Cross-Domain Argument Quality Estimation

Michael Fromm[1], Max Berrendorf[2], Evgeniy Faerman[2], Thomas Seidl[2]
[1]Fraunhofer IAIS, Germany
[2]Database Systems and Data Mining, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML)

Abstract

Argumentation is one of society's foundational pillars, and, sparked by advances in NLP, and the vast availability of text data, automated mining of arguments receives increasing attention. A decisive property of arguments is their strength or quality. While there are works on the automated estimation of argument strength, their scope is narrow: they focus on isolated datasets and neglect the interactions with related argument-mining tasks, such as argument identification and evidence detection. In this work, we close this gap by approaching argument quality estimation from multiple different angles: Grounded on rich results from thorough empirical evaluations, we assess the generalization capabilities of argument quality estimation across diverse domains and the interplay with related argument mining tasks. We find that generalization depends on a sufficient representation of different domains in the training part. In zero-shot transfer and multi-task experiments, we reveal that argument quality is among the more challenging tasks but can improve others. We publish our code at https://github.com/fromm-m/acl-cross-domain-aq.

1 Introduction

The argumentation process is one of the cornerstones of society, as it allows the exchange of opinions and reaching a consensus together. Fueled by advances in natural language processing, recent years have witnessed the advent of Argument Mining (AM), i.e., the field of automated discovery and organization of arguments. AM is helpful over various scenarios, reaching from legal reasoning (Wyner et al., 2010; Walker et al., 2014; Poudyal et al., 2020; Villata, 2020) to supporting the decision-making process of politicians (Haddadan et al., 2019; Duthie et al., 2016; Menini et al., 2017; Lippi and Torroni, 2016; Awadallah et al., 2012). Thus, there is a flurry of works on identification of arguments from text (Stab et al., 2018b; Fromm et al., 2019; Trautmann et al., 2020) and retrieval of them (Wachsmuth et al., 2017c; Fromm et al., 2021; Dumani and Schenkel, 2019; Dumani et al., 2020; Stab et al., 2018a). Since arguments often have to be weighed against each other, a central property of arguments is their Argument Quality (AQ) or *convincingness*, i.e., their (perceived) strength. While the ancient Greeks (Rapp, 2002) already discussed the constituents of strong arguments, automated estimation is a relatively uncharted field. Due to the high subjectivity of argument strength (Swanson et al., 2015; Gretz et al., 2020; Toledo et al., 2019; Habernal and Gurevych, 2016b; Stab et al., 2018b), obtaining high-quality annotations is challenging, cf. Section 1. In this light, a legitimate question is the reliability and robustness of the existing approaches for estimating AQ and their applicability in real-life scenarios. Existing AQ benchmark datasets are often restricted to a single domain (Wachsmuth et al., 2016; Persing and Ng, 2017) or/and make different assumptions about factors impacting the AQ. Thus, enabling transfer between sources and datasets appears especially appealing, but existing works (Gretz et al., 2020; Toledo et al., 2019; Swanson et al., 2015; Habernal and Gurevych, 2016b) cease to provide detailed studies thereupon.

In this work, we thus investigate for the first time the automatic evaluation of the quality of arguments from a holistic perspective, bringing together various aspects. First, we evaluate whether AQ models can generalize across datasets and domains, a crucial feature for deployment in the diverse environments encountered in relevant real-world applications. Next, we investigate the hypothesis of whether models for related argument mining tasks inherently learn the concept of argument strength without being explicitly trained to do so by evaluating their zero-shot performance for estimating AQ. A

In summary, our contributions are as follows:

Findings of the Association for Computational Linguistics: ACL 2023, pages 13435–13448

Topic: Polygamy Legalization	**Score**
"Polygamy makes for unhappy relationships and is patriarchal."	0.66
"Polygamy makes child-raising easier by spreading the needs of children across more people."	0.84

Table 1: Two example arguments from the studied datasets with Argument Quality score.

- As far as we know we are the first to study the generalization capabilities of AQ prediction models across different datasets and AQ notions.

- Since we determine the size of the dataset as one of the decisive performance factors, we further investigate a zero-shot setting of transferring from related Argument Mining tasks.

2 Related Work

2.1 Argument Quality

Argument Quality (AQ), sometimes also called Argument Strength, is a sub-task of Argument Mining (AM) that is one of the central research topics among argumentation scholars (Walton et al., 2008; Toulmin, 2003; Van Eemeren and Grootendorst, 1987). Due to its highly subjective nature, there is no single definition of AQ. As a result, there are various proposals for different factors that can affect the quality of an argument, such as the *convincingness* of an argument (Habernal and Gurevych, 2016a). There are several ways to express the strength of an argument. Some works take an absolute continuous score, while others argue that strength estimation works better in (pairwise) relation to other arguments. To the best of our knowledge, we are the first to evaluate how AQ estimators trained on different corpora, AQ notions, and AQ tasks correlate with each other.

One of the first relatively large corpora was presented by Swanson et al. (2015). The SwanRank corpus contains over 5k arguments, where each argument is labeled with a continuous score that describes the interpretability of an argument in the context of a topic. They propose several methods based on linear regression, ordinary kriging, and SVMs as regression algorithms to automatically estimate the strength from an input text encoded by hand-crafted features. Other corpora have followed, using relative- and/or absolute *convincingness* (Habernal and Gurevych, 2016b; Potash et al., 2019) as an annotation criterion. The works proposed AQ estimators based on SVMs or BiLSTMs combined with GloVe embeddings (Pennington et al., 2014). Gleize et al. (2019) provide a dataset, *IBM-EviConv*, that focuses on ranking the *evidence* convincingness. They used a Siamese network based on a BiLSTM with attention and trainable Word2Vec embeddings. Gretz et al. (2020), and Toledo et al. (2019) created their corpora by asking annotators whether they would recommend a friend to use the argument in a speech supporting or disputing the topic, regardless of their own opinion. Both use a fine-tuned BERT (Devlin et al., 2019) model for the absolute AQ regression task.

The shared evaluation practice in the previous works is to evaluate methods on each dataset independently. Gretz et al. (2020) use their newly introduced dataset for pre-training of their model. The authors then investigate the strength of their models by applying them on two related datasets UKPConv and SwanRank. By finetuning the model on the training part of two datasets, they investigate if the pretraining is helpful for the target corpora. Our work proposes to advance the evaluation and advocate for an accurate *cross-dataset* evaluation without additional fine-tuning on the evaluation dataset to estimate the model's applicability in challenging *real-life* scenarios.

As a common understanding of AQ is still lacking, Wachsmuth et al. (2017a,b) investigated different dimensions of AQ. Based on a survey paper of existing argument quality theories (Wachsmuth et al., 2017a), they developed a taxonomy that aims to capture *all* aspects of AQ. In their work, they present a small corpus of 320 arguments annotated for 15 dimensions and explore the correlations between the different dimensions. Thus, their work presents a different view that rather focuses on the argumentation theory than on multiple corpora and the generalization of AQ estimators.

Lauscher et al. (2020); Ng et al. (2020) created a cross-domain corpus (Q&A forums, debate forums, and review forums) with 5,295 arguments using the annotation scheme of Wachsmuth et al. (2017a). They conclude that, in most cases, models benefit from the inclusion of out-of-domain training

data. However, they do not perform a cross-corpora study of their architectures, which limits the generalizability and impact of their experiments.

3 Generalization across Argument Quality Corpora

High-level applications such as Argument Retrieval (Wachsmuth et al., 2017c; Fromm et al., 2021; Dumani and Schenkel, 2019; Dumani et al., 2020; Stab et al., 2018a) and autonomous debating systems (Slonim et al., 2021) require reliable Argument Quality (AQ) models to select strong arguments among the relevant ones. The research community has identified this gap and proposed and evaluated different automated models for AQ estimation (Gretz et al., 2020; Toledo et al., 2019; Swanson et al., 2015; Habernal and Gurevych, 2016b). However, AQ is often captured differently due to its high subjectivity, e.g., absolutely as a continuous score or relative to other arguments by pairwise comparison. Consequently, many publications also introduced their own corpus with individual annotation schemes capturing different notions of AQ. While they have compared multiple AQ estimators against each other *within* a single corpus, there is a lack of *cross-corpora* empirical evaluations. Thus, the robustness of predictions across datasets remains largely unexplored, which poses a severe challenge for reliable real-world applications integrating diverse data sources. To assess the generalizability capability of AQ estimation models, we designed a series of experiments across all four major AQ datasets to answer the following research questions:

1. How well do AQ models perform across datasets if annotations schema and domain of the arguments do not change?

2. How does the corpora size affect generalization?

3. How well do models generalize across different text domains?

4. How does the AQ quality notion affect generalization?

5. Does the AQ model become more robust if it is trained with a combined dataset containing data from different domains and labeling assumptions also vary?

3.1 Datasets and Evaluation Setting

We briefly describe the four AQ datasets used in our empirical study, which all capture AQ on a sentence level. They are also summarized in Table 2.

1. Swanson et al. (2015) constructed the dataset `SwanRank` with 5,375 arguments whose quality is labeled in the range of $[0, 1]$, where 1 indicates that an argument can be *easily interpreted*. It consists of four controversial topics taken from the debate portal CreateDebate[1].

2. Habernal and Gurevych (2016b) annotated a large corpus of 16k argument pairs and investigated which argument from the pair is more *convincing*. Based on the argument pair annotations, they created an argument graph and used PageRank to calculate absolute scores for the individual arguments. The dataset is called `UKPConvArgRank` (here shortly UKP-Conv) and contains 1,052 arguments. It consists of 32 topics extracted from the debate portals CreateDebate[2] and ProCon[3].

3. Toledo et al. (2019) created their corpora `IBM-ArgQ` of 5,300 arguments by asking (1) debate club members (from novice to experts) and (2) a broad audience of people attending the experiments, if they would *recommend a friend to use the argument* in a speech supporting or contesting the topic regardless of their personal opinion. They modeled the quality of each individual argument as a real value in the range of $[0, 1]$, by calculating the fraction of 'yes' answers.

4. Gretz et al. (2020) created their corpora of 30k arguments by asking crowd contributor the same question as Toledo et al. (2019). Gretz et al. (2020) further introduce new scoring methods that consider the annotators' credibility without removing them entirely from the labeled data, as done in Toledo et al. (2019). The new scoring functions and the broader annotator selection presumably better represent the general population compared to Toledo et al. (2019).

[1] `http://www.createdebate.com`
[2] `http://www.createdebate.com`
[3] `https://www.procon.org`

Name	Sentences	Topics	Domain	Quality notion
UKPConv	1,052	32	Debate Portal	Convincingness
SwanRank	5,375	4	Debate Portal	Interpretability
IBM-ArgQ	5,300	11	Crowd Collection	Recommendableness
IBM-Rank	30,497	71	Crowd Collection	Recommendableness

Table 2: Overview of the different Argument Quality (AQ) datasets with their number of arguments, the number of distinct topics, the different source domains, and the AQ notion used for annotation.

As some of the corpora did not provide official train-validation-test splits and differed in the number of topics and the formulated task (in-topic vs. cross-topic), we decided to do our *own split* based on the topics of the arguments. Contrary to the original topic splits in UKPConv, IBM-ArgQ and IBM-Rank, we treat the supporting and opposing arguments from a certain topic as *one* topic because they have very great similarities. Whereas in their work, e.g. the topics *"We should abandon cryptocurrency"* and *"We should adopt cryptocurrency"* are represented as two topics. We perform 10-fold cross-topic cross-validation, where each fold is a 60%/20%/20% train-validation-test split, and we additionally ensure that no topic occurs in more than one split. By the latter requirement and the topic merge, we ensure an inductive setting where the AQ estimation can not rely on similar arguments in the training corpus and therefore provides a more challenging but more realistic task.

3.2 Model and Training

Since transfer learning achieves state-of-the-art Argument Mining (AM) results on different corpora and tasks (Reimers et al., 2019; Fromm et al., 2019; Trautmann et al., 2020), we also apply it to our AQ estimation task. We use a bert-base model, pre-trained on masked-language-modeling, and fine-tune it to predict absolute AQ scores on the respective datasets, cf. Section 3.1. As an input, we used the arguments from the respective datasets and concatenated the topic information, separated by the BERT specific $[SEP]$ limiter, similar to other work in AM (Fromm et al., 2019; Reimers et al., 2019; Gretz et al., 2020). We concatenate the last four layers (as Gretz et al. (2020); Toledo et al. (2019) did it) of the fine-tuned BERT model output to obtain an embedding vector of the size $4 \cdot 768 = 3,072$. For the regression task, we stack a Multi-Layer Perceptron (MLP) with two hidden layers, one with 100 neurons and a ReLU activation, followed by the second hidden layer and a sigmoid activation function. We train the architecture end-to-end, with SGD with a weight decay of 0.35 and a learning rate of 9.1×10^{-6}. The MLP uses dropout with a rate of 10%.

3.3 Results

Table 3 summarizes our results. We report the Pearson correlation score between the predicted- and ground-truth absolute AQ evaluated on a hold-out test set. Contrary to the original topic splits in UKPConv, IBM-ArgQ and IBM-Rank we treated the supporting and opposing topics as *one* topic. The task is therefore more challenging, as topic information from the contrary stance can not be used during training. However, the task is also more realistic, as one can not expect to have arguments from all topics in the training set.

3.3.1 Evaluation on Similar Datasets and Importance of Training Set Size

First, we evaluate the performance of the model on similar datasets and the dependency on the size of the training dataset. We can observe that models perform very well on other datasets from a similar domain labeled with a similar quality notion, i.e., IBM-ArgQ and IBM-Rank (both are *crowd collected* and annotated based on *recommendableness*. Furthermore, we can notice that the size of the dataset is crucial for performance: a model trained on the largest IBM-Rank dataset achieves the best score also on IBM-ArgQ. This insight gives us a solid foundation for the next steps.

3.3.2 Generalization Across Domains and Quality Notions

Next, we investigate whether a transfer across domains is possible. Recall that the four datasets cover two different domains: the sentences from UKPConv and SwanRank have been extracted from debate portals, while IBM-Rank and IBM-ArgQ have been collected from the crowd.

		Size	Evaluation UKPConv	SwanRank	IBM-ArgQ	IBM-Rank
	Related Work Split	-	35.1%	-	41.0%	48.0%
Training	UKPConv	1,052	19.0%	42.5%	15.2%	3.0%
	SwanRank	5,375	18.9%	**47.5%**	17.1%	8.0%
	IBM-ArgQ	5,300	23.3%	27.8%	34.2%	38.9%
	IBM-Rank	30,497	**26.2%**	37.0%	**38.3%**	**48.1%**
	all except UKPConv	41,172	23.3%	45.8%	31.6%	46.6%
	all except SwanRank	36,849	**25.0%**	**49.1%**	35.0%	46.6%
	all except IBM-ArgQ	36,924	23.0%	43.6%	**38.4%**	**47.5%**
	all except IBM-Rank	12,224	20.4%	42.0%	35.0%	46.5%

Table 3: The models are evaluated by the Pearson correlation between ground truth and predicted Argument Quality on the respective test sets. The first row corresponds to the respective correlation values reported in the original work. In SwanRank (Swanson et al., 2015) the authors evaluated their approach with Root Mean Squared Error (RMSE), Pearson correlation was not measured in their work. The first four rows correspond to models trained on a single dataset, whereas for the last four rows, *all but one* dataset, have been used for training, i.e., following a leave-one-out scheme. **Bold** numbers indicate the best results for each column within the two groups.

Compared to in-domain generalization, we observe a considerably worse generalization between domains: For example, trained on the crowd dataset IBM-ArgQ, we can achieve a correlation of 38.9% on the crowd dataset IBM-Rank, while training on the debate datasets SwanRank and UKPConv results in negligibly low correlations of 8% and 3%, respectively. Conversely, when evaluated on the debate portal dataset SwanRank, we obtain a correlation of 42.5% when using a model trained on the other debate portal dataset UKPConv, while the crowd-collected datasets IBM-ArgQ and IBM-Rank only achieves 27.8% and 37.0%, respectively. The smaller difference compared to the first comparison can be explained by the larger training datasets.

Surprisingly, we observe a completely different picture for generalization across quality notions. We see only a moderate drop in performance for a fixed domain but a different quality notion. For instance, the model trained on SwanRank performs relatively well on the UKPConv dataset. Vice-versa, we observe a more considerable performance drop, which can be explained by the smaller size of the UKPConv dataset.

3.3.3 Multi-Domain and Multi-Quality Notion Training

To investigate whether a single model can grasp various dimensions of quality and work on arguments from various domains, we designed another set of "leave-one-out" experiments. We train on the training sentences of all but one AQ corpus and evaluate the performance on all test sets. The four rows "all except" define the three training sets, e.g. "all except UKPConv" consists of the training sets of (SwanRank, IBM-ArgQ, and IBM-Rank). The entries on the diagonal thus show how well the models perform when evaluated on an unseen corpus.

For evaluation on the unseen IBM-Rank dataset after training on the remaining ones, we can obtain a correlation of 46.5%, which nearly reaches the correlation of 48.1% we obtained when training and evaluating on IBM-Rank. For SwanRank, IBM-ArgQ and UKPConv, we can even surpass the correlation on the respective test set by training on all other training sets instead of the one from the respective corpus.

3.3.4 Cross-Corpora Generalization Conclusion

To summarize, we conclude that in our analysis the available datasets and models for AQ are reliable. Our most important insight is that AQ notions do not contradict each other, and a single model can estimate the AQ of text from different domains. Therefore, the practical recommendation for real-life application is to combine all available datasets across different domains and AQ notions.

4 Zero-Shot-Learning in Argument Mining

In this section, we investigate whether explicit Argument Quality (AQ) corpora are a necessity or whether the task of AQ can also be solved by transferring from other related argument mining tasks such as Argument Identification (AId) or Evidence Detection (ED), In contrast to the relatively new task of automatic AQ estimation, other Argument Mining (AM) tasks already offer a broad range of large datasets that cover different domains and annotation schemes. Moreover, the agreement between the annotators is higher on the other tasks, as AQ is highly subjective (Swanson et al., 2015; Gretz et al., 2020; Toledo et al., 2019; Habernal and Gurevych, 2016b; Stab et al., 2018b). Therefore, a successful transfer from related tasks to the target task of AQ would represent a significant advance in the field. To this end, we investigate the zero-shot capability of AM models across different corpora *and* different AM tasks. To the best of our knowledge, we are the first to compare AM task similarity by providing a first study on how individual tasks can benefit from each other.

In particular, we aim to answer the following guiding research questions:

1. Can we achieve satisfactory performance by zero-shot transfer from related AM tasks, i.e., without fine-tuning the respective task?

2. Is there a difference in transferring from different tasks, i.e., is one task more suited than the other?

While not a primary focus of this work, for completeness, we also provide experimental results for the reverse direction of transferring *from* AQ estimation *to* the other tasks.

4.1 Datasets and Tasks

This section provides an overview of the three different AM corpora and tasks we used in our experiments. They are also summarized in Table 4.

1. UKP-Sentential (Stab et al., 2018b) contains over 25k arguments distributed across eight controversial topics. It is annotated for AId, where each sentence is labeled as either *argumentative* or *non-argumentative* in the context of a topic.

2. The IBM-Evidence (Ein-Dor et al., 2020) corpus includes nearly 30k sentences from Wikipedia articles. All sentences are annotated with a score in the range of $[0, 1]$, denoting the confidence that the sentence is evidence (either expert or study evidence) of the article's topic.

3. IBM-Rank (Gretz et al., 2020) is the largest of the four AQ datasets, which has also been used in the previous Section 3. The corpus annotation is in the range of $[0, 1]$, where 1 indicates a strong argument and a score of 0 indicates a weak argument.

We split all three datasets into the train, validation, and test sets (70%/10%/20%). Similar to Section 3.1, we designed the splits such that no topic in the training set also occurs in the test set, which is often called the "cross-topic" scenario in AM and corresponds to a more interesting, but also more challenging task, which requires a sufficient degree of generalization to unseen topics.

4.2 Evaluation Setting

We use a standard BERT large model (Devlin et al., 2019) pre-trained on the masked-language-modeling task to evaluate the zero-shot generalization capability. As an input for the fine-tuning, we use the sentences from the respective datasets and concatenate the topic information, separated by the BERT specific `[SEP]` limiter, similar to Section 3.2. We develop three different zero-shot evaluation strategies for the different transfer settings:

- **AId $\rightarrow$ Regression Tasks**: We use the BERT encoder output as input to a linear layer with a dropout that predicts the classes. Cross-entropy serves as training loss. The probabilities between 0 and 1 indicate if a sentence is argumentative or not. The predicted probability of the positive class, i.e., whether it is argumentative, is then directly used as a score for ED and AQ on the respective corpora. We use Spearman rank correlation instead of Pearson correlation as an evaluation measure to account for the difference in scale.

- **Regression Tasks $\rightarrow$ AId**: ED and AQ use the BERT representations in a single hidden layer that scores the sentences according to their absolute quality or the probability of containing evidence. Since we train on regression

Name	Sentences	Topics	Domain	Task
IBM-Rank	30,497	71	Crowd Collection	Argument Quality (AQ)
UKP-Sentential	25,492	8	Web Documents	Argument Identification (AId)
IBM-Evidence	29,429	221	Wikipedia	Evidence Detection (ED)

Table 4: Overview of the different Argument Mining (AM) datasets, we used for the zero-shot experiments, with their size in terms of the number of sentences, the number of covered topics, the source domain and the AM task.

tasks, we use the Mean Squared Error loss during training. We then apply the trained models to AId. We select an optimal decision threshold α among all possible thresholds on UKP-Sentential's validation set according to Macro F_1. By choosing the validation set, we avoid an unfair leakage to the model. This model is then evaluated on the UKP-Sentential test set.

- **Regression Task $\leftrightarrow$ Regression Task**: For the evaluation between the two regression models, we calculate the Spearman correlation coefficient directly on their respective outputs.

4.3 Results

Table 5 shows the results of our experiments. We train three models with different random seeds for each *source task* and report the mean and standard deviation of the evaluation on all *target tasks*.

We generally observe, unsurprisingly, that training on the same task as evaluating yields the best results with Spearman correlations of $\approx 77.90\%$ for ED $\rightarrow$ ED and $\approx 47.45\%$ for AQ $\rightarrow$ AQ.

A notable exception is AId, where a model trained on ED achieves $\approx 75.16\%$ Macro F_1 and thus can slightly surpass the performance of a model directly trained on AId of $\approx 73.51\%$, although within the range of one standard deviation. Exceeding the in-task performance is a strong result, as the model has never explicitly been trained for the task. We generally observe almost perfect zero-shot transfer towards AId, as also the model trained on AQ achieves a performance of $\approx 71.27\%$, which is only 2% points behind the $\approx 73.53\%$ from AId to AId. Thus, models capable of predicting whether a sentence provides evidence (ED) or capable of predicting the AQ of an argument inherently learn concepts that enable the detection of whether a sentence is argumentative or not (AId). To further give context to the zero-shot performance, the BiCLSTM approach trained on the AId task from (Stab et al., 2018b) obtained a Macro F_1 of 64.14%, i.e., worse results than the zero-shot transfer despite explicitly being trained on the task, which underlines the remarkable zero-shot performance, and may indicate that AId is a simpler task than the other two, ED and AQ.

For ED, we achieve the best performance of $\approx 77.90\%$ Spearman correlation by directly training on this task. The model trained on AId obtains the closest zero-shot transfer result with a rank correlation of $\approx 55.53\%$, which still represents a considerable correlation, despite being $\approx 22\%$ points behind. The model trained for AQ shows the worst transfer from the studied tasks with a correlation of $\approx 43.50\%$. Overall, we note that the challenging zero-shot transfer is still possible with an acceptable loss in performance. Models trained on detecting whether a sentence is argumentative or not (AId) transfer better than those trained for predicting the argumentative strength of a sentence AQ to the target task of predicting the confidence in whether a sentence provides evidence (ED).

For AQ, the main focus of our paper, we achieve the best performance of $\approx 47.45\%$ Spearman correlation by directly training on this task. When transferring from related AM tasks in a zero-shot setting, we have to tolerate decreases in performance to $\approx 28.66\%$ for transfer from ED, and $\approx 27.49\%$ for transfer from AId, respectively. Both zero-shot models are better at the prediction of AQ than models directly trained on the *same* target tasks but on another corpus (previous section UKPConv could achieve 3.0% and SwanRank 8.0% on the Gretz dataset). Models capable of detecting whether a sentence is argumentative (AId) are slightly less applicable to predicting the sentence's argumentative strength than the models for predicting a level of supporting evidence (ED). One factor here may be that ED is also a regression task as opposed to the classification task of AId.

To summarize, the results suggest that the tasks of AId, i.e., classifying whether a sentence is argumentative, and ED, i.e., predicting a numeric

Train	Evaluation AId	ED	AQ
AId	$73.51\% \pm 3.37\%$	$55.53\% \pm 1.17\%$	$27.49\% \pm 1.54\%$
ED	$75.16\% \pm 0.71\%$	$77.90\% \pm 0.24\%$	$28.66\% \pm 0.92\%$
AQ	$71.27\% \pm 0.74\%$	$43.50\% \pm 3.10\%$	$47.45\% \pm 1.16\%$
Metric:	Macro F_1	ρ	ρ

Table 5: Zero-Shot performance of the Argument Mining models. The evaluation measure is Macro F_1 for Argument Identification (AId), and the Spearman correlation for Evidence Detection (ED) and Argument Quality (AQ).

Train	Evaluation AId	ED	AQ
AQ	-	-	$47.45\% \pm 1.16\%$
AQ/AId	$80.07\% \pm 1.16\%$	-	$47.46\% \pm 0.58\%$
AQ/ED	-	$78.07\% \pm 0.45\%$	$46.84\% \pm 0.25\%$
AQ/AId/ED	$78.91\% \pm 3.17\%$	$78.40\% \pm 0.03\%$	$48.39\% \pm 1.12\%$
Metric:	Macro F_1	ρ	ρ

Table 6: Performance of multi-task models trained on different Argument Mining task combinations, including Argument Identification (AId) and Evidence Detection (ED). The performance is measured by Macro F_1 for AId, and the Spearman correlation for ED and AQ.

level of supporting evidence, are closer to each other than to the more difficult task of assessing the argumentative strength, as witnessed by worse zero-shot transfer results from and to AQ. Nevertheless, in principle, a transfer in the highly challenging zero-shot setting is possible; for closer related tasks, it can even lead to similar scores as training directly on the target task.

4.4 Multi-Task Learning for Argument Quality

As shown in the last section, the AM tasks are sufficiently close to each other to enable successful zero-shot transfer. An interesting question from this observation is whether the performance in AQ estimation further improves by multi-task learning. To this end, we developed a multi-task model that involves a shared BERT encoder and separated linear layers for the respective tasks. We trained the architecture with weighted loss functions, ensuring that each task is weighted equally. Our results are shown in Table 6. Focusing on the right-most column first, we can see that the performance in terms of Spearman correlation only marginally improves by multi-task learning. A possible explanation is that we already observed that the other two tasks are seemingly less challenging and more closely related to each other than to AQ. As additional supporting evidence, ED slightly and AId considerably benefit from multi-task learning with AQ.

5 Conclusion

We see this work as a fundamental step towards a more holistic view of Argument Quality (AQ). We have shown that for good generalization across individual AQ corpora, a match between the source and target domain of the arguments is essential. In contrast, diversity in AQ concepts does not hinder generalization but rather enriches it. The target domain has a minor impact with a sufficiently broad coverage of different domains and adequate size. This insight is directly applicable to practical applications: the advantages of different AQ notions allow the direct integration of different data sources, which is a prerequisite for handling the input from different domains encountered, e.g., by general-purpose argument retrieval engines.

Moreover, we were able to elucidate the relationship between AQ's and other Argument Mining (AM) tasks, such as Evidence Detection (ED) and Argument Identification (AId). Our zero-shot transfer experiments showed that the concepts learned for one of the tasks are sufficient to solve the other to some extent without explicitly being trained for it. By comparing the results obtained, we con-

clude that AId and ED are more closely related to each other than to AQ and are per se also easier to transfer to. The multitask experiment further emphasized this, where AQ could gain less from the other tasks than vice-versa. Thus, an important open question is how to enable a more successful transfer to AQ, extending beyond the three tasks we studied in this work.

6 Limitations

1. Our investigation in the zero-shot experiment is not exhaustive, we focused on the interplay between the three main tasks that also provide datasets of similar size: argument identification, evidence detection, and argument quality. However, there are other tasks, such as stance classification (deciding whether an argument supports or opposes a particular issue) or argument structure identification (identifying argumentative discourse units, such as claims and premises). Other tasks might be better source tasks for estimating argument quality.

2. Our experiments are based on the most popular datasets in argument mining and argument quality and may not generalize to other more specialized text domains, such as law or politics.

3. Using only English datasets limits the generalizability of the results to other languages and cultures. The ability to identify and evaluate the quality of arguments may be different in other languages and cultures, and the annotators may not be able to accurately capture these differences. This may lead to a lack of robustness and reliability of the results.

7 Ethics Statement

The BERT architectures are pre-trained on a large corpus of text data, which may contain biases in terms of language, content, and social issues. These biases could be transferred to our work, which could lead to inaccurate or unfair results.

8 Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and by the Deutsche Forschungsgemeinschaft (DFG) within the project Relational Machine Learning for Argument Validation (ReMLAV), Grant Number SE 1039/10-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). We further want to thank our students *Johanna Reiml* and *Siddharth Bhargava* who supported us in this work. The authors of this work take full responsibilities for its content.

References

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Harmony and dissonance: Organizing the people's voices on political controversies. In *Proc. of the Fifth ACM Int. Conf. on Web Search and Data Mining*, WSDM '12, page 523–532.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Lorik Dumani, Patrick J. Neumann, and Ralf Schenkel. 2020. A framework for argument retrieval. In *Advances in IR*, pages 431–445.

Lorik Dumani and Ralf Schenkel. 2019. A systematic comparison of methods for finding good premises for claims. In *Proc. of the 42nd Int. SIGIR*, pages 957–960.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Computational Models of Argument: Proc. from the Sixth Int. Conference on Computational Models of Argument (COMMA)*, pages 299–310.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining—a working solution. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 34, pages 7683–7691.

Michael Fromm, Max Berrendorf, Sandra Obermeier, Thomas Seidl, and Evgeniy Faerman. 2021. Diversity aware relevance learning for argument search. In *Advances in IR - 43rd European Conf. on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proc., Part II*, volume 12657, pages 264–271.

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: topic and context aware argument mining. In *2019 IEEE/WIC/ACM Int. Conf. on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*, pages 99–106.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. 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. *AAAI Conf*, 34(05):7805–7813.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

Ivan Habernal and Iryna Gurevych. 2016b. 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.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *AAAI*, volume 16, pages 2979–2985.

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944, Copenhagen, Denmark. Association for Computational Linguistics.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Isaac Persing and Vincent Ng. 2017. Lightly-supervised modeling of argument persuasiveness. In *Proc. of the Eighth Int. Joint Conference on Natural Language Processing*, pages 594–604.

Peter Potash, Adam Ferguson, and Timothy J. Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155, Florence, Italy. Association for Computational Linguistics.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

C. Rapp. 2002. *Rhetorik*. Number Bd. 1 in Aristoteles Werke in deutscher Übersetzung.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018a. Argumentext: Searching for arguments in heterogeneous sources. In *NAACL*, pages 21–25.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018b. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

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. *arXiv preprint*.

S Toulmin. 2003. The uses of argument cambridge university press.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *AAAI*.

Frans H Van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation*, 1(3):283–301.

Serena Villata. 2020. Using argument mining for legal text summarization. In *IOS Press*, volume 334, page 184.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *COLING*, pages 1680–1691.

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. 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. Association for Computational Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017c. Building an argument search engine for the web. In *Proc. of the 4th Workshop on Argument Mining*, pages 49–59.

Vern Walker, Karina Vazirova, and Cass Sanford. 2014. Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 1–10, Baltimore, Maryland. Association for Computational Linguistics.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Adam Z. Wyner, Raquel Mochales Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036, pages 60–79.

A Computing & Software Infrastructure

All experiments were conducted on a Ubuntu 20.04 system with an AMD Ryzen Processor with 32 CPU-Cores and 128 GB memory. We used Python 3.7, PyTorch 1.4, and the Huggingface-Transformer library (4.15.0). For the experiments in Chapter 3, we used four NVIDIA RTX 2080 TI GPUs with 11 GB memory. The models in Chapters 4 and 5 were trained on a single NVIDIA Tesla V100. The default parameters from the Huggingface-Transformer library [4] were used for all hyperparameters not specified in the following sections.

B Generalization across Argument Quality Corpora

In Section 3, we trained bert-base-uncased models with a batch size of 64. The learning rate was set to $9.1 \cdot 10^{-6}$. A weight decay of 0.31 was used. We calculated the 95th percentile based on the four AQ validation sets and truncated longer sentences to that length. The losses in the multi-dataset setting were equally weighted for each of the four datasets. We used early stopping on the validation MSE loss, with a patience value of five epochs, as a regularization technique to avoid overfitting.

C Zero-Shot-Learning in Argument Mining

For Section 4, we trained bert-large-uncased architectures with a batch size of 64. The learning rate was set to $1 \cdot 10^{-5}$, and a warm-up period was used for the first 0.1 epochs. We opt for evaluations every 0.1 epoch in our training configuration, resulting in 10 evaluations per epoch. Our train/validation/test split is based on a reasonably standard 70%/10%/20% split. Furthermore, we calculate the 99th percentile of the max length of all sentences inside the validation split and truncate them to that length. This further decreases the required learning time due to a reduced input dimension without losing significant information. We used a dropout rate of 0.1 for the dropout layer in the **AId** $\rightarrow$ **Regression Tasks** setting. The losses in the multi-dataset and multi-task setting were equally weighted for each of the three argument mining datasets. Finally, to further reduce variance in training, we use three seeds for our experiments and calculate the mean and standard deviation for all of our results.

[4] https://huggingface.co/docs/transformers/master/en/main_classes/trainer#transformers.TrainingArguments

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3 and 4

☑ B1. Did you cite the creators of artifacts you used?
3 and 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Coverage of domains, languages and, linguistic phenomena, demographic groups represented are already discussed in their original publications

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

3 and 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A, B, C

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A, B, C

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3 and 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes in the code files, there is a requirements.txt

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.