Henning Wachsmuth* Nona Naderi** Ivan Habernal*** Yufang Hou**** Graeme Hirst** Iryna Gurevych*** Benno Stein*

* Bauhaus-Universität Weimar www.webis.de
** University of Toronto www.cs.toronto.edu/compling
*** Technische Universität Darmstadt www.ukp.tu-darmstadt.de
**** IBM Research Ireland ie.ibm.com

Argumentation Quality Assessment: Theory vs. Practice

Empirical comparison of the theoretical and practical views of argumentation quality

Argumentation quality assessment is critical for applications built upon argument mining.

Theoretical and practical views of argumentation quality differ considerably:

- Most theories suggest absolute quality ratings of normative dimensions.
- Practitioners object that arguments can only be judged in comparison to others.

Sample arguments for ”advancing the common good better than personal pursuit“:

A. ”While striving to make advancements for the common good you can change the world forever. A lot of people have succeeded 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.“

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 your hand out in help then you holding your own hand.“

In previous work, A was judged as more credible and thought through than B by lay annotators. That resembles the theoretical quality dimension cogency.

This paper studies empirically to what extent the different views actually match.

Expert ratings reflecting theory

320 arguments from Wachsmuth et al. (2017)
all also contained in the data reflecting practice

Absolute ratings of experts from 1 (worst) to 3 (best)

Normative guidelines for 15 predefined dimensions

Crowd judgments reflecting practice

736 argument pairs from Habernal & Gurevych (2016)
those also contained in the data reflecting theory

Relative judgments of crowd workers

No guidelines but 17+1 resulting reasons

B is generally weak
B is off-topic
B has less reasoning
A is thought through
B has no credible facts
A is crisp or well-written
B is hard to follow
A sticks to the topic
A more convincing than B because
B is nonsense
A is more credible
B is only an opinion
A makes you think
B has irrelevant reasons
A has better reasoning
B has language issues
A is more objective
B is attacking or abusive

High correlations between absolute ratings and reasonable judgments

Kendall's τ rank correlation when given a reason	attacking, abusive	language issues	hard to follow	no credible facts	less reasoning	irrelevant reasons	only an opinion	nonsense	off-topic	generally weak	better reasoning	more objective	more credible	crisp, well-written	sticks to the topic	makes you think	thought through	**more convincing**
Cogency	.86	.74	.67	.66	.85	.43	.81	.83	.84	.75	.59	.58	.62	.70	.67	.64	**.75**	.59
Local acceptability	.92	.77	.86	.49	.90	**.80**	.86	.89	.89	.74	.58	.43	.73	.64	.67	.56	.73	.58
Local relevance	.87	.77	.86	.70	**.95**	.45	.84	.92	**.95**	.73	.61	.56	.68	.69	.65	.70	.66	.62
Local sufficiency	.79	.69	.67	.68	.74	.38	.85	.92	.84	**.79**	.63	.67	.54	.64	.52	**.78**	.70	.61
Effectiveness	.84	.71	.67	.66	.85	.62	**.87**	.92	.84	.71	.59	.57	.65	.66	.58	**.78**	.72	.59
Credibility	.78	.69	.71	.52	**.95**	**.80**	.66	.81	.67	.57	.51	.44	.66	.60	.71	.39	.62	.50
Emotional appeal	.80	.50	.59	.55	.70	**.80**	.70	.80	.67	.60	.36	.35	.41	.30	.42	.73	.50	.38
Clarity	.61	.70	**.91**	.41	**.95**	.58	.61	.87	.67	.60	.41	.40	.41	.68	.71	.56	.58	.44
Appropriateness	.94	**.86**	**.91**	.50	**.95**	.45	.87	.74	.36	**.79**	.57	.59	.69	.72	**.79**	.53	.57	.59
Arrangement	.81	.75	.86	.67	.85	.40	.78	.77	.67	.68	.60	**.73**	.64	**.73**	.73	**.78**	.72	.62
Reasonableness	.92	**.86**	.67	**.73**	.90	.49	.85	**.94**	.84	.73	.64	.56	.70	.69	.65	**.78**	.64	.63
Global acceptability	**1.00**	.80	.82	.65	.76	.62	**.87**	.86	**.95**	.71	.63	.62	**.75**	.59	.67	.72	.68	.63
Global relevance	.97	**.86**	.82	.63	.82	.71	.86	.82	**.95**	.75	.61	.51	.49	.66	.46	.72	.57	.61
Global sufficiency	.77	.57	.59	.62	.85	.47	.75	.72	.71	.64	.59	.69	.46	.53	.39	.71	.61	.56
Overall quality	.94	.85	.79	.71	.90	.53	.85	.92	.84	.72	**.65**	.58	.69	.72	.61	.73	.73	**.64**
Pairs with the reason	34	55	18	115	11	16	64	37	10	50	536	79	72	86	34	26	39	**736**

Highest τ value in each column marked bold. Selected values colored that are highly intuitive (cyan) or rather unintuitive (red).

Agreement of experts and the crowd similar

Krippendorff's α agreement	crowd$_{1..5}$ / experts	crowd$_{6..10}$ / experts	**crowd / experts**	**experts**
Cogency	.38	.05	.27	.44
Local acceptability	.49	.30	.49	.46
Local relevance	.41	.26	.42	.47
Local sufficiency	.34	–.04	.18	.44
Effectiveness	.27	–.06	.13	.45
Credibility	.43	.22	.41	.37
Emotional appeal	.41	.25	.45	.26
Clarity	.39	.29	.42	.35
Appropriateness	.48	**.43**	**.54**	.36
Arrangement	.49	.35	.53	.39
Reasonableness	.42	.09	.33	.50
Global acceptability	**.53**	.33	**.54**	.44
Global relevance	.50	.22	.44	.42
Global sufficiency	.00	–.27	–.17	.27
Overall quality	.43	.28	.43	**.51**

Highest α value in each column marked bold.

Theory should guide practical assessment — Practice should guide simplification of theory

Selected findings from the study:

- Generally high correlations (.30 – 1.00).
- Most correlations very intutive, very few unintuitive.
- Some theoretical dimensions hard to separate.
- More convincing correlates most with overall quality (.64).
- Thought through shows the highest ratings (overall quality 1.8).
- Off-topic shows the lowest ratings (overall quality 1.1).

Conclusions on argumentation quality assessment:

- Most reasons for quality differences observed in practice are adequately represented by theory.
- Lay annotators achieve similar agreement as experts.
- Theory-based quality assessment remains complex.
- The comprehensive theory may guide assessment in practice.
- The most important reasons indicate where to simplify theory.