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[4], AXIS[7] Juxtapeer[1]

2.2 Argument Quality & Convincingness

- [3] Predicting convincingness, reducing noise in annotations by building an acyclic argument graph
- [5] Gaussian Process Preference Learning
- [6] Assessment of argument quality, with a dataset that has both individual scores and pairwise-ranked data
- [2] Evidence quality, predicted using a Siamese network architecture

3 Methods

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 observations where we have at least 3 records per chosen explanation. 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 7918 observations, spanning 3076 learner annotators having worked on on average 3.0 items each, from a total of 1029 items across three disciplines.

Table 1. 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

discipline	transition	N
Biology	rr	1800
	ww	630
	wr	403
	rw	121
Chemistry	rr	1232
	ww	702
	wr	390
	rw	82
Physics	rr	1479
	ww	618
	wr	358
	rw	103

Table 1 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 18 % 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 studnets to actually move away from the correct answer choice (*"rw"*)?

3.2 Models

The first baseline model we compare to is where students simply choose the longer explanation of the pair, while the second is based solely on a Bag of

Words model trained on all of the words used by students for this item, and the words in

4 Results

model	accuracy	AUC
Argument Length		
SVM-R		
GPPL		
BiLSTM		
Siamese		
BERT		

5 Discussion

6 Future Work

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. 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/
3. 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
4. Khosravi, H., Kitto, K., Williams, J.J.: Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522 (2019)
5. 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
6. 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

7. 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