Automatic Argument Quality Assessment - New Datasets and Methods

Assaf Toledo*, Shai Gretz*, Edo Cohen-Karlik*, Roni Friedman*, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov and Noam Slonim
IBM Research

Abstract

We explore the task of automatic assessment of argument quality. To that end, we actively collected $6.3k$ arguments, more than a factor of five compared to previously examined data. Each argument was explicitly and carefully annotated for its quality. In addition, $14k$ pairs of arguments were annotated independently, identifying the higher quality argument in each pair. In spite of the inherent subjective nature of the task, both annotation schemes led to surprisingly consistent results. We release the labeled datasets to the community. Furthermore, we suggest neural methods based on a recently released language model, for argument ranking as well as for argument-pair classification. In the former task, our results are comparable to state-of-the-art; in the latter task our results significantly outperform earlier methods.

1 Introduction

Computational argumentation has been receiving growing interest in the NLP community in recent years (Reed, 2016). With this field rapidly expanding, various methods have been developed for subtasks such as argument detection (Lippi and Torroni, 2016; Levy et al., 2014; Rinott et al., 2015), stance detection (Bar-Haim et al., 2017) and argument clustering (Reimers et al., 2019).

Recently, IBM introduced *Project Debater*, the first AI system able to debate humans on complex topics. The system participated in a live debate against a world champion debater, and was able to mine arguments, use them for composing a speech supporting its side of the debate, and also rebut its human competitor.[1] The underlying technology is intended to enhance decision-making.

More recently, IBM also introduced *Speech by Crowd*, a service which supports the collection of free-text arguments from large audiences on debatable topics to generate meaningful narratives. A real-world use-case of Speech by Crowd is in the field of civic engagement, where the aim is to exploit the wisdom of the crowd to enhance decision making on various topics. There are already several public organizations and commercial companies in this domain, e.g., Decide Madrid[2] and Zencity.[3] As part of the development of Speech by Crowd, $6.3k$ arguments were collected from contributors of various levels, and are released as part of this work.

An important sub-task of such a service is the automatic assessment of argument quality, which has already shown its importance for prospective applications such as automated decision making (Bench-Capon et al., 2009), argument search (Wachsmuth et al., 2017b), and writing support (Stab and Gurevych, 2014). Identifying argument quality in the context of Speech by Crowd allows for the top-quality arguments to surface out of many contributions.

Assessing argument quality has driven practitioners in a plethora of fields for centuries from philosophers (Aristotle et al., 1991), through academic debaters, to argumentation scholars (Walton et al., 2008). An inherent difficulty in this domain is the presumably subjective nature of the task. Wachsmuth et al. (2017a) proposed a taxonomy of quantifiable dimensions of argument quality, comprised of high-level dimensions such as *cogency* and *effectiveness*, and sub-dimensions such as *relevance* and *clarity*, that together enable the assignment of a holistic quality score to an argument.

Habernal and Gurevych (2016b) and Simpson

[*] These authors equally contributed to this work.

[1] For more details: https://www.research.ibm.com/artificial-intelligence/project-debater/live/

[2] https://decide.madrid.es

[3] https://zencity.io

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5625–5635,
Hong Kong, China, November 3–7, 2019.

and Gurevych (2018) take a relative approach and treat the problem as relation classification. They focus on *convincingness* – a primary dimension of quality – and determine it by comparing pairs of arguments with similar stance. In this view, the convincingness of an individual argument is a derivative of its relative convincingness: arguments that are judged as more convincing when compared to others are attributed higher scores. These works explore the labeling and automatic assessment of argument convincingness using two datasets introduced by Habernal and Gurevych (2016b): *UKPConvArgRank* (henceforth, *UKPRank*) and *UKPConvArgAll*, which contain $1k$ and $16k$ arguments and argument-pairs, respectively.

Gleize et al. (2019) also take a relative approach to argument quality, focusing on ranking convincingness of *evidence*. Their solution is based on a Siamese neural network, which outperforms the results achieved in Simpson and Gurevych (2018) on the *UKP* datasets, as well as several baselines on their own dataset, *IBM-ConvEnv*.[4]

Here, we extend earlier work in several ways: (1) introducing a large dataset of actively collected arguments, carefully annotated for quality; (2) suggesting a method for argument-pair classification, which outperforms state-of-the-art accuracy on available datasets; (3) suggesting a method for individual argument ranking, which achieves results comparable to the state of the art.

Our data was collected actively, via a dedicated user interface. This is in contrast to previous datasets, which were sampled from online debate portals. We believe that our approach to data collection is more controlled and reduces noise in the data, thus making it easier to utilize it in the context of learning algorithms (see Section 7).

Moreover, we applied various cleansing methods to ensure the high quality of the contributed data and the annotations, as detailed in Section 3.

We packaged our data in the following datasets, which are released to the research community[5]:

- *IBM-ArgQ-6.3kArgs* - the full dataset, comprised of all $6.3k$ arguments that were collected and annotated with an individual quality score in the range $[0, 1]$.
- *IBM-ArgQ-14kPairs* - $14k$ argument pairs annotated with a relative quality label, indicating which argument is of higher quality.
- *IBM-ArgQ-5.3kArgs* - the subset of $5.3k$ arguments from *IBM-ArgQ-6.3kArgs* that passed our cleansing process. This set is used in the argument ranking experiments in Section 9.2, henceforth: *IBMRank*.
- *IBM-ArgQ-9.1kPairs* - the subset of $9.1k$ argument pairs from *IBM-ArgQ-14kPairs* that passed our cleansing process, used in the argument-pair classification experiments in Section 9.1. Henceforth: *IBMPairs*.

The dataset *IBMRank* differs from *UKPRank* in a number of ways. Firstly, *IBMRank* includes $5.3k$ arguments, which make it more than 5 times larger than *UKPRank*. Secondly, the arguments in *IBMRank* were collected actively from contributors. Thirdly, *IBMRank* includes *explicit* quality-labeling of all individual arguments, which is absent from earlier data, enabling us to explore the potential of training quality-prediction methods on top of such labels, presumably easier to collect.

Finally, with the abundance of technologies such as automated personal assistants, we envision automated argument quality assessment expanding to applications that include oral communication. Such use-cases pose new challenges, overlooked by prior work, that mainly focused on written arguments. As an initial attempt to address these issues, in the newly contributed data we guided annotators to assess the quality of an argument within the context of using the argument as-is to generate a persuasive *speech* on the topic. Correspondingly, we expect these data to reflect additional quality dimensions – e.g., a quality premium on efficiently phrased arguments, and low tolerance to blunt mistakes such as typos that may lead to poorly stated arguments.

2 Argument Collection

As part of the development of Speech by Crowd, online and on-site experiments have been conducted, enabling to test the ability of the service to generate a narrative based on collected arguments. Arguments were collected from two main sources: (1) debate club members, including all levels, from novices to experts; and (2) a broad audience of people attending the experiments.

[4] As this work is relatively recent and was published after our submission, we were not able to compare to it.

[5] `https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml#ArgumentQuality`

For the purpose of collecting arguments, we first selected 11 well known controversial concepts, common in the debate world, such as *Social Media*, *Doping in Sports* and *Flu vaccination*. Using debate jargon, each concept is used to phrase two "motions", by proposing two specific and opposing policies or views towards that concept. For example, for the concept *Autonomous Cars*, we suggested the motions *We should promote Autonomous Cars* and *We should limit Autonomous Cars*.[6] The full list of motions appears in Table 1 with the number of arguments collected for each.[7]

Guidelines Contributors were invited to a dedicated user interface in which they were guided to contribute arguments per concept, using the following concise instructions:

> You can submit as many arguments as you like, both pro and con, using original language and no personal information (i.e. information about an identifiable person).

In addition, to exemplify the type of arguments that we expect to receive, contributors were shown an example of one argument related to the motion, provided by a professional debater. The arguments collected had to have $8 - 36$ words, aimed at obtaining efficiently phrased arguments (longer/shorter arguments were rejected by the UI). In total, we collected $6,257$ arguments.

3 Argument Quality Labeling

We explored two approaches to labeling argument quality: (a) labeling individual arguments (absolute approach): each individual argument is directly labeled for its quality; and (b) labeling argument pairs (relative approach): each argument pair is labeled for which of the two arguments is of higher quality. In this section we describe the pros and cons of each approach as well as the associated labeling process.

Approaches to Argument Quality Labeling

The effort in labeling individual arguments scales linearly with the number of arguments, compared to the quadratic scaling of labeling pairs (within the same motion); thus, it is clearly more feasible when considering a large number of arguments. However, the task of determining the quality of arguments in isolation is presumably more challenging; it requires to evaluate the quality of an argument without a clear reference point (except for the motion text). This is where the relative approach has its strength, as it frames the labeling task in a specific context of two competing arguments, and is expected to yield higher inter-annotator agreement. Indeed, a comparative approach is widely used in many NLP applications, e.g. in Chen et al. (2013) for assessing reading difficulty of documents and in Aranberri et al. (2017) for machine translation. In light of these considerations, here we decided to investigate and compare both approaches. We used the Figure Eight platform[8], with a relatively large number of $15 - 17$ annotators per instance, to improve the reliability of the collected annotations.

Motion	**#Args**
Flu vaccination should be mandatory	204
Flu vaccination should not be mandatory	174
Gambling should be banned	342
Gambling should not be banned	382
Online shopping brings more harm than good	198
Online shopping brings more good than harm	215
Social media brings more harm than good	879
Social media brings more good than harm	686
We should adopt cryptocurrency	172
We should abandon cryptocurrency	160
We should adopt vegetarianism	221
We should abandon vegetarianism	179
We should ban the sale of vvg to minors	275
We should allow the sale of vvg to minors	240
We should ban fossil fuels	146
We should not ban fossil fuels	116
We should legalize doping in sport	212
We should ban doping in sport	215
We should limit autonomous cars	313
We should promote autonomous cars	480
We should support information privacy laws	355
We should discourage information privacy laws	93

Table 1: Motion list and statistics on data collection.

3.1 Labeling Individual Arguments

The goal of this task is to assign a quality score for each individual argument. Annotators were presented with the following binary question per argument:

> Disregarding your own opinion on the topic, would you recommend a friend preparing a speech supporting/contesting the topic to use this argument *as is* in the speech? (yes/no)

All arguments that were collected as described in

[6] Habernal and Gurevych (2016b) uses the term *topic* for what we refer to as *motion*.

[7] In Table 1, *vvg* stands for *violent video games*.

[8] `http://figure-eight.com/`

Section 2 were labeled in this task. We model the quality of each individual argument as a real value in the range of $[0, 1]$, by calculating the fraction of ‘yes’ answers. To ensure the annotators will carefully read each argument, the labeling of each argument started with a test question about the stance of the argument towards the concept (*pro* or *con*). The annotators’ performance on these test questions was used in the quality control process described in Section 4, and also in determining which pairs of arguments to label.

3.2 Labeling Argument Pairs

In this task, annotators were presented with a pair of arguments, having the same stance towards the concept (to reduce bias due to the annotator’s opinion), and were asked the following:

> Which of the two arguments would have been preferred by most people to support/contest the topic?

Table 2 presents an example of such an argument pair in which the annotators unanimously preferred the first argument.

Argument 1	**Argument 2**
Children emulate the media they consume and so will be more violent if you don’t ban them from violent video games	These are less fun and more harmful games but specifically violent games are played in groups and exclude softer souls

Table 2: An example of an argument pair for the motion *We should ban the sale of violent video games to minors*. The first argument was unanimously preferred by all annotators.

As mentioned, annotating all pairs in a large collection of arguments is often not feasible. Thus, we focused our attention on pairs that are presumably most valuable to train a learning algorithm. Specifically, we annotated $14k$ randomly selected pairs, that satisfy the following criteria:

1. At least 80% of the annotators agreed on the stance of each argument, aiming to focus on clearly stated arguments.

2. Individual quality scores in each pair differ by at least 0.2, aiming for pairs with a relatively high chance of a clear winner.

3. The length of both arguments, as measured by number of tokens, differs by $\leq 20\%$, aiming to focus the task on dimensions beyond argument length.

4 Quality Control

To monitor and ensure the quality of collected annotations, we employed the following analyses:

Kappa Analysis –

1. Pairwise Cohen’s kappa (κ) (Cohen, 1960) is calculated for each pair of annotators that share *at least* 50 common argument/argument pairs judgments, and based only on those common judgments.

2. *Annotator-κ* is obtained by averaging all pairwise κ for this annotator as calculated in Step 1, and if and only if this annotator had ≥ 5 pairwise κ values estimated. This is used to ignore annotators as described later.

3. Averaging all *Annotator-κ*, calculated in Step 2, results in *Task-Average-κ*.[9]

Test Questions Analysis – Hidden embedded test questions, based on ground truth, are often valuable for monitoring crowd work. In our setup, at least one fifth of the judgments provided by each annotator are on test questions. When annotators fail a test question, they are alerted. Thus, beyond monitoring annotator’s quality, test questions also provide annotators feedback on task expectations. In addition, an annotator that fails more than a pre-specified fraction (e.g., 20%) of the test questions is removed from the task, and his judgments are ignored.

High Prior Analysis – An annotator that always answers ’yes’ to a particular question should obviously be ignored; more generally, we discarded the judgments contributed by annotators with a relatively high prior to answer positively on the presented questions.

Note, if an annotator is discarded due to failure in any of the above analyses, he is further discarded from the estimation of *Annotator-κ* and *Task-Average-κ*.

[9] It is noted that some annotators remain without valid *Annotator-κ* and cannot be filtered out based on their κ. Similarly, those annotators do not contribute to the *Task-Average-κ*. However, in both annotation tasks, those annotators contributed only $0.01 - 0.03$ of the judgments collected.

5 Data Cleansing

5.1 Cleansing of Individual Arguments Judgments

To enhance the quality of the collected data, we discard judgments by annotators who (1) failed $\geq 20\%$ of the test-questions[10]; and/or (2) obtained *Annotator-*$\kappa \leq 0.35$ in the stance judgment task; and/or (3) answered 'yes' for $\geq 80\%$ of the quality judgment questions. Finally, we discarded arguments that were left with less than 7 valid judgments. This process left us with $5.3k$ arguments, each with 11.4 valid annotations on average. The *Task-Average-*κ was 0.69 on the stance question and 0.1 on the quality question. We refer to the full, unfiltered, set as *IBM-ArgQ-6.3kArgs*, and to the filtered set as *IBM-ArgQ-5.3kArgs* (*IBMRank*).

For completeness, we also attempted to utilize an alternative data cleansing tool, MACE (Hovy et al., 2013). We ran MACE with a threshold k, keeping the top k percent of arguments according to their entropy. We then re-calculated *Task-Average-*κ on the resulting dataset. We ran MACE with k=0.95, as used in Habernal and Gurevych (2016b), and with k=0.85, as this results in a dataset similar in size to *IBMRank*. The resulting *Task-Average-*κ is 0.08 and 0.09, respectively, lower than our reported 0.1. We thus maintain our approach to data cleansing as described above.

The low average κ of 0.1 for quality judgments is expected due to the subjective nature of the task, but nonetheless requires further attention. Based on the following observations, we argue that the labels inferred from these annotations are still meaningful and valuable: (1) the high *Task-Average-*κ on the stance task conveys the annotators carefully read the arguments before providing their judgments; (2) we report high agreement of the individual quality labels with the argument-pair annotations, to which much better κ values were obtained (see Section 6.1); (3) we demonstrate that the collected labels can be successfully used by a neural network to predict argument ranking (see section 9.2), suggesting these labels carry a real signal related to arguments' properties.

5.2 Cleansing of Argument Pair Labeling

To enhance the quality of the collected pairwise data, we discard judgments by annotators who (1) failed $\geq 30\%$ of the test-questions; and/or (2) obtained *Annotator-*$\kappa \leq 0.15$ in this task. Here, the test questions were directly addressing the (relative) quality judgment of pairs, and not the stance of the arguments. In initial annotation rounds the test questions were created based on the previously collected individual arguments labels - considering pairs in which the difference in individual quality scores was ≥ 0.6.[11] In following annotation rounds, the test questions were defined based on pairs for which $\geq 90\%$ of the annotators agreed on the winning pair. Following this process we were left with an average of 15.9 valid annotations for each pair, and with *Task-Average-*κ of 0.42 on the quality judgments – a relatively high value for such a subjective task. As an additional cleansing, for training the learning algorithms, we considered only pairs for which $\geq 70\%$ of the annotators agreed on the winner, leaving us with a total of $9.1k$ pairs for training and evaluation. We refer to the full, unfiltered, set as *IBM-ArgQ-14kPairs*, and to the filtered set as *IBM-ArgQ-9.1kPairs*.

6 Data Consistency

6.1 Consistency of Labeling Tasks

Provided with both individual and pairwise quality labeling, we estimated the consistency of these two approaches. For each pair of arguments, we define the *expected winning argument* as the one with the higher *individual* argument score, and compare that to the *actual winning argument*, namely the argument preferred by most annotators when considering the pair directly. Overall, in 75% of the pairs the actual winner was the expected one. Moreover, when focusing on pairs in which the individual argument scores differ by > 0.5, this agreement reaches 84.3% of pairs.

6.2 Reproducibility Evaluation

An important property of a valid annotation is its reproducibility. For this purpose, a random sample of 500 argument pairs from the *IBMPairs* dataset was relabeled by the crowd. This relabeling took place a few months after the main annotation tasks, with the exact task and data cleansing methods that were employed originally. For measuring correlation, the following A_score was defined: the fraction of valid annotations selecting

[10] Since quality judgments are relatively subjective, we focused the test questions on the stance question.

[11] Note, that annotators have the option of contesting problematic test questions, and thus unfitting ones were disabled during the task by our team.

"argument A" in an argument pair (A, B) as having higher quality, out of the total number of valid annotations. Pearson's correlation coefficient between A_score in initial and secondary annotation of the defined sample was 0.81.

A similar process was followed with the individual arguments quality labeling. Instead of relabeling, we split existing annotations to two even groups. We chose only individual arguments in which at least 14 valid annotations remained after data cleansing $(1,154$ such arguments). This resulted in two sets of labels for the same data, each based on at least 7 annotations. Pearson's correlation coefficient between quality scores of the two sets was 0.53. We then divided the quality score, which ranges between 0 to 1, to 10 equal bins. The bin frequency counts between the two sets are displayed in the heatmap in Figure 1.

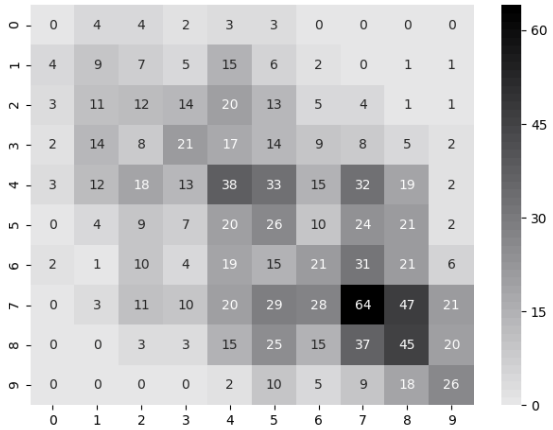

Figure 1: Counts of quality score bins between two equally sized sets of annotators.

6.3 Transitivity Evaluation

Following Habernal and Gurevych (2016b), we further examined to what extent our labeled pairs satisfy *transitivity*. Specifically, a triplet of arguments (A, B, C) in which A is preferred over B, and B is preferred over C, is considered transitive if and only if A is also preferred over C. We examined all 892 argument triplets for which all pair-wise combinations were labeled, and found that transitivity holds in 96.2% of the triplets, further strengthening the validity of our data.

7 Comparison of *IBMRank* and *UKPRank*

A distinctive feature of our *IBMRank* dataset is that it was collected actively, via a dedicated user interface with clear instructions and enforced length limitations. Correspondingly, we end up with cleaner texts, that are also more homogeneous in terms of length, compared to the *UKPRank* that relies on arguments collected from debate portals.

Text Cleanliness

We counted tokens representing a malformed span of text in *IBMRank* and *UKPRank*. These are HTML markup tags, links, excessive punctuation[12], and tokens not found in GloVE vocabulary (Pennington et al., 2014). Our findings show that 94.78% of *IBMRank* arguments contain no malformed text, 4.38% include one such token, and 0.71% include two such tokens. In the case of *UKPRank*, only 62.36% of the arguments are free of malformed text, 17.59% include one such token, and 20.05% include two or more tokens of malformed text.

Text Length

As depicted in Figure 2, the arguments in *IBMRank* are substantially more homogeneous in their length compared to *UKPRank*. A potential drawback of the length limitation is that it possibly prevents any learning system from being able to model long arguments correctly. However, by imposing this restriction we expect our quality labeling to be less biased due to argument length, holding greater potential to reveal other properties that contribute to argument quality. We confirmed this intuition with respect to the argument pair labeling as described in Section 9.1.

Data Size and Individual Argument Labeling

Finally, *IBMRank* covers $5,298$ arguments, compared to $1,052$ in *UKPRank*. In addition, in *UKPRank* no individual labeling is provided, and individual quality scores are inferred from pairs labeling. In contrast, for *IBMRank* each argument is individually labeled for quality, and we explicitly demonstrate the consistency of these individual labeling with the provided pairwise labeling.

8 Methods

In this section we describe neural methods for predicting the individual score and the pair-wise

[12] Sequences of three or more punctuation characters, e.g. *"?!?!?!"*

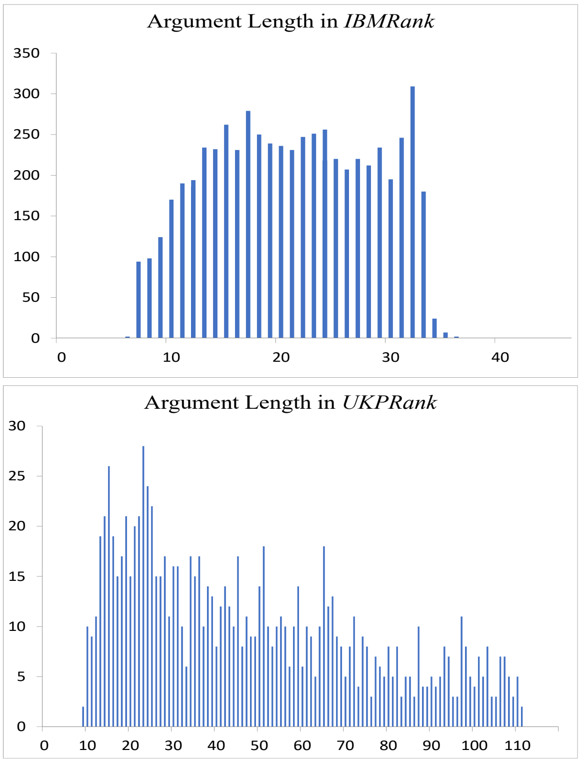

Figure 2: Histograms of argument length in *IBMRank* and *UKPRank*. X-axis: length (token count). Y-axis: the number of arguments at that length.

classification of arguments. We devise two methods corresponding to the two newly introduced datasets. Our methods are based upon a powerful language representational model named Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) which achieves state-of-the-art results on a wide range of tasks in NLP (Wang et al. (2018), Rajpurkar et al. (2016, 2018)). BERT has been extensively trained over large corpora to perform two tasks: (1) *Masked Language Model* - randomly replace words with a predefined token, `[MASK]`, and predict the missing word. (2) *Next Sentence Prediction* - given a pair of sentences `A` and `B`, predict whether sentence `B` follows sentence `A`. Due to its bidirectional nature, BERT achieves remarkable results when fine-tuned to different tasks without the need for specific modifications per task. For further details refer to Devlin et al. (2018).

8.1 Argument-Pair Classification

We fine-tune BERT's Base Uncased English pre-trained model for a binary classification task.[13] The fine-tuning process is initialized with weights from the general purpose pre-trained model and a task specific weight matrix $W_{out} \in R^{768 \times 2}$ is added to the 12-layer base network. Following standard practice with BERT, given a pair of arguments `A` and `B`, we feed the network with the following sequence '`[CLS]A[SEP]B`'. The `[SEP]` token indicates to the network that the input is to be treated as a pair and `[CLS]` is a token which is used to obtain contextual embedding for the entire sequence. The network is trained for 3 epochs with a learning rate of 2^{-5}. We refer to this model as *Arg-Classifier*.

8.2 Argument Ranking

For training a model to output a score between $[0, 1]$ we obtain contextual embeddings from the Arg-Classifier fine-tuned model. We concatenate the last 4 layers of the model output to obtain an embedding vector of size $4 \times 768 = 3072$. The embedding vectors are used as input to a neural network with a single output and one hidden layer with 300 neurons. In order for the network to output values in $[0, 1]$, we use a sigmoid activation, $\sigma_{sigmoid}(x) = \frac{1}{1+e^{-x}}$. Denote the weight matrices $W_1 \in R^{3072 \times 300}$ and $W_2 \in R^{300 \times 1}$, the regressor model, f_R, is a 2-layered neural network with $\sigma_{relu}(x) = \max\{0, x\}$ activation. f_R can be written as:[14]

$$f_R(x) = \sigma_{sigmoid}\left(W_2^T \sigma_{relu}(W_1^T x)\right)$$

where $x \in R^{3072}$ is the embedding vector representing an argument. We refer to this regression model as *Arg-Ranker*.

9 Experiments

9.1 Argument-Pair Classification

In this section we evaluate the methods described in Section 8. First, we evaluate the accuracy of *Arg-Classifier* on our *IBMPairs* dataset and on *UKPConvArgStrict* (henceforth, *UKPStrict*), the filtered argument pairs dataset of Habernal and Gurevych (2016b), in k-fold cross-validation.[15] We calculate accuracy and ROC area under curve (AUC) for each fold, and report the weighted averages over all folds. We also evaluate Simpson and Gurevych (2018)'s *GPPL* median heuristic method with *GloVe + ling* features in cross-

[13] Initial experiments with BERT's Large model showed only minor improvements, so for the purpose of the experiments detailed in Section 9 we used the Base model.

[14] We omit bias terms for readability.

[15] 22 and 32 folds respectively.

validation on our *IBMPairs* dataset. For completeness, we quote Simpson and Gurevych (2018)'s figures of *GPPL opt.* and *GPC* on *UKPStrict*.[16] We add a simple baseline classifying arguments based on their token count (*Arg-Length*).

	IBMPairs		
	Arg-Length	Arg-Classifier	GPPL
Acc.	.55	**.80**	.71
AUC	.59	**.86**	.78

	UKPStrict				
	Arg-Length	Arg-Classifier	GPPL	GPPL opt.	GPC
Acc.	.76	**.83**	.79	.80	.81
AUC	.78	**.89**	.87	.87	**.89**

Table 3: Accuracy and AUC on *IBMPairs* and *UKPStrict*.

As can be seen in Tables 3, *Arg-Classifier* improves on the *GPPL* method on both datasets ($p \ll .01$ using two-tailed Wilcoxon signed-rank test).[17] We note that *Arg-Classifier*'s accuracy on the *UKPStrict* set is higher than all methods tested on this dataset in Habernal and Gurevych (2016b); Simpson and Gurevych (2018). Interestingly, all methods reach higher accuracy on *UKPStrict* compared to *IBMPairs*, presumably indicating that the data in *IBMPairs* is more challenging to classify. With regards to *Arg-Length*, we can see that it is inaccurate on *IBMPairs* but achieves a respectable result on *UKPStrict*. This is in agreement with Habernal and Gurevych (2016a) who analyzed the reasons that annotators provided for their labeling. In most cases the reason indicated preference for arguments with more information – which is what longer arguments tend to be better at. This further strengthens the value of creating *IBMPairs* and *IBMRank* as much more homogeneous datasets in terms of argument length.

9.2 Argument Ranking

We proceed to evaluate the *Arg-Ranker* on the *IBMRank* and *UKPRank* datasets in k-fold cross-validation, and report weighted correlation measures. We also evaluate the *Arg-Ranker* by feeding it vanilla BERT embeddings, instead of the fine-tuned embeddings generated by the *Arg-Classifier* model. We refer to this version as *Arg-Ranker-base*. In both *Arg-Ranker* and *Arg-Ranker-base* evaluations we report the mean of 3 runs.[18]

	IBMRank		UKPRank		
	Arg-Ranker-base	Arg-Ranker	Arg-Ranker-base	Arg-Ranker	GPPL
r	.41	**.42**	.44	**.49**	.45
ρ	.38	**.41**	.57	.59	**.65**

Table 4: Pearson's (r) and Spearman's (ρ) correlation of *Arg-Ranker-base*, *Arg-Ranker* and *GPPL* on the *IBMRank* and *UKPRank* datasets.

As can be seen in Table 4, on the *UKPRank* dataset, *Arg-Ranker* is slightly better than *GPPL* for Pearson's correlation, but slightly worse for Spearman's correlation. Additionally, using direct BERT embeddings provides worse correlation[19] than using the *Arg-Classifier* embeddings for both datasets, justifying its use. Finally, similarly to the findings in the argument-pair classification task, the *IBMRank* dataset is harder to predict.[20]

10 Error Analysis

We present a qualitative analysis of examples that the *Arg-Classifier* and *Arg-Ranker* models did not predict correctly. For each of the argument-pair and ranking tasks, we analyzed $50 - 100$ arguments from three motions on which the performance of the respective model was poor. For each motion we selected the arguments in which the model was most confident in the wrong direction.

A prominent insight from this analysis, common to both models, is that the model tends to fail when the argument persuasiveness outweighs its delivery quality (such as bad phrasing or typos). An example of this is shown in row 1 of Table 5. In this case, Argument2 is labeled as having a higher quality, even though it contains multiple typos, and thus is typical to arguments that the model was trained to avoid selecting.

Another phenomenon that both our models fail to address is arguments that are off-topic, too provocative or not grounded. An example of this, from the argument-pair task, is shown in row 2 - Argument2 is presumably considered harsh by annotators, even though it is fine in terms of gram-

[16] We were unable to reproduce the results reported in Simpson and Gurevych (2018) by running the *GPPL opt.* and *GPC* algorithms on the *UKPStrict* dataset. We have approached the authors and reported the issue, which was not solved by the time this paper was published, and hence we only quote the figures as reported there.

[17] The results per fold in both tasks are included in the supplementary material.

[18] The *GPPL* regressor of Simpson and Gurevych (2018) relies on pair-wise (relative) labeling of arguments and as a result it cannot be used for predicting the individual (absolute) labeling of arguments, as in *IBMRank*.

[19] Significantly for the *IBMRank* data on both measures, and for the *UKPRank* on Pearson's correlation, $p \ll .05$.

[20] For the experiments on *IBMRank*, we included by mistake a small fraction of arguments which actually should have been filtered. The effect on the results is minimal.

Motion	Type	Argument1	Argument2
We should ban fossil fuels	Impact over delivery	*the only way to provide any space for energy alternatives to enter the market is by artificially decreasing the power of fossil fuels through a ban.*	**fossil fuels are bad for the environment, they have so2 in them that is the thing that maks acid rain and it is today harming the environment and will only be wors.**
Flu vaccination should not be mandatory	Provocative or not grounded	**the only responsible persons for kids are their parents. if they dont think that their kids should get the vaccine its their own decision.**	*the body has an automatic vaccination due to evolution, those who got sick and died are the weakest link and we are better off without them*
We should abandon vegetarianism	Consistent annotator preference	**it's harder to get all the things you need for a balanced diet while being vegetarian.**	*animals deserve less rights than humans, and it is legitimate for humans to prioritize their enjoyment over the suffering of animals.*

Table 5: Examples of argument pairs for which there is a high difference between the argument selected by the annotators, marked in bold, and the argument predicted to be of higher quality by the model, marked in italics.

matical structure and impact on the topic. These types of arguments are becoming more important to recognize, especially in the "fake-news" era. We leave dealing with them for future work.

Finally, we also notice certain arguments were consistently preferred by annotators, regardless of the quality of the opposing argument. This is a pattern relevant only to the *Arg-Classifier* model, shown in row 3.

11 Conclusions and Future Work

A significant barrier in developing automatic methods for estimating argument quality is the lack of suitable data. An important contribution of this work is a newly introduced data composed of $6.3k$ carefully annotated arguments, compared to $1k$ arguments in previously considered data. Another barrier is the inherent subjectivity of the *manual* task for determining argument quality. To overcome this issue, we employed a relatively large set of crowd annotators to consider each instance, associated with various measures to ensure the quality of the annotations associated with the released data. In addition, while previous work focused on arguments collected from web debate portals, here we collected arguments via a dedicated interface, enforcing length limitations, and providing contributors with clear guidance. Moreover, previous work relied solely on annotating pairs of arguments, and used these annotations to *infer* the individual ranking of arguments; in contrast, here, we annotated all individual arguments for their quality, and further annotated $14k$ pairs. This two–fold approach allowed us, for the first time, to explicitly examine the relation between relative (pairwise) annotation and explicit (individual) annotation of argument quality. Our analysis suggests that these two schemes provide relatively consistent results. In addition, these annotation efforts may complement each other. As pairs of arguments with a high difference in individual quality scores appear to agree with argument-pair annotations, one may deduce the latter from the former. Thus, it may be beneficial to dedicate the more expensive pair-wise annotation efforts to pairs in which the difference in individual quality scores is small, reminiscent of active learning (Settles, 2009). In future work we intend to further investigate this approach, as well as explore in more detail the low fraction of cases where these two schemes led to clearly different results.

The second contribution of this work is suggesting neural methods, based on Devlin et al. (2018), for argument ranking as well as for argument-pair classification. In the former task, our results are comparable to state-of-the-art; in the latter task they significantly outperform earlier methods (Habernal and Gurevych, 2016b).

Finally, to the best of our knowledge, current approaches do not deal with argument pairs of relatively *similar* quality. A natural extension is to develop a ternary-class classification model that will be trained and evaluated on such pairs, as we intend to explore in future work.

Acknowledgements

We thank Tel Aviv University Debating Society, Ben Gurion University Debating Society, Yale Debate Association, HWS Debate Team, Seawolf Debate Program of the University of Alaska, and many other individual debaters.

References

Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza, and Kepa Sarasola. 2017. Ebaluatoia: crowd evaluation for english–basque machine translation. *Language Resources and Evaluation*, 51(4):1053–1084.

Aristotle, G.A. Kennedy, and G.A. Kennedy. 1991. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

T Bench-Capon, K Atkinson, and Peter McBurney. 2009. *Altruism and agents: an argumentation based approach to designing agent decision mechanisms*, pages 1073 – 1080. Unknown Publisher.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 193–202, New York, NY, USA. ACM.

Cohen. 1960. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, pages 37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Chris Reed. 2016. Proceedings of the third workshop on argument mining (argmining2016). In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *CoRR*, abs/1906.09821.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Edwin D Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510. ACL.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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. 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. 2017b. Building an argument search engine for the web. In *ArgMining@EMNLP*, pages 49–59. Association for Computational Linguistics.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.