataset. Table 1 shows the sample sizes for essays on the given topics Cybercrime (CyberCri) with 44, Autonomous Vehicles (AutoV) with 42, and CryptoCurrencies (CrypCurr) with 37. In the second assignment, there are 101 essays about AI. Table 2 shows that the second assignment had a higher average of tokens per sentence across summary, argument and overall. The vocabulary size of the whole dataset is 5,923.

| Assignment1 | | | Assignment2 |
|---|---|---|---|
| CyberCri | AutoV | CrypCurr | AI |
| 44 | 42 | 37 | 101 |

Table 1: Sample size given each assignment and topic; the total number of essays is 224.

In contrast to other data sets investigated for argument mining, here the assignments are from a single course with the same set of students. The size of our data set is comparable to one used in (Ghosh et al., 2016) (TOEFL essays), but smaller than those used in (Stab and Gurevych, 2014; Klebanov et al., 2016). In addition, the data set we collected has multiple essays for four topics, based on source readings. This gives us the opportunity to investigate students' reliance on source material in their argumentation.

| Sum | CyberCri | AutoV | CrypCurr | AI |
|---|---|---|---|---|
| Sents | 7.43 | 7.24 | 8.62 | 7.30 |
| Tk/Sents | 34.21 | 32.84 | 28.52 | 36.08 |
| Arg | CyberCri | AutoV | CrypCurr | AI |
| Sents | 17.36 | 19.66 | 19.51 | 19.05 |
| Tk/Sents | 31.96 | 32.32 | 33.20 | 34.79 |
| Overall | CyberCri | AutoV | CrypCurr | AI |
| Sents | 24.78 | 27.04 | 28.14 | 26.34 |
| Tk/Sent | 32.46 | 32.90 | 32.04 | 36.17 |
| Vocabulary Size | | | 5923 | |

Table 2: Dataset statistics of average numbers of sentences (Sents) and average tokens per sentence (Tk/Sents) from summaries (Sum) and arguments (Arg) across topics. The total vocabulary size is also given.

## 5 Rubric Reliability

Educational intervention studies where researchers investigate the potential benefit of a proposed intervention apply rubrics whose reliability has been assessed. For example, in their meta-analysis of educational interventions, Graham and Perin (2007) exclude interventions whose reliability is below 0.60. We also test the reliability of the rubric used in the classroom assignments discussed here. The reliability study we present has two purposes. First, it provides insight into the difficulty of graders' use of a multi-dimensional content and argument rubric under ordinary classroom conditions where there is time pressure to assign grades. Second, it provides a measure of the quality of the gold standard against which to evaluate the automated NLP techniques we will develop. Our reliability study addressed the content and argument dimensions of the rubric, and achieved high inter-rater reliability.

Two advanced undergraduates were recruited for the reliability study. They were trained by a team consisting of the collaborators and their research assistants during a period of seven weeks, with each rater devoting 10 hours per week. Each rater then graded half the essays (apart from six used in training).

The raters' training consisted of activities in which they learned about the structure of argument writing and carried out the assignment, used the rubric to assess different topics and writers, and participated in webinars where they received feedback and further training. Figure 1 shows the seven-week training regimen. During weeks 1-2, they became familiar with all three writing prompts through their own essay writing, and grading the other rater's essays. There were two rounds in which they independently assessed three Cryptocurrency essays (weeks 4, 6; six distinct essays), with webinar feedback in between (weeks 5, 7). The week 5 webinar involved all three researchers, the two raters, and a PhD student on the project. All other webinars with the raters were presented by the instructor co-author. A final webinar (week 7) pointed to minor differences between the two assignments and rubrics pertaining to use of open-ended external sources in Essay 1 and not Essay 2.

For the assessment tasks in weeks 4 and 6, three Cryptocurrency essays representing high, medium and low students' scores were selected. The raters did not know there was a difference in the students' original grades, and no one on the project other than the instructor knew how the three were originally graded. As a result of the discussion from the week 5 webinar, a consensus was reached on the three initial Cryptocurrency essays. The raters were instructed to use these as a model for applying the rubric in a consistent manner.

Rater agreement was assessed using Pearson correlation on the content and argument components of the rubric. Content quality, content coverage, and content coherence were each independently rated on a six-point scale (0 to 5). Argument was rated on an eleven-point scale (0 to 10). After the raters applied the rubric to the first three Cryptocurrency essays, their correlations with each other and with the assigned grades varied widely, from negative correlation to high correlation. After the second round of three essays, the correlation between the two raters was perfect on two, and poor on the third. After a brief discussion, we decided that this was sufficient agreement for the raters to work indepen-

| Week | Activity |
|---|---|
| 1 | Webinar to review argument writing, assignment #1, and rubric #1 |
| 2 | Each rater writes one essay on AV, and one on Crypto or Cyber; raters apply rubric to the other rater's two essays |
| 3 | Webinar on their essays and assessments |
| 4 | Raters each assess the same three Crypto |
| 5 | Webinar on their first round of assessment with detailed discussion among raters and all researchers |
| 6 | Raters each assess three additional Crypto |
| 7 | Feedback on the second round of assessment; webinar on assignment and rubric #2 |

Figure 1: Seven-week rater training

| Essay | CQual | CCov | CCoher | Arg | Conven | Ref | Cont/Arg | Total |
|---|---|---|---|---|---|---|---|---|
| AutoV | 0.38 | 0.52 | 0.29 | 0.32 | 0.19 | 0.56 | 0.63 | 0.52 |
| Crypto | 0 | 0.12 | 0 | 0.48 | 0.47 | 0.56 | 0.72 | 0.36 |
| Cyber | 0.36 | 0.33 | 0.12 | 0.44 | 0.30 | 0.84 | 0.59 | 0.50 |
| All Essay 1 | 0.23 | 0.40 | 0.14 | 0.41 | 0.31 | 0.69 | 0.62 | 0.47 |
| Essay 2 (AI and Ed) | 0.32 | 0.44 | 0.42 | 0.36 | 0.44 | 0.61 | 0.54 | 0.47 |
| All | 0.25 | 0.38 | 0.25 | 0.34 | 0.38 | 0.69 | 0.55 | 0.42 |

Table 3: Correlations of the reliable grades with the tutors' grades

dently to apply the rubric to the remaining essays.

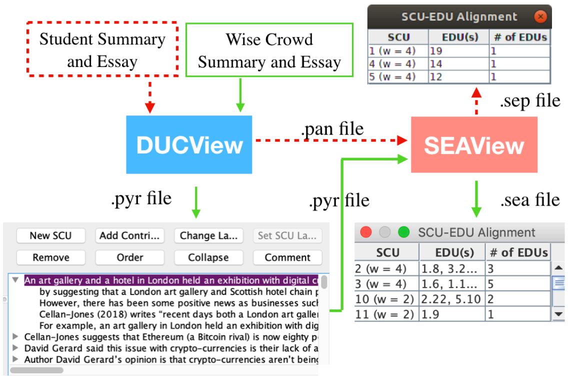
To complete the gold standard rubric scores, each rater worked on 28 essays per week for three weeks, and 31 in the fourth week. A random selection of 10 essays were assessed by both raters. The correlation for the content and argument dimensions on the ten essays ranged from one low outlier of -0.52 to 1.0. The average was 0.75, and on all but the outlier it was 0.89.

The reliability study shows that the rubric can be applied reliably, but also highlights the difficulty of incorporating a fine-grained rubric into classroom use, where tutors have little time to engage in training. The assessments from the reliable raters generally have moderate correlation with the tutors' grades, ranging from 0.72 for the Cryptocurrency essays to 0.54 for the AI Education essays for the complete rubric. Table 3 gives the correlations between the raters and the tutors who did the grading on each component of the rubric, the sum of the four content and argument dimensions, and the total. In addition to providing the correlations for all the essays as a whole, the table also gives the breakdown for each assignment, and for the three topics in assignment 1.

## 6 Annotation of Essay Content

Here we describe the annotation of the content of the summary portion and argument portion of students' essays. This comprises three annotation tasks: identification of summary content units (SCUs) in the summary; identification of elementary discourse units (EDUs) in the argument; and alignment of EDUs with SCUs.

To annotate the summary content, we use Pyramid annotation (Nenkova et al., 2007), a summary content annotation that has been shown to correlate with a main ideas rubric used in an educational intervention with community college students (Passonneau et al., 2018). As in that study, we collect five reference summaries written by more advanced students, referred to as a wise crowd. The wise crowd summaries are first annotated with DUCView[1] to create a list of summary content units (SCUs) (see Figure 2), where each SCU appears in at least one wise crowd summary and at most in all five. An SCU is roughly a proposition, but need not be expressed as a full clause. SCUs are ranked by their frequency in the wise crowd summaries to provide an importance measure of the SCU. Content scores given to student summaries are based on matches from the student summary to SCUs in the pyramid. Pyramid scores measure the inherent quality of a student's summary (relative proportion of high-weighted SCUs), and the content coverage (proportional representation of average SCU weights in wise crowd summaries). Pyramid annotation has been found to be highly reliable (Passonneau, 2010). Agreement measured by Krippendorff's alpha (scores in [-1,1]) on ten pairs of pyramids created by different annotators, for five topics from each of two distinct datasets, ranged from 0.75 to 0.89. For sixteen peer summarizers on three topics each, average alpha for annotation of pyramid SCUs in summaries was 0.78. Due to the exten-



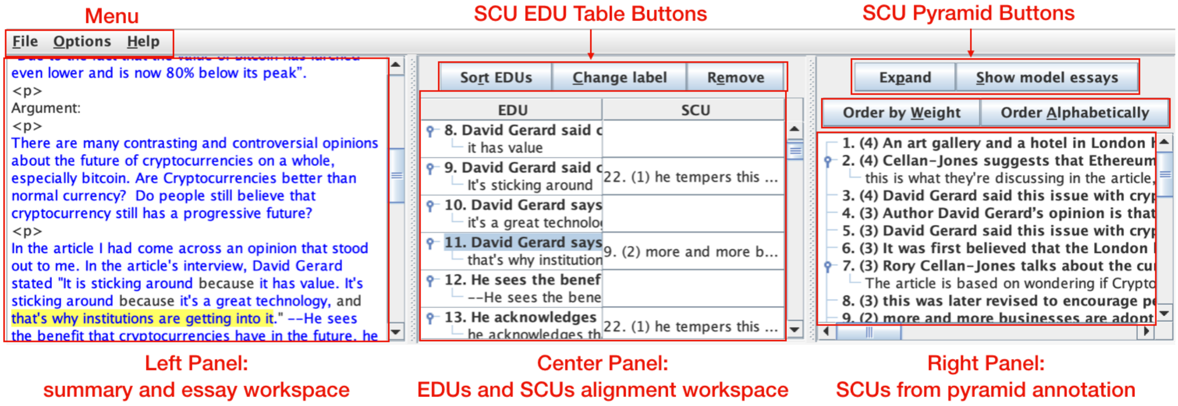
| SCU | EDU(s) | # of EDUs |
|---|---|---|
| 1 (w = 4) | 19 | 1 |
| 4 (w = 4) | 14 | 1 |
| 5 (w = 4) | 12 | 1 |

An art gallery and a hotel in London held an exhibition with digital c
by suggesting that a London art gallery and Scottish hotel chain p
However, there has been some positive news as businesses such
Cellan-Jones (2018) writes "recent days both a London art galler
For example, an art gallery in London held an exhibition with dig
Cellan-Jones suggests that Ethereum (a Bitcoin rival) is now eighty p
David Gerard said this issue with crypto-currencies is their lack of a
Author David Gerard's opinion is that crypto-currencies aren't being

| SCU | EDU(s) | # of EDUs |
|---|---|---|
| 2 (w = 4) | 1.8, 3.2... | 3 |
| 3 (w = 4) | 1.6, 1.1... | 5 |
| 10 (w = 2) | 2.22, 5.10 | 2 |
| 11 (w = 2) | 1.9 | 1 |

Figure 2: Workflow diagram for content annotation: from DUCView to SEAView. The green box and arrows indicate the flow of the wise crowd summaries and essays, and the box and arrows in dashed red lines are show the flow of a student summary and essay.

[1] We made some modifications to the original DUCView; the new version is available from https://github.com/psunlpgroup/DUCView.



even lower and is now 80% below its peak".
<p>
Argument:
<p>
There are many contrasting and controversial opinions about the future of cryptocurrencies on a whole, especially bitcoin. Are Cryptocurrencies better than normal currency? Do people still believe that cryptocurrency still has a progressive future?
<p>
In the article I had come across an opinion that stood out to me. In the article's interview, David Gerard stated "It is sticking around because it has value. It's sticking around because it's a great technology, and that's why institutions are getting into it." --He sees the benefit that cryptocurrencies have in the future, he

| EDU | SCU |
|---|---|
| 8. David Gerard said c<br>it has value | |
| 9. David Gerard said c<br>It's sticking around | 22. (1) he tempers this ... |
| 10. David Gerard says<br>it's a great technolo | |
| 11. David Gerard says<br>that's why institution | 9. (2) more and more b... |
| 12. He sees the benef<br>--He sees the bene | |
| 13. He acknowledges<br>he acknowledges th | 22. (1) he tempers this ... |

1. (4) An art gallery and a hotel in London h
2. (4) Cellan-Jones suggests that Ethereum
this is what they're discussing in the article,
3. (4) David Gerard said this issue with cryp
4. (3) Author David Gerard's opinion is that
5. (3) David Gerard said this issue with cryp
6. (3) It was first believed that the London
7. (3) Rory Cellan-Jones talks about the cu
The article is based on wondering if Crypto
8. (3) this was later revised to encourage pe
9. (2) more and more businesses are adopt

Figure 3: SEAView interface. Left panel as the workspace of summary and essay: users could select the span of text, drag and drop to the center panel as an EDU then change the label; center panel for EDUs and SCUs alignments; right panel for displaying list of SCUs from manual pyramid. The text highlighted by yellow color in left panel is the current selected EDU annotated and labeled as blue highlighted text in EDU and SCU alignment table, as shown in center panel, with EDU ID as 11 and a match of SCU weighted 2 and ID as 9, corresponding to the pyramid in right panel.

sive reliability measures in past work on pyramid annotation, we did not re-assess its reliability here.

To annotate the content of the argument portion of the essay, we identify all distinct Elementary Discourse Units (EDU). Identifying (segmenting) EDUs from text and representing their meanings play a key role in discourse parsing (Marcu, 2000; Li et al., 2016; Braud et al., 2017). Definitions of EDUs vary, thus Prasad et al. (2008) consider the full range of clause types, including verb arguments and non-finite clauses. To simplify the annotation, we restrict EDUs to propositions derived from tensed clauses that are not verb arguments (such as *that* complements of verbs of belief). In (Gao et al., 2019), we report the iterative development of reliable annotation guidelines for untrained annotators.[2] Annotators first identify the start and end of tensed clauses, omitting discourse connectives from the EDU spans, which can be discontinuous. Annotators then provide a paraphrase of the EDU span as an independent simple sentence. The EDU annotation is illustrated in a subsection below along with the annotation tool developed for this purpose.

### 6.1 Content annotation workflow

To follow the principles of pyramid annotation applied to education (Passonneau et al., 2018), we collected wise crowd essays written by sophomores who took the academic skills course in the previous year and by the trained raters on the project (advanced undergraduates), to constitute five references per topic. We used the guidelines from DUC 2006 (Passonneau, 2006), and an enhanced annotation tool (see above). As shown in Figure 2, the annotation workflow begins with pyramid content annotation, which takes wise crowd summaries as input to DUCView. The annotator creates SCUs and exports the pyramid XML file (*.pyr). A pyramid file and a student summary are then the input for the annotator to match phrases in the student summary to pyramid SCUs, which is also exported as XML (*.pan).

### 6.2 SEAView

We designed a tool to annotate EDUs in the wise crowd essays and the student essays, and to align EDUs with SCUs.[3] As shown in Figure 3, SEAView (for **S**CU and **E**DU **A**lignment) takes as input two-part essays that contain a summary and an argument, where the summary has already been annotated in DUCView. To annotate the wise crowd essays, a .pyr file is loaded into SEAView. The input files must contain document separator lines between the essays, and another header line between the summary and argument of each essay. The annotator identifies EDUs in each of the wise crowd essays. To annotate a student essay, a .pan file is loaded into SEAView. Annotators perform the annotation in two steps: identification of all the EDUs in the argument text; alignment of EDUs with any SCUs that share the same meaning. The final output from SEAView includes a list

[2] Currently in submission to another venue.

[3] Available for download: https://github.com/psunlpgroup/SEAView.

of EDUs, a list of SCUs matched with the EDUs, and an alignment table. Depending on the type of input, SEAView will generate SCU-EDU alignment for wise crowd essays (.sea files), or SCU-EDU alignments for student essays (.sep files).

## 7 Initial Content Annotation Results

We first present preliminary content annotation results on topic Cryptocurrencies and Autonomous Vehicle. Two manual pyramids are annotated, with statistics shown in Table 4. The total number of SCUs are 34 and 41 for Cryptocurrencies and Autonomous Vehicle respectively. Neither topic has found SCUs weighted 5 (number of wise crowd). Both found 8 SCUs that are weighted 4 and 3, and a long tail distribution of low-weighted SCUs (26 for Cryptocurrencies; 33 for Autonomous Vehicles).

Table 5 presents statistics of content annotation of essays, from both wise crowd submissions and students submissions, on EDU-SCU alignment between manual pyramid and essays. In wise crowds, the average weight of SCUs matched in essays is 2.60 (Cryptocurrencies) and 2.37 (Autonomous Vehicle). Autonomous Vehicle has more EDUs on average (N=35.00) than Cryptocurrencies (N=23.80), while Cryptocurrencies has longer length of EDU than Autonomous Vehicle, respectively 17.39 and 15.18 words. Finally, the SCU weights normalized by the total number of EDUs are 0.11 and 0.7, and by the number of matched EDUs are 1.07 and 0.62, for Cryptocurrencies and Autonomous Vehicles, respectively. For student submissions, the Autonomous Vehicles set has slightly higher numbers except for total EDUs, which is 36.70 for Autonomous Vehicle and 36.76 for Cryptocurrencies. Autonomous Vehicle has 2.75 as average weight of SCUs and Cryptocurrencies has 2.07. Cryptocurrencies has shorter length of EDUs compared to Autonomous Vehicle, as 13.62 and 14.00. For the normalized SCU by total number of EDUs and number of matched EDUs, Autonomous Vehicle shows more with 0.08 and 0.84, while Cryptocurrencies has 0.06 and 0.70. This indicates that more important content is mentioned in Autonomous Vehicle submissions than Cryptocurrencies.

| Topic | Total SCUs | w=5 | w=4 | w=3 | w $\leq$ 2 |
|---|---|---|---|---|---|
| CrypCurr | 34 | 0 | 3 | 5 | 26 |
| AutoV | 41 | 0 | 6 | 2 | 33 |

Table 4: Distributions of SCUs with weights from manual pyramids annotation of Cryptocurrencies and Autonomous Vehicle

Table 5 also lists the average (reliable) total grade, and the breakdown for content quality and content coverage. Students' grades on Autonomous Vehicle and Cryptocurrencies are similar in all three aspects, as 23.48, 3.68 and 3.83 for Autonomous Vehicle, and 23.16, 3.81 and 3.39 in Cryptocurrencies.

## 8 Annotation of Argument Structure

To annotate the argumentative part of the essays, we used the coarse-grained argumentative structure proposed by Stab and Gurevych (2014): argument components (major claim, claim, premises) and argument relations (support/attack). Similar to Hidey et al. (2017), we took as annotation unit the proposition instead of the clause, given that premises are frequently propositions that conflate multiple clauses. For this pilot annotation task we labeled the 37 Cryptocurrency essays and used two expert annotators with background in linguistics and/or argumentation. We used Brat as annotation tool.[4] The set contains 36 main claims, 559 claims, 277 premises, 560 support relations and 101 attack relations.

Ghosh et al. (2016) proposed a set of argumentative features and showed that they correlate well with the holistic essay scores (low, medium and high) when applied to TOEFL persuasive essays: 1) features related to *argument components* (AC) such as the number of claims, number of premises, fraction of sentences containing argument components; 2) features related to *argument relations* (AR), such as the number and percentage of supported claims, and the number and percentage of dangling claims (i.e., claims with no supporting premises), the number of attack relations and attacks against the major claim; and 3) features related to the *typology of argument structures* (TS) such as the number of argument chains, number of argument trees. In this study, we wanted to see whether these proposed features correlate well with the 6 scale rubric that rate the "quality" of the argument. The scored used were the one obtained in our reliability study. Table 6 summarizes the features (for details see (Ghosh et al., 2016)).

Given the manual annotation of the essays, to measure the effectiveness of the argumenta-

[4] https://brat.nlplab.org.

| Stat | CrypCurr$_{Wise}$ | AutoV$_{Wise}$ | CrypCurr$_{Peer}$ | AutoV$_{Peer}$ |
|---|---|---|---|---|
| Weight$_{SCU}$ | 2.60 | 2.37 | 2.07 | 2.75 |
| Total EDUs | 23.80 | 35.00 | 36.76 | 36.70 |
| Tokens per EDU | 17.39 | 15.18 | 13.62 | 14.00 |
| NormSCUEDU$_{Total}$ | 0.11 | 0.07 | 0.06 | 0.08 |
| NormSCUEDU$_{Matched}$ | 1.07 | 0.62 | 0.70 | 0.84 |
| Final Scores$_{Rubric}$ | - | - | 23.16 | 23.48 |
| Content Quality$_{Rubric}$ | - | - | 3.81 | 3.68 |
| Content Coverage$_{Rubric}$ | - | - | 3.39 | 3.83 |

Table 5: Statistics of annotated wise crowd summaries and essays form Cryptocurrencies (CrypCurr$_{Wise}$) and Autonomous Vehicle (AutoV$_{Wise}$), and student submissions (CrypCurr$_{Peer}$ and AutoV$_{Peer}$): average matched SCU weights (Weight$_{SCU}$), average numbers of EDUs (Total EDUs), average tokens per EDU (Tokens per EDU), weighted SCU normalized by total number of EDUs (NormSCUEDU$_{Total}$), weighted SCU normalized by the number of matched EDUs (NormSCUEDU$_{Matched}$). We also provide the scores from rubrics here (bottom of the table): Final scores across 6 categories (Final Scores$_{Rubric}$), content quality (Content Quality$_{Rubric}$) and content coverage (Content Coverage$_{Rubric}$).

| Feature Group | Argumentation Feature Description |
|---|---|
| AC | # of Claims<br># of Premises<br># and fraction of sentences containing argument components |
| AR | # and % of supported Claims<br># and % of dangling Claims<br># of Claims supporting Major Claim<br># of total Attacks and Attacks against Major Claim |
| TS | # of Argument Chains<br># of Argument Trees (hight=1 or >1) |

Table 6: Argumentation Features

| Features | Correlations |
|---|---|
| bl | 0.15 |
| AC | 0.27 |
| AR | 0.35 |
| TS | 0.17 |
| bl + AC | 0.21 |
| bl + AR | 0.26 |
| bl + TS | 0.33 |
| AC + AR + TS | 0.41 |
| bl + AC + AR + TS | 0.26 |

Table 7: Correlation of LR (5 fold CV) with argument quality scores.

tive features in predicting the quality of argument scores (0-5) we use Logistic Regression (LR) learners and evaluate the learners using quadratic-weighted kappa (QWK) against the human scores, a methodology generally used for essay scoring (Farra et al., 2015; Ghosh et al., 2016). QWK corrects for chance agreement between the system prediction and the human prediction, and it takes into account the extent of the disagreement between labels. Since the number of essays is very small we did a five-fold cross validation. Table 7 reports the performance for the three feature groups as well as their combination. The baseline feature (bl) is the number of sentences in the essay, since essay length has been shown to be generally highly correlated with essay scores (Chodorow and Burstein, 2004).

As seen in Table 7 out of the individual features groups the higher correlation is obtained by the Argument Relation group. The best correlation is obtained when using all the argumentative features (AC+Ar+TS). Unlike Ghosh et al. (2016), we found that adding the baseline feature to the argument features did not help, except when combining with the typology of argument structure features. We also looked at what features are associated with different rubric scores based on the the regression coefficients. As expected, the tree structure features (TS) correlated with high score essays (4 and 5). In addition, high scoring essays (5) have a higher number of ”attack” relations to the main claim, showing that the essays contain counterarguments (presenting both sides of the issue). Number of supported claims correlated negatively with lower scoring essays (meaning that the low scoring essays has more unsupported claims). Moreover, number of claims supporting the main claim was negatively correlated with low scoring essays. In those essays, the students, although advancing arguments, they failed to connect them to their main claim. Looking at the different between high scoring essays (4 vs. 5) we noticed an interesting aspect: for the essays scored with 5 the ratio of argumentative sentences w.r.t total number of sentence was higher than for the essays with a 4 score, while the essays with a 4 scores tend to be longer. In general our correlations scores were much lower than the ones reported by Ghosh et al. (2016). There are several explanations for that. First, the number of essays is smaller (37 compared to 107) and we have a 6-point scale rather than a 3 point scale. In addition, our scale reflected the argument quality and not a holistic essays score; looking just at argumentative discourse structure might not be enough, we need to look both at structure and the semantics of arguments (content) to more reliably distinguish essays based on their argument quality (Klebanov et al., 2016). Our annotation of content and argument will allow us to pursue this line of inquiry in our future work.

## 9 Conclusion

We have presented the collection of a rich data set of essays written by college freshman in an academic skills class. We conducted a reliability assessment of the rubric used to explain the assignment expectations. The moderate correlation of the raters' scores with the grades assigned by tutors, combined with the lengthy investment in time to train the raters, shows that high reliability can be achieved at a cost that cannot be sustained in ordinary classrooms. One of the questions our future work will address is the degree to which rubrics could be partly automated using NLP techniques. Partial automation could free instructors from the demands of managing a team of graders, and potentially lead to greater consistency in student feedback.

Our future work will investigate the interdependence of the content and argument annotations presented here, and the ramifications for student learning. Two teams of annotators working completely independently performed the content annotation (SCUs, EDUs) and the argument annotation. We will investigate the correspondence between EDUs and argument components, both of which are simple propositions. Depending on how well they correspond, it is possible that providing EDU annotation as input to argument annotation could improve the argument annotation reliability. Ultimately we aim to help instructors provide students with better feedback on their ability to summarize the main ideas of source material, the role that these ideas play in their arguments, and the overall quality of their essays.

## Acknowledgments

This work is supported by NSF IIS 1847842. We thank two undergraduate REUs, Connor Heaton and Annie Qin Sui, who participated in the reliability studies and other project activities. Finally, we thank the reviewers for the valuable feedback.

## References

Tamara Sladoljev Agejev and Jan Šnajder. 2017. Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in l2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 181–186.

Simon Black. 2018. *Crack the Essay: Secrets of Argumentative Writing Revealed*. Gramercy House Publishing.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 292–304.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: Evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74, Denver, Colorado. Association for Computational Linguistics.

Ralph P. Ferretti, Charles A. MacArthur, and Nancy S. Dowdy. 2000. The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, 92(4):694–702.

H. A. Gallagher, K. R. Woodworth, and N. L. Arshan. 2015. Impact of the national writing projects college-ready writers program on teachers and students.

Yanjun Gao, Kenneth Huang, and Rebecca J. Passonneau. 2019. AESOP: Annotated elementary discourse units from student opinion essays. In submission.

Libby Gerard, Marcia C Linn, and Jacquie Madhok. 2016. Examining the impacts of annotation and automated guidance on essay revision and science learning. Singapore: International Society of the Learning Sciences.

Libby F Gerard and Marcia C Linn. 2016. Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education*, 27(1):111–129.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 549–554.

Amy Gillespie, Steve Graham, Sharlene Kiuhara, and Michael Hebert. 2014. High school teachers use of writing to support students learning: a national survey. *Reading and Writing*, 27(6):1043–1072.

Steve Graham, Andrea Capizzi, KarenR Harris, Michael Hebert, and Paul Morphy. 2014. Teaching writing to middle school students: a national survey. *Reading and Writing*, 27(6):1015–1042.

Steve Graham, Jill Fitzgerald, Linda D. Friedrich, Katie Greene, James S. Kim, and Carol Booth Olson. 2016. Teaching secondary students to write effectively. Technical Report NCEE 2017-4002, National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Steve Graham and Dolores Perin. 2007. A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99:445476.

Murat Gunel, Brian Hand, and Vaughan Prain. 2007. Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education*, 5(4):615–637.

Brian Hand, Lori A. Norton-Meier, Murat Gunel, and Recai Akkus. 2015. Aligning teaching to learning: A three-year study examining the embedding of language and argumentation into elementary science classrooms. *International Journal of Science and Mathematics Education*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*.

D. H. Jonassen. 2008. Instructional design as design problem solving: An iterative process. *Educational Technology*, 48(3):21–26.

Anders Jonsson and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2):130 – 144.

Sharlene A. Kiuhara, Steve Graham, and Leanne S. Hawken. 2009. Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, 101(1):136–160.

Beata Beigman Klebanov, Nitin Madnani, Jill Burstein, and Swapna Somasundaran. 2014. Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 247–252.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75.

Perry D. Klein and Mary A. Rose. 2010. Teaching argument and explanation to prepare junior students for writing to learn. *Reading Research Quarterly*, 45(4):433–461.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2).

Stephen P. Norris and Linda M. Phillips. 2003. How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87(2):224–240.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. *ACL 2014*, page 24.

Rebecca J Passonneau. 2006. Pyramid annotation guide: Duc 2006.

Rebecca j. Passonneau. 2010. Formal and functional assessment of the pyramid method for summary content evaluation*. *Nat. Lang. Eng.*, 16(2):107–131.

Rebecca J. Passonneau, Ananya Poddar, Gaurav Gite, Alisa Krivokapic, Qian Yang, and Dolores Perin. 2018. Wise crowd content assessment and educational rubrics. *International Journal of Artificial Intelligence in Education*, 28(1):29–55.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse Treebank 2.0. In *LREC*.

Y. Maline Reddy and Heidi Andrade. 2010. A review of rubric use in higher education. *Assessment Evaluation in Higher Education*, 35(4):435–448.

Victor Sampson, Patrick Enderle, Jonathon Grooms, and Shelbie Witte. 2013. Writing to learn by learning to write during the school science laboratory: Helping middle and high school students develop argumentative writing skills as they learn core ideas. *Science Education*, 97(5):643–670.

Miguel Santamaría Lancho, Mauro Hernández, Ángeles Sánchez-Elvira Paniagua, José María Luzón Encabo, and Guillermo de Jorge-Botana. 2018. Using semantic technologies for formative assessment and scoring in large courses and MOOCs. *Journal of Interactive Media in Education*, 2018(1).

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510.

# A Essay 1 Rubric

| Points | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **CONTENT: quality** | Most of the ideas in the summary and argument are either not central to the topic, not expressed clearly, or are vague or repetitive. | Many of the ideas in the summary and argument relate to the topic, but only a few them are central to the topic, which may be due to vagueness, repetition, lack of clarity, or failure to express central ideas. | About half the ideas in the summary and argument are expressed clearly and are central to the topic, but about half the ideas do not meet the combined criteria of clarity and centrality, which may be due to vagueness, repetition, lack of clarity, or failure to identify central ideas. | About half the ideas in the summary and argument are expressed clearly and are central to the topic, all ideas are related to the topic, and there is little to no vagueness or repetition. However, about half the ideas are either not central to the topic or unclear. | Most of the ideas in the summary and argument are expressed clearly and are central to the topic, and there is little to no vagueness or repetition. However, some ideas are either unclear, or not central to the topic, or some combination. | All or nearly all the ideas in the summary and argument are related to the topic, most of them are central to the topic, and all or nearly all are expressed clearly, with little or no vagueness or repetition. |
| **CONTENT: coverage** | Most of the central ideas from the article(s) are not expressed clearly in the summary and argument, and ideas from the article(s) that are included are expressed in a way that is unclear, vague or repetitive. | Some of the central ideas from the article(s) are expressed clearly in the summary and argument, but many of the central ideas from the article(s) are missing, or are expressed in a way that is unclear, vague, or repetitive. Most ideas from the article(s) that are expressed clearly in the summary and argument are not central to the topic. | Many of the central ideas from the articles(s) are expressed clearly in the summary and argument, but many of the central ideas from the article(s) are missing, or are expressed in a way that is unclear, vague or repetitive. Many ideas from the article that are expressed clearly in the summary and argument are not central to the topic. | Most of the central ideas from the article(s) are expressed clearly in the summary and argument. The remaining ideas from the article that are expressed in the summary and argument are either not central, not clear, vague or repetitive. | Most of the central ideas from the article(s) are expressed clearly in the summary and argument. Nearly all ideas from the article(s) expressed in the summary and argument are related to the topic, and are expressed clearly, with little vagueness or repetition. | All or nearly all of the central idea(s) from the article are expressed clearly in the summary and argument. Very few of the ideas from the article(s) that are expressed in the summary and argument are not central to the topic, and very few are expressed in a way that is unclear, vague or repetitive. |
| **CONTENT: coherence** | The ideas expressed in the summary and argument are not easy to follow, and do not relate well to one another. | Some of the ideas expressed in the summary and argument relate well to one another, but most of the ideas do not relate well to one another, and are not easy to follow. | Many of the ideas expressed in the summary and argument relate well to one another, making it fairly easy to follow much of the discussion. But many of the ideas expressed in the summary and argument do not relate well to one another, so it is difficult to form a coherent understanding of the whole. | Most of the ideas expressed in the summary and argument relate well to one another, making it fairly easy to follow most of the discussion. Some of the ideas, however, do not relate well and as a result, part of the discussion is hard to follow. | Most of the ideas expressed in the summary and argument relate well to one another, and the discussion as a whole is fairly easy to follow. A few ideas seem out of place or less well integrated into the overall organization. | All or nearly all of the ideas expressed in the summary and argument relate well to one another, making it easy to follow the discussion as a whole. Ideas flow well from one to the next, and the overall organization is very coherent. |
| **ARGUMENT** | Essay responds to the topic in some way but does not state a claim on the issue. | Essay states a claim, but no reasons are given to support the claim, or the reasons given are unrelated to or inconsistent with the claim, or they are incoherent. | Essay states a clear claim and gives one or two reasons to support the claim, but the reasons are not explained or supported in any coherent way. The reasons may be of limited plausibility, and inconsistencies may be present. | Essay states a claim and gives reason(s) to support the claim, plus some explanation or elaboration of the reasons. The reasons are generally plausible though not enough information is provided to convince a reader. There may be some inconsistencies, irrelevant information, or problems with organization and clarity. | Essay states a clear claim and gives reasons to support the claim. The reasons are explained clearly and elaborated using information that is convincing. Organisation of the essay is generally good but it is missing a concluding statement, or there are inconsistencies or irrelevancies that weaken the argument. | Meets the criteria for previous level. In addition, The essay is generally well organized and includes a concluding statement. The writing is free of inconsistencies and irrelevancies that would weaken the argument. |
| **CONVENTIONS: lexis & grammar** | Unacceptable number of mistakes in spelling and/or grammar. Word choice is awkward, vague or unclear. | Significant number of mistakes in spelling and/or grammar. Word choice is awkward, vague or unclear. | Few mistakes in spelling and grammar. Word choice is awkward, vague or unclear. | No mistakes in spelling and grammar, but poor choice of words. | No mistakes in spelling and grammar, and good choice of words. | No mistakes in spelling and grammar, and excellent choice of words. |
| **REFERENCING: sources & citation** | Does not cite sources. | Less than two post-2015 sources with good reliability. Does not cite all data obtained from other sources. Citation style is either inconsistent or incorrect. | Two post-2015 very reliable sources. Cites all data obtained from other sources. Harvard citation style is used in both text and reference list. | Three post-2015 sources with good reliability. Cites all data obtained from other sources. Citation style is either inconsistent or incorrect. | Three post-2015 sources, with good reliability. Cites all data obtained from other sources. Harvard citation style is used in both text and reference list. | Three or more post-2015 very reliable sources. Cites all data obtained from other sources. Harvard citation style is used in both text and reference list. |