Automatic Explanation Quality Assessment in Online Learning Environments

Sameer Bhatnagar[1], Amal Zouaq[1], Michel C. Desmarais[1], and Elizabeth Charles[2]

[1] Ecole Polytechnique Montreal
{sameer.bhatnagar,amal.zouaq,michel.desmarais}@polymtl.ca
[2] Dawson College echarles@dawsoncollege.qc.ca

Abstract. *150-250 words*

Keywords: Argument mining · Learnersourcing · Peer Instruction

1 Introduction

2 Related Work

2.1 Learnersourcing & Comparative Peer Assessment

Ripple[7], AXIS[15] Juxtapeer[1]

2.2 Argument Quality & Convincingness

Conventional argument-mining pipelines include several successive components, starting with the automatic detection of argumentative units, classification of these units into types (e.g. major claim, minor claim, premise), and identification of argumentative relations (which evidence units support which claim). Such pipelines are essential in question-answering systems [8] and are at the heart if the IBM Project Debater initiative.

Work in the area of automatic evaluation of argument quality finds its roots in detecting evidence in legal texts[9], but has accelerated in recent years as more datasets become available in everyday contexts, and focus shifts to modelling more qualitative measures, such as *convincingness*.

Some of the earlier efforts included work on automatically scoring of persuasive essays [11] and modelling persuasiveness in online debate forums [13]. However, evaluating argument *convincingness* with an absolute score can be challenging, which has led to significant work in adopting a pairwise approach, where data consists of pairwise observations of two arguments, labelled with which of the two is most convincing.

In [5], the authors propose a feature-rich support vector machine, as well as an end-to-end neural approach based on pre-trained Glove vectors and a bidirectional Long-Short-Term Memory network for the pairwise classification task.

rationale	chosen_rationale
The graph shows constant positive acceleration and then constant negative acceleration. This means that the velocity-time graph should have a positive slope and then a negative slope, and graph C is the only option that satisfies those requirements.	1st, acceleration is +, indicating an increase in velocity.\r\nthen, acceleration is suddenly and without warning negative, and velocity is reduced.

Table 1. Example of instance in pairwise comparison task, where two students explanations are compared, and one is chosen as more convincing

This is extended in [4], where the authors build a Siamese network architecture, where each leg is a BiLSTM, taking as input the pair of explanations as Glove embeddings [10], in order to detect which of argument in a pair has the most convincing evidence. Finally, based on the success of transformer models such as BERT[3], the authors of [14] release a dataset of argument pairs and show that these models accurately predict the most convincing argument in a pair.

– [14] Assessment of argument quality, with a dataset that has both individual scores and pairwise-ranked data

3 Methods

Feature Type	Feature	τ
Lexical	Uni+BiGrams	
	Spelling Errors	
	Equations	
	Type Token Ratio	
	Punctuation	
	Readability scores	
Syntactic	PoS Uni+Bigrams	
	Dependancy Tree Depth	
	Conjunctions	
	Modal Verbs	
	Tree Production Rules	
Semantic	LSA Similarity	
	LSA Similarity Question	
	Likelihood (Textbook)	

Table 2. Features used in experiments with Linear SVM, annotated with kendall τ correlation and sigificance with target label

3.1 Data

The dataset is comprised of pairs of student explanations for a particular answer choice to a given question. The first explanation is always the one written by the learner-annotator, while the second is an alternative which they either chose as more convincing, or not. The data is filtered so as to only keep observations where the explanations are within half a standard deviation in length of each other. To ensure internal reliability, we only keep explanations that were chosen at least 1 times. To ensure that the explanations in each pair are of comparable length, we keep only those with word counts that are within 0.5 standard deviations or 10 words of each other. This leaves us a dataset with 7646 observations, spanning 1894 learner annotators having completed, on average, 4.0 items each, from a total of 104 items across three disciplines, with at least 50 explanation-pairs per item.

Table 3. Observations of students choosing a peer explanation as more convincing than their own, or not, aggregated by discipline and whether they started and finished with the correct answer

		N
discipline	transition	
Biology	rr	3441
	ww	919
	wr	704
	rw	164
Chemistry	rr	434
	ww	211
	wr	139
	rw	38
Physics	rr	1104
	ww	287
	wr	146
	rw	59

Table 3 highlights one key difference between the modelling task of this study, and related work in argument mining, where annotators are presented pairs of arguments that are always for the same stance, in order to limit bias due to their opinion on the motion when evaluating which argument is more convincing. In a *Peer Instruction* learning environment, other pairings are possible and pedagogically relevant. In this dataset, the majority of students keep the same answer choice between the two steps of the prompt, and so they are comparing two explanations that are either both correct (*"rr"*) or incorrect (*"wr"*). However, there is 16 % of the observations in this dataset are for students who not only choose an explanation more convincing than their own, but also switch answer choice, either from the incorrect to correct, or the reverse . These pairs add a different level of complexity to the model, but are very pertinent in the pedagogical context: what are the argumentative features which can help students remediate an initial wrong answer choice (*"wr"*)? What are the features that might be responsible for getting students to actually move away from the correct answer choice (*"rw"*)?

3.2 Models

	BoW	+/-	Longest
Biology	0.74	0.13	0.64
Chemistry	0.64	0.14	0.57
Physics	0.66	0.12	0.62

Table 4. Baseline models per discipline

In pairwise classification tasks so 50% is the baseline performance. In Table 4 we begin by comparing a *longest* model baseline, where we predict that students simply choose the longer explanation of the pair, and one based on a simple *Bag of Words* model, where the words are taken from a open source textbook from the corresponding discipline.

For our experiments, we begin by following the line of work proposed by [5], and experiment with a feature-rich linear SVM classifier for the pairwise classification task. We use a similar feature set, which we categorize as **lexical**, **syntactic**, and **semantic**, as described in Table 2. we begin by computing the feature vector for each explanation, and compute the difference for each pairwise ranking instance as per the well established SVM-Rank algorithms [6], training the model to learn which of the pair is the more convincing argument.

4 Results

model	accuracy	AUC
SVM		
BiLSTM		
BERT		

5 Discussion

6 Future Work

In this study we do not ever infer which are, overall, the most convincing student explanations for any given item. Inferring a gold standard of global rankings, starting from these pairwise preference data can be accomplished using research from the information retrieval community[2]. Work on deriving point wise scores for argument pairs is proposed as a Gaussian Process Preference Learning task by [12]. Seeing the lack of pointwise labels for overall convincingness, [14] released a dataset where they collect this data as well. A comparable source of data inside the myDALITE platform are the feedback scores teachers can optionally provide to students on thier explanations.

References

1. Cambre, J., Klemmer, S., Kulkarni, C.: Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. pp. 1–13. ACM Press, Montreal QC, Canada (2018). https://doi.org/10.1145/3173574.3173868
2. Chen, X., Bennett, P.N., Collins-Thompson, K., Horvitz, E.: Pairwise ranking aggregation in a crowdsourced setting. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 193–202 (2013)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., Slonim, N.: Are you convinced? choosing the more convincing evidence with a Siamese network. arXiv preprint arXiv:1907.08971 (2019), `https://www.aclweb.org/anthology/P19-1093/`
5. Habernal, I., Gurevych, I.: 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). pp. 1589–1599. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1150, `https://www.aclweb.org/anthology/P16-1150`
6. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 133–142 (2002)
7. Khosravi, H., Kitto, K., Williams, J.J.: Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522 (2019)
8. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Transactions on Internet Technology (TOIT) **16**(2), 10 (2016)
9. Moens, M.F., Boiy, E., Palau, R.M., Reed, C.: Automatic detection of arguments in legal texts. In: Proceedings of the 11th international conference on Artificial intelligence and law. pp. 225–230 (2007)

10. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
11. Persing, I., Ng, V.: End-to-End Argumentation Mining in Student Essays. In: HLT-NAACL. pp. 1384–1394 (2016)
12. Simpson, E., Gurevych, I.: Finding Convincing Arguments Using Scalable Bayesian Preference Learning. Transactions of the Association for Computational Linguistics **6**, 357–371 (2018), `https://www.aclweb.org/anthology/Q18-1026`
13. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. arXiv:1602.01103 [physics] pp. 613–624 (2016). https://doi.org/10.1145/2872427.2883081, `http://arxiv.org/abs/1602.01103`, arXiv: 1602.01103
14. Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., Slonim, N.: Automatic Argument Quality Assessment-New Datasets and Methods. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5629–5639 (2019), `https://www.aclweb.org/anthology/D19-1564.pdf`
15. Williams, J.J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K.Z., Lasecki, W.S., Heffernan, N.: AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16. pp. 379–388. ACM Press, Edinburgh, Scotland, UK (2016). https://doi.org/10.1145/2876034.2876042