Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments

Isaac Persing and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{persingq,vince}@hlt.utdallas.edu

Abstract

Recent work on argument persuasiveness has focused on determining *how* persuasive an argument is. Oftentimes, however, it is equally important to understand *why* an argument is unpersuasive, as it is difficult for an author to make her argument more persuasive unless she first knows what errors made it unpersuasive. Motivated by this practical concern, we (1) annotate a corpus of debate comments with not only their persuasiveness scores but also the errors they contain, (2) propose an approach to persuasiveness scoring and error identification that outperforms competing baselines, and (3) show that the persuasiveness scores computed by our approach can indeed be explained by the errors it identifies.

1 Introduction

The recent surge of interest in computational argumentation stems in part from the prevalence of arguments in both formally and informally written texts in a variety of genres, ranging from persuasive student essays to posts and comments in online debate forums (see van Eemeren *et al.* [2014], Baroni *et al.* [2016], and Saint-Dizier and Stede [2016] for a general overview of the field). While traditional work on argument mining in the natural language processing (NLP) community has focused on extracting argument components (e.g., claims, premises) and determining the relationships (e.g., support, attack) between them, researchers have begun work on a variety of tasks that involve the *persuasiveness* of arguments.

Recent work on argument persuasiveness has focused on determining *how* persuasive an argument is. One fundamental yet under-investigated task that falls in this category is *argument persuasiveness scoring*. Given a text consisting of an argument written for a particular topic, the goal of argument persuasiveness scoring is to assign a score to the text that indicates how persuasive the argument is. To better understand how difficult this task is, consider the debate argument shown in Table 1. Written in response to a debate motion, this argument consists of an assertion and its justification. While it is fairly easy for a human to determine that this argument is not persuasive, the same is not true for a machine because none of the words in the argument provide suggestive evidence of its persuasiveness. To illustrate the practical imporance of this task, consider the case of an online debate, where an author's primary goal is to convince others of the argument expressed in her comment(s). It is similarly important in persuasive essay writing for an author to construct convincing arguments. Hence, an argument persuasiveness scoring system could provide useful feedback in many situations.

Oftentimes, however, it is important to determine not only how (un)persuasive an argument is, but also *why* it is not persuasive. Without knowing what errors contribute to its lack of persuasiveness, it is difficult for its author to understand what revisions are needed to make her argument more persuasive.

Our goal in this paper is to fill this gap in current research on argument persuasiveness. Given a debate argument, we seek to not only score its persuasiveness but also determine the errors that adversely affect its persuasiveness. Specifically, our work makes the following contributions. First, we identify five major classes of error that could negatively impact persuasiveness, annotate a corpus of 1,208 debate comments with argument persuasiveness scores and the errors they contain, and empirically show using these manual annotations that unpersuasiveness can be largely explained by these five errors. Second, we present an automatic approach to persuasiveness scoring and error identification that significantly outperforms recently proposed baselines. Third, we show that these automatically computed persuasiveness scores can be largely explained by the errors we automatically identify from the corresponding arguments. This is important because, if our system assigns a low persuasiveness score to an argument, but the score cannot be adequately explained by the errors the system identifies, the argument's author cannot get a clear understanding of why the score is low and how to improve it. Finally, we make our annotated dataset publicly available.[1] Given the difficulty of obtaining annotated data for this task, we believe our dataset will be a valuable resource to the NLP community.

2 Related Work

There have been several attempts to address tasks related to argument persuasiveness. Habernal and Gurevych [2016a; 2016b] *rank* a pair of arguments w.r.t. persuasiveness, but

[1] See `http://www.hlt.utdallas.edu/~persingq/Debate/` for a complete list of our annotations.

Motion	This House would ban teachers from interacting with students via social networking websites.
Assertion	Acting as a warning signal for children at risk.
Justification	It is very difficult for a child to realize that he is being groomed; they are unlikely to know the risk. After all, a teacher is regarded as a trusted adult. But, if the child is aware that private electronic contact between teachers and students is prohibited by law, the child will immediately know the teacher is doing something he is not supposed to if he initiates private electronic contact. This will therefore act as an effective warning sign to the child and might prompt the child to tell a parent or another adult about what is going on.

Table 1: The motion, assertion, and justification text of a debate argument.

ranking alone cannot tell us *how* persuasive an argument is. Persing and Ng [2015] score a student essay based on whether it makes a (un)convincing argument for its thesis. Using the conversations in the ChangeMyView subreddit, Tan *et al.* [2016] study factors affecting whether a challenger can successfully persuade a commenter to change the view she expressed in her original post. Both Tan et al. and Persing and Ng perform feature analyses that could help understand which features correlate with (un)persuasive arguments. However, as we will see, our results show their features are insufficient for our argument persuasiveness scoring task.

Like us, Wei *et al.* [2016] predict the persuasiveness of debate posts, but differ from us in several aspects. First, many of their debate posts are written in response to a preceding comment in the conversation. Hence, it is not uncommon to see emotional rather than logical arguments or even insults and personal attacks. In addition, it may not always be possible to understand what the argument is and why the author made a particular argument without understanding the (preceding) context. On the other hand, the debate comments in our corpus are written in response to a given debate topic. In other words, each comment is written independently of the other comments and therefore can be understood without them.

3 Corpus and Annotation

We use as our corpus a randomly selected subset of 165 debates obtained from the International Debate Education Association (IDEA) website[2]. These debates cover a wide range of topics including politics, economics, religion, and science. Each debate consists of a *M*otion, which expresses a stance on the debate's topic, and an average of 7.3 arguments, each of which either agrees or disagrees with the motion's stance. Each of the 1,208 arguments consists of an *A*ssertion, which expresses in one sentence why the author agrees or disagrees with the motion, and a *J*ustification, which explains in an average of 6.9 sentences why the author believes her assertion.

We ask two native speakers of English to annotate each argument w.r.t. (1) its persuasiveness score, and (2) whether its author made any of five errors that could have a negative impact on its persuasiveness. These five errors are motivated by theoretical research on argument persuasiveness. Below we define the persuasiveness score and these errors in detail.

Argument Persuasiveness (AP) We ask our annotators to score each argument's persuasiveness on a scale of $1-6$, with a score of 6 indicating a very persuasive argument, and a score of 1 indicating an unclear or missing argument, according to the scoring rubric shown in Table 2. The example argument in Table 1 gets an AP score of 2 because it is not clear what the author is trying to argue.

Grammar Error (GE) Connor and Lauer [1985] noted that grammar errors can interrupt the flow of discourse in persuasive essays, thereby reducing their coherence. For this reason, we give arguments a GE score of 1 if they contain GEs that are severe enough to make the argument hard to understand, and 0 otherwise. The example argument gets a GE score of 0 because it contains no severe GEs.

Lack of Objectivity (LO) Oktavia and Yasin [2014] identify it as a fallacy when students flatly state their personal opinions as evidence for a claim they are trying to support in argumentative writing. For this reason, we give arguments a LO score of 1 if they display an inappropriate lack of objectivity, and 0 otherwise. So the example argument receives a LO score of 1 because the author weaves a scenario in which she repeatedly speculates on what a child thinks or will do.

Inadequate Support (IS) Petty and Cacioppo [1984] and Maddux and Rogers [1980] found that arguments with more support are more persuasive, so we give arguments an IS score of 0 if they offer adequate support, 1 if they don't offer enough support to justify their assertion, or 2 if they offer almost no support. The example argument gets an IS score of 2 because the author's scenario is completely unsupported.

Unclear Assertion (UA) In Connor's [1990] criteria for judging assertions in persuasive writing, the lowest score is assigned to essays which did not clearly assert the problem they address. So we give an argument a UA score of 1 if it is not clear how the assertion is related to the motion without also reading the justification, or 2 if the assertion is incomprehensible without also reading the justification. It receives a score of 0 otherwise. So the example argument receives a UA score of 1.

Unclear Justification (UJ) Because a smooth flow of ideas throughout an argument is important to its persuasiveness, Connor [1990] also evaluated persuasive essays' coherence. Since it is not clear what an incoherent argument is arguing for, we give it a UJ score of 1 if it does not concisely justify the assertion, or 2 if the justification appears unrelated to the assertion. The example argument gets a UJ score of 0, as it is easy to understand the author's point in the justification.

Table 3 shows the distributions of scores for argument persuasiveness and each error. To measure inter-annotator agreement, we select a subset of 69 arguments and ask both annotators to score them w.r.t. argument persuasiveness and all the error classes. We measure the average difference between the annotator-assigned scores to obtain the disagreement levels shown in Table 4. For the sake of our experiments, when annotators disagree on a score, we average their annotations

[2] http://idebate.org/

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.

	0	1	2	3	4	5	6
GE	98	02					
LO	76	24					
IS	49	35	16				
UA	57	32	11				
UJ	58	39	03				
AP		03	12	20	21	20	24

Table 3: Distribution of error and argument persuasiveness scores as percentages.

GE	LO	IS	UA	UJ	AP
.029	.087	.319	.580	.391	.899

Table 4: Average difference between the argument persuasiveness and error scores assigned by two annotators.

GE	LO	IS	UA	UJ	Bias
−0.9	−1.0	−0.9	−0.9	−1.0	5.9

Table 5: Relative importance of error classes to argument persuasiveness score.

together, rounding up to the nearest whole number to obtain the gold score.

To gain insights into the relative importance the different types of errors play in reducing an argument's persuasiveness score, we construct a training set out of our corpus, representing each argument as an instance whose label is the argument's gold argument persuasiveness score. Five of its features are the argument's gold error scores, and its last feature is a bias feature. We then train a linear support vector regressor (SVR) on these instances using the LIBSVM software package [Chang and Lin, 2001] with default parameters.

As we can see from Table 5, the five errors have similarly high (negative) impacts on argument persuasiveness. The fact that the bias is close to 6 suggests that the regressor successfully learned that an argument with no errors should have a perfect score, and so the five error classes account for most of the variance in persuasiveness scores.

4 Approach

We cast the task of predicting an argument's error and argument persuasiveness scores as six independent regression problems. Given an error or argument persuasiveness problem, each argument in the training set is represented as an instance whose label is the argument's gold score for that problem. Each instance is in turn represented by 11 feature types. After creating training instances, we train a linear SVR implemented in LIBSVM with default parameters. We then use the resulting regressor to score the test set arguments. Test instances are created in the same way as the training instances. Below we describe the 11 feature types used to represent each training/test instance.

1. #grammar error Our first feature encodes the per sentence grammar error frequency in an argument's justification. To detect these grammar errors, we use the LanguageTool proofreading program[3]. This feature would be useful for predicting GEs.

2. #subjectivity indicators This group of features encodes the frequencies of the words "morally", "certain", and "perhaps" per token in the justification, as arguments that are too concerned with the author's morality or in which the author seems too certain of herself are likely to display a lack of objectivity.

3. #definite articles This feature encodes the count of definite articles (i.e. "the") appearing in the justification. This feature makes sense since an article with few definite articles usually lacks specificity, and thus may also be too subjective.

4. #first person plural pronouns This feature encodes the count of first person plural pronouns appearing in an argument's justification. The rationale for this feature is that justifications that lack objectivity often rely on stories about the writer's personal experiences. We use plural pronouns to capture this rather than singular ones because thesis statements (which aren't inappropriately subjective) often begin with "I believe" or "I think".

5. #citations In our corpus, most arguments cite sources for the claims they make. An argument that cites no sources is likely to inadequately support its assertion. For this reason, we employ a count of citations in a justification as a feature.[4]

6. #content lemmas only in justification Justifications with adequate support usually make a variety of points to support their assertion, as opposed to making a sequence of flat statements about the topic (e.g., on the topic "Homework is a waste of time", one arguer wrote, "Time is valuable. We all

[3] `https://languagetool.org/`

[4] We use heuristics to extract references from the justification.

need some time to ourselves. School already takes up a lot of time and it is necessary to have time which does not involve concentrating on learning..."). For this reason, we encode as a feature the number of content lemmas (nouns, pronouns, verbs, adjectives, adverbs) appearing in the justification that do not appear in the motion.

7. Assertion length UAs typically consist of very short sentence fragments (e.g. "Europe"). For this reason, we incorporate as a feature the assertion's length in words.

8. #content lemmas only in assertion We encode as a feature the number of content lemmas that appear in the assertion but not the justification. This feature estimates whether the assertion is topically similar to the argument being made in the justification. If not, the assertion could be unclear.

9. Justification length As with UAs, UJs are often very short. For this reason, we encode the count of sentences in the justification as a feature.

10. #subject matches in discourse relation We employ as a feature the number of times words that lemmatically match the assertion's subject appear in the first argument of a contingency-cause discourse relation in the justification. A justification that discusses its assertion's topic's effects frequently is likely to be very topically coherent, making the justification clearer. We identify subjects and discourse relations using Stanford CoreNLP [Manning *et al.*, 2014] and Lin et al.'s [2014] PDTB-style discourse parser, respectively.

11. #strong thesis statements The presence of a strong thesis statement can make a justification clearer. For this reason, we construct a feature that counts the frequency of statements wherein the writer states that she cognizes, speaks, perceives, or believes something, identifying the existence of the corresponding semantic frames using SEMAFOR [Das *et al.*, 2010].

5 Evaluation

In this section, we evaluate our approach to error and persuasiveness prediction.

5.1 Scoring Metrics

We employ three evaluation metrics for error and persuasiveness scoring, namely E, ME, and PC. The simplest metric, E, measures the frequency at which a system predicts the wrong score. When evaluating by E, we round predicted scores to the nearest valid score (e.g., $1-6$ at one-point increments for persuasiveness). ME measures the mean distance between a system's prediction and the gold score. The formulas below illustrate how we calculate E and ME:

$$\frac{1}{N}\sum_{A_j \neq E'_j} 1, \quad \frac{1}{N}\sum_{j=1}^{N} |A_j - E_j|,$$

where A_j, E_j, and E'_j are the annotator assigned, system predicted, and rounded system predicted scores respectively for argument j, and N is the number of arguments.

The last metric, PC, computes Pearson's correlation coefficient between a system's predicted scores and the annotator assigned scores. A positive (negative) PC implies that the two sets of predictions are positively (negatively) correlated.

Note that E and ME are *error* metrics, so lower scores on them imply better performance. In contrast, PC is a *correlation* metric, so higher correlation implies better performance.

5.2 Baseline Systems

We employ six baseline systems. All baselines employ error and persuasion prediction SVRs differing from those described in Section 4 only in terms of the features used by the learner.

Bag of words (BOW) In the first baseline, we use as features the bag of words (BOW) extracted from the argument's assertion and justification.

Word n-grams (WNG) The second baseline uses word n-grams (n=1,2,3) extracted from the argument's assertion and justification as features.

Bag of part-of-speech tags (BOPOS) Our third baseline employs as features the bag of part-of-speech (POS) tags in the argument's assertion and justification.

Style Our fourth baseline captures aspects of an argument's style. It employs four types of features that are motivated by Tan et al.'s [2016] Style baseline, as described below.

Length-based features encode the length in tokens and sentences of an argument's assertion and justification.

Word category-based features encode the absolute count and frequency per token in an argument's justification for each of the following categories of words/tokens: (1) definite and indefinite articles and first and second person pronouns, both of which can be useful for detecting lack of objectivity; (2) question marks and quotations, which indicate how an argument is structured; (3) positive and negative sentiment words as determined by Mohammad and Yang [2011] since excessive emotion can also signal a lack of objectivity; (4) URLs, since these may be another way of citing evidence; (5) hedge words[5], which can be used to express argument uncertainty; and (6) phrases that indicate the author is giving an example ("e.g.", "for instance", "for example").

Word complexity features capture the justification's complexity of word choice, namely its word entropy, type-token ratio, and grade level [Kincaid *et al.*, 1975].

Word score-based features indicate the average concreteness, arousal, valence, and dominance of content words in an argument's justification as described in Warriner *et al.* [2013] and Brysbaert *et al.* [2014]. They are intended to capture how abstract, intensely emotional, pleasant, and vulnerability-evoking an argument is.

Duplicated Tan et al. (Tan) As our fifth baseline, we employ our re-implementation of Tan et al.'s [2016] system. Their feature set comprises all the features described in the Style, BOW, and BOPOS baselines. Their system additionally employs a set of word score-based features exactly like those described above, except that they involve first quartering the justification, then calculating the word scores on each quarter of the text. These are useful because, for example, successful arguments begin by using calmer words.[6]

[5] From `http://english-language-skills.com/item/177writing-skills-hedge-words.html`

[6] Tan et al. employ two types of features that are inapplicable to our corpus. First, their interaction features capture the interaction

	System	GE	LO	IS	UA	UJ	AP
E	WNG	**.022**	**.242**	.650	**.429**	**.426**	.786
	BOW	**.022**	**.242**	.650	**.429**	**.426**	.786
	BOPOS	**.022**	**.242**	.593	**.429**	**.426**	.786
	Style	**.022**	**.242**	.515	**.465**	**.427**	.748
	Tan	**.022**	**.242**	.494	**.456**	**.425**	.744
	P&N	**.022**	**.242**	.531	**.435**	**.441**	.785
	OUR	**.022**	**.242**	**.439**	**.469**	**.431**	**.721**
ME	WNG	.118	.294	.653	**.550**	**.472**	1.218
	BOW	.117	.294	.654	**.551**	**.473**	1.218
	BOPOS	.118	.294	.620	**.551**	**.472**	1.217
	Style	**.106**	**.283**	.547	**.563**	**.476**	1.102
	Tan	**.103**	**.282**	.537	**.517**	**.478**	1.109
	P&N	.115	.293	.607	**.546**	**.476**	1.198
	OUR	.115	.291	**.472**	**.561**	**.474**	**1.036**
PC	WNG	**.006**	.033	.113	.042	.029	.063
	BOW	**-.009**	.082	.124	.060	.036	.073
	BOPOS	**-.070**	.007	.242	.084	.003	.089
	Style	**-.044**	**.221**	.412	**.124**	**.187**	.408
	Tan	**.028**	**.234**	.439	**.169**	**.171**	.398
	P&N	**.034**	.085	.206	.086	**.116**	.252
	OUR	**.004**	**.222**	**.595**	**.241**	**.205**	**.488**

Table 6: System performances on AP and the five errors (GE, LO, IS, UA, and UJ) as measured by the three scoring metrics (E, ME, and PC).

Persing and Ng (P&N) Our last baseline is Persing and Ng's [2015] system. P&N employs five types of features: (1) POS n-grams (n=1,2,3), which capture the syntactic generalizations of an argument's justification; (2) frame-semantic features, which capture the semantic generalizations of the justification; (3) features computed based on the frequency of transitional phrases in the justification, which encode its degree of coherence; (4) topic relevance features, which capture the relevance of the justification to its motion based on the number of overlapping entities; and (5) argument label features, which are n-grams of sentence-based argument labels (e.g., CLAIMS, SUPPORT) derived from the justification.[7]

5.3 Results and Discussion

Five-fold cross-validation[8] results of the six baselines and OUR approach on the AP scoring task and the five error severity tasks as measured by E, ME, and PC are shown in the three subtables of Table 6.[9]

between different users in a conversation. Since each argument in our corpus is written independently of other arguments, interaction features are not applicable to our arguments. Second, they capture formatting features (e.g., bullet lists) which our corpus lacks.

[7]We exclude their features based on argument component predictions because (1) our arguments are much shorter than student essays, and (2) our "major claims" are explicitly stated outside the body of the argument (in the assertion field).

[8]To ensure generalizability across new topics, we distribute arguments into folds based on the motions they respond to. So we never train on an argument and test on another argument written in response to the same motion.

[9]Boldfaced results are considered the best in their column, as they are not significantly different than the best result in their column (paired t-tests with $p > .05$).

	E		ME		PC	
System	SF	EF	SF	EF	SF	EF
WNG	.786	.786	1.218	1.218	.063	.060
BOW	.786	.786	1.218	1.220	.073	.073
BOPOS	.786	.785	1.217	1.224	.089	.093
Style	.748	.735	1.102	1.094	.408	.426
Tan	.744	.730	1.109	1.106	.398	.410
P&N	.785	.783	1.198	1.195	.252	.259
OUR	**.721**	**.708**	**1.036**	**1.027**	**.488**	**.495**

Table 7: Difference in persuasiveness performance when using a system's normal features (SF) and when using error prediction features (EF).

Consider first the AP scoring results. While BOW and WNG serve as strong baselines for many NLP tasks, the same is not true for AP scoring: they, together with BOPOS, are among the worst baselines. This is perhaps not surprising given the discussion of the running example in the introduction: in many cases an argument's persuasiveness and errors cannot be determined solely from its words and phrases. This is further reinforced by the results on the baselines developed explicitly for AP scoring tasks, Style, Tan, and P&N. These baselines fare much better, outperforming all the more general baselines by most metrics. OUR approach, by contrast, significantly outperforms the best baseline on the AP scoring task as measured by all three scoring metrics.

Next, consider the results on the five error severity prediction tasks. As we can see, OUR system's results are consistently strong. In fact, except when GE and LO were evaluated w.r.t. the ME metric, OUR system offered the strongest performance. Even in the two cases when OUR system is outperformed, its ME score on GE is lower than the best score only by 0.012, and its ME score on LO is lower than the best score only by 0.009.

5.4 Additional Experiments

Explanatory Power of the Error Classes

So far we have predicted the AP scores and the errors independently of each other. This means that it is possible that no errors are identified in an argument that is predicted to have a low AP score. This is not ideal because the possible lack of correlation between the predicted errors and the AP scores defeats our original goal of using the errors to help the user understand why an AP score is low.

To address this problem, we perform a set of experiments to determine how well the *predicted* errors can score persuasiveness. For each of these experiments, we train a persuasiveness SVR that employs as features only the error severity predictions outputted by a set of 5 error prediction SVRs.[10] So for example, rather than training a persuasiveness SVR on word n-grams, as is done in row 1 columns 1, 3, and 5 (the columns labeled SF) of Table 7, we train 5 SVRs for predicting the 5 errors using word n-gram features. Then, for each argument, we extract the predicted error values as outputted by the error regressors. We employ these predicted values as

[10]We obtain error severity predictions on the training arguments by performing five-fold cross-validation over the training set.

	E		ME		PC	
%Per	SF	EF	SF	EF	SF	EF
10	.723	.732	1.089	1.101	.427	.418
20	.730	.729	1.068	1.059	.455	.464
30	.729	.723	1.043	1.037	.485	.494
40	.729	.731	1.048	1.052	.482	.476
50	.726	.723	1.054	1.041	.470	.484
60	.722	.716	1.035	1.033	.488	.486
70	.725	.715	1.037	1.038	.485	.485
80	.724	.713	1.036	1.032	.491	.491
90	.722	.712	1.034	1.030	.493	.492
100	.721	.708	1.036	1.027	.488	.495

Table 8: Learning curves for OUR system's performance on persuasiveness prediction when using its normal features (SF) and when using its error prediction features (EF) on 10% to 100% of the available training data.

GE	LO	IS	UA	UJ	Bias
−.016	−.154	−1.253	−.822	−1.191	6.009

Table 9: Weights learned by OUR linear SVR when trained on error class predictions (previously EF).

5 features for a new persuasiveness regressor, whose results are shown in columns 2, 4, and 6 (the columns labeled EF).[11]

Results using each of the seven different feature sets as a base w.r.t. the three scoring metrics are shown in Table 7. We can see by comparing the original persuasiveness results from the SF columns with the new persuasiveness results from the EF columns that, regardless of whether a feature set is used to directly predict persuasiveness, or whether it is used to predict error severities which in turn are used to predict persuasiveness, the score remains roughly unchanged. In fact, for the better systems (Tan, Style, and OUR), the EF scores are better than the corresponding SF scores. This suggests that the predicted errors capture all information relevant to persuasiveness prediction from any feature set.

Learning Curves

Table 8 shows two learning curves for AP scoring. For each of these experiments, we train a persuasiveness SVR using either OUR features (SF) or error features predicted by error SVRs which in turn use OUR features (EF).

With just one exception (10%, E), the EF scores are nearly as good (within 0.010 points) as the corresponding SF scores. These results imply that, no matter how much training data was used to train the error predictors and the persuasiveness predictors, the predicted errors capture all the information pertinent to AP scoring from the original (i.e., SF) features.

Comparing the second line of results to the baseline results in Table 7, we also notice that OUR system outperforms all baselines regardless of what setting (SF or EF) or scoring metric is used even when it is trained on only 20% of persuasiveness-annotated data.

[11]Since the error regressors can predict non integral severity levels, we round each prediction to the nearest valid severity level.

Error	E	ME	PC
GE	.708	1.027	.495
LO	.708	1.027	.495
IS	.719	1.063	.448
UA	.729	1.047	.478
UJ	.717	1.028	.494
–	.708	1.027	.495

Table 10: Persuasiveness results when OUR system is trained using all but one error feature (previously EF). First column shows which error feature is removed. The last row shows the results of OUR system when no features are removed.

Feature Analysis

In Table 9, we see the weights assigned by the SVR to each of the error features for the EF version of OUR system when trained on all of the labeled persuasiveness data. From the bias feature weight, we can tell that the SVR successfully learned that an argument with no errors should have a perfect persuasiveness score. The bias feature weight also suggests that persuasiveness can nearly be completely accounted for by the five predicted errors. We also notice that the weight the regressor assigns an error is greatest (in absolute value) when the error appears frequently in the text (over 40% of arguments contain IS, UA, or UJ errors), and to a lesser extent when OUR error prediction SVR is better at predicting it (OUR PC score on IS is greatest, and the weight the regressor puts on the IS feature is the greatest of all error weights).

Another way we analyze the contribution of each predicted error feature to OUR system's EF performance is with feature ablation experiments. That is, for each row of Table 10, we train a regressor on all of OUR EF features except the prediction corresponding to the error shown in the leftmost column. We display the system's performance w.r.t. the three error metrics in the row. From these results, we see that the error prediction features whose removal hurts OUR system's performance the most w.r.t. all metrics (IS and UA) are also the systems on which OUR error prediction SVRs have the greatest PC performance. This suggests that if we could improve the predictiveness of our error classifiers on the remaining errors, OUR EF performance might improve as well.

6 Conclusion and Future Work

Unlike previous work on argument persuasiveness, we examined not only *how* unpersuasive an argument is but also *why* an argument is unpersuasive. Results on 1,208 arguments showed that our approach significantly outperformed six baselines w.r.t. the persuasiveness scoring task and most of the error severity prediction tasks. To stimulate research on these tasks, we make our annotated data publicly available. In future work, we plan to conduct a user study to determine how useful the errors identified by our system are for helping authors understand how to make their arguments persuasive.

Acknowledgments

We thank the three reviewers and our annotators, Dino Occhialini and Christopher Knoll. This work was supported in part by NSF Grants IIS-1219142 and IIS-1528037.

References

[Baroni *et al.*, 2016] Pietro Baroni, Thomas F. Gordon, Tatjana Scheffler, and Manfred Stede, editors. *Computational Models of Argument – Proceedings of COMMA 2016*. IOS Press, Amsterdam, 2016.

[Brysbaert *et al.*, 2014] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014.

[Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Connor and Lauer, 1985] Ulla Connor and Janice Lauer. Understanding persuasive essay writing: Linguistic/rhetorical approach. *Text-Interdisciplinary Journal for the Study of Discourse*, 5(4):309–326, 1985.

[Connor, 1990] Ulla Connor. Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, pages 67–87, 1990.

[Das *et al.*, 2010] Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. Probabilistic frame-semantic parsing. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 948–956, 2010.

[Habernal and Gurevych, 2016a] Ivan Habernal and Iryna Gurevych. 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.

[Habernal and Gurevych, 2016b] Ivan Habernal and Iryna Gurevych. 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.

[Kincaid *et al.*, 1975] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for Navy enlisted personnel. Technical report, DTIC Document, 1975.

[Lin *et al.*, 2014] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184, 2014.

[Maddux and Rogers, 1980] James E. Maddux and Ronald W. Rogers. Effects of source expertness, physical attractiveness, and supporting arguments on persuasion: A case of brains over beauty. *Journal of Personality and Social Psychology*, 39(2):235, 1980.

[Manning *et al.*, 2014] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[Mohammad and Yang, 2011] Saif Mohammad and Tony Yang. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 70–79, 2011.

[Oktavia and Yasin, 2014] Witri Oktavia and Anas Yasin. An analysis of students' argumentative elements and fallacies in students' discussion essays. *English Language Teaching*, 2(3), 2014.

[Persing and Ng, 2015] Isaac Persing and Vincent Ng. 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, 2015.

[Petty and Cacioppo, 1984] Richard E. Petty and John T. Cacioppo. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46(1):69, 1984.

[Saint-Dizier and Stede, 2016] Patrick Saint-Dizier and Manfred Stede, editors. *Proceedings of the COMMA Workshop on Foundations of the Language of Argumentation*. University of Potsdam, Germany, 2016.

[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. In *Proceedings of the 25th International World Wide Web Conference*, pages 613–624, 2016.

[van Eemeren *et al.*, 2014] Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, Francisca A. Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. Chapter 11: Argumentation and artificial intelligence. In *Handbook of Argumentation Theory*. Springer, Dordrecht, 2014.

[Warriner *et al.*, 2013] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

[Wei *et al.*, 2016] Zhongyu Wei, Yang Liu, and Yi Li. 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, 2016.