Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Learning to Give Feedback: Modeling Attributes Affecting Argument Persuasiveness in Student Essays

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

Abstract

Argument persuasiveness is one of the most important dimensions of argumentative essay quality, yet it is little studied in automated essay scoring research. Using a recently released corpus of essays that are simultaneously annotated with argument components, argument persuasiveness scores, and attributes of argument components that impact an argument's persuasiveness, we design the first set of neural models that predict the persuasiveness of an argument and its attributes in a student essay, enabling useful feedback to be provided to students on *why* their arguments are (un)persuasive in addition to *how* persuasive they are.

1 Introduction

Recent work on automated essay scoring has largely focused on *holistic* scoring, which summarizes the quality of an essay with a single score. There are at least two reasons for this focus. First, corpora manually annotated with holistic scores such as the one used in the Kaggle-sponsored ASAP competition[1] are publicly available, facilitating the training and evaluation of holistic essay scoring engines. Second, holistic scoring technologies have large commercial values: being able to successfully automate the scoring of the millions of essays written for aptitude tests such as SAT, GRE, and GMAT every year can save a lot of manual grading effort.

Though useful for scoring essays written for aptitude tests, holistic essay scoring technologies are far from adequate for use in classroom settings, where providing students with feedback on how to improve their essays is of utmost importance. Specifically, merely returning a low holistic score to an essay provides essentially no feedback to its author on which aspect(s) of the essay contributed to the low score and how it can be improved. Recently, researchers have attempted to score 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]. Automated systems that provide instructional feedback along multiple dimensions of essay quality such as Criterion [Burstein *et al.*, 2004] have also begun to emerge. Providing scores along different dimensions of essay quality could help an author identify which aspects of her essay need improvements.

One may argue that the feedback provided by these dimension-specific scores is still limited: if a student receives a low score along a particular dimension, she may still not know why her score is low. Our goal is to address this concern by developing computational models that can explain why an essay receives a particular score along a given dimension of essay quality. In this paper, we focus on a dimension that is largely ignored in existing automated essay scoring research despite being one of the most important dimensions of essay quality: argument persuasiveness. To our knowledge, Persing and Ng's [2015] (P&N) work is the only attempt to date on argument persuasiveness scoring in student essays. However, their system does *not* explain why an argument is not persuasive if its score is low, and is therefore rather undesirable from a feedback perspective.

Developing computational models for providing feedback on scoring argument persuasiveness is by no means easy, however. The difficulty stems in part from the scarcity of persuasiveness-annotated corpora. For this reason, we have recently annotated and made publicly available a corpus of persuasive student essays [Carlile *et al.*, 2018], wherein we not only score the persuasiveness of *each* argument in each essay (rather than simply the persuasiveness of the overall argument as in P&N), but also identify a set of attributes that can explain an argument's persuasiveness and annotate each argument with the values of these attributes. To our knowledge, this is the first corpus of essays that are simultaneously annotated with argument components, argument persuasiveness scores, and related attributes.

Using this corpus, we train the first set of neural models that predict the persuasiveness score of an argument in a student essay as well as the scores of its various attributes. Unlike previous persuasiveness scoring models, our models could 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.

[1] https://www.kaggle.com/c/asap-aes

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

Table 1: Corpus statistics.

	1	2	3	4	5	6
MC	3	62	60	28	17	15
C	82	278	84	74	39	10
P	8	112	145	249	123	70

Table 3: Score distribution of persuasiveness.

2 Corpus

The corpus we use is composed of 102 essays randomly chosen from the Argument Annotated Essays corpus [Stab and Gurevych, 2014a]. These essays were taken from *essayforum*[2], 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?". Below we describe the two types of annotations associated with each essay.

Argument trees. Each essay is annotated by Stab and Gurevych [2014a] with an argument tree. Each argument tree is composed of three types of tree nodes that correspond to argument components. The three 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 component statement. The two relation types include: **Support**, which indicates that one component supports another, and **Attack**, which indicates that one 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 supports or attacks its (premise) parent.

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 collected from the resulting argument trees are shown in Table 1.

Persuasiveness-related attributes. Recently, we have annotated each argument in each argument tree with (1) its persuasiveness score and (2) the attributes that potentially impact persuasiveness [Carlile *et al.*, 2018]. By definition, an argument consists of a conclusion that may or may not be supported/attacked by a set of evidences [van Eemeren *et al.*, 2014]. 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. More specifically, an argument that we consider can be composed of (1) a major claim and a set of supporting/attacking claims; (2) a claim and a (possibly empty) set of supporting/attacking premises; or (3) a premise and a (possibly empty) set of supporting/attacking premises.[3]

As noted above, each argument is scored w.r.t. its persuasiveness (see Table 2 for the rubric for scoring persuasiveness and Table 3 for the resulting score distribution), and each of its components is annotated with a set of predefined attributes that could impact the argument's persuasiveness. Owing to space limitations, we will only provide a high-level overview of the subset of attributes that we use to train our neural models (see Table 4 for a summary of these attributes) below.[4]

Each component type (Premise, Claim, MajorClaim) 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.

MajorClaim Since the major claim represents the entire argument of the essay, it is in this component that we annotate the persuasive strategies employed (i.e., *Ethos, Pathos* and *Logos*). These three attributes are not inherent to the text identifying the major claim but instead summarize the child components in the argument tree. Different people are persuaded in different ways: some are persuaded by logic (*logos*), some by emotion (*pathos*), and some by trust in a higher authority (*ethos*). In order to appeal to the broadest audience, usage of multiple persuasive strategies tends to improve persuasiveness.

Claim A claim possesses all of the attributes of a major claim in addition to a *ClaimType*. The ClaimType can be *value* (e.g., something is good or bad, important or not important, etc.), *fact* (e.g. something is true or false), or *policy* (claiming that some action should or should not be taken).

[2] www.essayforum.com

[3] Because of our conversion of subjective premises to claims, in a small number of cases a claim can be supported/attacked by another claim and a premise could be supported/attacked by a claim.

[4] For the full set of attributes, the rubrics for scoring/annotating each attribute and its resulting score/class distribution, we refer the reader to Carlile *et al.* [2018] for details.

Score	Description of Argument Persuasiveness
6	A **very persuasive**, **clear** argument. It would persuade most previously uncommitted readers and is devoid of problems that might detract from its persuasiveness or make it difficult to understand.
5	A **persuasive**, or only **pretty clear** argument. It would persuade most previously uncommitted readers, but may contain some minor problems that detract from its persuasiveness or understandability.
4	A **decent**, or only **fairly clear** argument. It could persuade some previously uncommitted readers, but problems detract from its persuasiveness or understandability.
3	A **poor**, or only **mostly understandable** argument. It might persuade readers who are already inclined to agree with it, but contains severe problems that detract from its persuasiveness or understandability.
2	A **very unpersuasive** or **very unclear** argument. It is unclear what the author is trying to argue or the argument is just so riddled with problems as to be completely unpersuasive.
1	The author **does not make an argument** or it is **unclear what the argument is**. It could not persuade any readers because there is nothing to be persuaded of.

Table 2: Descriptions of argument 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
ClaimType	Value,Fact,Policy	C	The category of what is being claimed
PremiseType	see Section 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 4: 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.

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. *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 components 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.

3 Persuasiveness Scoring Models

3.1 Baseline Model

Since one of the goals of our evaluation is to determine the usefulness of the automatically predicted attributes for persuasiveness scoring, we design a Baseline model that scores persuasiveness *without* using any attributes.

Baseline takes as input an argument, which corresponds to a node in an argument tree and its n children, if any, and scores the argument's persuasiveness. Baseline relies on bidirectional long short term memory networks (biLSTMs) [Schuster and Paliwal, 1997]. Recall that biLSTMs use both the previous and future context by processing the input sequence in two directions. The final representation is the concatenation of the (last timesteps of each of the) forward and backward steps.

Specifically, Baseline first uses n+1 biLSTMs: one for creating a representation of the parent's word sequence and the remaining n for creating representations of its n children. It then concatenates these n+1 representations. The resulting vector first goes through a dense layer, which reduces the vector's dimension to 150 (with Leaky ReLU as the activation function), then goes through another dense layer for scoring (again with Leaky ReLU as the activation function). To represent the words, we use the 300-dimensional Facebook FastText pre-trained word embeddings [Bojanowski *et al.*, 2017]. To handle out-of-word vocabulary words, we create random word vectors and map each of them to the same random vector. The network is trained to minimize mean absolute error. Early stopping is used to choose the best epoch. Specifically, training stops when the loss on development data stops improving after 20 epochs.

In addition, we evaluate two extensions to Baseline.

Attention mechanism. In order for the model to focus on the relevant parts of the n+1 representations created by the biLSTMs, we apply an attention mechanism to each of these representations separately. In our attention mechanism, we determine the importance weighting α_t for each hidden state h_t of a biLSTM as follows:

$$e_t = \tanh(Wh_t + b); \quad \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)};$$

where W (the kernel weight matrix) and b (the bias) are tunable parameters of the mechanism, and e_t is the hidden rep-

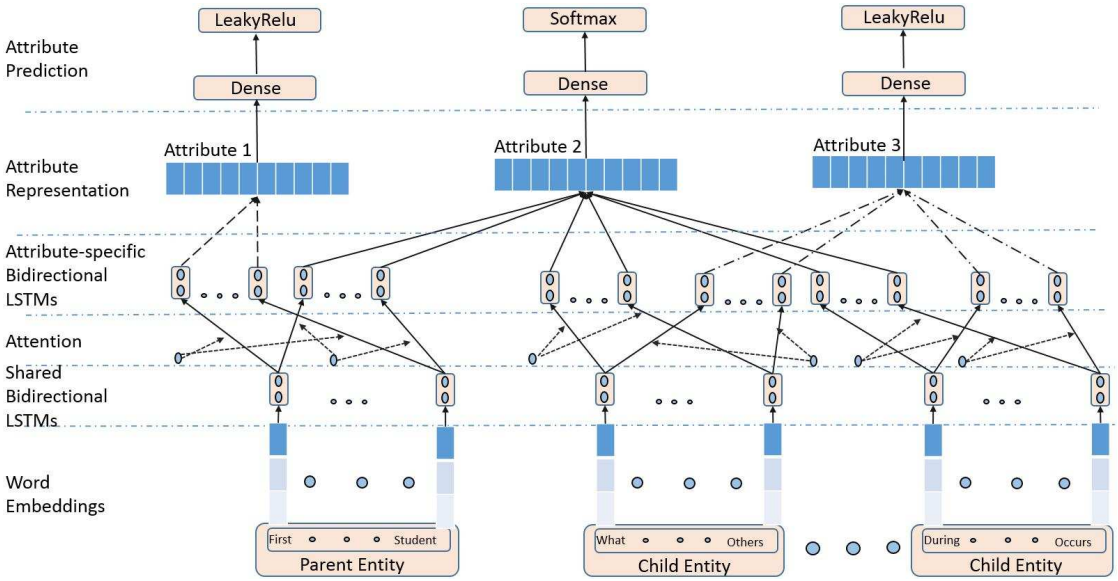

Figure 1: Neural network architecture for joint persuasiveness scoring and attribute prediction.

resentation of h_t. Using α_t, we determine the context vector c_t at timestep t as follows:

$$c_t = \alpha_t e_t$$

This yields a sequence of context vectors $c_1 c_2 \ldots c_T$ for each of the n+1 biLSTMs. These n+1 sequences then serve as the inputs for a second set of n+1 biLSTMs, whose output vectors are concatenated and passed to the first dense layer.

Additional features. We determine whether incorporating additional features can improve Baseline's performance. Specifically, we employ 17 of the linguistic features originally used by Tan *et al.* [2016] for a task related to argument persuasiveness. The 17 features, which are defined on the argument under consideration, include: #words, #definite/indefinite articles, #positive/negative words, #1st/2nd person pronouns, #1st person plural pronouns, #hedges, #examples, #quotations, #sentences, #quantifiers, #children, type-token ratio, fraction of definite articles, and fraction of positive words. When applied, these features will be concatenated with the representations created by the n+1 biLSTMs.

Note that the attention mechanism and the additional features can be applied in isolation and in combination. When used in combination, the additional features will be concatenated with the vectors created by the second (rather than the first) set of n+1 biLSTMs described above.

3.2 Pipeline Model

Our Pipeline model operates in two steps. First, it predicts each attribute in Table 4 independently of other attributes. Then, it uses the predicted attributes to score persuasiveness. Below we describe these two steps in detail.

Step 1: Some attributes, including Specificity, Eloquence, Strength, ClaimType, and PremiseType, are defined on an argument *component* (as opposed to an argument, which involves more than one argument component). To predict each of these attributes, we employ a network whose architecture is the same as that of Baseline except that it uses one biLSTM (rather than n+1 biLSTMs) because the input is composed solely of the word sequence appearing in the argument component whose attributes are to be predicted. If a discrete-(rather than a real-)valued attribute is to be predicted (e.g., ClaimType), we need to replace Leaky ReLU with softmax as the activation function and mean absolute error (ME) with cross entropy as the objective function in the dense layers.

Evidence, in contrast, is defined on the n children of an argument. So, to predict evidence, we use the Baseline network architecture but with n biLSTMs, each of which creates a representation of one of the children.

Each of the remaining attributes (Logos, Ethos, Pathos) is defined on an entire argument and will be predicted using a network that has the same architecture as that of Baseline.

Finally, note that the attention mechanism can be applied to the networks described in this step in the same way as in the Baseline.

Step 2: To score the persuasiveness of an argument, we feed the vector of attributes predicted in Step 1 that involve all of its components into a dense layer that has Leaky ReLU as its activation function. This network is trained on vectors of gold attribute values. Note that the additional features described in Baseline can be applied to train this network simply by concatenating them with the vector of attributes.

3.3 Joint Model

Figure 1 shows the Joint model, which is a neural network that simultaneously scores the persuasiveness of an argument and predicts its attributes. Note that the Baseline network is a special case of this network where only one attribute is predicted, namely persuasiveness. Similarly for the networks used in Step 1 of the Pipeline model: each of them is a special case of this network where only one attribute is predicted.

We also note that Figure 1 is a simplified view of the Joint network. Specifically, while the output layer seems to sug-

		Baseline				Pipeline				Joint			
	System	MC	C	P	Avg	MC	C	P	Avg	MC	C	P	Avg
PC	U	.115	.093	.207	.158	.035	.136	.263	.196	**.165**	.074	.199	.163
	UF	−.041	.204	.221	.177	**.155**	.025	.303	.210	.060	**.207**	.276	**.226**
	UA	.022	.181	.214	.179	.088	**.169**	**.343**	**.259**	.054	.182	.081	.104
	UFA	**.165**	**.245**	**.293**	**.261**	.047	.082	.204	.149	.052	.087	**.304**	.209
ME	U	**1.028**	1.099	1.058	1.065	**.991**	1.035	1.015	1.017	**1.021**	1.060	1.063	1.056
	UF	1.150	1.039	1.047	1.062	1.272	1.418	.977	1.135	1.097	1.008	.989	1.011
	UA	1.107	**.968**	**1.016**	**1.018**	1.081	**.970**	**.948**	**.975**	1.141	**.969**	**.965**	**.993**
	UFA	1.178	.996	1.032	1.046	1.229	1.034	.957	1.019	1.217	1.016	.999	1.036

Table 5: Persuasiveness scoring results of the four variants (U, UF, UA, UFA) of the three models (Baseline, Pipeline, and Joint) on the development set as measured by the two scoring metrics (PC and ME).

		Baseline				Pipeline				Joint			
	System	MC	C	P	Avg	MC	C	P	Avg	MC	C	P	Avg
PC	Best	.034	.145	.269	.205	.038	.138	.353	.248	.148	.163	.290	.236
ME	Best	1.280	1.036	1.056	1.086	1.363	1.237	1.041	1.147	1.220	1.032	.983	1.035

Table 6: Persuasiveness scoring results on the test set obtained by employing the variant that performs the best on the development set w.r.t. the scoring of MC/C/P's persuasiveness.

gest that the network predicts only three things, in reality there is one output node for each of the attributes of the argument under consideration and its persuasiveness. One representation will be created for each such attribute as well as persuasiveness in the Attribute Representation layer. An attribute belongs to one of three *types*. A Type 1 attribute, which includes Eloquence, Specificity, Strength, ClaimType, and PremiseType, can be computed using a single argument component (i.e., either the parent or a child), as exemplified by Attribute 1 in the Attribute Representation layer of Figure 1. A Type 2 attribute, which includes Persuasiveness, Logos, Ethos, and Pathos, is computed using both the parent and all of its children, as exemplified by Attribute 2 in the figure. A Type 3 attribute, which includes Evidence, is computed using all of the children, as exemplified by Attribute 3.

Each of the representations in the Attribute Representation layer is created using an attribute-specific biLSTM in the Attribute-specific Bidirectional LSTMs layer of Figure 1. For instance, to predict the parent's Eloquence, one biLSTM will be created specifically for it in this layer.

Like other networks for multi-task learning, this network has n+1 "shared" biLSTMs (see the Shared Bidirectional LSTMs layer) that create representations for the parent and its n children that are shared by multiple prediction tasks.

Like the Baseline and Pipeline models, the attention mechanism and the additional features can be optionally applied in the Joint model. If the attention mechanism is not applied, the outputs of the biLSTMs in the lower layer will directly become the inputs for the biLSTMs in the upper layer. If additional features are used, they will be concatenated with each of the vectors in the Attribute Representation layer.

4 Evaluation

4.1 Experimental Setup

We first randomly partition our 102 essays into five folds, each of which contains 20–21 essays, and then conduct five-fold cross-validation experiments. In each fold experiment, we employ three folds for training, one fold for development, and one fold for testing. Given a training/development/test set, we first divide the available arguments into three subsets depending on whether the argument's parent node is a MajorClaim, Claim, or Premise. We then train one model on each of the three subsets of training arguments and apply each model to classify the corresponding development/test arguments.

For persuasiveness scoring, we employ two evaluation metrics, PC and ME. PC computes the Pearson's Correlation Coefficient between a model's predicted scores and the annotator assigned scores. In contrast, ME measures the mean absolute distance between a system's prediction and the gold score. Note that PC is a *correlation* metric, so higher correlation implies better performance. In contrast, ME is an *error* metric, so lower scores imply better performance.

4.2 Results and Discussion

Persuasiveness scoring results of the three models on the *development* set obtained via five-fold cross validation are shown in Table 5. Four variants of each model are evaluated. The U variants are trained *without* the 17 additional features and attention; the UF variants are trained with the 17 features but without attention; the UA variants are trained with attention but without the 17 features; and the UFA variants are trained with both attention and the 17 features. Results, expressed in terms of the PC and ME scoring metrics, are first computed separately for arguments whose parent nodes correspond to a MajorClaim (MC), Claim (C), and Premise (P) before being micro-averaged (Avg). The strongest result in each column w.r.t. each metric is boldfaced.

First, to determine whether automatically computed attributes are useful for persuasiveness scoring, we compare Baseline, which does not use attributes, with Pipeline and Joint, both of which predict attributes. W.r.t. ME, we see that Joint's Avg scores are consistently better than those of Baseline, and Pipeline's Avg scores are better than those of Baseline on all but one variant (UF). The best Avg score is achieved using Pipeline-UA. If we examine the MC, C, and P results, we see some consistency across the three models. Specifically, the best MC results come from the U variant,

and the best C and P results both come from the UA variant. These results seem to suggest that scoring a MC's persuasiveness does not benefit from the addition of features and the use of an attention mechanism, whereas scoring a C or P's persuasiveness benefits from applying attention in the absence of additional features. W.r.t. PC, the Avg results are somewhat mixed. Specifically, Joint outperforms Baseline on two of the four variants (U and UF), whereas Pipeline outperforms Baseline on all but the UFA variant. Examining the MC, C, and P results, we no longer see any consistency across the three models. Specifically, for Baseline, the best MC, C, and P results all come from the UFA variant; for Pipeline, the best MC come from UF, and the best C and P results come from UA; and for Joint, the best MC, C, and P results are from different variants. These results suggest that while Baseline consistently benefits from employing attention and the additional features, the same is not true for Pipeline and Joint. For instance, Pipeline benefits from attention when scoring a C or P's persuasiveness and from using additional features when scoring a MC's persuasiveness.

Given these development results, we hypothesize that our three persuasiveness scoring models could be improved by scoring a MC, C, and P's persuasiveness using different variants. To test this hypothesis, we conduct the following experiment. For each of the three models, we score a MC/C/P's persuasiveness on the *test* set using the variant that achieved the best performance on the development set w.r.t. a particular scoring metric. Doing so gives each model the flexibility to use different variants when scoring MC's, C's, and P's persuasiveness. For instance, it is possible for Pipeline to use UF when scoring a MC's persuasiveness and UA when scoring a C's persuasiveness, and such choices can change depending on the scoring metric.

Five-fold cross-validation results of the aforementioned experiment are shown in Table 6. As mentioned before, these are results on the *test* set. A few points deserve mention. First, Joint consistently beats Baseline, outperforming it on MC, C, P, and Avg w.r.t. both scoring metrics. Second, while Pipeline outperforms Baseline w.r.t. Avg PC (primarily because of its superior performance on P), it underperforms Baseline w.r.t. Avg ME (primarily because of its inferior performance on MC and C). Finally, Joint consistently outperforms Pipeline w.r.t. ME, but underperforms it w.r.t. Avg PC only because of its inferior performance on P. Hence, we may be able to obtain further gains by creating an "ensemble" model where we apply Pipeline when scoring a P's persuasiveness w.r.t. PC and use Joint otherwise. Overall, given that Joint has consistently superior performance to Baseline, we conclude that automatically computed attributes are useful for persuasiveness scoring. Nevertheless, the usefulness of these automatically computed attributes depends in part on how they are used, as shown by the difference in Pipeline's and Joint's results.

To gain additional insights into the usefulness of these attributes, we conduct an oracle experiment where we use *gold* attribute values for persuasiveness scoring by training the same neural network that was used in the second step of the Pipeline model. Cross-validation results, which are shown in Table 7, provide strong evidence that these attributes are very useful for persuasiveness scoring.

	MC	C	P	Avg
PC	.969	.945	.942	.952
ME	.150	.250	.251	.217

Table 7: Persuasiveness scoring using gold attributes.

Finally, we report attribute prediction performance on the test set in Table 8 where each attribute's results are micro-averaged over its respective argument component types. Real- and discrete-valued attributes are evaluated using PC and $F1$ respectively. As we can see, Joint outperforms Pipeline on predicting Strength, Pathos and Ethos, but the two yield similar results otherwise. Overall, attribute prediction performance is rather mediocre. Comparing the results in Tables 5, 6, and 8, we see that while the attributes are useful for persuasiveness scoring, their usefulness in our models is limited by the accuracy with which they are computed.

5 Related Work

While argument mining research has traditionally focused on determining the argumentative structure of a text document [Stab and Gurevych, 2014b; 2017a; Eger *et al.*, 2017], researchers have recently begun to study new argument mining tasks. Below we give an overview of these tasks.

Persuasiveness-related tasks. Most related to our study is work involving argument persuasiveness. For instance, Habernal and Gurevych [2016] 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, but they do not examine *why* an argument is (un)persuasive. 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. Hidey *et al.* [2017] examine the different semantic types of claims and premises. Higgins and Walker [2012] investigate persuasion strategies (i.e., ethos, pathos, logos). 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 focuses on selecting the correct warrant that explains reasoning of an argument that consists of a claim and a reason.[5]

6 Conclusion

We designed the first set of neural models for predicting the persuasiveness of an argument and its attributes in a student

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

	Pipeline									Joint								
	Evid.	**Eloq.**	**Spec.**	**Stre.**	**Log.**	**Path.**	**Eth.**	**CT**	**PT**	**Evid.**	**Eloq.**	**Spec.**	**Stre.**	**Log.**	**Path.**	**Eth.**	**CT**	**PT**
U	.353	.126	.284	.132	.281	.045	.045	.692	.200	.335	.106	.243	.697	.281	.863	.967	.695	.207
UF	.309	.206	.417	.132	.281	.045	.045	.692	.210	.340	.178	.409	.697	.281	.829	.961	.676	.189
UA	.347	.148	.306	.132	.281	.045	.045	.692	.308	.366	.168	.332	.697	.281	.967	.967	.692	.268
UFA	.368	.202	.414	.132	.281	.045	.045	.692	.239	.341	.173	.429	.697	.281	.868	.967	.692	.312

Table 8: Attribute prediction results of different variants of Pipeline and Joint on the test set.

essay. While the results are promising, they also suggest that the performance of our models could be substantially improved by improving attribute prediction.

References

[Al Khatib *et al.*, 2016] Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. *COLING*, 2016.

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[Burstein *et al.*, 2004] Jill Burstein, Martin Chodorow, and Claudia Leacock. Automated essay evaluation: The Criterion online writing evaluation service. *AI Magazine*, 25(3):27–36, 2004.

[Carlile *et al.*, 2018] Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. *ACL*, 2018.

[Eger *et al.*, 2017] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. *ACL*, 2017.

[Habernal and Gurevych, 2016] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. *ACL*, 2016.

[Hidey *et al.*, 2017] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. *Fourth Workshop on Argument Mining*, 2017.

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

[Higgins *et al.*, 2004] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating multiple aspects of coherence in student essays. *HLT-NAACL*, 2004.

[Lukin *et al.*, 2017] Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. *EACL*, 2017.

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

[Persing and Ng, 2013] Isaac Persing and Vincent Ng. Modeling thesis clarity in student essays. *ACL*, 2013.

[Persing and Ng, 2014] Isaac Persing and Vincent Ng. Modeling prompt adherence in student essays. *ACL*, 2014.

[Persing and Ng, 2015] Isaac Persing and Vincent Ng. Modeling argument strength in student essays. *ACL/IJCNLP*, 2015.

[Persing and Ng, 2017] Isaac Persing and Vincent Ng. Why can't you convince me? Modeling weaknesses in unpersuasive arguments. *IJCAI*, 2017.

[Persing *et al.*, 2010] Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. *EMNLP*, 2010.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Stab and Gurevych, 2014a] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. *COLING*, 2014.

[Stab and Gurevych, 2014b] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. *EMNLP*, 2014.

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

[Stab and Gurevych, 2017b] Christian Stab and Iryna Gurevych. Recognizing insufficiently supported arguments in argumentative essays. *EACL*, 2017.

[Tan *et al.*, 2016] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *WWW*, 2016.

[van Eemeren *et al.*, 2014] F. H. van Eemeren, B. Garssen, E. C. W. Krabbe, F. A. Snoeck Henkemans, B. Verheij, and J. H. M. Wagemans. In *Handbook of Argumentation Theory*. Springer, Dordrecht, 2014.

[Wachsmuth *et al.*, 2017] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. *EACL*, 2017.

[Wei *et al.*, 2016] Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? Ranking argumentative comments in online forum. *ACL*, 2016.