Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays

Winston Carlile Nishant Gurrapadi Zixuan Ke Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{winston,zixuan,vince}@hlt.utdallas.edu,Nishant.Gurrapadi@utdallas.edu

Abstract

While argument persuasiveness is one of the most important dimensions of argumentative essay quality, it is relatively little studied in automated essay scoring research. Progress on scoring argument persuasiveness is hindered in part by the scarcity of annotated corpora. We present the first corpus of essays that are simultaneously annotated with argument components, argument persuasiveness scores, and attributes of argument components that impact an argument's persuasiveness. This corpus could trigger the development of novel computational models concerning argument persuasiveness that provide useful feedback to students on *why* their arguments are (un)persuasive in addition to *how* persuasive they are.

1 Introduction

The vast majority of existing work on automated essay scoring has focused on *holistic* scoring, which summarizes the quality of an essay with a single score and thus provides very limited feedback to the writer (see Shermis and Burstein (2013) for the state of the art). While recent attempts address this problem by scoring a particular dimension of essay quality such as coherence (Miltsakaki and Kukich, 2004), technical errors, relevance to prompt (Higgins et al., 2004; Persing and Ng, 2014), organization (Persing et al., 2010), and thesis clarity (Persing and Ng, 2013), argument persuasiveness is largely ignored in existing automated essay scoring research despite being one of the most important dimensions of essay quality.

Nevertheless, scoring the persuasiveness of arguments in student essays is by no means easy. The difficulty stems in part from the scarcity of persuasiveness-annotated corpora of student essays. While persuasiveness-annotated corpora exist for other domains such as online debates (e.g., Habernal and Gurevych (2016a; 2016b)), to our knowledge only one corpus of persuasiveness-annotated student essays has been made publicly available so far (Persing and Ng, 2015).

Though a valuable resource, Persing and Ng's (2015) (P&N) corpus has several weaknesses that limit its impact on automated essay scoring research. First, P&N assign only *one* persuasiveness score to each essay that indicates the persuasiveness of the argument an essay makes for its *thesis*. However, multiple arguments are typically made in a persuasive essay. Specifically, the arguments of an essay are typically structured as an argument tree, where the major claim, which is situated at the root of the tree, is supported by one or more claims (the children of the root node), each of which is in turn supported by one or more premises. Hence, each node and its children constitute an argument. In P&N's dataset, only the persuasiveness of the overall argument (i.e., the argument represented at the root and its children) of each essay is scored. Hence, any system trained on their dataset cannot provide any feedback to students on the persuasiveness of any arguments other than the overall argument. Second, P&N's corpus does not contain annotations that explain *why* the overall argument is not persuasive if its score is low. This is undesirable from a feedback perspective, as a student will not understand why her argument is not persuasive if its score is low.

Our goal in this paper is to annotate and make publicly available a corpus of persuasive student essays that addresses the aforementioned weaknesses via designing appropriate annotation schemes and scoring rubrics. Specifically, not only do we score the persuasiveness of *each* ar-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 621–631
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

gument in each essay (rather than simply the persuasiveness of the overall argument), but we also identify a set of attributes that can explain an argument's persuasiveness and annotate each argument with the values of these attributes. These annotations enable the development of systems that can provide useful feedback to students, as the attribute values predicted by these systems can help a student understand why her essay receives a particular persuasiveness score. To our knowledge, this is the first corpus of essays that are simultaneously annotated with argument components, persuasiveness scores, and related attributes.[1]

2 Related Work

While argument mining research has traditionally focused on determining the argumentative structure of a text document (i.e., identifying its major claim, claims, and premises, as well as the relationships between these argument components) (Stab and Gurevych, 2014b, 2017a; Eger et al., 2017), researchers have recently begun to study new argument mining tasks, as described below.

Persuasiveness-related tasks. Most related to our study is work involving argument persuasiveness. For instance, Habernal and Gurevych (2016b) and Wei et al. (2016) study the persuasiveness *ranking* task, where the goal is to rank two internet debate arguments written for the same topic w.r.t. their persuasiveness. As noted by Habernal and Gurevych, ranking arguments is a relatively easier task than scoring an argument's persuasiveness: in ranking, a system simply determines whether one argument is more persuasive than the other, but not *how much more* persuasive one argument is than the other; in scoring, however, a system has to determine how persuasive an argument is on an absolute scale. Note that ranking is not an acceptable evaluation setting for studying argument persuasiveness in the essay domain, as feedback for an essay has to be provided *independently* of other essays.

In contrast, there are studies that focus on factors affecting argument persuasiveness in internet debates. For instance, Lukin et al. (2017) examine how audience variables (e.g., personalities) interact with argument style (e.g., factual vs. emotional arguments) to affect argument persuasiveness. Persing and Ng (2017) identify factors that *negatively* impact persuasiveness, so their factors, unlike ours, cannot explain what makes an argument persuasive.

Other argument mining tasks. Some of the attributes that we annotate our corpus with have been studied. For instance, Hidey et al. (2017) examine the different semantic types of claims and premises, whereas Higgins and Walker (2012) investigate persuasion strategies (i.e., ethos, pathos, logos). Unlike ours, these studies use data from online debate forums and social/environment reports. Perhaps more importantly, they study these attributes independently of persuasiveness.

Several argument mining tasks have recently been proposed. For instance, Stab and Gurevych (2017b) examine the task of whether an argument is sufficiently supported. Al Khatib et al. (2016) identify and annotate a news editorial corpus with fine-grained argumentative discourse units for the purpose of analyzing the argumentation strategies used to persuade readers. Wachsmuth et al. (2017) focus on identifying and annotating 15 logical, rhetorical, and dialectical dimensions that would be useful for automatically accessing the quality of an argument. Most recently, the Argument Reasoning Comprehension task organized as part of SemEval 2018 has focused on selecting the correct warrant that explains reasoning of an argument that consists of a claim and a reason.[2]

3 Corpus

The corpus we chose to annotate is composed of 102 essays randomly chosen from the Argument Annotated Essays corpus (Stab and Gurevych, 2014a). This collection of essays was taken from *essayforum*[3], a site offering feedback to students wishing to improve their ability to write persuasive essays for tests. Each essay is written in response to a topic such as "should high school make music lessons compulsory?" and has already been annotated by Stab and Gurevych with an argument tree. Hence, rather than annotate everything from scratch, we annotate the persuasiveness score of each argument in the already-annotated argument trees in this essay collection as well as the attributes that potentially impact persuasiveness.

Each argument tree is composed of three types of tree nodes that correspond to argument compo-

[1] Our annotated corpus and annotation manual are publicly available at the website http://www.hlt.utdallas.edu/∼zixuan/EssayScoring.

[2] https://competitions.codalab.org/competitions/17327

[3] www.essayforum.com

Essays: 102	Sentences: 1462	Tokens: 24518
Major Claims: 185	Claims: 567	Premises: 707
Support Relations: 3615	Attack Relations: 219	

Table 1: Corpus statistics.

nents. The three annotated argument component types include: **MajorClaim**, which expresses the author's stance with respect to the essay's topic; **Claims**, which are controversial statements that should not be accepted by readers without additional support; and **Premises**, which are reasons authors give to persuade readers about the truth of another argument component statement. The two relation types include: **Support**, which indicates that one argument component supports another, and **Attack**, which indicates that one argument component attacks another.

Each argument tree has three to four levels. The root is a major claim. Each node in the second level is a claim that supports or attacks its parent (i.e., the major claim). Each node is the third level is a premise that supports or attacks its parent (i.e., a claim). There is an optional fourth level consisting of nodes that correspond to premises. Each of these premises either supports or attacks its (premise) parent. Stab and Gurevych (2014a) report high inter-annotator agreement on these annotations: for the annotations of major claims, claims, and premises, the Krippendorff's α values (Krippendorff, 1980) are 0.77, 0.70, and 0.76 respectively, and for the annotations of support and attack relations, the α values are both 0.81.

Note that Stab and Gurevych (2014a) determine premises and claims by their position in the argument tree and not by their semantic meaning. Due to the difficulty of treating an opinion as a non-negotiable unit of evidence, we convert all subjective premises into claims to demonstrate that they are subjective and require backing. At the end of this process, several essays contain argument trees that violate the scheme used by Stab and Gurevych, due to some premises supported by opinion premises, now converted to claims. Although the ideal argument should not violate the canonical structure, students attempting to improve their persuasive writing skills may not understand this, and mistakenly support evidence with their own opinions.

Statistics of this corpus are shown in Table 1. Its extensive use in argument mining research in recent years together with its reliably annotated argument trees makes it an ideal corpus to use for our annotation task.

4 Annotation

4.1 Definition

Since persuasiveness is defined on an argument, in order to annotate persuasiveness we need to define precisely what an argument is. Following van Eemeren et al. (2014), we define an argument as consisting of a conclusion that may or may not be supported/attacked by a set of evidences. Given an argument tree, a non-leaf node can be interpreted as a "conclusion" that is supported or attacked by its children, which can therefore be interpreted as "evidences" for the conclusion. In contrast, a leaf node can be interpreted as an unsupported conclusion. Hence, for the purposes of our work, an argument is composed of a node in an argument tree and all of its children, if any.

4.2 Annotation Scheme

Recall that the goal of our annotation is to score each argument w.r.t. its persuasiveness (see Table 2 for the rubric for scoring persuasiveness) and annotate each of its components with a set of predefined attributes that could impact the argument's persuasiveness. Table 3 presents a summary of the attributes we annotate. The rest of this subsection describes these attributes.

Each component type (MajorClaim, Claim, Premise) has a distinct set of attributes. All component types have three attributes in common: Eloquence, Specificity, and Evidence. *Eloquence* is how well the author uses language to convey ideas, similar to clarity and fluency. *Specificity* refers to the narrowness of a statement's scope. Statements that are specific are more believable because they indicate an author's confidence and depth of knowledge about a subject matter. Argument assertions (major claims and claims) need not be believable on their own since that is the job of the supporting evidence. The *Evidence* score describes how well the supporting components support the parent component. The rubrics for scoring Eloquence, Evidence, Claim/MajorClaim Specificity, and Premise Specificity are shown in Tables 4, 5, 6, and 7 respectively.

MajorClaim Since the major claim represents the overall argument of the essay, it is in this component that we annotate the persuasive strategies employed (i.e., *Ethos, Pathos* and *Logos*). These

Score	Description
6	A very strong, clear argument. It would persuade most readers and is devoid of errors that might detract from its strength or make it difficult to understand.
5	A strong, pretty clear argument. It would persuade most readers, but may contain some minor errors that detract from its strength or understandability.
4	A decent, fairly clear argument. It could persuade some readers, but contains errors that detract from its strength or understandability.
3	A poor, understandable argument. It might persuade readers who are already inclined to agree with it, but contains severe errors that detract from its strength or understandability.
2	It is unclear what the author is trying to argue or the argument is poor and just so riddled with errors as to be completely unpersuasive.
1	The author does not appear to make any argument (e.g. he may just describe some incident without explaining why it is important). It could not persuade any readers because there is nothing to be persuaded of. It may or may not contain detectable errors, but errors are moot since there is not an argument for them to interfere with.

Table 2: Description of the Persuasiveness scores.

Attribute	Possible Values	Applicability	Description
Specificity	1–5	MC,C,P	How detailed and specific the statement is
Eloquence	1–5	MC,C,P	How well the idea is presented
Evidence	1–6	MC,C,P	How well the supporting statements support their parent
Logos/Pathos/Ethos	yes,no	MC,C	Whether the argument uses the respective persuasive strategy
Relevance	1–6	C,P	The relevance of the statement to the parent statement
ClaimType	value,fact,policy	C	The category of what is being claimed
PremiseType	see Section 4.2	P	The type of Premise, e.g. statistics, definition, real example, etc.
Strength	1–6	P	How well a single statement contributes to persuasiveness

Table 3: Summary of the attributes together with their possible values, the argument component type(s) each attribute is applicable to (**MC**: MajorClaim, **C**: Claim, **P**: Premise), and a brief description.

Score	Description
5	Demonstrates mastery of English. There are no grammatical errors that distract from the meaning of the sentence. Exhibits a well thought out, flowing sentence structure that is easy to read and conveys the idea exceptionally well.
4	Demonstrates fluency in English. If there are any grammatical or syntactical errors, their affect on the meaning is negligible. Word choice suggests a broad vocabulary.
3	Demonstrates competence in English. There might be a few errors that are noticeable but forgivable, such as an incorrect verb tense or unnecessary pluralization. Demonstrates a typical vocabulary and a simple sentence structure.
2	Demonstrates poor understanding of sentence composition and/or poor vocabulary. The choice of words or grammatical errors force the reader to reread the sentence before moving on.
1	Demonstrates minimal eloquence. The sentence contains errors so severe that the sentence must be carefully analyzed to deduce its meaning.

Table 4: Description of the Eloquence scores.

Score	Description
6	A very strong, very persuasive argument body. There are many supporting components that have high Relevance scores. There may be a few attacking child components, but these components must be used for either concession or refuting counterarguments as opposed to making the argument indecisive or contradictory.
5	A strong, persuasive argument body. There are sufficient supporting components with respectable scores.
4	A decent, fairly persuasive argument body.
3	A poor, possibly persuasive argument body.
2	A totally unpersuasive argument body.
1	There is no argument body for the given component.

Table 5: Description of the Evidence scores.

three attributes are not inherent to the text identifying the major claim but instead summarize the child components in the argument tree.

Claim The claim argument component possesses all of the attributes of a major claim in addition to a *Relevance* score and a *ClaimType*. In order for an argument to be persuasive, all supporting components must be relevant to the component that they support/attack. The scoring rubric for Relevance is shown in Table 8. The ClaimType can be *value* (e.g., something is good or bad, important or not important, etc.), *fact* (e.g. something

Score	Description
5	The claim summarizes the argument well and has a qualifier that indicates the extent to which the claim holds true. Claims that summarize the argument well must reference most or all of the supporting components.
4	The claim summarizes the argument very well by mentioning most or all of the supporting components, but does not have a qualifier indicating the conditions under which the claim holds true. Alternatively, the claim may moderately summarize the argument by referencing a minority of supporting components and contain qualifier.
3	The claim has a qualifier clause or references a minority of the supporting components, but not both.
2	The claim does not make an attempt to summarize the argument nor does it contain a qualifier clause.
1	Simply rephrases the major claim or is outside scope of the major claim (argument components were annotated incorrectly: major claim could be used to support claim).

Table 6: Description of the Claim and MajorClaim Specificity scores.

Score	Description
5	An elaborate, very specific statement. The statement contains numerical data, or a historical example from the real world. There is (1) both a sufficient qualifier indicating the extent to which the statement holds true and an explanation of why the statement is true, or (2) at least one real world example, or (3) a sufficient description of a hypothetical situation that would evoke a mental image of the situation in the minds of most readers.
4	A more specific statement. It is characterized by either an explanation of why the statement is true, or a qualifier indicating when/to what extent the statement is true. Alternatively, it may list examples of items that do not qualify as historical events.
3	A sufficiently specific statement. It simply states a relationship or a fact with little ambiguity.
2	A broad statement. A statement with hedge words and without other redeeming factors such as explicit examples, or elaborate reasoning. Additionally, there are few adjectives or adverbs.
1	An extremely broad statement. There is no underlying explanation, qualifiers, or real-world examples.

Table 7: Description of the Premise Specificity scores.

Score	Description
6	Anyone can see how the support relates to the parent claim. The relationship between the two components is either explicit or extremely easy to infer. The relationship is thoroughly explained in the text because the two components contain the same words or exhibit coreference.
5	There is an implied relationship that is obvious, but it could be improved upon to remove all doubt. If the relationship is obvious, both relating components must have high Eloquence and Specificity scores.
4	The relationship is fairly clear. The relationship can be inferred from the context of the two statements. One component must have a high Eloquence and Specificity scores and the other must have lower but sufficient Eloquence and Specificity scores for the relationship to be fairly clear.
3	Somewhat related. It takes some thinking to imagine how the components relate. The parent component or the child component have low clarity scores. The two statements are about the same topic but unrelated ideas within the domain of said topic.
2	Mostly unrelated. It takes some major assumptions to relate the two components. A component may also receive this score if both components have low clarity scores.
1	Totally unrelated. Very few people could see how the two components relate to each other. The statement was annotated to show that it relates to the claim, but this was clearly in error.

Table 8: Description of the Relevance scores.

is true or false), or *policy* (claiming that some action should or should not be taken).

Premise The attributes exclusive to premises are *PremiseType* and *Strength*. To understand Strength, recall that only premises can persuade readers, but also that an argument can be composed of a premise and a set of supporting/attacking premises. In an argument of this kind, Strength refers to how well the parent premise contributes to the persuasiveness independently of the contributions from its children. The scoring rubric for Strength is shown in Table 9. PremiseType takes on a discrete value from one of the following: real_example, invented_instance, analogy, testimony, statistics, definition, common_knowledge, and warrant. Analogy, testimony, statistics, and definition are self-explanatory. A premise is labeled *invented_instance* when it describes a hypothetical situation, and *definition* when it provides a definition to be used elsewhere in the argument. A premise has type *warrant* when it does not fit any other type, but serves a functional purpose to explain the relationship between two entities or clarify/quantify another statement. The *real_example* premise type indicates that the statement is a historical event that actually occurred, or something that is verfiably true about the real world.

Score	Description
6	A very strong premise. Not much can be improved in order to contribute better to the argument.
5	A strong premise. It contributes to the persuasiveness of the argument very well on its own.
4	A decent premise. It is a fairly strong point but lacking in one or more areas possibly affecting its perception by the audience.
3	A fairly weak premise. It is not a strong point and might only resonate with a minority of readers.
2	A totally weak statement. May only help to persuade a small number of readers.
1	The statement does not contribute at all.

Table 9: Description of the Strength scores.

Attribute	Value	MC	C	P
Specificity	1	0	80	64
	2	73	259	134
	3	72	155	238
	4	32	59	173
	5	8	14	98
Logos	Yes	181	304	
	No	4	263	
Pathos	Yes	67	59	
	No	118	508	
Ethos	Yes	16	9	
	No	169	558	
Relevance	1		1	5
	2		33	45
	3		58	59
	4		132	145
	5		97	147
	6		246	306
Evidence	1	3	246	614
	2	62	115	28
	3	57	85	12
	4	33	80	26
	5	16	35	15
	6	14	6	12
Eloquence	1	3	23	24
	2	19	106	97
	3	116	320	383
	4	42	102	154
	5	5	16	49
ClaimType	fact		368	
	value		145	
	policy		54	
PremiseType	real_example			93
	invented_instance			53
	analogy			2
	testimony			4
	statistics			15
	definition			3
	common_know.			493
	warrant			44
Persuasiveness	1	3	82	8
	2	62	278	112
	3	60	84	145
	4	28	74	249
	5	17	39	123
	6	15	10	70

Table 10: Class/Score distributions by component type.

Attribute	MC	C	P
Persuasiveness	.739	.701	.552
Eloquence	.590	.580	.557
Specificity	.560	.530	.690
Evidence	.755	.878	.928
Relevance		.678	.555
Strength			.549
Logos	1	.842	
Pathos	.654	.637	
Ethos	1	1	
ClaimType		.589	
PremiseType			.553

Table 11: Krippendorff's α agreement on each attribute by component type.

4.3 Annotation Procedure

Our 102 essays were annotated by two native speakers of English. We first familiarized them with the rubrics and definitions and then trained them on five essays (not included in our corpus). After that, they were both asked to annotate a randomly selected set of 30 essays and discuss the resulting annotations to resolve any discrepancies. Finally, the remaining essays were partitioned into two sets, and each annotator received one set to annotate. The resulting distributions of scores/classes for persuasiveness and the attributes are shown in Table 10.

4.4 Inter-Annotator Agreement

We use Krippendorff's α to measure inter-annotator agreement. Results are shown in Table 11. As we can see, all attributes exhibit an agreement above 0.5, showing a correlation much more significant than random chance. Persuasiveness has an agreement of 0.688, which suggests that it can be agreed upon in a reasonably general sense. The MajorClaim components have the highest Persuasiveness agreement, and it declines as the type changes to Claim and then to Premise. This would indicate that persuasiveness is easier to articulate in a wholistic sense, but difficult to explain as the number of details involved in the explanation increases.

The agreement scores that immediately stand out are the perfect 1.0's for Logos and Ethos. The perfect Logos score is explained by the fact that every major claim was marked to use logos. Although ethos is far less common, both annotators easily recognized it. This is largely due to the indisputability of recognizing a reference to an accepted authority on a given subject. Very few authors utilize this approach, so when they do it is extremely apparent. Contrary to Persuasiveness, Evidence agreement exhibits an upward trend as the component scope narrows. Even with this pattern, the Evidence agreement is always higher than Persuasiveness agreement, which suggests that it is not the only determiner of persuasiveness.

In spite of a rubric defining how to score Eloquence, it remains one of the attributes with the lowest agreement. This indicates that it is difficult to agree on exact eloquence levels beyond basic English fluency. Additionally, Specificity produced unexpectedly low agreement in claims and major claims. Precisely quantifying how well a claim summarizes its argument turned out to be a complicated and subjective task. Relevance agreement for premises is one of the lowest, partly because there are multiple scores for high relevance, and no examples were given in the rubric.

All attributes but those with the highest agreement are plagued by inherent subjectivity, regardless of how specific the rubric is written. There are often multiple interpretations of a given sentence, sometimes due to the complexity of natural language, and sometimes due to the poor writing of the author. Naturally, this makes it difficult to identify certain attributes such as Pathos, ClaimType, and PremiseType.

Although great care was taken to make each attribute as independent of the others as possible, they are all related to each other to a minuscule degree (e.g., Eloquence and Specificity). While annotators generally agree on what makes a persuasive argument, the act of assigning blame to the persuasiveness (or lack thereof) is tainted by this overlapping of attributes.

4.5 Analysis of Annotations

To understand whether the attributes we annotated are indeed useful for predicting persuasiveness, we compute the Pearson's Correlation Coefficient (PC) between persuasiveness and each of the attributes along with the corresponding p-values. Results are shown in Table 12. Among the correlations that are statistically significant at the $p < .05$ level, we see, as expected, that Persuasiveness is positively correlated with Specificity, Evidence, Eloquence, and Strength. Neither is it surprising that support provided by a premise in the form of statistics and examples is positively correlated with Persuasiveness. While Logos and invented_instance also have significant correlations with Persuasiveness, the correlation is very weak.

Attribute	PC	p**-value**
Specificity	.5680	0
Relevance	$-.0435$	.163
Eloquence	.4723	0
Evidence	.2658	0
Strength	.9456	0
Logos	$-.1618$	0
Ethos	$-.0616$	.1666
Pathos	$-.0835$	.0605
ClaimType:fact	.0901	.1072
ClaimType:value	$-.0858$	.1251
ClaimType:policy	$-.0212$	.7046
PremiseType:real_example	.2414	0
PremiseType:invented_instance	.0829	.0276
PremiseType:analogy	.0300	.4261
PremiseType:testimony	.0269	.4746
PremiseType:statistics	.1515	0
PremiseType:definition	.0278	.4608
PremiseType:common_knowledge	$-.2948$	1.228
PremiseType:warrant	.0198	.6009

Table 12: Correlation of each attribute with Persuasiveness and the corresponding p-value.

	MC	**C**	**P**	**Avg**
PC	.9688	.9400	.9494	.9495
ME	.0710	.1486	.0954	.1061

Table 13: Persuasiveness scoring using gold attributes.

Next, we conduct an oracle experiment in an attempt to understand how well these attributes, when used together, can explain the persuasiveness of an argument. Specifically, we train three linear SVM regressors (using the SVM^{light} software (Joachims, 1999) with default learning parameters except for C (the regularization parameter), which is tuned on development data using grid search) to score an argument's persuasiveness using the *gold* attributes as features. The three regressors are trained on arguments having MajorClaims, Claims, and Premises as parents. For instance, to train the regressor involving MajorClaims, each instance corresponds to an argument represented by all and only those attributes involved in the major claim and all of its children.[4]

Five-fold cross-validation results, which are

[4] There is a caveat. If we define features for each of the children, the number of features will be proportional to the number of children. However, SVMs cannot handle a variable number of features. Hence, all of the children will be represented by one set of features. For instance, the Specificity feature value of the children will be the Specificity values averaged over all of the children.

Prompt: Government budget focus, young children or university?

Education plays a significant role in a country's long-lasting prosperity. It is no wonder that governments throughout the world lay special emphasis on education development. As for the two integral components within the system, elementary and advanced education, there's no doubt that a government is supposed to offer sufficient financial support for both.

Concerning that elementary education is the fundamental requirement to be a qualified citizen in today's society, government should guarantee that all people have equal and convenient access to it. So a lack of well-established primary education goes hand in hand with a high rate of illiteracy, and this interplay compromises a country's future development. In other words, if countries, especially developing ones, are determined to take off, one of the key points governments should set on agenda is to educate more qualified future citizens through elementary education.

...

Table 14: An example essay. Owing to space limitations, only its first two paragraphs are shown.

		P	E	S	Ev	R	St	Lo	Pa	Et	cType	pType
M1	government is supposed to offer sufficient financial support for both	3	4	2	3			T	F	F		
C1	if countries, especially developing ones, are determined to take off, one of the key points governments should set on agenda is to educate more qualified future citizens through elementary education	4	5	4	4	6		T	F	F	policy	
P1	elementary education is the fundamental requirement to be a qualified citizen in today's society	4	5	3	1	6	4					A
C2	government should guarantee that all people have equal and convenient access to it	2	3	1	1	6		F	F	F	policy	
P2	a lack of well-established primary education goes hand in hand with a high rate of illiteracy, and this interplay compromises a country's future development	4	5	3	1	6	4					C

Table 15: The argument components in the example in Table 14 and the scores of their associated attributes: **P**ersuasiveness, **E**loquence, **S**pecificity, **Ev**idence, **R**elevance, **St**rength, **Lo**gos, **Pa**thos, **Et**hos, **c**laim**Type**, and **p**remise**Type**.

shown in Table 13, are expressed in terms of two evaluation metrics, PC and ME (the mean absolute distance between a system's prediction and the gold score). Since PC is a *correlation* metric, higher correlation implies better performance. In contrast, ME is an *error* metric, so lower scores imply better performance. As we can see, the large PC values and the relatively low ME values provide suggestive evidence that these attributes, when used in combination, can largely explain the persuasiveness of an argument.

What these results imply in practice is that models that are trained on these attributes for persuasiveness scoring could provide useful feedback to students on *why* their arguments are (un)persuasive. For instance, one can build a pipeline system for persuasiveness scoring as follows. Given an argument, this system first predicts its attributes and then scores its persuasiveness using the predicted attribute values computed in the first step. Since the persuasiveness score of an argument is computed using its predicted attributes, these attributes can explain the persuasiveness score. Hence, a student can figure out which aspect of persuasiveness needs improvements by examining the values of the predicted attributes.

4.6 Example

To better understand our annotation scheme, we use the essay in Table 14 to illustrate how we obtain the attribute values in Table 15. In this essay, Claim **C1**, which supports MajorClaim **M1**, is supported by three children, Premises **P1** and **P2** as well as Claim **C2**.

After reading the essay in its entirety and acquiring a holistic impression of the argument's strengths and weaknesses, we begin annotating the atomic argument components bottom up, starting with the leaf nodes of the argument tree. First, we consider **P2**. Its Evidence score is 1 because it is a leaf node with no supporting evidence. Its Eloquence score is 5 because the sentence has no serious grammatical or syntactic errors, has a flowing, well thought out sentence structure, and uses articulate vocabulary. Its Specificity score is 3 because it is essentially saying that poor primary education causes illiteracy and consequently inhibits a country's development. It does not state why or to what extent, so we cannot assign a score of 4. However, it does explain a simple relationship with little ambiguity due to the lack of hedge words, so we can assign a score of 3. Its PremiseType is *common_knowledge* because it is reasonable to assume most people would agree that poor primary education causes illiteracy, and also that illiteracy inhibits a country's development. Its Relevance score is 6: its relationship with its parent is clear because the two components exhibit coreference. Specifically, **P2** contains a reference to primary/elementary education and shows how this affects a country's inability to transition from developing to developed. Its Strength is 4: though eloquent and relevant, **P2** is lacking substance in order to be considered for a score of 5 or 6. The PremiseType is *common_knowledge*, which is mediocre compared to statistics and real_example. In order for a premise that is not grounded in the real world to be strong, it must be very specific. **P2** only scored a 3 in Specificity, so we assign a Strength score of 4. Finally, the argument headed by **P2**, which does not have any children, has a Persuasiveness score of 4, which is obtained by summarizing the inherent strength of the premise and the supporting evidence. Although there is no supporting evidence for this premise, this does not adversely affect persuasiveness due to the standalone nature of premises. In this case the persuasiveness is derived totally from the strength.

Next, the annotator would score **C2** and **P1**, but for demonstration purposes we will examine the scoring of **C1**. **C1**'s Eloquence score is 5 because it shows fluency, broad vocabulary, and attention to how well the sentence structure reads. Its ClaimType is *policy* because it specifically says that the government should put something on their agenda. Its Specificity score is 4: while it contains information relevant to all the child premises (i.e., creating qualified citizens, whose role it is to provide the education, and the effect of education on a country's development), it does not contain a qualifier stating the extent to which the assertion holds true. Its Evidence score is 4: **C1** has two premises with decent persuasiveness scores and one claim with a poor persuasiveness score, and there are no attacking premises, so intuitively, we may say that this is a midpoint between many low quality premises and few high quality premises. We mark Logos as true, Pathos as false, and Ethos as false: rather than use an emotional appeal or an appeal to authority of any sort, the author attempts to use logical reasoning in order to prove their point. Its Persuasiveness score is 4: this score is mainly determined by the strength of the supporting evidence, given that the assertion is precise and clear as determined by the specificity and eloquence. Its Relevance score is 6, as anyone can see how endorsement of elementary education in **C1** relates to the endorsement of elementary and university education in its parent (i.e., **M1**).

After all of the claims have been annotated in the bottom-up method, the annotator moves on to the major claim, **M1**. **M1**'s Eloquence score is 4: while it shows fluency and a large vocabulary, it is terse and does not convey the idea exceptionally well. Its persuasion strategies are obtained by simply taking the logical disjunction of those used in its child claims. Since every claim in this essay relied on logos and did not employ pathos nor ethos, **M1** is marked with Logos as true, Pathos as false, and Ethos as false. Its Evidence score is 3: in this essay there are two other supporting claims not in the excerpt, with persuasiveness scores of only 3 and 2, so **M1**'s evidence has one decently persuasive claim, one claim that is poor but understandable, and one claim that is so poor as to be completely unpersuasive (in this case it has no supporting premises). Its Specificity score is 2 because it does not have a quantifier nor does it attempt to summarize the main points of the evidence. Finally, its Persuasiveness score is 3: all supporting claims rely on logos, so there is no added persuasiveness from a variety of persuasion strategies, and since the eloquence and specificity are adequate, they do not detract from the Evidence score.

5 Conclusion

We presented the first corpus of 102 persuasive student essays that are simultaneously annotated with argument trees, persuasiveness scores, and attributes of argument components that impact these scores. We believe that this corpus will push the frontiers of research in content-based essay grading by triggering the development of novel computational models concerning argument persuasiveness that could provide useful feedback to students on why their arguments are (un)persuasive in addition to how persuasive they are.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-1219142 and IIS-1528037.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, Francisca A. Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. In *Handbook of Argumentation Theory*. Springer, Dordrecht.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

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*, pages 11–21.

Colin Higgins and Robyn Walker. 2012. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. *Accounting Forum*, 36:194-208.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 185–192.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage commtext series. Sage, Thousand Oaks, CA.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

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.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? Modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4082–4088.

Mark D. Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge Chapman & Hall.

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

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for*

Computational Linguistics: Volume 1, Long Papers, pages 980–990.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200.