Computational Argumentation Quality Assessment in Natural Language

Henning Wachsmuth  Bauhaus-Universität Weimar
Nona Naderi  University of Toronto
Yonatan Bilu  IBM Research – Haifa
Yufang Hou  IBM Research – Ireland
Vinodkumar Prabhakaran  Stanford University
Tim Alberdingk Thijm  University of Toronto
Graeme Hirst  University of Toronto
Benno Stein  Bauhaus-Universität Weimar
Motivation

- **Argument mining**
  - Identifies arguments in natural language text
  - Does not assess quality

- **Argumentation quality assessment**
  - Critical for any application built upon argument mining

- **Challenges**
  - Several quality dimensions at different granularities
  - Some highly subjective
  - How *should* we argue vs. how *do* we argue

"Everyone has an inalienable human right to life, even those who commit murder; sentencing a person to death and executing them violates that right."
Background

- **Debating Technologies at Dagstuhl**
  - Common understanding of argumentation quality missing
  - Working group to coordinate research

- **Research questions**
  - Can we unify the different views of quality?
  - Can we provide a common ground for quality assessment?
Core result

Argumentation quality

- cogency
- reasonableness
- effectiveness

- local relevance
- global relevance
- local acceptability
- global acceptability
- local sufficiency
- global sufficiency
- clarity
- arrangement
- appropriateness
- emotional appeal
- credibility
Starting point

Deliberation
e.g., Steenbergen et al. (2003)

Persuasion research
e.g., Zhao et al. (2011)

Critical thinking
e.g., Freeley and Steinberg (2009)

Probabilistic logic
e.g., Pfeifer (2013)

Formal logic
e.g., Dung (1995)

Argument mining
e.g., Duthie et al. (2016)

Deliberative quality
e.g., Gold et al. (2015)

Review helpfulness
e.g., Liu et al. (2008)

Review deception
e.g., Ott et al. (2011)

Wikipedia quality flaws
e.g., Anderka et al. (2012)

Readability
e.g., Pitler and Nenkova (2008)

Essay overall quality
e.g., Burstein et al. (1998)

Evidence types
e.g., Rinott et al. (2015)

Argumentation schemes
e.g. Feng and Hirst (2011)

Assessment approaches

Argumentation theory

Argumentation quality
### Survey of existing research

<table>
<thead>
<tr>
<th>Logic</th>
<th>Dialectic</th>
<th>Argumentation quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habernal and Gurevych (2016)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Three main quality aspects

"An argument is cogent if its premises are relevant to its conclusion, individually acceptable, and together sufficient to draw the conclusion."

Blair (2012)

"A dialectical discussion derives its reasonableness from a dual criterion: problem validity and intersubjective validity."

van Eemeren (2015)

"In making a speech, one must study three points: the means of producing persuasion, the style or language to be used, and the proper arrangement of the various parts."

Aristotle (2007)
Unification of views

- focus on theory
  - validity
  - soundness
  - level of support
  - amount of evidence
  - sufficiency
  - well-formedness
  - global coherence
  - thesis clarity
  - prompt adherence
  - credibility
  - evaluability

- focus on accepted
  - premise acceptability
  - cogency
  - local/probative relevance
  - premise sufficiency
  - amount of rebuttal
  - strength

- prefer general
  - intersubjective acceptability
  - global/dialectical relevance
  - argument relevance
  - global/dialectical sufficiency

- unify names
  - argument acceptability
  - argument persuasiveness
  - arrangement
  - organization
  - emotional appeal

Argumentation quality

- Logic
  - fallaciousness
  - reasonableness
  - dialectical sufficiency

- Dialectic
  - convincingness
  - winning side

- Rhetoric
  - effectiveness
  - appropriateness of style
  - persuasiveness
A taxonomy of computational argumentation quality

- cogency
  - local relevance
  - global relevance
  - local sufficiency
  - global sufficiency
  - clarity
  - arrangement
- reasonableness
  - argument acceptability
  - argument prominence
  - argument relevance
- effectiveness
  - winning side
  - convincingness
  - emotional appeal
  - appropriateness
  - crediblity

- Thesis clarity: Persing and Ng (2013)
- Prompt adherence: Persing and Ng (2014)
- Global coherence: Feng et al. (2014)
- Evaluability: Park et al. (2015)
- Level of support: Braunstain et al. (2016)
- Amount of evidence: Rahimi et al. (2014)
- Sufficiency: Stab and Gurevych (2017)
- Organization: Persing et al. (2010)
- Rahimi et al. (2015)
- Argument strength: Persing and Ng (2015)
- Persuasiveness: Tan et al. (2016)
- Winning side: Zhang et al. (2016)
- Convincingness: Habernal and Gurevych (2016)
### The Dagstuhl-15512 ArgQuality corpus

#### Corpus based on the taxonomy
- 320 debate portal arguments (Habernal and Gurevych, 2016)
- 10 per issue/stance pair
- 3 annotators per argument
- Score from [1, 3] for all 15 dimensions

#### Agreement
- Krippendorff’s alpha limited
- Majority agreement very high

#### Correlations
- Overall quality correlates most with reasonableness (.86), cogency (.84), and effectiveness (.81)
- Several other intuitive correlations

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean</th>
<th>Alpha</th>
<th>Maj.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cogency</td>
<td>1.6</td>
<td>.44</td>
<td>92%</td>
</tr>
<tr>
<td>local acceptability</td>
<td>1.9</td>
<td>.46</td>
<td>91%</td>
</tr>
<tr>
<td>local relevance</td>
<td>2.3</td>
<td>.47</td>
<td>92%</td>
</tr>
<tr>
<td>local sufficiency</td>
<td>1.5</td>
<td>.44</td>
<td>93%</td>
</tr>
<tr>
<td>effectiveness</td>
<td>1.4</td>
<td>.45</td>
<td>94%</td>
</tr>
<tr>
<td>credibility</td>
<td>1.7</td>
<td>.37</td>
<td>96%</td>
</tr>
<tr>
<td>emotional appeal</td>
<td>1.9</td>
<td>.26</td>
<td>94%</td>
</tr>
<tr>
<td>clarity</td>
<td>2.1</td>
<td>.35</td>
<td>90%</td>
</tr>
<tr>
<td>appropriateness</td>
<td>2.1</td>
<td>.36</td>
<td>88%</td>
</tr>
<tr>
<td>arrangement</td>
<td>1.8</td>
<td>.39</td>
<td>93%</td>
</tr>
<tr>
<td>reasonableness</td>
<td>1.6</td>
<td>.50</td>
<td>96%</td>
</tr>
<tr>
<td>global acceptability</td>
<td>1.9</td>
<td>.44</td>
<td>95%</td>
</tr>
<tr>
<td>global relevance</td>
<td>2.0</td>
<td>.42</td>
<td>90%</td>
</tr>
<tr>
<td>global sufficiency</td>
<td>1.2</td>
<td>.27</td>
<td>98%</td>
</tr>
<tr>
<td>overall quality</td>
<td>1.6</td>
<td>.51</td>
<td>94%</td>
</tr>
</tbody>
</table>

Available at [www.arguana.com](http://www.arguana.com)
Contributions and outlook

**Contributions**

Comprehensive survey

Unifying taxonomy

Annotated corpus

**Outlook**

Reliable assessment

Target audience

Granularity levels

Theory vs. practice

---

© Henning Wachsmuth  henning.wachsmuth@uni-weimar.de  Webis group, Weimar  www.webis.de