Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing

Anne Lauscher[1], Lily Ng[2], Courtney Napoles[2], Joel Tetreault[3]

[1]Data and Web Science Group, University of Mannheim, Germany
[2]Grammarly [3]Dataminr, Inc.
[1]anne@informatik.uni-mannheim.de,
[2]first.last@grammarly.com, [3]jtetreault@dataminr.com

Abstract

Though preceding work in computational argument quality (AQ) mostly focuses on assessing overall AQ, researchers agree that writers would benefit from feedback targeting individual dimensions of argumentation theory. However, a large-scale theory-based corpus and corresponding computational models are missing. We fill this gap by conducting an extensive analysis covering three diverse domains of online argumentative writing and presenting `GAQCorpus`: the first large-scale English multi-domain (community Q&A forums, debate forums, review forums) corpus annotated with theory-based AQ scores. We then propose the first computational approaches to theory-based assessment, which can serve as strong baselines for future work. We demonstrate the feasibility of large-scale AQ annotation, show that exploiting relations between dimensions yields performance improvements, and explore the synergies between theory-based prediction and practical AQ assessment.

1 Introduction

Providing relevant and sufficient justifications for a claim and using clear language to express reasoning are important features of everyday writing. These are components of *Argument Quality (AQ)*, which has been studied in many domains, such as student essays (Wachsmuth et al., 2016), news editorials (El Baff et al., 2018), and debate forums (Lukin et al., 2017).

Preceding work in natural language processing (NLP) and computational linguistics (CL) has mostly focused on practical AQ assessment[1], considering either the *overall quality* of arguments (Toledo et al., 2019; Gretz et al., 2020, inter alia) or a single specific conceptualization of AQ, e.g., *argument strength* (Persing and Ng, 2015), *convincingness* (Habernal and Gurevych, 2016), and *relevance* (Wachsmuth et al., 2017c). However, Gretz et al. (2020) note the need to predict quality in terms of fine-grained aspects. Fine-grained prediction enables a deeper understanding of argumentation and offers specific feedback to authors aiming to improve their argumentative writing skills. For instance, authors might want to know whether their premises are *sufficient* with regard to their claim(s) or whether their language is *appropriate*. Wachsmuth et al. (2017b) surveyed and synthesized theory-based dimensions of AQ into a taxonomy of three main dimensions: Cogency (Logic), Effectiveness (Rhetoric), and Reasonableness (Dialectic). Their initial annotation study showed that assessing these dimensions is challenging, even for experts, but that crowd workers can handle the task comparably well if the guidelines and task are simplified.

Given the feasibility of annotation and the recognized need for fine-grained dimensions in AQ assessment, it is surprising that no further efforts in NLP and CL have been made. There is no large scale annotated corpus and, consequently, no computational model. In this work, we aim to fill this research gap by conducting an in-depth analysis of theory-based AQ assessment covering overall AQ and the three dimensions (logic, rhetoric, and dialectic) of the Wachsmuth et al. taxonomy, and three diverse domains of online argumentative writing (Q&A forums, debate forums, and review forums).

Drawing on existing AQ theories, we address five research questions (**RQs**) to inform and fuel future AQ annotation studies and computational AQ research:

[1]We adopt the terminology of Wachsmuth et al. (2017a) who refer to task-driven approaches, which often also focus on the *relative* assessment of AQ, as "practical".

Proceedings of the 28th International Conference on Computational Linguistics, pages 4563–4574
Barcelona, Spain (Online), December 8-13, 2020

RQ1: *Can we develop a large-scale theory-based AQ corpus?* We conduct an extensive annotation study with trained linguists and crowd workers on 5,295 arguments from three domains to create the Grammarly Argument Quality Corpus (`GAQCorpus`), the first large-scale multi-domain English corpus annotated with theory-based AQ scores.

RQ2: *Are we able to develop computational models that can do theory-based AQ assessment in varying domains?* Based on `GAQCorpus`, we are the first to propose computational approaches to theory-based AQ assessment and show that it is possible to develop models for this task. Our models can serve as strong baselines for future research and enable the field to investigate follow-up research questions.

RQ3: *Can the interrelations between the different AQ dimensions be exploited in a computational setup?* Inspired by the hierarchical structure of the taxonomy, we explore whether the relationships between dimensions can be computationally exploited. In addition to simple single-task learning approaches, we study the effect of jointly predicting AQ dimensions in two variants (*flat* vs. *hierarchical*) and find that combining the training signals of all four aspects benefits theory-based AQ assessment.

RQ4: *Does the corpus support training a single unified model for multi-domain evaluation?* When enough data from a single domain is available, training on in-domain data is typically preferred over multi-domain. However, larger amounts of data are especially useful for complex model architectures currently prominent in NLP (e.g., BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019)). We study these two mutually opposing effects on `GAQCorpus` and show that our corpus supports training a single unified model across all three domains, with improved performances in individual domains.

RQ5: *Can we empirically substantiate the idea that theory-based and practical AQ assessment can learn from each other?* Wachsmuth et al. (2017a) suggest that both the practical and the theory-based views can learn from each other, but so far, this has been only tested manually. Employing our models, we go one step further and conduct a bi-directional experiment employing a practical AQ corpus. We demonstrate two concrete ways how theory-based and practical AQ research can profit from their combination.

Structure. After discussing related work in §2, we describe our annotation study and resulting corpus (§3). §4 describes the computational approaches which we employ in the experiments (§5). Last, we conclude our work and give potential directions for future work (§6).

2 Related Work

Earlier work in computational AQ assessment can be divided into practical and theory-based approaches.

Practical approaches. Recently, the field of computational AQ research has been mostly driven by practical approaches that each target an individual domain. Accordingly, past approaches tackle either overall quality (Toledo et al., 2019) or specific subqualities of argumentation, such as convincingness (Habernal and Gurevych, 2016) and relevance (Wachsmuth et al., 2017c). The popularity of practical approaches can partly be attributed to the relative simplicity of crowd-sourcing annotations.

Much prior work has focused on aspects of student essays, including essay clarity (Persing and Ng, 2013), organization (Persing et al., 2010), prompt adherence (Persing and Ng, 2014), and argument strength (Persing and Ng, 2015). Later, Wachsmuth et al. (2016) present an approach driven by detecting argumentative units, thereby demonstrating the usefulness of argument mining techniques to the problem. Similarly, Stab and Gurevych (2016) predict the absence of opposing arguments and Stab and Gurevych (2017) predict insufficient premise support in arguments. Another well-studied domain is web debates. Wachsmuth et al. (2017c) adapt PageRank to identify argument relevance. Pairwise comparison of the convincingness of debate arguments has been conducted (Habernal and Gurevych, 2016). Persing and Ng (2017) additionally predict why an argument receives a low persuasive power score. By explaining flaws in argumentation, they highlight the importance of explainability and specific author feedback.

Other approaches take into account properties of the source, i.e., the author (Durmus and Cardie, 2019) or the audience (El Baff et al., 2018; Durmus and Cardie, 2018). In contrast, we assume that a system may not have much knowledge about the authors or audience and thus our models operate solely on the text. Toledo et al. (2019) and Gretz et al. (2020) present large corpora of crowd-sourced arguments and their quality. These corpora cover a variety of topics, but only within single domains. The authors emphasize

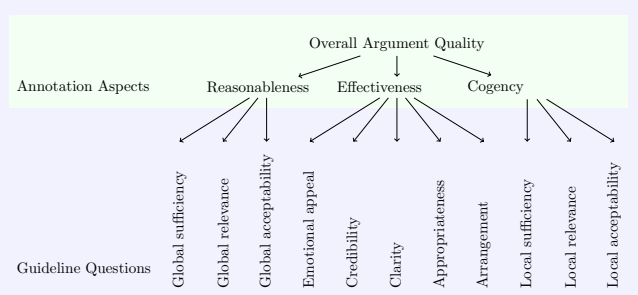

Figure 1: Taxonomy of theory-based AQ (Wachsmuth et al., 2017b). Questions related to each aspect guided annotators in assessing higher level dimensions.

Title: *Should 'blogging' be a capital crime? Iran is considering it...*
Stance: *A government has the right to censor speech (...)*
Text: *My government doesn't give me freedom of speech, so I have to argue for this side. Freedom of speech is bad because ... um ... then Our Leader's beliefs could be challenged. No one wants that. I mean, if everyone would just say and believe what Our Leader says to, we wouldn't need those firing squads altogether! Everyone wins.*

	Cogency	Effectiveness	Reasonableness	Overall
Annotator 1	4	1	1	2
Annotator 2	4	5	3	4
Annotator 3	2	2	2	2

Figure 2: Example text from our annotation pilot. Linguistic expert annotators highly disagree on scoring the effectiveness dimension.

that research on theory-based approaches could further advance the field of computational AQ.

Theory-based approaches. Rooted in classic argumentation theory, the works can according to Wachsmuth et al. (2017b), be categorized based on whether they related to the *logical* (Johnson and Blair, 2006; Hamblin, 1970), *rhetorical* (Aristotle, 2007), or *dialectical* (Chaïm Perelman and Weaver, 1969; Van Eemeren et al., 2004) properties of an argument.

Wachsmuth et al. (2017b) were the first to survey and highlight the importance of the theory-based approach to computational AQ and synthesized the argumentation-theoretic literature into a taxonomy. Wachsmuth et al. (2017a) conducted a study in which crowd workers annotated 304 arguments for all 15 quality dimensions following Wachsmuth et al. (2017b), and demonstrated that the theory-based and practical AQ assessment match to a large extent and that the two views can learn from each other, for instance, when it comes to more practical annotation processes for theory-based AQ annotations.

However, until now, no further research on computational theory-based AQ assessment in NLP has been conducted, no larger-scale annotated corpus has been presented, and thus no computational model that would allow further investigation into the concrete synergies between the two perspectives exists.

3 Annotation Study

Wachsmuth et al. (2017a) suggest that large-scale annotation of theory-based AQ dimensions is possible. We test this finding and take it one step further by asking whether we can develop a large-scale theory-based AQ corpus (**RQ1**). This section presents `GAQCorpus`, the result of the first study annotating theory-based dimensions, including 5,285 arguments from three diverse domains of real-world argumentative writing.

3.1 Annotation Scheme

Our annotation scheme is based on the Wachsmuth et al. (2017a) taxonomy of argumentation quality depicted in Figure 1. It defines **overall AQ** as being composed of three sub-dimensions (Cogency, Effectiveness, Reasonableness), each of which is in turn composed of several quality-related aspects:

- **Cogency** relates to the logical aspects of AQ. High cogency indicates that an argument's premises are acceptable as well as relevant and sufficient with regard to the argument's conclusion.
- **Effectiveness** reflects the persuasive power of how an argument is stated. Important aspects of an effective argument include its arrangement, clarity, appropriateness in a given context, emotional appeal, and author's credibility.
- **Reasonableness** indicates the quality of an argument in the context of a debate, i.e., its relevance, its acceptability and the way it is stated as a whole, and its sufficiency toward the resolution of the issue.

Starting from the guidelines of Wachsmuth et al. (2017b), we developed our annotation guidelines through a series of pilot studies with four *expert annotators* who are all fluent or native English speakers with advanced degrees in linguistics. Wachsmuth et al. (2017a) recommend simplifying the task and guidelines, and based on the findings of our pilots, we made the following modifications under consultation with our experts: Since the annotators noted difficulties distinguishing between the 15 fine-grained aspects, we

collapse the scheme to Overall AQ and the three higher level dimensions and represent the finer-grained sub-dimensions as questions to guide the annotators' judgments. We additionally use a five-point scale (very low, low, medium, high, very high, plus "cannot judge"), which simplifies the task according to feedback from our expert annotators and previous findings (Cox III, 1980). We experimented with both the five-point and the original three-point rating scale (low, medium, high) used by Wachsmuth et al. (2017b), and found that switching the scales did not negatively affect inter-annotator agreement.

Ng et al. (2020) describe the annotation design and guidelines in more detail.

3.2 Data

We investigate different domains to obtain a deeper understanding of real-world AQ and the feasibility of the annotation scheme in different settings. We include three domains in our study: Community questions and answers forum posts (*CQA*), debate forum posts (*Debates*), and business review forum posts (*Reviews*). Figures 2 and 3 display an example text for each domain.

CQA. We include 2,088 arguments from Yahoo! Answers,[2] a community questions and answers forum where users ask questions and answer questions posted by others. Not enforcing strict debating rules or topics, the argumentative posts are diverse and therefore interesting for our study. While not a dedicated debate forum, we found that some categories contain a relatively high proportion of argumentative posts, like *Politics & Government* → *Law & Ethics*, from which we select posts. We only include posts marked as best answer for a question and exclude posts containing uniform resource identifiers or media content. From these, we select posts that were labeled as argumentative by a majority of 10 raters in a secondary experiment (see Appendix).

Debates. To reflect online debate forums-style argumentation, we include 1,337 arguments from Change My View (CMV) (Tan et al., 2016) and 766 from the Internet Argument corpus V2 (IAC) (Abbott et al., 2016) resulting in 2,103 arguments in total. CMV is an internet forum in which users post their opinion and ask others to challenge their beliefs on the topic. The IAC is composed of posts retrieved from three online debate forums, and in this study we include only arguments from the ConvinceMe subset. We try to restrict the sample to instances that do not require much background knowledge or thread-level context. From CMV, we include original posts only and for ConvinceMe, we include the first post reacting to the topic. From CMV we also exclude posts tagged `[MOD]`, which indicate moderator posts.

Reviews. Yelp is an online platform where users publish business reviews and rate their experience from 1 (poor) to 5 (excellent) stars. From the Yelp-Challenge-Dataset[3], we sampled 1,104 arguments reviewing restaurants. While the review texts often do not appear as "classic" arguments, i.e., with a dedicated claim and premises supporting this claim, the texts can indeed be considered argumentative (Wachsmuth et al., 2014; Wachsmuth et al., 2015); The star rating corresponds to a claim a user is making about the business and the review text is intended to support this claim.

For Debate and Review posts, we include the star rating and stance (if provided) with the text. Across all domains, we filter for posts with text length between 70 and 200 words. To ensure high quality annotations, we first ran 13 pilot studies in two flavors: (1) with three of the linguistic expert annotators (§3.1), and (2) with a crowd-sourced workforce of 24 contributors from Appen.[4] Both groups used the same annotation guidelines and interface, which we iteratively improved based on feedback from each pilot. Table 1 shows the number of judgments per instance per domain as well as the number of instances that were annotated by each group. For each domain, up to 538 arguments were annotated by both experts and crowd workers.

We provide and use a standard split for each domain, which is composed as follows: The training and development sets consist of the instances which were *either* annotated by our linguistic experts or the crowd workers. In contrast, the test sets encompass only instances scored by *both* experts and the crowd. For each instance and group, we obtain a single score by averaging the annotators' votes. In addition to

[2] `https://answers.yahoo.com/`

[3] `https://www.yelp.com/dataset`

[4] Formerly *Figure Eight*, `https://www.appen.com/`

Question: *should juveniles be trialed as adults?*

Answer: *It all depends on the crime. For the most part i believe if your grown enough to go and do an adult crime then you need to do the adult time. If we continue to let the youth get away with serious crimes then older crimebodies will continue to get our youth in trouble. We must raise our children correctly so they want end up in some prison but there are certain things that is morally wrong no matter if your 15 or 35 and those are the crimes our young "adult" should be charged for.*

(a) Community Q&A Forums.

Title: *Business name: Little Shanghai. City: Pittsburgh. Categories: Restaurants, Chinese*
Stars: *5.0*

Review: *Little Shanghai has the best Chinese food that I've been able to find in the city. The steamed flounder with bean curd is great. It comes in 2 fillets for $13.95. I loved the texture of the crispy tofu in the spinach with garlic and tofu dish. The broth of the noodle soup with spare ribs has a wonderful flavor and the dish is more than enough to fill up one person. I wish the restaurant had better loose leaf tea (they use a tea bag) but the food is excellent. I would highly recommend this restaurant.*

(b) Review Forums.

Figure 3: Example texts and quality trends provided by our linguistic experts.

	Crowd	Experts			Overlap
# Annotators	10	1	2	3	11–13
CQA	1,334	626	–	625	500
Debates	1,438	600	–	600	538
Reviews	600	200	400	–	100

Table 1: Number of annotators per instance and total instances annotated by Experts and the Crowd, and the number of overlapping instances by domain.

Domain	Total	Train	Dev.	Test
CQA	2,085	1,109	476	500
Debates	2,100	1,093	469	538
Reviews	1,100	700	300	100
All	5,285	2,902	1,245	1,138

Table 2: Number of instances in the train, development, and test sets of `GAQCorpus`.

	Cogency	Effectiveness	Reasonableness	Overall
Ours	**0.46**	**0.48**	**0.48**	**0.55**
TvsP	0.27	0.38	0.13	0.43

Table 3: Agreement between expert annotations from Wachsmuth et al. (2017b) and crowd-sourced annotations from two sources: `GAQCorpus` (Ours) and TvsP on 200 randomly sampled instances.

	Cogency	Effectiveness	Reasonableness	Overall
CQA	0.42	0.52	0.52	0.53
Debates	0.14	0.11	0.21	0.19
Reviews	0.32	0.32	0.31	0.33

Table 4: IAA between the Expert and Crowd scores for Cogency, Effectiveness, Reasonableness, and Overall AQ.

the group-specific annotations (*expert* and *crowd*), we also compute a *mix* score which is the average of the two group-specific scores. This way, we train on a mix of expert and crowd annotations (where the dominant portion comes from the crowd) and test on overlapping instances, enabling us to compare model performance to both expert and crowd ratings on a static set of instances. The numbers of instances in each portion of `GAQCorpus` are given in Table 2.

3.3 Data Analysis

Inter-annotator Agreements (IAA). To assess the quality of our crowd-sourced annotations and our simplified guidelines, we employ the Dagstuhl-ArgQuality-Corpus-V2 (DS, originally from UKPConvArgRank (Habernal and Gurevych, 2016)) and conduct a comparative study against the crowd-sourced Wachsmuth et al. (2017a) annotations (TvsP). We take "gold" ratings from the original, author-produced annotations of Wachsmuth et al. (2017b). DS was presented in combination with the taxonomy of theory-based AQ described above and consists of 304 web debate arguments annotated with all 15 AQ aspects. We randomly sample 200 arguments and crowd-source annotations on Appen with our revised guidelines.[5] For each instance and AQ dimension, we collect 10 votes and average them to obtain the

[5]Here, we stuck to the original 3-point scale to match the original expert annotations we compare with.

group vote (Mean). We measure IAA between the group vote and the DS expert vote with Krippendorff's α (Krippendorff, 2007). The results are depicted in Table 3. The agreement does not exceed 0.55, which is not surprising for a task of this subjectivity, and generally, the agreement scores of our crowd ratings surpass the agreement scores reported by TVSP. we therefore conclude that our guidelines and user interface support the task and confirm the suitability of our crowd annotators.

Next we consider the agreement between experts and crowd workers on the overlapping portions of GAQCorpus using the mean scores (Table 4). For debate forums, Krippendorff's α is up to 0.21, while for the Q&A forums, the agreement is higher – up to 0.53. These results suggest that the difficulty of the task is highly dependent on the domain.

Analysis of Disagreements. We noticed disagreements among the annotators along all stages of the annotation process, especially for arguments which were of sarcastic or ironic nature or included rhetorical questions. Consider the argument given in Figure 2 as an example.

This example on the topic of *freedom of speech* seems to support the stance that a government has the right to censor speech. However, several linguistic cues indicate that the argument might be ironic: (a) Punctuation: Ellipsis indicates thinking/searching for justifications; similarly, (b) the filler *um*; (c) Capitalization: The noun phrase *Our Leader* is capitalized, indicating hyperbolic apotheosis; and finally, (d) the phrase *(...) so I have to argue for this side.* acts like an apologia, which is put in front of the actual argument. In discussion with our expert annotators, it became clear that Annotators 1 and 2 based their judgments on an interpretation of this text that related to the estimated degree of irony in the post. While Annotator 1 did not perceive irony and judged the argument as *very weak* in *Effectiveness*, Annotator 2 considered it to be highly effective as in their view, the irony positively underlined the perceived stance. Annotator 3 gave medium scores across the board but was leaning more towards Annotator 2's opinion. Such disagreements were regularly discussed and usually revealed that multiple opinions may exist according to how the texts were interpreted, which highlights the high subjectivity of the task.

Disagreements can also be observed across different domains. Debates and CQA are dialectic by nature, but original posts (or top answers in the case of CQA) are relatively straightforward to assess in isolation. In contrast, business reviews are monologues and cite experiences as justifications for a claim, e.g., *My meal was delicious*. Given that experience is subjective, evaluating reviews presents unique challenges.

4 Models

Having developed GAQCorpus to enable computational AQ assessment (**RQ1**), we address the remaining research questions by experimenting with AQ models. To determine whether we can develop a computational theory-based AQ model (**RQ2**), we employ a naive length baseline, three different Support Vector Regression (SVR) models, and a BERT-based (Devlin et al., 2019) model. We next investigate whether the interrelations between AQ dimensions can be exploited in a computational setup (**RQ3**), employing two multi-task BERT-based models. For the BERT-based models, we transform each argument into a "BERT-compatible" format, i.e., into a sequence of WordPiece (Wu et al., 2016) tokens and prepend the sequence with BERT's start token (`[CLS]`). The pooled hidden representation of the latter corresponds to the aggregated document representation. The specific details of each model are described below.

Argument Length (ARG LENGTH). To estimate the task difficulty and to measure a potential length bias, our naive baseline is the correlation between the argument's character length and quality scores.

SVR with Lexical Features (SVR_{tfidf}). We run a simple SVR with tf–idf feature vectors.

SVR with Semantic Features (SVR_{embd}). We represent each argument as the average of the fastText (Bojanowski et al., 2017) embedding[6] representation of each word in the argument.

Feature-rich SVR (WACHSMUTH$_{CFS}$). We reimplement the approach of Wachsmuth et al. (2016), who employ standard features (token n-grams, part-of-speech tags, etc.) and higher-level features (sentiment

[6] https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M-subword.vec.zip

	Model	CQA	Debates	Reviews
Overall	ARG LENGTH	0.406	0.420	0.365
	SVR_{tfidf}	0.389	0.265	0.450
	SVR_{embd}	0.278	0.388	0.265
	WACHSMUTH_{CFS}	0.492	0.432	0.533
	BERT ST	**0.652**	**0.511**	**0.605**
	BERT MT_{flat}	**0.667**	**0.537**	0.588
	BERT MT_{hier}	0.661	0.494	0.593
Cogency	ARG LENGTH	0.420	0.437	0.340
	SVR_{tfidf}	0.444	0.257	0.384
	SVR_{embd}	0.261	0.333	0.103
	WACHSMUTH_{CFS}	0.503	0.429	0.464
	BERT ST	**0.587**	**0.503**	**0.554**
	BERT MT_{flat}	0.633	**0.541**	**0.561**
	BERT MT_{hier}	**0.638**	0.474	0.541

	Model	CQA	Debates	Reviews
Effectiveness	ARG LENGTH	0.390	0.399	0.372
	SVR_{tfidf}	0.411	0.120	0.340
	SVR_{embd}	0.293	0.403	0.187
	WACHSMUTH_{CFS}	0.523	0.450	0.432
	BERT ST	**0.612**	**0.542**	**0.555**
	BERT MT_{flat}	**0.671**	**0.570**	0.514
	BERT MT_{hier}	0.670	0.532	0.486
Reasonableness	ARG LENGTH	0.396	0.377	0.405
	SVR_{tfidf}	0.457	0.247	0.452
	SVR_{embd}	0.379	0.258	0.234
	WACHSMUTH_{CFS}	0.476	0.399	0.432
	BERT ST	**0.665**	**0.418**	**0.609**
	BERT MT_{flat}	0.644	**0.473**	0.610
	BERT MT_{hier}	0.626	0.408	**0.611**

Table 5: Pearson correlations of our model predictions with the annotation scores on the mix test annotations when training on in-domain data. Numbers in bold indicate best performances.

flows, argumentative discourse units etc.). We run correlation-based feature selection on the training set and include only the most predictive features.

Single Task Learning Setting (BERT ST). For each AQ dimension, we train an individual regressor. Our AQ predictor is a simple linear regression layer in which we feed the pooled document representation. The loss L_t is then simply the mean squared error (MSE) over k instances in the training batch.

Flat Multi-Task Learning Setting (BERT MT_{flat}). We explore whether a joint training setup would improve the individual score predictions. For each quality dimension, we employ an individual prediction layer as described above and compute an individual task loss. We then define the total training loss as the sum of the task losses.

Hierarchical Multi-Task Learning Setting (BERT MT_{hier}). We propose a hierarchical multi-task learning setting to exploit the hierarchical relationship between the scores. Similar to above, we first learn jointly the lower-level tasks (Cogency, Effectiveness, Reasonableness) resulting in three scores $\hat{y}_{\text{Cog}}$, $\hat{y}_{\text{Eff}}$ and $\hat{y}_{\text{Rea}}$. Next, we employ these scores for informing the overall AQ predictor by concatenating these with the hidden document representation $\mathbf{h}_D$: $\mathbf{h}_{informed} = \mathbf{h}_D^{\frown}[\hat{y}_{\text{Cog}}, \hat{y}_{\text{Eff}}, \hat{y}_{\text{Rea}}]$. The resulting vector $\mathbf{h}_{informed}$ serves as input to the overall AQ predictor as defined in before.

5 Experiments

We employ the proposed architectures above to answer research questions RQ2–RQ5.

5.1 RQ2: Computational theory-based AQ assessment

To test whether our corpus supports the development of theory-based AQ assessment models, this experiment employs all single-task models presented in Section 4 (ARG LENGTH, SVR_{tfidf}, SVR_{embd} WACHSMUTH_{CFS}, and BERT ST). We train and predict on the domain-specific data sets and report the results on the *mix* test set per AQ dimension for each domain.[7] Details on the hyperparameter optimization can be found in the appendix.

Results. The respective Pearson correlation scores for AQ dimensions on the three domain-specific test sets are shown in Table 5. Generally, we reach medium to high Pearson correlation scores of up to nearly 0.67. However, like the IAA, performance varies across domains: On Debates, the best model, BERT ST, achieves a correlation coefficient with the annotation scores for reasonableness of 0.42 and on the CQA forums, it achieves a performance of 0.67. The BERT-based regressor outperforms the other methods, showing that we can successfully utilize a large-scale corpus with theory-based AQ dimensions

[7]Trends for the other evaluation sets (crowd and expert) are similar. Full results can be found in the supplementary material.

	Model	CQA	Debates	Reviews
Overall	Best in-domain	0.667	0.537	0.605
	BERT ST	0.676	0.545	0.596
	BERT MT_{flat}	**0.681**	**0.562**	**0.633**
	BERT MT_{hier}	0.665	**0.562**	0.622
Cogency	Best in-domain	0.638	0.541	0.561
	BERT ST	0.608	0.515	0.563
	BERT MT_{flat}	**0.653**	0.542	0.570
	BERT MT_{hier}	0.638	**0.552**	**0.599**
Effectiveness	Best in-domain	0.671	0.570	0.555
	BERT ST	**0.686**	**0.598**	0.601
	BERT MT_{flat}	0.670	0.578	**0.603**
	BERT MT_{hier}	0.653	0.592	0.576
Reasonableness	Best in-domain	**0.665**	0.473	0.611
	BERT ST	0.635	**0.487**	0.603
	BERT MT_{flat}	0.657	0.486	0.631
	BERT MT_{hier}	0.633	0.483	**0.643**

Table 6: Pearson correlations of the model predictions with the annotation scores when training on the joint training sets of all domains. We compare with the best result of the in-domain setting.

Domain	Dimension	r	ρ
BERT IBM	–	0.492	0.456
Gretz et al. (2020)	–	0.52	0.48
All	Overall	**0.313**	**0.303**
	Cogency	0.311	0.300
	Effectiveness	**0.313**	**0.303**
	Reasonableness	0.304	0.298
CQA Forums	Overall	0.258	0.224
	Cogency	**0.269**	**0.228**
	Effectiveness	0.262	0.225
	Reasonableness	0.262	0.226
Debate Forums	Overall	**0.336**	**0.326**
	Cogency	0.331	0.321
	Effectiveness	**0.336**	**0.326**
	Reasonableness	0.333	0.319
Review Forums	Overall	0.150	0.145
	Cogency	0.139	0.138
	Effectiveness	**0.152**	**0.151**
	Reasonableness	0.149	0.148

Table 7: Performance of BERT MT_{flat} trained on `GAQCorpus`, predicting on IBM-Rank-30k evaluated against the weighted average score.

to train models for automatic AQ assessment (**RQ2**). Note that ARG LENGTH is relatively high across all domains and properties and often outperforms SVR_{tfidf} and SVR_{embd}, indicating a slight length bias in the corpus. However, BERT ST outperforms this baseline in all cases by a large margin, demonstrating this model's ability to capture useful information beyond pure length.

5.2 RQ3: Effect of AQ dimension interrelations

Next we seek to determine whether it is possible to exploit the interrelations between the three dimensions and the overall AQ by conducting experiments on `GAQCorpus`. We compare the multi-task learning architectures, BERT MT_{flat} and BERT MT_{hier}, against the results of the BERT ST model, the best performing single-task model. Again, we train and predict on the domain-specific data splits.

Results. Table 5 shows the respective Pearson correlation scores for the four AQ dimensions on each domain. The multi-task learning models outperform the single-task model in 9 out of 12 cases,which suggests that the interrelations between the AQ dimensions and overall AQ can be exploited to improve model performance (**RQ3**). More specifically, the best method is BERT MT_{flat}, which outperforms the other methods in 7 cases. BERT ST and BERT MT_{hier} are best in 3 and 2 cases, respectively.

5.3 RQ4: Unified multi-domain model

We examine whether our corpus supports training a unified multi-domain model. To this end, we train the BERT-based models on the joint training set covering all domains and test performance on each individual domain, thereby including out-of-domain data in training. Similarly, we optimize the hyperparameters on the joint development set. We compare with the best in-domain score from Table 5.

Results. The respective results for the four argument quality dimensions can be seen in Table 6. In 11 out of 12 cases, training on all domains increases the performance compared to the best in-domain model. While the models are less domain-specific, the increased amount of data leads to better convergence and lead to gains up to 5 percentage points.

5.4 RQ5: Synergies between practical and theory-driven AQ

To empirically test the hypothesis that synergies exist between practical and theory-based AQ assessment, we conduct a bi-directional experiment with the recently released IBM-Rank-30k (Gretz et al., 2020).

		CQA	Debates	Reviews
Overall	BERT IBM	0.392	0.317	0.154
	BERT IBM MT$_{flat}$	0.666	0.543	0.568
	BERT MT$_{flat}$	**0.681**	**0.562**	**0.633**
Cogency	BERT IBM	0.368	0.274	0.149
	BERT IBM MT$_{flat}$	0.639	0.518	0.541
	BERT MT$_{flat}$	**0.653**	**0.542**	**0.570**

		CQA	Debates	Reviews
Effectiveness	BERT IBM	0.426	0.378	0.195
	BERT IBM MT$_{flat}$	**0.678**	**0.594**	0.545
	BERT MT$_{flat}$	0.670	0.578	**0.603**
Reasonableness	BERT IBM	0.348	0.246	0.151
	BERT IBM MT$_{flat}$	0.637	0.465	0.581
	BERT MT$_{flat}$	**0.657**	**0.486**	**0.631**

Table 8: Pearson correlations on GAQCorpus when predicting with BERT IBM (trained on IBM-Rank-30k) and BERT IBM MT$_{flat}$ trained on IBM-Rank-30k in STILT setup fine-tuned on GAQCorpus in comparison to BERT MT$_{flat}$.

Experimental setup. IBM-Rank-30k consists of 30,497 crowd-sourced arguments relating to 71 topics, where each argument is restricted to 35–210 characters. The corpus has binary judgments indicating whether raters would recommend the argument to a friend. Based on these ratings, a score for each argument was computed, either using MACE or weighted average of all ratings. Compared to GAQCorpus, IBM-Rank-30k is much larger but the arguments are much shorter and more artificial than real world texts. Manual inspection revealed that the nature of the texts substantially differs from each those in GAQCorpus, i.e., arguments mainly cover reasons for higher-level claims. For example, in IBM-Rank-30k for the topic *"We should end racial profiling"*, a highly rated argument is *"racial profiling unfairly targets minorities and the poor"*.

We conduct three experiments in two directions: (E1) train on GAQCorpus, then predict on IBM-Rank-30k, (E2) train on IBM-Rank-30k, then predict on GAQCorpus, and finally, (E3) train on IBM-Rank-30k, next, train on GAQCorpus, and then, predict on GAQCorpus.

For **experiment (E1)**, we take the (already trained) BERT MT$_{flat}$ models trained on each domain of GAQCorpus and predict on the test portion of IBM-Rank-30k. This enables us to determine which one of our domains and dimensions are closest to the data and annotations in IBM-Rank-30k. We compare against the best score reported in the Gretz et al. (2020) as well as against our own reimplementation using BERT$_{BASE}$, dubbed BERT IBM.[8] We optimize the BERT IBM baseline by grid searching for the learning rate $\lambda \in \{2e-5, 3e-5\}$ and the number of training epochs $\in \{3,4\}$ on the IBM-Rank-30k development set. For the already trained models from Sections 5.2 and 5.3, no further optimization is necessary. In **experiment (E2)**, we reverse the direction of (E1): We train a BERT-based regressor as defined before on the MACE-P aggregated annotations of IBM-Rank-30k.[9] We predict on GAQCorpus and correlate the scores with our annotations. Finally for **experiment (E3)**, in order to flatten out expected losses from the zero-shot domain transfer, inspired by Phang et al. (2018) we use IBM-Rank-30k in the Supplementary Training on Intermediate Labeled Tasks-setup (STILT), i.e., we take the trained BERT IBM encoder and continue training the model as BERT IBM MT$_{flat}$ in the all-domain setup. We compare both models from (2) and (3) with the BERT MT$_{flat}$.

Results. In experiment (E1) (Table 7), as expected, the zero-shot domain transfer results in a large drop compared to training on the associated train set of IBM-Rank-30k. Quite surprisingly, the model trained on the debate forums reaches the highest correlation scores – even higher than the model trained on *all-domains*. Further, in most cases, the effectiveness predictions correlate best with the annotations provided by Gretz et al. (2020). This is in-line with the authors' observations, who manually had to annotate the data for the theory-based scores.

Table 8 displays the results of (E2)–(E3). Experiment (E2), draws a similar picture: zero-shot domain transfer using BERT IBM results in a huge loss in performance compared to BERT MT$_{flat}$. Finally, the results in (E3) indicate potential for using resources drawn from practical approaches in a theory-based AQ assessment scenario: When reusing the encoder in the STILT setup, BERT IBM MT$_{flat}$, the losses originating from the zero-shot domain transfer can be flattened out – in some cases even outperforming

[8]Note that Gretz et al. (2020) do not indicate whether they employ BERT$_{BASE}$ or BERT$_{LARGE}$.

[9]This corresponds to our BERT IBM baseline from before.

BERT MT$_{flat}$. This is especially the case when correlating the predictions with our annotations for the effectiveness dimensions. To sum up, our experiment (E1)–(E3) yield the following findings: (1) Large-scale predictions, obtained from a theory-based AQ model on a large (practical) AQ data set, correlate mostly with the Effectiveness dimension. (2) The transferred knowledge obtained in the STILT-setup on IBM-Rank-30k in BERT IBM MT$_{flat}$ improves the performance on `GAQCorpus` for Effectiveness the most. These two facts match Gretz et al. (2020)'s hypothesis that their annotations mostly captured Effectiveness. We empirically substantiate the idea (without any manual effort) that a theory-based approach can inform practical AQ research and increase interpretability of practically-driven research outcomes and, on the other hand, the practical approach can increase the efficacy of theory-based AQ models when targeting a certain domain and dimension.

6 Conclusion and Future Work

Specific assessment of the rhetorical, logical, and dialectical perspectives on argumentative texts can inform researchers, e.g., about phenomena captured with annotations, and help people improve their writing skills. However, the field of computational AQ assessment has been almost exclusively driven by practical approaches. Aiming to fill this gap, in this work we advance theory-based computational AQ research with the following contributions:

We performed a large-scale annotation study on English argumentative texts covering debate forums, Q&A forums, and business review forums. We thereby presented `GAQCorpus`, the largest and first multi-domain corpus annotated with theory-based AQ scores (**RQ1**).[10] Next, we proposed the first computational theory-based AQ models (**RQ2**) and demonstrated that jointly predicting AQ scores can improve the performance of the models (**RQ3**) and that in most cases, models benefit from including out-of-domain training data (**RQ4**). Finally, we investigated concrete synergies between the practical and the theory-based approach to AQ assessment in a bi-directional experimental setup (**RQ5**). The theory-based models can help to increase the interpretability of practical approaches, and practical approaches can be employed to increase performance of the theory-based models. In the future, we would like to deploy the models and study to what extent users can actually improve their argumentative writing by getting theory-based AQ feedback. Further, we will seek to develop ways of adding even finer-grained aspect scores at scale; this remains an open problem.

Acknowledgements

The work of Anne Lauscher is supported by the Eliteprogramm of the Baden-Württemberg Stiftung (AGREE grant). We thank our linguistic expert annotators for providing interesting insights and discussions as well as the anonymous reviewers for their helpful comments. We also thank Henning Wachsmuth for consulting us w.r.t. his previous work and Yahoo! for granting us access to their data.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4445–4452, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, Oxford, UK. Translated by George A. Kennedy.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

John Wilkinson Chaïm Perelman, Lucie Olbrechts-Tyteca and Purcell Weaver. 1969. The new rhetoric: A treatise on argumentation. Notre Dame, IN. University of Notre Dame Press.

[10] Available from `https://github.com/grammarly/gaqcorpus` with annotation guidelines and interface.

Eli P. Cox III. 1980. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4):407–422.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana, June. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *The World Wide Web Conference*, WWW '19, page 2701–2707, New York, NY, USA. Association for Computing Machinery.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium, October. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Preoceedings of AAAI2020*.

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, August. Association for Computational Linguistics.

Charles L Hamblin. 1970. *Fallacies*. Methuen, London, UK.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. International Debate Education Association, New York.

Klaus Krippendorff. 2007. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania, Annenberg School for Communication.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain, April. Association for Computational Linguistics.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. AQCorpus: A domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining (ArgMining 2020)*.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria, August. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland, June. Association for Computational Linguistics.

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)*, pages 543–552.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4082–4088. AAAI Press.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 229–239, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain, April. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*, pages 613–624.

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)*, pages 5629–5639.

Frans Van Eemeren, Frans H Van Eemeren, and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 115–127. Springer.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment flow-a general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada, July. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. 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*, pages 176–187, Valencia, Spain, April. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain, April. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.