Using Argument Mining to Assess the Argumentation Quality of Essays

The first study of argument mining for argumentation quality assessment
Argument mining determines the argumentative structure of texts. The benefit of this structure has rarely been evaluated.

Argumentation quality assessment is needed for envisaged applications such as argumentative writing support.

Argumentative writing support for persuasive essays:
1. Mining of an essay's argumentative structure.
2. Assessment of argumentation quality dimensions.
3. Synthesis of suggestions for improvements [future work].

We score persuasive essays based on the output of mining for four argumentation-related quality dimensions:
- Organization (Persing et al., EMNLP 2010)
- Thesis clarity (Persing and Ng, ACL 2013)
- Prompt adherence (Persing and Ng, ACL 2014)
- Argument strength (Persing and Ng, ACL 2013–2015)

Main contributions of our work:
- The first study of the benefit of argument mining for argumentation quality assessment.
- Statistical insights into essay argumentation.
- The new state of the art for two quality dimensions.

Statistical insights into argumentation based on the output of mining
Modeling of an essay as a flow of paragraph-level arguments with sentence-level argumentative discourse units (ADUs).

Learning of mining four ADU types using standard features on the Argument Annotated Essays corpus (Stab and Gurevych, COLING 2014)

Arguments of a mining approach Accuracy F-score
Majority baseline 0.525 0.361
State-of-the-art baseline (Stab and Gurevych, EMNLP 2014) 0.773 0.726
Our approach 0.745 0.745

Application of mining on all 6085 student essays from the International Corpus of Learner English (Granger et al., 2009).

Experimental set-up exactly as in the papers of the (former) state-of-the-art approaches.

Essay scoring with several supervised approaches:
- Average score baseline
- State-of-the-art baseline (Persing et al. EMNLP 2010, Persing and Ng ACL 2013–2015)
- Content: Token n-grams, prompt similarities
- POS: Part-of-speech n-grams
- Flows: Sentiment flow patterns (Wachsmuth et al., COLING 2014, EMNLP 2015)
- Our approach: ADU flows, n-grams, and compositions

Analysis of common ADU change flows in all ICLE paragraphs.

Mean squared errors in 5-fold cross-validation:

<table>
<thead>
<tr>
<th>Essay scoring approach</th>
<th>Organization</th>
<th>Thesis clarity</th>
<th>Prompt adherence</th>
<th>Argument strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score baseline</td>
<td>0.349</td>
<td>0.469</td>
<td>0.291</td>
<td>0.266</td>
</tr>
<tr>
<td>State-of-the-art baseline</td>
<td>0.175</td>
<td>0.369</td>
<td>0.197</td>
<td>0.244</td>
</tr>
<tr>
<td>Content</td>
<td>0.336</td>
<td>0.425</td>
<td>0.231</td>
<td>0.236</td>
</tr>
<tr>
<td>POS</td>
<td>0.326</td>
<td>0.461</td>
<td>0.231</td>
<td>0.333</td>
</tr>
<tr>
<td>Flows</td>
<td>0.228</td>
<td>0.481</td>
<td>0.257</td>
<td>0.259</td>
</tr>
<tr>
<td>Our approach</td>
<td>0.184</td>
<td>0.470</td>
<td>0.241</td>
<td>0.242</td>
</tr>
<tr>
<td>ADU flows</td>
<td>0.234</td>
<td>0.461</td>
<td>0.247</td>
<td>0.242</td>
</tr>
<tr>
<td>ADU n-grams</td>
<td>0.225</td>
<td>0.466</td>
<td>0.265</td>
<td>0.243</td>
</tr>
<tr>
<td>ADU compositions</td>
<td>0.194</td>
<td>0.457</td>
<td>0.239</td>
<td>0.239</td>
</tr>
<tr>
<td>Our approach + POS Flows</td>
<td>0.164</td>
<td>0.496</td>
<td>0.232</td>
<td>0.246</td>
</tr>
<tr>
<td>ADU compositions + Content</td>
<td>0.178</td>
<td>0.435</td>
<td>0.216</td>
<td>0.225</td>
</tr>
</tbody>
</table>

Henning Wachsmuth, Khalid Al-Khatib, Benno Stein

Mean squared errors in green significantly improve the state of the art with a confidence of over 90%.