Efficient Pairwise Annotation of Argument Quality

Lukas Gienapp[1] **Benno Stein**[2] **Matthias Hagen**[3] **Martin Potthast**[1]

[1] Leipzig University
[2] Bauhaus-Universität Weimar
[3] Martin-Luther-Universität Halle-Wittenberg

Abstract

We present an efficient annotation framework for argument quality, a feature difficult to be measured reliably as per previous work. A stochastic transitivity model is combined with an effective sampling strategy to infer high-quality labels with low effort from crowdsourced pairwise judgments. The model's capabilities are showcased by compiling Webis-ArgQuality-20, an argument quality corpus that comprises scores for rhetorical, logical, dialectical, and overall quality inferred from a total of 41,859 pairwise judgments among 1,271 arguments. With up to 93% cost savings, our approach significantly outperforms existing annotation procedures. Furthermore, novel insight into argument quality is provided through statistical analysis, and a new aggregation method to infer overall quality from individual quality dimensions is proposed.

1 Introduction

For a broad variety of tasks, such as argument mining, argument retrieval, argumentation generation, and question answering, compiling labeled data for argument quality remains an important prerequisite, yet, also a difficult problem. Most commonly, human assessors have been presented with one argument at a time and then asked to assign labels on a graded quality scale $\langle 0, 1, 2 \rangle$ with label descriptions such as (0) "low quality", (1) "medium quality" and (2) "high quality" for guidance.

In previous work, this was usually done concurrently for multiple orthogonal sub-dimensions of argument quality; judging the overall quality of an argument has been deemed complex (Wachsmuth et al., 2017). But on closer inspection, even the more specialized quality dimensions considered are difficult to be assessed as evidenced by the low reliability scores reported. Especially crowdsourcing suffers from assessors often having different reference frames to base their judgments on and task instructions being nondescript and therefore unhelpful in ensuring consistency. Employing experts, however, not only comes at a significantly higher cost per label; despite their expertise, even experts did not achieve more reliable judgments.

We pursue an alternative approach: stochastic transitivity modeling based on pairwise judgments of arguments. This enables the employment of laymen; the decisions required from them remain comparably simple and expect neither prior knowledge nor a common reference frame, while the labels that can be derived from their judgments still exhibit a high accuracy and informativeness. Though pairwise judgment has already been considered for assessment of argument quality (Habernal and Gurevych, 2016; Toledo et al., 2019), its significant cost overhead has hindered widespread application.

We explore the lower bound of effort needed to infer labels of sufficient quality. We combine a pairwise model with a highly effective offline sampling strategy to minimize the set of needed pairwise judgments, saving up to 93% of the effort of an exhaustive comparison. As part of this work, we release the Webis Argument Quality Corpus 2020, which includes a total of 41,859 pairwise judgments between 1,271 arguments across the three dimensions of rhetorical, logical, and dialectical quality. Further, inferred scalar scores for the three dimensions as well as overall quality and topic relevance are provided, alongside a reference implementation of our model.[1]

Carrying out a first analysis of the statistical properties of the corpus, we validate both the new annotation method and the corpus by drawing comparisons to previous work. Since judging overall quality by itself is a difficult task, based on our statistical analysis, we find that euclidean vector length adequately combines scores from the three aforementioned specialized quality dimensions into a single overall quality score.

[1] Resources: https://webis.de/publications.html?q=ACL+2020
Corpus: https://zenodo.org/record/3780049
Code base: https://github.com/webis-de/ACL-20

2 Related Work

Wachsmuth et al. (2017) surveyed many facets of argument quality that are distinguished in argumentation theory, organizing them within three major dimensions: logical quality (the argument's structure and composition), rhetorical quality (persuasive effectiveness, vagueness, and style), and dialectical quality (contribution to the discourse). Further, they built the first comprehensive argument quality corpus, tasking three experts with annotating arguments with respect to all 15 (sub-)dimensions in their taxonomy. Each dimension has been annotated on a scale from 1 (low) to 3 (high), reaching Krippendorff's α values between 0.26 and 0.51, depending on the quality dimension. Despite a rigorous setup, the low agreement is evidence that even experts have difficulties to reliably judge argument quality.

Potthast et al. (2019) explore the use of argument quality as an evaluation criterion beyond relevance for argument retrieval, thus needing to collect quality judgments for their evaluation task. Based on the taxonomy of Wachsmuth et al., they had the three major dimensions annotated on graded scales ranging from 1 (low) to 4 (high), reproducing the findings of Wachsmuth et al. They recruited highly educated students of at least bachelor's level education from a national foundation for gifted students who have a strong interest in societal issues. Still, reliable annotation was difficult to achieve due to the highly subjective, complex, and nuanced nature of argument quality.

Each study operates on rather small amounts of data; they only annotate 320 (Wachsmuth et al., 2017) and 437 (Potthast et al., 2019) individual arguments. Both setups become nonviable for larger annotation tasks, since the associated labor costs in such (semi-)expert studies are usually high.

This is not an issue in crowdsourced settings, where judgments can be collected in abundance for a comparatively cheap price. However, the problem of annotation quality is more severe here: argument quality might be even more difficult to judge without prior domain-specific knowledge, creating the need for annotation frameworks that can still maintain a sufficiently high data quality. Judging from the agreement scores given by Wachsmuth et al. and Potthast et al., obtaining reliable data using classic graded scales proves infeasible, an effect that should be even more pronounced in a crowdsourced setting.

Swanson et al. (2015) measure an arguments' quality as the amount of context or inference required for it to be understood, describing an annotation setup where assessors judge seven individual quality dimensions on a 0-1-slider. Recruiting assessors on Amazon Mechanical Turk (MTurk), they use intra-class correlation to estimate inter-rater agreement, with an average value of 0.42 over all topics, thus also indicating a poor reliability (Portney et al., 2009). They further observe a correlation with sentence length, prompting them to remove all sentences shorter than four words.

All three studies indicate that absolute rating (i.e., having assessors label a single argument on a given absolute scale without the context of other arguments) performs unfavorably. This rating method, also known as Likert scale or Mean Opinion Score, is known to have two major drawbacks (Ye and Doermann, 2013): (1) Absolute rating is often treated as if it produces data on an interval scale. However, assessors rarely perceive labels as equidistant, thus producing only ordinal data. This leads to a misuse of statistical tests and results in low statistical power of subsequent analyses. (2) Absolute rating is difficult for assessors without prior domain knowledge, since they may be unsure which label to assign. This results in noisy, inconsistent, and unreliable data.

As an alternative, preference rating (i.e., a relative comparison by showing two arguments to an assessor and letting them declare their preference towards one of them) has been considered by Habernal and Gurevych (2016), who compile an exhaustive set of pairwise comparisons to infer labels for argument convincingness. For 1,052 arguments on 32 issues, each of the over 16,000 total comparisons was annotated by five different crowd workers on MTurk. While no α statistics are provided, the authors do conclude that preference ratings in a crowdsourced setting are sufficiently accurate, since the best-ranked rater for each pair achieves 0.935 accuracy compared to a gold label.

The indicated reliability of pairwise annotation for argument quality is further corroborated by Toledo et al. (2019), who compile a large dataset of about 14,000 annotated argument pairs, and absolute ratings in the 0-1-range for about 6,300 arguments. Pairwise annotations were made in regard to the overall quality of arguments, operationalized as "Which of the two arguments would have been preferred by most people to support/contest the

topic?" Using a strict quality control, they show that the annotated relations consistently reproduce the direction implied by absolute ratings. Yet, annotating quality as a single feature is problematic, since (1) it is hard to capture the multi-facet nature of argument quality in that way (Wachsmuth et al., 2017) and the chosen operationalization is similar to the facet of dialectical quality, neglecting the other major two; and (2) scores for individual quality dimensions are warranted for in-depth training and evaluation for a broad range of argumentation technology (Potthast et al., 2019).

Although preference rating seems promising based on the reported reliability, it creates the need for a model that infers score labels from the collected comparison data. Habernal and Gurevych propose the use of PageRank (Page et al., 1999). This is problematic, since cycles in the comparison graph may form rank sinks, distorting the latent rankings. Habernal and Gurevych deal with this problem by constructing a directed acyclic graph (DAG) from the collected data prior to applying PageRank, assuming that argument convincingness exhibits the property of total order. However, no prior evidence for this property is apparent. Simpson and Gurevych (2018) note further problems with PageRank and propose the use of Gaussian process preference learning instead, demonstrating a high scalability.

However, for a practical approach, an effective strategy to minimize the number of needed comparisons is warranted, since, to build the DAG, exhaustive comparison data is required. This is inefficient; at worst $\binom{n}{2}$ comparisons have to be obtained for n arguments. Also, no data was collected on how the PageRank method performs on incomplete or sparse comparison data.

Chen et al. (2013) also propose an online sampling strategy based on the Bradley-Terry model (Bradley and Terry, 1952). They implement an online Bayesian updating scheme, which, contrary to previous work such as presented by Pfeiffer et al. (2012), does not require retraining the whole model when new comparisons are added. After each comparison added to the total set of annotated pairs, they identify the next pair to be compared by calculating which comparisons would reduce the overall model uncertainty the most. Simpson and Gurevych (2018) opt for a similar active learning approach, but note that that it is prone to overfitting, causing accuracy to decrease.

While online learning uses an approximately minimal amount of comparisons, additional drawbacks besides overfitting can be noted: (1) The updating scheme diminishes the reusability of the collected data, since such a specific method of choosing pairs introduces data bias for other applications. (2) Online sampling is complicated to implement on a crowdsourcing platform, preventing multiple workers from making judgments in parallel. (3) In the case of Chen et al. (2013), the model is not equipped to handle comparison ties, i.e., an assessor declaring no preference. Yet, ties frequently occur in real-world annotation tasks.

Overall, the Bradley-Terry model appears to be a promising candidate for our purposes: its robustness and statistical properties have been studied in great detail (Hunter, 2004), and it can be efficiently computed (Chen et al., 2013). However, an alternative offline sampling method has to be formulated, which we introduce in the following section.

3 Pairwise Quality Annotation

In this section, we define a model to aggregate pairwise judgments into scalar ranking scores and combine different sampling strategies to form a highly efficient annotation framework.

3.1 The Bradley-Terry Model

Let $D = \{d_1, \dots, d_n\}$ denote a set of n items (e.g., arguments) for which a latent ranking is assumed according to a scale-invariant set $\Gamma = \{\gamma_1, \dots, \gamma_n\}$ of real-valued "merits", where the i-th item d_i has merit γ_i. When independently comparing pairs of items (d_i, d_j) from D, the probability of item d_i beating item d_j is defined as follows:

$$P(d_i \succ d_j) = \frac{\gamma_i}{\gamma_i + \gamma_j}. \tag{1}$$

Using exponential score functions $p_i = e^{\gamma_i}$ reduces the model to a logistic regression on pairs of individuals (Agresti, 2003):

$$P(d_i \succ d_j) = \frac{p_i}{p_i + p_j}. \tag{2}$$

The merits Γ can thus be inferred with maximum likelihood optimization (Hunter, 2004) and the following log-likelihood equation for a pool of pairwise comparisons C, a multiset of pairs (i, j), where i and j are drawn from $[1, n]$:

$$\mathcal{L}(\Gamma, C) = \sum_{(i,j) \in C} \log P(d_i \succ d_j). \tag{3}$$

3.2 Incorporating Ties

Pairs of items (d_i, d_j) may exist whose merit difference is below a threshold τ so that assessors cannot decide which is better. Rao and Kupper (1967) incorporate such ties into the model as follows:

$$P(d_i \succ d_j) = \frac{p_i}{p_i + p_j\theta} \tag{4}$$

for the probability of preference of d_i over d_j, and

$$P(d_i \approx d_j) = \frac{p_i p_j(\theta^2 - 1)}{(p_i + p_j\theta)(p_i\theta + p_j)} \tag{5}$$

for the probability of no preference between the two, where $\theta = e^\tau$. For $\tau = 0$, i.e., assessors being able to differentiate every item pair, these equations reduce to the standard Bradley-Terry model.

3.3 Regularization

The maximization is guaranteed to converge to the unique maximum likelihood estimator in finite steps under the assumption that in every possible partition of the items into two nonempty subsets, some subject in the second set beats some subject in the first set at least once (Hunter, 2004). Thus, a pairwise comparison experiment is restricted in two ways: (i) The matrix formed by the comparisons must construct a strongly connected graph; (ii) The comparisons between the partitions cannot all be won by subjects from the same group, i.e., no item has losses or wins exclusively.

Even though the adherence becomes asymptotically likely given an appropriate experiment design (Yan et al., 2011), the problem can be regularized to increase robustness. The regularization term

$$\mathcal{R}(\Gamma) = \sum_{i=1}^{n} \left[\log\left(\frac{e^1}{e^1 + p_i}\right) + \log\left(\frac{p_i}{p_i + e^1}\right) \right], \tag{6}$$

weighted by a regularization parameter λ, is added to model a dummy item d_0 with merit $\gamma_0 = e^1$, which is defined to compare against every item with exactly one win and one loss (Chen et al., 2013). Convergence is now ensured as the graph is guaranteed to be strongly connected. Additionally, the merits Γ are no longer scale-invariant, since the merit of the dummy item is fixed at 1.

3.4 Log-Likelihood Maximization

The log-likelihood equation, with regularization parameter λ and merit threshold τ takes the form

$$\mathcal{L}(\Gamma, \tau, \lambda, C) = \sum_{(i,j)\in C} \log \left[\begin{cases} P(d_i \succ d_j) & \text{if } d_i \succ d_j \\ P(d_i \approx d_j) & \text{if } d_i \approx d_j \end{cases} \right] + \lambda\mathcal{R}(\Gamma). \tag{7}$$

Γ is initialized with 1's by convention. Chen et al. (2013) propose $\lambda \in [0.1, 10]$, inferring rankings similar to the unregularized problem for sufficiently small values, while regularized rankings for larger λ values often outperform the baseline for a broad range of applications. The maximization was solved using BFGS optimization.

3.5 Sparsification

Sampling strategies are needed to reduce the amount of comparisons as obtaining an exhaustive set of $\binom{n}{2}$ comparisons becomes infeasible with larger item counts. Nevertheless, sampling strategies should preserve a high annotation quality.

Burton (2003) describes a strategy where items are arranged in a cyclical way. A main feature is that each item is required to appear in the same number of pairs in order to gain the same amount of information about each item. For a random permutation of the items in D, the i-th is compared with the $(i+1)$-th item for $i < n$, and item d_n with item d_1, thus completing the cycle. This can be generalized to higher step sizes s: for instance, if $s = 2$, all items that are separated by two positions around the ring are compared. However, this strategy suffers from the major drawback that for some step sizes, the resulting graph has multiple unconnected components, thus violating the restriction that the comparison matrix must form a strongly connected graph. Therefore, complex combinations of different step sizes are needed, resulting in needlessly complicated experimental setups.

Alternatively, Yan et al. (2011) proposed a method of sparse grouped comparisons, where the set of all items D is partitioned into m equisized disjoint subsets D_k, where $k \in [1, m]$, so that the following constraints hold true:

(i) for each D_k, $(i, j) \in C$ when $d_i, d_j \in D_k, i \neq j$, and

(ii) $(i, j) \in C$ when $d_i \in D_k, d_j \in D_{k+1}$ for $k = 1, \ldots, m-1$.

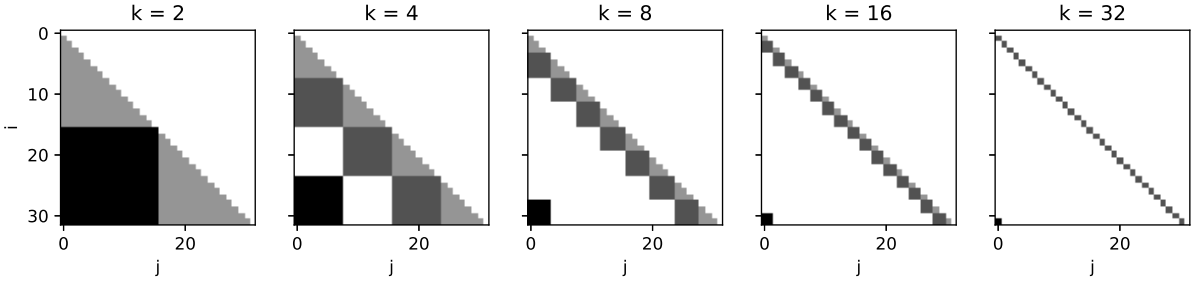


Figure 1: Comparison matrices for $n = 32$ and different values of k. Variables i and j denote the matrix indices as used in the constraints.

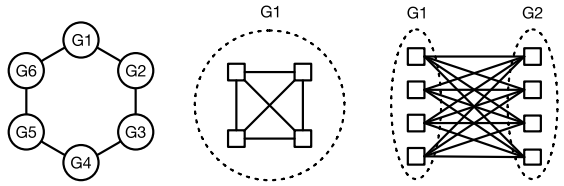


Figure 2: Example cyclic group design for six groups of four items.

However, in this approach not every item has the same amount of comparisons. To make the strategy of Yan et al. (2011) consistent with this requirement, both strategies can be combined to derive a cyclical group by also including comparisons between group D_1 and group D_k, as reflected in the additional constraint

(iii) $(i, j) \in C$ when $d_i \in D_1, d_j \in D_m$.

This way, the design adheres to the principle that every item should have the same number of comparisons but the overall construction of the experiment remains simple. All combinations of items in the same group, and the Cartesian product of adjacent groups are included. Therefore $k \cdot \binom{n/k}{2}$ intra-group comparisons and $k \cdot \left(\frac{n}{k}\right)^2$ inter-group comparisons are needed. Thus, the total amount of comparisons c is

$$c = k\left(\left(\frac{n}{k}\right)^2 + \binom{n/k}{2}\right) = \frac{3n^2}{2k} - \frac{n}{2}. \quad (8)$$

If multiple independent judgments per unique comparison are collected, as is frequently done in crowdsourcing, c has to be multiplied by a factor x denoting how many unique judgments are collected per comparison.

Figure 2 shows an exemplary cyclic group design for six groups of four items. Shown on the left is the overall design, with intra-group comparisons (Constraint (i)) depicted in the middle, and inter-group comparisons between adjacent groups (Constraints (ii) and (iii)) on the right.

Example comparison matrices for $n = 32$ and different values of k are shown in Figure 1 to provide a visual intuition of the sampling process. Although the comparison matrix is inherently symmetric, to reflect the true count of comparisons, only the lower half is depicted. All comparisons introduced by Constraint (i) are colored in light gray, Constraint (ii) in medium gray, and dark gray for Constraint (iii). Note that for the special case of $k = 2$, Constraints (ii) and (iii) are equal.

3.6 Model Evaluation

To test the accuracy trade-off between exhaustive comparison and sparse comparison on real-world data, twenty topics of 32 arguments each were randomly selected from the UKPConvArg1 corpus (Habernal and Gurevych, 2016), which includes an exhaustive pairwise comparison set for argument convincingness. With each comparison having five independent annotations, a baseline was established by fitting the model on the full set. Then, different values for k and x were used to sample a subset of the comparisons and the proposed model was fitted with each of the sampled comparison sets. The obtained merit ranking was compared against the baseline ranking using Pearson's ρ, with confidence intervals calculated using bootstrapping ($n = 10,000$). For each of the resulting rankings, the amount of used comparisons and the correlation with the baseline ranking are detailed in Table 1.

The following interesting properties are apparent: (1) Collecting multiple judgments per unique comparison, as is usual practice in methods relying on graded scales is not sufficiently beneficial. Increasing the annotation effort by factor 5 (from $x = 1$ to $x = 5$) results in only a minimal gain in ranking accuracy of 0.06 for $k = 4$. For higher sampling rates, decreasing k yields a larger net increase

x	k	Our approach				
		Judgments	**Judgments %**	$\bar{\rho}$	**95% CI**	
5	2	2480	100	1.00	1.00	1.00
	4	1904	76	0.99	0.99	1.00
	8	944	38	0.96	0.95	0.97
	16	464	18	0.88	0.85	0.90
	32	224	9	0.67	0.61	0.72
4	4	1520	61	0.99	0.99	0.99
	8	752	30	0.95	0.94	0.96
	16	368	14	0.86	0.83	0.88
	32	176	7	0.64	0.60	0.69
3	2	1488	60	0.99	0.99	0.99
	4	1136	45	0.98	0.98	0.99
	8	560	22	0.93	0.82	0.95
	16	272	10	0.82	0.79	0.85
	32	128	5	0.65	0.60	0.69
2	2	992	40	0.98	0.97	0.99
	4	752	30	0.97	0.96	0.97
	8	368	14	0.91	0.89	0.93
	16	176	7	0.78	0.75	0.81
	32	80	3	0.59	0.56	0.63
1	2	496	20	0.95	0.94	0.96
	4	368	14	0.92	0.90	0.93
	8	176	7	0.82	0.79	0.86
	16	80	3	0.66	0.61	0.71
	32	32	1	0.47	0.40	0.54

Table 1: Our model's performance under sparsification. x denotes the number of judgments collected per unique comparison, k the group factor along the resulting total number of judgments in absolute and relative terms, compared to an exhaustive comparison and corresponding ρ-correlations with the baseline ranking.

in ranking accuracy than increasing x, at the same cost. By example, going from $x = 1, k = 16$ to $x = 1, k = 4$ ends up at the same number of comparisons as $x = 2, k = 8$, but has a slightly higher ranking accuracy. Therefore, it is more economical to increase the sampling rate until the required accuracy is met than collecting multiple judgments. (2) The proposed model and comparison strategy are able to produce near-perfect rankings ($x = 1, k = 4, \bar{\rho} = 0.92 \pm 0.02$) using only 14%, and acceptable rankings ($x = 1, k = 8, \bar{\rho} = 0.82 \pm 0.04$) using only 7% of the full comparison set. This significant reduction is a promising sign for employing our model in crowdsourced studies at scale. However, the specific choice of k depends on scale and domain of the data as well as trustworthiness of comparisons. Therefore, we refrain from making a general suggestion for the choice of k. Thus, if the model is to be adapted to drastically different domains or item counts, exploratory studies are advised to estimate the quality tradeoff for a specific use case.

4 The Webis Argument Quality Corpus

To maximize its usefulness, the sample of arguments for our argument quality corpus was drawn from the recently published args.me corpus (Stein and Wachsmuth, 2019), a collection of 387,606 arguments crawled from various debate portals. To ensure some topic diversity and relevance of the arguments to the topic, while keeping the amount of judgments within our budget limits, we (1) indexed the args.me corpus using three retrieval models from the Terrier information retrieval library (Ounis et al., 2006) (namely BM25, DPH, and DirichletLM), (2) retrieved texts for 20 topic queries at a depth of 50 texts per topic per model, (3) and pooled all 3000 retrieved texts to remove overlap between the different models. In total, 1,610 unique spans of text remained for the 20 topics.

Using Amazon's Mechanical Turk, in a first preprocessing step, we tasked crowd workers with deciding whether or not a given retrieved item actually contained argumentative text. Each text was judged by five crowd workers, using majority vote as a decision rule. To ensure quality, we recruited only workers for the task with an approval rate of at least 95%, like Swanson et al. (2015).

Of the 1,610 input texts, 339 were flagged as non-arguments. Most of these texts are noise resulting from using debate platforms as data source; examples include statements of acceptance for a debate (*"I accept this debate on [Topic] and will go first [...]"*), statements with no argumentative value (*"I think [Topic] is good."*), definitions, and in some cases even jokes or personal attacks. 1,271 arguments remained for quality annotation.

In a second step, all remaining arguments were annotated for argument quality via a sample of pairwise judgments on which we applied our model. For each pair of arguments, a crowd worker was tasked to select the one that exhibits a higher quality compared to the other with regard to a given description of the respective quality dimensions. The annotation was repeated separately for each of 20 topics and each of the three aforementioned quality dimensions. To make the task accessible to workers without prior knowledge of argumentation theory, the quality dimensions were operationalized as follows: "Which text has the better logical structure?" (logical quality), "Which text has the better style of speech?" (rhetorical quality), and "Which text would be more useful in a debate?" (dialectical quality). Examples were given to annotators as guidance. Table 2 shows exemplary arguments for each of the three quality dimensions alongside a brief explanation why this argument lacks the specified quality. In each task, one such

Dimension	Argument	Explanation
Rhetorical quality	"Gender is a social construct cuse we are told when we are first born by a dude what gender but if he didnt tell us that we woudnt have a gender its only cuse he told us that gender that we are that gender."	This argument is of low rhetorical quality, as it lacks proper sentence structure, uses informal speech, has typos, and its use of ellipsis makes it hard to follow.
Logical quality	"I support an abortion ban. We must not forget that abortion opposes the principle of sanctity of life. Women are blessed with the gift of giving birth to another life and hence, should accept it with responsibility."	Even though this argument has a clearly stated claim, the evidence used to support it is insufficient. Key concepts are not defined (What is 'sanctity of life'? Why does it apply to unborn fetuses?) and the conclusion ('Women should accept it with responsibility.') does not necessarily follow from the evidence.
Dialectical quality	"Banning abortion would mean that there is more people in the world. This leads to overpopulation which is a major problem and 842 million people are undernourished every year. More people only causes more problems."	This argument is not very convincing since the evidence (overpopulation) presented in support of the conclusion (abortion ban) is not very relevant to the issue. It can easily be invalidated by, for example, offering better solutions to overpopulation than abortion. Thus, the argument does not make a meaningful contribution to resolving the debate conflict.

Table 2: Example arguments from the args.me corpus with accompanying explanation for why each argument lacks the specified quality.

negative example as well as one positive example was provided with explanations, ensuring no topic overlap between annotated material and example.

Five comparisons were presented together as one task. The comparison sets of five were compiled randomly to minimize order effects. A cyclic group comparison strategy as described in Section 3.5 with $k = 8$ was employed, with each pair annotated by one worker. On average, a topic pooling consists of $n = 64$ unique arguments, with $c_{\text{sampled}} = 698$ and $c_{\text{exhaustive}} = 2,043$. The mean sample rate compared to the exhaustive comparison set therefore is 0.342. Erring on the safe side, we chose to maximize correlation with what can be expected from an exhaustive comparison as guided by Table 1. In total, 2,797 HITs were carried out. A reward of \$0.08 per HIT was paid, amounting to \$268.54 per quality dimension and \$805.54 total while ensuring an hourly rate of at least \$8.

In comparison to the setup of Habernal and Gurevych (2016), who carried out an exhaustive comparison with a factor of $x = 5$ crowd workers per pair, the annotation effort for our study could be reduced by 93.17% based on our model. Nevertheless, if a higher accuracy is deemed necessary in future experiments, our comparison set can easily be extended by adding additional votes per comparison or by increasing the group size.

Compared to a traditional annotation setup using graded scales, having five workers annotate each item on a scale from 0 (low) to 4 (high) would put the total annotation cost for one quality dimension at around \$150.00, supposing a reward of \$0.08 per HIT. Although cheaper, the annotation quality would be much lower as per the reliabilities reported in previous work. Moreover, the highly increased level of detail in the quality scores produced by our new approach is worth the extra cost. In cases where annotation quality is not as important, sampling at higher values of k still achieves acceptable correlation scores, rendering our method even cheaper than the traditional approach.

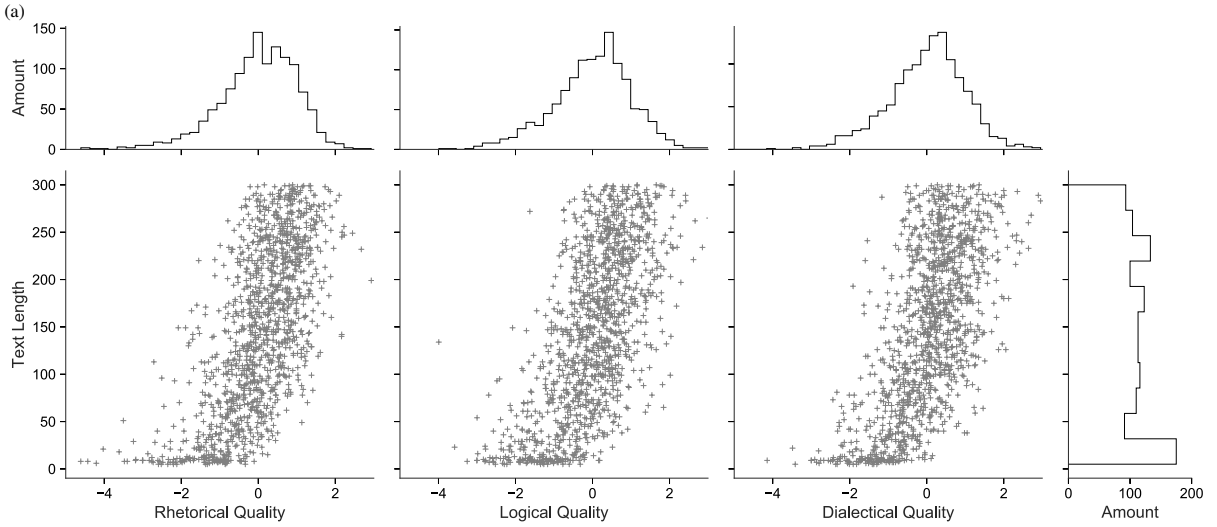
5 Corpus Analysis

In this section, we carry out a statistical analysis of our new corpus. First, we study the distribution and the correlation effects between the different quality dimensions, and between the quality dimensions and text length, to draw comparisons to the prior work of Wachsmuth et al. (2017). Then, we explore the hypothesis of overall quality being a latent variable by analyzing the influence of quality dimensions.

5.1 Distribution

Distributions of scores for all three different quality dimensions are shown in Table 3a. Additionally, the distribution of text length in the corpus as well as scatterplots for text length and quality are given. All three dimensions exhibit a similar distribution, centered at zero. Given that the dummy item of the model, i.e., an item that is defined to have exactly one win and one loss against every other item, has a fixed score of 1, this indicates that the majority

(a)



(b)

Dimension	Rhetorical		Logical		Dialectical	
	Our	Expert	Our	Expert	Our	Expert
Rhetorical			**0.63**	**0.81**	**0.61**	0.75
Logical	**0.63**	**0.81**			0.55	**0.78**
Dialectical	0.61	0.75	0.55	0.78		
	Pro	Con	Pro	Con	Pro	Con
Rhetorical			**0.62**	**0.63**	**0.60**	**0.61**
Logical	**0.62**	**0.63**			0.60	0.51
Dialectical	0.60	0.61	0.60	0.51		

(c)

Dimension	Overall	$l > 100$
Rhetorical	0.65	0.37
Logical	0.64	0.37
Dialectical	0.63	0.28

(d)

Step	Variance	Rhetorical	Logical	Dialectical
1	0.73	-0.5866	-0.5715	-0.5738
2	0.15	0.1050	0.6489	-0.7536
3	0.12	-0.8031	0.5023	0.3206

(e)

Sample Sizes	
Overall	$l > 100$
1271	869
Pro	Con
675	596
Args.	Non-args.
1271	339

Table 3: (a) Distributions and scatterplots for quality scores and text length in the corpus. (b) Pearson ρ correlation coefficient cross-tabulation for different attribute combinations, full set and per stance. Expert values are taken from Wachsmuth et al. (2017), Table 3. Maximum per column in bold. (c) Correlation between quality dimensions and text length, full and only for texts longer than 100 words. (d) Component vectors and explained variance for PCA steps on argument quality. (e) Sample sizes for (b) and (c).

of texts in our corpus are only of mediocre argumentative quality. Also, all three distributions are slightly asymmetric, with the lower end extending more. The distribution of text lengths is fairly similar across all lengths, with only one apparent spike around 0-50 words.

5.2 Correlation

Table 3b shows correlation coefficients for the three quality dimensions when compared to each other, and when compared to argument stance. The inter-quality correlation as given by Wachsmuth et al. is also included, and found to be commensurate with their figures. Although the correlation is lower in total, which is expected given the much bigger sample size and different annotation methodology, the general pattern is reproduced, with one slight deviation: dialectical quality appears to correlate slightly more with rhetorical quality in our corpus, but with logical quality in their data. However, given that the two correlation coefficients are nearly equal in the data of Wachsmuth et al., and the value differences between the different quality dimensions in our data are also too small to draw any conclusions regarding whether two of the three are more intertwined than the third, this effect is not problematic.

The correlation between the quality dimensions being fairly high hints at them being dependent on a latent variable, which could be the overall argumentation quality. When computing the scores separately for each stance, no systematic difference is apparent. Following the reasoning of Potthast et al. (2019), this may indicate that the scoring method is not prone to assessor bias.

A correlation of quality and text length (measured as word count), as also noted by Swanson et al. (2015), is evident (Table 3c). While this could hint at a data bias, with crowd workers just voting for longer texts in the comparison but not actually reading all of it, the effect is much less pronounced when only measuring the correlation in texts longer than 100 words ($n = 869$). Thus, much of the pronounced effect can be explained by short texts receiving justified low scores rather

than longer texts being voted higher regardless of content. From a qualitative point of view, a correlation effect between length and quality would also be expected, since a solid argumentative reasoning (claim and justification) usually requires at least some amount of text. The scatterplots in Table 3a additionally corroborate the correlation effects: only a very minor trend is apparent, with longer text receiving slightly higher scores. Given the accumulation of texts towards the lower end of the length spectrum, and these receiving lower scores further explains the lower overall correlation when only measuring in texts over 100 words.

5.3 Overall Argument Quality

For some applications a scalar value for overall argument quality is warranted. As it has been argued that an overall argument quality is hard to measure, the three different explored quality dimensions could be combined to derive such a rating. The high correlation of the different quality dimensions implies such a latent variable. As a working hypothesis, the overall argument quality could be interpreted as a three-dimensional vector in a space spanned by the three quality dimensions. Based on this, two essential questions have to be explored: (1) Are the different dimensions equally influential on the overall argument quality? (2) How can a scalar quality value for overall quality be derived from such a vector?

To address the first question, principal component analysis (PCA) was carried out to measure the influence of each quality dimension on the hypothesized latent variable. Results are given in Table 3d. The first step of the PCA accounts for 73% of the data variance, and is equally influenced by all three quality dimensions. Therefore, evidence is given towards the hypothesis. As for how to derive a numerical value for this overall argument quality, since the influence of all dimensions is equal, the euclidean vector length is proposed. However, since the quality scores derived in this work are positive as well as negative, the length of a vector is the same as that of its negative counterpart. To account for this, the score distributions are equally shifted into the positive domain. Thus, a standardized scalar value for overall argument quality can be calculated.

6 Conclusion

A novel approach for annotating argument quality based on stochastic transitivity modeling has been proposed, outperforming existing approaches in terms of annotation effort and annotation detail, while maintaining a high annotation quality. The overall workload in comparison to previous approaches within the same class of approaches was reduced by 93.17% through an efficient sampling method. Sampling at even higher rates is possible, resulting in the new framework operating at the same cost as the traditional approach relying on graded scales.

The collected data and a reference implementation of our model are made available in form of the Webis-ArgQuality-20 corpus, one of the largest and most detailed corpora for pairwise argument quality. The collected corpus can be used for a multitude of purposes—especially in the emerging field of argument retrieval, it is suitable as basis for retrieval evaluation, or to train new learning to rank models. A second field of application is debate systems, where a dataset can be of use for training a system to formulate new arguments. The developed annotation approach is also not only limited to rate argument quality: it can easily be transferred to other questions or criteria that can be rated by comparison. Even though the annotation cost can be slightly higher compared to the traditional absolute rating approach, the derived data is much more detailed and allows for conclusions with higher statistical power.

Insight into argument quality was derived on a larger scale than in previous studies. It has been shown that the three quality dimensions can be successfully annotated by laymen when using the described annotation procedure. The correlation patterns found in previous studies were reproduced, showing the quality dimensions to be equally correlating with each other. This is likely due to them being dependent on a latent overall quality, a hypothesis that was supported using a PCA analysis of derived quality vectors. A procedure to derive a scalar value for overall quality was introduced, proposing Euclidean vector length to combine the different dimension scores.

References

Alan Agresti. 2003. *Categorical data analysis*, 2 edition. John Wiley & Sons, Hoboken, NY.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: The method of paired comparison. *Biometrika*, 39(3-4):324–345.

Michael L Burton. 2003. Too many questions? The uses of incomplete cyclic designs for paired comparisons. *Field Methods*, 15(2):115–130.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on web search and data mining*, WSDM '13, pages 193–202, New York, NY. ACM.

Ivan Habernal and Iryna Gurevych. 2016. 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, Berlin, Germany. Association for Computational Linguistics.

David R. Hunter. 2004. MM algorithms for generalized Bradley-Terry models. *The annals of statistics*, 32(1):384–406.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web.

Thomas Pfeiffer, Xi Alice Gao, Yiling Chen, Andrew Mao, and David G Rand. 2012. Adaptive polling for information aggregation. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Leslie Gross Portney, Mary P Watkins, et al. 2009. *Foundations of clinical research: Applications to practice*, volume 892. Pearson/Prentice Hall, Upper Saddle River, NJ.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument Search: Assessing Argument Relevance. In *42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019)*. ACM.

P.V. Rao and Lawrence L. Kupper. 1967. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204.

Edwin D. Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Trans. Assoc. Comput. Linguistics*, 6:357–371.

Benno Stein and Henning Wachsmuth, editors. 2019. *6th Workshop on Argument Mining (ArgMining 2019) at ACL*. Association for Computational Linguistics, Berlin Heidelberg New York.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - New datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5624–5634. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 176–187.

Ting Yan, Jinfeng Xu, and Yaning Yang. 2011. Grouped sparse paired comparisons in the Bradley-Terry model. *arXiv preprint*.

Peng Ye and David Doermann. 2013. Combining preference and absolute judgements in a crowd-sourced setting. In *ICML Workshop*.