ARGANALYSIS35K - A LARGE SCALE DATASET FOR ARGUMENT QUALITY DETECTION

Omkar Joshi, Priya Pitre, Dr. Y. V. Haribhakta

COEP Technological University

Abstract

Argument Quality Detection is an emerging field in NLP which has seen significant recent development. However, existing datasets in this field suffer from a lack of quality, quantity and diversity of topics and arguments, specifically the presence of vague arguments that are not persuasive in nature. In this paper, we leverage a combined experience of 10+ years of Parliamentary Debating to create a dataset that covers significantly more topics and has a wide range of sources to capture more diversity of opinion. With 35k high-quality arguments, this is also the largest dataset of its kind to our knowledge. In addition to this contribution, we introduce an innovative argument scoring system based on instance-level annotator reliability and propose a quantitative model of scoring the relevance of arguments to a range of topics.

Dataset Creation

We used the following process to construct the dataset.

1. **Argument Collection** - We collected arguments from a variety of sources with the following distribution.
   - Contributions from active debaters (60%)
   - Argument extraction from speech transcripts at major debating tournaments (40%)
2. **Argument Annotation Collection** - We asked 200 debaters to answer the following questions as part of the annotation process.
   - Is the argument something you would recommend a friend use as-is in a speech supporting/opposing a topic, regardless of personal opinion?
   - Would you recommend a friend use the analysis to defend the argument as it is?

   A hidden question was also included in the survey to filter out annotators that weren't paying attention. The filtered results were taken through to the next stage.
3. **Annotator Reliability** - In this stage, we calculated the task-average κ to be **0.89** as opposed to IBM30K's value of **0.83**.
4. **Scoring Function** - However in order to make our dataset usable and interfaceable with others in the field , we need to convert these annotations to a quality score. In order to do this, we used MACE-P, Weighted Average and Instance-based Annotator Reliability as scoring functions. Each scoring function generates two scores per argument analysis pair ($Score_{arg}$,$Score_{analys}$) that are aggregated as follows:

$$Score_{pair} = Score_{arg} * Score_{analys} \tag{1}$$

Creation of the Relevance Model

1. In order to build our relevance model, we first generated a list of 24 topics considering inputs from our experts, analysis of trends in debating and classification of motions that we had presented to our annotators in order to generate our arguments.
2. In order to get more nuance on these topics, we asked 50 annotators to come up with a list of 5 keywords (also referred to as subtopics) per topic. The annotators chosen for this task were the ones scoring the highest in the previous tasks we set.
3. The keywords were then aggregated for similarity and reduced to the simplest representation. The keywords with the most agreement between the annotators ($< 70\%$ of annotators having included the keyword) were then collected.
4. The list of keywords was then sent to the experts who were asked to classify them into two bins: one bin containing keywords that they perceived to be highly relevant to the topic and one bin containing keywords that they perceived to be not as relevant. The weight of the keyword was taken to be the percentage of experts placing the keyword in the high relevance bin.
5. In total we saw that approximately 15% of keywords generated were attached to more than one topic. In this case they were assigned different weights for the different topics depending on the percentage of experts that placed the word in the high relevance bin for that particular topic. This created a set of 84 unique keywords with different weights for different topics.
6. Then, we found the probability of each argument-analysis pair belonging to the subtopics that we had generated in the form of keywords. This was done by the application of W2V and BERT to classify text by word vector similarity. This generated a list of scores per argument-analysis pair and subtopic, which indicates the probability of the pair belonging to that topic.
7. These scores are then combined via a simple Weighted Linear Combination as follows to generate the overall relevance score of a particular argument-analysis pair to the main topic.

$$\frac{\sum_{i=1}^{n} \alpha_{percentage} * Prob_{BERT}}{\sum_{i=1}^{n} \alpha_{percentage}} \tag{2}$$

Validation of the Relevance Model

In order to validate the relevance model we performed a simple experiment. The hypothesis was that as the delta of relevance scores increases, it would be easier for annotators to identify which of the pair of arguments is more relevant to the given topic. To make the comparisons fairer, we randomly selected a topic for which the relevance scores would be considered. We placed argument-analysis pairs into four bins based on the delta of their relevance scores to the selected topic. We then randomly sampled 150 pairs and send them for pairwise annotations to a set of 50 people (highest scoring annotators and experts). Each annotator was asked to pick the more relevant argument for the given topic and the percentage of annotators picking the higher ranked argument was noted as the precision. If sufficient agreement ($> 80\%$) between annotators was not achieved, the pair was dropped. This procedure was followed for two more randomly sampled topics to ensure coverage of the dataset and the agreements with the relevance scores were tabulated as follows.

Topic	Delta	Filtered Pairs	Precision
Art	< 0.25	14%	0.72
Art	0.25-0.5	10%	0.77
Art	0.5-0.75	5%	0.84
Art	0.75+	2%	0.96

Relevance Model Validation

We found that all three topics showed similar trends in terms of agreeing with the annotator scoring. Annotator scoring also showed a high correlation with our relevance model for high deltas. This validates the relevance model as it satisfies the basic requirement of a quantitative score: bigger differences are more easily recognized.

Relationship between scores and argument length

We notice an interesting trend when looking at analysis length with comparison to the IA score they receive (Figure 1). Analysis scores reach a peak score at 180 characters, following which they drop, giving a slight resemblance to a normal curve. This proves that less characters are insufficient to express a point in a persuasive manner, but having more characters than necessary is also not considered persuasive, as the analysis becomes repetitive and less impactful.

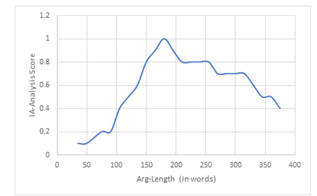

Fig. 1: IA-Analysis Scores Vs Arg-Length

Comparison with other datasets

Since WA had been used as a scoring function for ArgAnalysis35K as well as IBM-Rank30K, we are able to compare the scores of both datasets to compare argument quality. Out of the 5000 arguments ranked 1 in IBM-Rank30, we randomly sampled 200. We ran these arguments through our relevance model to find the topic in our dataset they are closest related to. The specified argument was only taken if it had a relevance score above 0.8 (that is, the argument strongly belongs to that category). From the ArgAnalysis35K dataset, we have randomly selected an argument-analysis pair from the same topic that had been scored 1. This pair of arguments were then sent to 500 random debaters where they were asked which argument they found more persuasive (similar to the question asked during the debate between Project Debater and Harish Natarajan). We then looked at the agreement between the different annotators on each of the pairs, similar to the experiment performed to compare the different scoring functions. We found that annotators preferred a ArgAnalysis35K argument 71% of the time, hence showing that the arguments in ArgAnalysis35K are more relevant in the context of parliamentary debating, and that an argument is more persuasive when followed by analysis.

Future Works

Creation of datasets in novel ways enables us to expand the field of computational argumentation. We can integrate this dataset with existing models to create a system that is able to debate more efficiently, be more persuasive, and as a result win debates more often. We can further use the scores of relevance, argument and analysis to qualitatively judge and access the winner of a parliamentary debate.