Argumentation Quality Assessment: Theory vs. Practice

Henning Wachsmuth [*] **Nona Naderi** [**] **Ivan Habernal** [***] **Yufang Hou** [****]
Graeme Hirst [**] **Iryna Gurevych** [***] **Benno Stein** [*]
[*] Bauhaus-Universität Weimar, Weimar, Germany, www.webis.de
[**] University of Toronto, Toronto, Canada, www.cs.toronto.edu/compling
[***] Technische Universität Darmstadt, Darmstadt, Germany, www.ukp.tu-darmstadt.de
[****] IBM Research, Dublin, Ireland, ie.ibm.com

Abstract

Argumentation quality is viewed differently in argumentation theory and in practical assessment approaches. This paper studies to what extent the views match empirically. We find that most observations on quality phrased spontaneously are in fact adequately represented by theory. Even more, relative comparisons of arguments in practice correlate with absolute quality ratings based on theory. Our results clarify how the two views can learn from each other.

1 Introduction

The assessment of argumentation quality is critical for any application built upon argument mining, such as debating technologies (Rinott et al., 2015). However, research still disagrees on whether quality should be assessed from a theoretical or from a practical viewpoint (Allwood, 2016).

Theory states, among other things, that a cogent argument has acceptable premises that are relevant to its conclusion and sufficient to draw the conclusion (Johnson and Blair, 2006). Practitioners object that such quality dimensions are hard to assess for real-life arguments (Habernal and Gurevych, 2016b). Moreover, the normative nature of theory suggests *absolute* quality ratings, but in practice it seems much easier to state which argument is more convincing—a *relative* assessment. Consider two debate-portal arguments for "advancing the common good is better than personal pursuit", taken from the corpora analyzed later in this paper:

Argument A *"While striving to make advancements for the common good you can change the world forever. Allot of people have succeded in doing so. Our founding fathers, Thomas Edison, George Washington, Martin Luther King jr, and many more. These people made huge advances for the common good and they are honored for it."*

Argument B *"I think the common good is a better endeavor, because it's better to give then to receive. It's better to give other people you're hand out in help then you holding your own hand."*

In the study of Habernal and Gurevych (2016b), annotators assessed Argument *A* as more convincing than *B*. When giving reasons for their assessment, though, they saw *A* as more credible and well thought through; that does not seem to be too far from the theoretical notion of cogency.

This paper gives empirical answers to the question of how different the theoretical and practical views of argumentation quality actually are. Section 2 briefly reviews existing theories and practical approaches. Section 3 then empirically analyzes correlations in two recent argument corpora, one annotated for 15 well-defined quality dimensions taken from theory (Wachsmuth et al., 2017a) and one with 17 reasons for quality differences phrased spontaneously in practice (Habernal and Gurevych, 2016a). In a crowdsourcing study, we test whether lay annotators achieve agreement on the theoretical quality dimensions (Section 4).

We find that assessments of overall argumentation quality largely match in theory and practice. Nearly all phrased reasons are adequately represented in theory. However, some theoretical quality dimensions seem hard to separate in practice. Most importantly, we provide evidence that the observed relative quality differences are reflected in absolute quality ratings. Still, our study underpins the fact that the theory-based argumentation quality assessment remains complex. Our results do not generally answer the question of what view of argumentation quality is preferable, but they clarify where theory can learn from practice and vice versa. In particular, practical approaches indicate what to focus on to simplify theory, whereas theory seems beneficial to guide quality assessment in practice.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 250–255
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2039

Quality Dimension	Short Description of Dimension
Cogency	Argument has (locally) acceptable, relevant, and sufficient premises.
Local acceptability	Premises worthy of being believed.
Local relevance	Premises support/attack conclusion.
Local sufficiency	Premises enough to draw conclusion.
Effectiveness	Argument persuades audience.
Credibility	Makes author worthy of credence.
Emotional appeal	Makes audience open to arguments.
Clarity	Avoids deviation from the issue, uses correct and unambiguous language.
Appropriateness	Language proportional to the issue, supports credibility and emotions.
Arrangement	Argues in the right order.
Reasonableness	Argument is (globally) acceptable, relevant, and sufficient.
Global acceptability	Audience accepts use of argument.
Global relevance	Argument helps arrive at agreement.
Global sufficiency	Enough rebuttal of counterarguments.
Overall quality	Argumentation quality in total.

Table 1: The 15 theory-based quality dimensions rated in the corpus of Wachsmuth et al. (2017a).

Polarity	Label	Short Description of Reason
Negative properties of Argument B	5-1	*B* is attacking / abusive.
	5-2	*B* has language/grammar issues, or uses humour or sarcasm.
	5-3	*B* is unclear / hard to follow.
	6-1	*B* has no credible evidence / no facts.
	6-2	*B* has less or insufficient reasoning.
	6-3	*B* uses irrelevant reasons.
	7-1	*B* is only an opinion / a rant.
	7-2	*B* is non-sense / confusing.
	7-3	*B* does not address the topic.
	7-4	*B* is generally weak / vague.
Positive properties of Argument A	8-1	*A* has more details/facts/examples, has better reasoning / is deeper.
	8-4	*A* is objective / discusses other views.
	8-5	*A* is more credible / confident.
	9-1	*A* is clear / crisp / well-written.
	9-2	*A* sticks to the topic.
	9-3	*A* makes you think.
	9-4	*A* is well thought through / smart.
Overall	**Conv**	*A* is more convincing than *B*.

Table 2: The 17+1 practical reason labels given in the corpus of Habernal and Gurevych (2016a).

2 Theory versus Practice

This section outlines major theories and practical approaches to argumentation quality assessment, including those we compare in the present paper.

2.1 Theoretical Views of Quality Assessment

Argumentation theory discusses logical, rhetorical, and dialectical quality. As few real-life arguments are logically sound, requiring true premises that deductively entail a conclusion, cogency (as defined in Section 1) is largely seen as the main logical quality (Johnson and Blair, 2006; Damer, 2009; Govier, 2010). Toulmin (1958) models the general structure of logical arguments, and Walton et al. (2008) analyze schemes of fallacies and strong arguments. A fallacy is a kind of error that undermines reasoning (Tindale, 2007). Strength may mean cogency but also rhetorical effectiveness (Perelman and Olbrechts-Tyteca, 1969). Rhetoric has been studied since Aristotle (2007) who developed the notion of the means of persuasion (logos, ethos, pathos) and their linguistic delivery in terms of arrangement and style. Dialectical quality dimensions resemble those of cogency, but arguments are judged specifically by their reasonableness for achieving agreement (van Eemeren and Grootendorst, 2004).

Wachsmuth et al. (2017a) point out that dialectical builds on rhetorical, and rhetorical builds on logical quality. They derive a unifying taxonomy from the major theories, decomposing quality hierarchically into cogency, effectiveness, reasonableness, and subdimensions. Table 1 lists all 15 dimensions covered. In Section 3, we use their absolute quality ratings from 1 (low) to 3 (high) annotated by three experts for each dimension of 304 arguments taken from the *UKPConvArg1* corpus detailed below.

2.2 Practical Views of Quality Assessment

There is an application area where absolute quality ratings of argumentative text are common practice: essay scoring (Beigman Klebanov et al., 2016). Persing and Ng (2015) annotated the argumentative strength of essays composing multiple arguments with notable agreement. For single arguments, however, all existing approaches that we are aware of assess quality in relative terms, e.g., Cabrio and Villata (2012) find accepted arguments based on attack relations, Wei et al. (2016) rank arguments by their persuasiveness, and Wachsmuth et al. (2017b) rank them by their relevance. Boudry et al. (2015) argue that normative concepts such as fallacies rarely apply to real-life arguments and that they are too sophisticated for operationalization.

Based on the idea that relative assessment is easier, Habernal and Gurevych (2016b) crowdsourced the *UKPConvArg1* corpus. Argument pairs *(A, B)* from a debate portal were classified as to which argument is more convincing. Without giving any guidelines, the authors also asked for reasons as to why *A* is more convincing than *B*. In a follow-up study (Habernal and Gurevych, 2016a), these reasons were used to derive a hierarchical annotation scheme. 9111 argument pairs were then labeled with one or more of the 17 reason labels in Table 2

		Negative Properties of Argument B										*Positive Properties of Argument A*							
Quality Dimension		**5-1**	**5-2**	**5-3**	**6-1**	**6-2**	**6-3**	**7-1**	**7-2**	**7-3**	**7-4**	**8-1**	**8-4**	**8-5**	**9-1**	**9-2**	**9-3**	**9-4**	**Conv**
Cog	**Cogency**	.86	.74	.67	.66	.85	.43	.81	.83	.84	.75	.59	.58	.62	.70	.67	.64	**.75**	.59
LA	Local acceptability	.92	.77	.86	.49	.90	**.80**	.86	.89	.89	.74	.58	.43	.73	.64	.67	.56	.73	.58
LR	Local relevance	.87	.77	.86	.70	**.95**	.45	.84	.92	**.95**	.73	.61	.56	.68	.69	.65	.70	.66	.62
LS	Local sufficiency	.79	.69	.67	.68	.74	.38	.85	.92	.84	**.79**	.63	.67	.54	.64	.52	**.78**	.70	.61
Eff	**Effectiveness**	.84	.71	.67	.66	.85	.62	**.87**	.92	.84	.71	.59	.57	.65	.66	.58	**.78**	.72	.59
Cre	Credibility	.78	.69	.71	.52	**.95**	**.80**	.66	.81	.67	.57	.51	.44	.66	.60	.71	.39	.62	.50
Emo	Emotional appeal	.80	.50	.59	.55	.70	**.80**	.70	.80	.67	.60	.36	.35	.41	.30	.42	.73	.50	.38
Cla	Clarity	.61	.70	**.91**	.41	**.95**	.58	.61	.87	.67	.60	.41	.40	.41	.68	.71	.56	.58	.44
App	Appropriateness	.94	**.86**	**.91**	.50	**.95**	.45	**.87**	.74	.36	**.79**	.57	.59	.69	.72	**.79**	.53	.57	.59
Arr	Arrangement	.81	.75	.86	.67	.85	.40	.78	.77	.67	.68	.60	**.73**	.64	**.73**	.73	**.78**	.72	.62
Rea	**Reasonableness**	.92	**.86**	.67	**.73**	.90	.49	.85	**.94**	.84	.73	.64	.56	.70	.69	.65	**.78**	.64	.63
GA	Global acceptability	**1.00**	.80	.82	.65	.76	.62	**.87**	.86	**.95**	.71	.63	.62	**.75**	.59	.67	.72	.68	.63
GR	Global relevance	.97	**.86**	.82	.63	.82	.71	.86	.82	**.95**	.75	.61	.51	.49	.66	.46	.72	.57	.61
GS	Global sufficiency	.77	.57	.59	.62	.85	.47	.75	.72	.71	.64	.59	.69	.46	.53	.39	.71	.61	.56
OQ	**Overall quality**	.94	.85	.79	.71	.90	.53	.85	.92	.84	.72	**.65**	.58	.69	.72	.61	.73	.73	**.64**
# Pairs with label x-y		34	55	18	115	11	16	64	37	10	50	536	79	68	86	34	26	39	736

Table 3: Kendall's τ rank correlation of each of the 15 quality dimensions of all argument pairs annotated by Wachsmuth et al. (2017a) given for each of the 17+1 reason labels of Habernal and Gurevych (2016a). Bold/gray: Highest/lowest value in each column. Bottom row: The number of labels for each dimension.

by crowd workers (*UKPConvArg2*). These pairs represent the practical view in our experiments.

3 Matching Theory and Practice

We now report on experiments that we performed to examine to what extent the theory and practice of argumentation quality assessment match.[1]

3.1 Corpus-based Comparison of the Views

Several dimensions and reasons in Tables 1 and 2 seem to refer to the same or opposite property, e.g., *clarity* and *5-3 (unclear)*. This raises the question of how absolute ratings of arguments based on theory relate to relative comparisons of argument pairs in practice. We informally state three hypotheses:

Hypothesis 1 The reasons for quality differences in practice are adequately represented in theory.

Hypothesis 2 The perception of overall argumentation quality is the same in theory and practice.

Hypothesis 3 Relative quality differences are reflected by differences in absolute quality ratings.

As both corpora described in Section 2 are based on the *UKPConvArg1* corpus and thus share many arguments, we can test the hypotheses empirically.

3.2 Correlations of Dimensions and Reasons

For Hypotheses 1 and 2, we consider all 736 pairs of arguments from Habernal and Gurevych (2016a) where both have been annotated by Wachsmuth et al. (2017a). For each pair *(A, B)* with *A* being more convincing than *B*, we check whether the ratings of *A* and *B* for each dimension (averaged over all annotators) show a concordant difference (i.e., a higher rating for *A*), a disconcordant difference (lower), or a tie. This way, we can correlate each dimension with all reason labels in Table 2 including *Conv*. In particular, we compute Kendall's τ based on all argument pairs given for each label.[2]

Table 3 presents all τ-values. The phrasing of a reason can be assumed to indicate a clear quality difference—this is underlined by the generally high correlations. Analyzing the single values, we find much evidence for Hypothesis 1: Most notably, label *5-1* perfectly correlates with *global acceptability*, fitting the intuition that abuse is not acceptable. The high τ's of *8-5 (more credible)* for *local acceptability* (.73) and of *9-4 (well thought through)* for *cogency* (.75) confirm the match assumed in Section 1. Also, the values of *5-3 (unclear)* for *clarity* (.91) and of *7-2 (non-sense)* for *reasonableness* (.94) as well as the weaker correlation of *8-4 (objective)* for *emotional appeal* (.35) makes sense.

Only the comparably low τ of *6-1 (no credible evidence)* for *local acceptability* (.49) and *credibility* (.52) seem really unexpected. Besides, the descriptions of *6-2* and *6-3* sound like *local* but cor-

[1] Source code and annotated data: http://www.arguana.com

[2] Lacking better options, we ignore pairs where a label is not given: It is indistinguishable whether the associated reason does not hold, has not been given, or is just not included in the corpus. Thus, τ is more "boosted" the fewer pairs exist for a label and, thus, its values are not fully comparable across labels. Notice, though, that *Conv* exists for all pairs. So, the values of *Conv* suggest the magnitude of τ without boosting.

Polarity	Label	Cog	LA	LR	LS	Eff	Cre	Emo	Cla	App	Arr	Rea	GA	GR	GS	OQ
Negative properties of Argument B	5-1	1.30	1.44	1.77	1.29	1.26	1.46	1.64	1.84	1.62	1.55	1.34	1.45	1.65	1.19	1.29
	5-2	1.51	1.73	1.97	1.39	1.41	1.66	1.82	1.96	1.89	1.72	1.55	1.72	1.74	1.21	1.48
	5-3	1.46	1.78	2.06	1.43	1.39	1.63	**1.96**	1.87	2.04	1.65	**1.63**	1.85	1.76	1.28	1.52
	6-1	1.54	**1.87**	2.22	1.43	1.44	**1.72**	1.85	**2.15**	2.12	1.79	1.62	**1.89**	1.89	1.27	1.55
	6-2	1.30	1.52	1.88	1.27	1.21	1.52	1.85	1.94	1.88	1.67	1.36	1.61	1.55	1.15	1.33
	6-3	**1.60**	1.85	**2.23**	**1.52**	**1.52**	1.65	1.79	2.00	**2.15**	**1.92**	**1.63**	1.85	**2.00**	**1.40**	**1.60**
	7-1	1.43	1.74	1.97	1.33	1.34	1.60	1.82	1.95	1.89	1.72	1.48	1.71	1.68	1.22	1.43
	7-2	1.45	1.68	1.97	1.41	1.39	1.53	1.86	1.84	1.95	1.67	1.53	1.68	1.70	1.25	1.48
	7-3	1.20	1.47	1.60	1.10	1.17	1.47	1.60	1.70	1.80	1.40	1.20	1.40	1.30	1.07	1.13
	7-4	1.43	1.71	2.02	1.37	1.34	1.71	1.79	1.95	1.97	1.65	1.55	1.75	1.75	1.23	1.46
Positive properties of Argument A	8-1	1.56	1.89	2.20	1.46	1.48	1.71	1.88	2.05	2.07	1.79	1.65	1.88	1.92	1.30	1.57
	8-4	1.65	1.97	2.27	1.53	1.61	1.73	1.86	2.12	2.14	1.89	1.73	1.92	1.96	1.37	1.64
	8-5	1.69	**2.07**	**2.39**	1.58	1.60	**1.81**	1.98	**2.19**	**2.25**	**1.99**	1.82	2.04	**2.11**	1.38	1.75
	9-1	1.54	1.86	2.22	1.49	1.43	1.67	1.84	2.09	2.03	1.74	1.63	1.85	1.92	1.30	1.54
	9-2	1.56	1.76	2.22	1.45	1.49	1.58	1.98	2.02	2.00	1.74	1.62	1.81	1.84	1.28	1.51
	9-3	1.55	1.78	2.31	1.42	1.49	1.68	**2.01**	2.18	2.10	1.79	1.63	1.83	1.97	1.27	1.50
	9-4	**1.78**	1.99	2.32	**1.64**	**1.68**	**1.81**	1.99	2.17	2.19	1.93	**1.86**	**2.05**	2.09	**1.44**	**1.79**
min(*Pos.*)−min(*Neg.*)		0.34	0.32	0.60	0.32	0.26	0.12	0.24	0.32	0.38	0.34	0.42	0.41	0.54	0.20	0.37
max(*Pos.*)−max(*Neg.*)		0.18	0.20	0.16	0.12	0.16	0.09	0.05	0.04	0.10	0.07	0.23	0.16	0.11	0.04	0.19

Table 4: The mean rating for each quality dimension of those arguments from Wachsmuth et al. (2017a) given for each reason label (Habernal and Gurevych, 2016a). The bottom rows show that the minimum maximum mean ratings are consistently higher for the positive properties than for the negative properties.

relate more with *global* relevance and sufficiency respectively. Similarly, *7-3 (off-topic)* correlates strongly with local *and* global relevance (both .95). So, these dimensions seem hard to separate.

In line with Hypothesis 2, the highest correlation of *Conv* is indeed given for *overall quality* (.64). Thus, argumentation quality assessment seems to match in theory and practice to a broad extent.

3.3 Absolute Ratings for Relative Differences

The correlations found imply that the relative quality differences captured are reflected in absolute differences. For explicitness, we computed the mean rating for each quality dimension of all arguments from Wachsmuth et al. (2017a) with a particular reason label from Habernal and Gurevych (2016a). As each reason refers to one argument of a pair, this reveals whether the labels, although meant to signal relative differences, indicate absolute ratings.

Table 4 compares the mean ratings of "negative labels" (*5-1* to *7-4*) and "positive" ones (*8-1* to *9-4*). For all dimensions, the maximum and minimum value are higher for the positive than for the negative labels—a clear support of Hypothesis 3.[3] Also, Table 4 reveals which reasons predict absolute differences most: The mean ratings of *7-3 (off-topic)* are very low, indicating a strong negative impact, while *6-3 (irrelevant reasons)* still shows rather high values. Vice versa, especially *8-5 (more credible)* and *9-4 (well thought through)* are reflected in high ratings, whereas *9-2 (sticks to topic)* does not have much positive impact.

[3] While the differences seem not very large, this is expected, as in many argument pairs from Habernal and Gurevych (2016a) both arguments are strong or weak respectively.

4 Annotating Theory in Practice

The results of Section 3 suggest that theory may guide the assessment of argumentation quality in practice. In this section, we evaluate the reliability of a crowd-based annotation process.

4.1 Absolute Quality Ratings by the Crowd

We emulated the expert annotation process carried out by Wachsmuth et al. (2017a) on *CrowdFlower* in order to evaluate whether lay annotators suffice for a theory-based quality assessment. In particular, we asked the crowd to rate the same 304 arguments as the experts for all 15 given quality dimensions with scores from 1 to 3 (or choose "cannot judge"). Each argument was rated 10 times at an offered price of $0.10 for each rating (102 annotators in total). Given the crowd ratings, we then performed two comparisons as detailed in the following.

4.2 Agreement of the Crowd with Experts

First, we checked to what extent lay annotators and experts agree in terms of Krippendorff's α. On one hand, we compared the mean of all 10 crowd ratings to the mean of the three ratings of Wachsmuth et al. (2017a). On the other hand, we estimated a reliable rating from the crowd ratings using MACE (Hovy et al., 2013) and compared it to the experts.

		(a) Crowd / Expert		(b) Crowd 1 / 2 / Expert		(c) Crowd 1 / Expert		(d) Crowd 2 / Expert	
Quality Dimension		**Mean**	**MACE**	**Mean**	**MACE**	**Mean**	**MACE**	**Mean**	**MACE**
Cog	**Cogency**	.27	.38	.24	.29	.38	.37	.05	.27
LA	Local acceptability	.49	.35	.37	.27	.49	.33	.30	.25
LR	Local relevance	.42	.39	.33	.28	.41	.39	.26	.25
LS	Local sufficiency	.18	.31	.21	.21	.34	.27	–.04	.19
Eff	**Effectiveness**	.13	.31	.19	.20	.27	.28	–.06	.20
Cre	Credibility	.41	.27	.31	.20	.43	.23	.22	.19
Emo	Emotional appeal	.45	.23	.32	.13	.41	.20	.25	.10
Cla	Clarity	.42	.28	.33	.23	.39	.27	.29	.20
App	Appropriateness	**.54**	.26	**.40**	.20	.48	.24	**.43**	.17
Arr	Arrangement	.53	.30	.36	.24	.49	.27	.35	.24
Rea	**Reasonableness**	.33	.40	.27	.31	.42	**.40**	.09	.29
GA	Global acceptability	**.54**	.40	.36	.29	**.53**	.37	.33	.28
GR	Global relevance	.44	.31	.31	.20	.50	.29	.22	.18
GS	Global sufficiency	–.17	.19	.04	.11	.00	.16	–.27	.11
OQ	**Overall quality**	.43	**.43**	.38	**.33**	.43	**.40**	.28	**.33**

Table 5: Mean and MACE Krippendorff's α agreement between (a) the crowd and the experts, (b) two independent crowd groups and the experts, (c) group 1 and the experts, and (d) group 2 and the experts.

Table 5(a) presents the results. For the mean ratings, most α-values are above .40. This is similar to the study of Wachsmuth et al. (2017b), where a range of .27 to .51 is reported, meaning that lay annotators achieve similar agreement to experts. Considering the minimum of mean and MACE, we observe the highest agreement for *overall quality* (.43)—analog to Wachsmuth et al. (2017b). Also, *global sufficiency* has the lowest agreement in both cases. In contrast, the experts hardly said "cannot judge" at all, whereas the crowd chose it for about 4% of all ratings (most often for global sufficiency), possibly due to a lack of training. Still, we conclude that the crowd generally handles the theory-based quality assessment almost as well as the experts.

However, the complexity of the assessment is underlined by the generally limited agreement, suggesting that either simplification or stricter guidelines are needed. Regarding simplification, the most common practical reasons of Habernal and Gurevych (2016a) imply what to focus on.

4.3 Reliability of the Crowd Annotations

In the second comparison, we checked how many crowd annotators are needed to compete with the experts. For this purpose, we split the crowd ratings into two independent groups of 5 and treated the mean and MACE of each group as a single rating. We then computed the agreement of both groups and each group individually against the experts.

The α-values for both groups are listed in Table 5(b). On average, they are a bit lower than those of all 10 crowd annotators in Table 5(a). Hence, five crowd ratings per argument seem not enough for sufficient reliability. Tables 5(c) and 5(d) reveal the reason behind, namely, the results of crowd group 1 and group 2 differ clearly. At the same time, the values in Table 5(c) are close to those in Table 5(a), so 10 ratings might suffice. Moreover, we see that the most stable α-values in Table 5 are given for *overall quality*, indicating that the theory indeed helps assessing quality reliably.

5 Conclusion

This paper demonstrates that the theory and practice of assessing argumentation quality can learn from each other. Most reasons for quality differences phrased in practice seem well-represented in the normative view of theory and correlate with absolute quality ratings. In our study, lay annotators had similar agreement on the ratings as experts. Considering that some common reasons are quite vague, the diverse and comprehensive theoretical view of argumentation quality may guide a more insightful assessment. On the other hand, some quality dimensions remain hard to assess and/or to separate in practice, resulting in limited agreement. Simplifying theory along the most important reasons will thus improve its practical applicability.

Acknowledgments

We thank Vinodkumar Prabhakaran and Yonatan Bilu for their ongoing participation in our research on argumentation quality. Also, we acknowledge financial support of the DFG (ArguAna, AIPHES), the Natural Sciences and Engineering Research Council of Canada, and the Volkswagen Foundation (Lichtenberg-Professorship Program).

References

Jens Allwood. 2016. Argumentation, activity and culture. In *6th International Conference on Computational Models of Argument (COMMA 16)*. Potsdam, Germany, page 3.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, translator). Clarendon Aristotle series. Oxford University Press.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Association for Computational Linguistics, pages 70–75. https://doi.org/10.18653/v1/W16-2808.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29(4):431–456.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 208–212. http://aclweb.org/anthology/P12-2041.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Wadsworth, Cengage Learning, 6th edition.

Trudy Govier. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, 7th edition.

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*. Association for Computational Linguistics, pages 1214–1223. http://aclweb.org/anthology/D16-1129.

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)*. Association for Computational Linguistics, pages 1589–1599. https://doi.org/10.18653/v1/P16-1150.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1120–1130. http://aclweb.org/anthology/N13-1132.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. Intern. Debate Education Association.

Chaïm Perelman and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric: A Treatise on Argumentation* (John Wilkinson and Purcell Weaver, translator). University of Notre Dame Press.

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)*. Association for Computational Linguistics, pages 543–552. https://doi.org/10.3115/v1/P15-1053.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence — An automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 440–450. https://doi.org/10.18653/v1/D15-1050.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Alberdingk Tim Thijm, Graeme Hirst, and Benno Stein. 2017a. 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*. Association for Computational Linguistics, pages 176–187. http://aclweb.org/anthology/E17-1017.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 1117–1127. http://aclweb.org/anthology/E17-1105.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

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)*. Association for Computational Linguistics, pages 195–200. https://doi.org/10.18653/v1/P16-2032.