Automatic Explanation Quality Assessment in Online Learning Environments: new datasets and methods

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. *Asynchronous Peer Instruction* is increasingly popular in online learning environments. It relies on the principle of leveraging content that is both *generated* and *evaluated* by novices to foster learning and self-reflection. Students first respond to a question item, but they must also provide an explanation for their reasoning. They are then presented alternative explanations as written by their peers, and given the opportunity to change their initial answer choice, based on those they find most convincing. The peer-explanations that students find most convincing represent valuable data, for teachers to better grasp their students' understanding, and for the learning environment itself, as higher quality explanations can be shown to students as examples to compare their work to. This study reports on the application of argument mining methods in the context of asynchronous peer instruction, with the objective of automatically identifying high quality student explanations. Our results offer the potential to inform the design of "learnersourcing" systems, such as asynchronous peer instruction. These design choices are critical as these systems scale, especially with respect to providing pedagogically insightful reports to teachers, and presenting engaging alternative explanations to students to promote higher order thinking.

Keywords: Argument mining · Learnersourcing · Peer Instruction

1 Introduction

As online learning environments scale in the number of question items, providing immediate and tailored feedback to students becomes intractable for activities that prompt for open ended responses. *Peer assessment* and *peer feedback* can address this issue, and have the added benefit of being an effective way to learn for the student who give the feedback as well [11]. However the drawback of such approaches often lie in the varying ability of novices to provide good feedback to their peers. Frameworks such as Adaptive Comparative Judgement [21], where teachers assess student submissions by simply choosing which is better from a pair, have been shown to be a reliable and valid alternative to absolute

grading. Thus it is not surprising that there are a growing number of online learning environments that extend this paradigm to students with *comparative peer assessment*, wherein, after students submit their own work, they are asked to compare and contrast a pair of two of their peers' submissions, and provide feedback [1] [2] [22].

A subset of these tools then use the comparative peer assessment data to inform the choice of how the pairs are constructed, and which students are assigned to which pairs, [13] [29] [23]. This leads to the classic trade-off from the field reinforcement learning: exploiting the student submissions for which we have reliable data and can estimate their quality, while exploring student work that is newer to the database, and needs to be shown and evaluated in order to get an estimate of its quality.

Automatic assessment of quality is especially crucial in systems where the content generated by students is actually a central part of the pedagogical script: in platforms where students generate explanations, hints, or even new question items, which are then shown to future students, selecting bad content can negatively impact the learning of future students. We focus our attention on learning environments that enable *peer instruction*[5], which are built on a two-stage script: (i) students are prompted to answer a multiple choice question, and provide a free-text explanation that justifies their answer; (ii) without revealing the correct answer, students are prompted to reconsider their answer, by presenting them a selection of explanations written by previous students [3]. Students can either decide that their own explanation is best, or indicate which of their peers' explanation was the most convincing. In this instance of *learnersourcing*[28], this "vote data" is valuable at two levels: it can then be used to determine what to present to future students, but also inform instructors of their students' understanding of the material.

We frame the explanations to multiple choice questions as *arguments* meant to persuade one's peers. As such, we explore methods and datasets from the *argument mining* research community, where there has been a growing body of work on automatically assessing argument quality, along the dimension of *convincingness*. The objective of this study is to determine whether we can automatically assess the *convincingness* of student explanations in peer-instruction based learning environments, starting from pairwise preference data generated by students in undergraduate level science courses. We apply our methods, and compare our results, across similar publicly available datasets of argument pairs, annotated for pairwise preference. We report on the effectiveness of vector space models, as well as state-of-the-art neural approaches, applied to the task of learning pairwise preferences for convincing arguments. Our findings suggest that the arguments generated in learning environments centred on undergraduate science topics present a more challenging variant of the task originally proposed in the argument mining community, and that classical approaches match neural models for performance on this task across datasets.

2 Related Work

2.1 Automated Short Answer Grading

Scalable Peer Assessment of open response items [20] Automated Essay Scoring [24]

2.2 Learnersourcing & Comparative Peer Assessment

Ripple [13], AXIS [29] Juxtapeer [2], ComPAIR [22] UBCPI [1] Peerwise [6]

2.3 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 [14] 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 [16], 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 [19] and modelling persuasiveness in online debate forums [26]. 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.

Table 1: A pair of arguments from the UKP dataset, both with the stance CON, for the prompt topic: "william-farquhar-ought-to-be-honoured-as-the-rightful-founder-of-singapore-yes-of-course-". Argument a1 is labelled as more convincing.

a1	a2
Farquar protected Singapore from Dutch attacks and attracted traders to Singapore while Raffles was away.	When Farquar was fired by Raffles, he was given a grander ceremony before he left compared to Raffles, which show how much the people loved him.

In [9], 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.

Table 2: A pair of arguments from the IBM dataset, both with the stance CON, for the prompt topic: "Online-shopping-brings-more-good-than-harm-(CON)". Argument a1 is labelled as more convincing.

a1	a2
it enables cheaper products by not needing to have physical store locations (which have rent), but only a warehouse.	it enables redistribution of wealth through rich people buying from poorer areas instead of their local, richer, stores and manufacturers.

This is extended in [8], where the authors build a Siamese network architecture, where each leg is a BiLSTM, taking as input the pair of explanations as Glove embeddings [18], 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 [7], the authors of [27] release a dataset of argument pairs and show that these models accurately predict the most convincing argument in a pair.

3 Data

Must be anonymized

One of the objectives of this study is to compare and contrast how argument mining methods for evaluating argument quality, specifically for argument *convincingness*, perform in an online learning environment with learner-generated and annotated arguments. Along with myDALITE data, we include in our study two publicly available datasets, each specifically curated for the task of automatic assessment of argument quality along the dimension *convincingness*. Table 3 summarizes some of the descriptive statistics that can be used to compare these sources, and potentially explain some of our experimental results.

Table 3: Descriptive statistic for each dataset of argument pairs, with last rows showing *dalite* split by discipline.

dataset	N_{pairs}	N_{topics}	N_{args}	N_{vocab}	$\overline{wc}$ (SD)	$\overline{\Delta wc}$ (SD)
IBM_ArgQ	9125	11	3474	6710	23 (7)	3 (2)
UKP	11650	16	1052	5170	49 (28)	30 (23)
dalite	8551	102	8942	5571	17 (15)	12 (7)
IBM_Evi	5274	41	1513	6755	29 (11)	3 (2)
dalite:Biology	3919	49	4116	3170	15 (14)	10 (6)
dalite:Chemistry	1666	24	1758	2062	20 (14)	12 (7)
dalite:Physics	2966	29	3068	2478	19 (15)	15 (7)

3.1 UKP & IBM

UKPConvArgStrict[9], hence forth referred to as **UKP**, was the first to propose the task of pairwise preference learning for argument convincingness. The dataset consists of just over 1k individual arguments, that support a particular stance for one of 16 topics. These arguments were distributed as 11.6k pairs to annotators on a crowd-sourcing platform, where the task was to choose which of the two arguments, for the same stance regarding the same topic, was more convincing.

More recently, a second similar dataset, **IBMArgQ-9.1kPairs**[27], henceforth referred to as **IBM_ArgQ**, which is made of 3.4k individual arguments for 11 topics, assembled into 9.1k pairs labelled for which is more convincing. One of the key differences between these two is that **IBM_ArgQ** data is more strongly curated with respect to the relative length of the arguments in each pair: in order to control for the possibility that annotators may make their choice of which argument in the pair is more *convincing* based merely on the length of the text, the mean difference in word count, $\overline{\Delta wc}$, is just 3 words across the entire dataset, which is 10 times more homogeneous than pairs in **UKP**.

Finally, we include a fourth dataset, **IBM_Evi**, consisting of 1.5k individual arguments, organized into 5.2k pairs annotated for pairwise preference for convincingness [8]. The important distinction here is that the arguments are actually extracted as evidence for their respective topic from Wikipedia, and hence represent cleaner well-formed text than our other reference datasets.

3.2 myDALITE

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. To ensure internal reliability, we only keep explanations that were chosen at least 5 times. To ensure that the explanations in each pair are of comparable length, we keep only those with word counts that are within 25 words of each other.

This leaves us a dataset with 8551 observations, spanning 2216 learner annotators having completed, on average, 4.0 items each, from a total of 109 items across three disciplines, with at least 50 explanation-pairs per item.

Table 4 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 17 % of the observations in this dataset are for students who not only choose an explanation more convincing than their own, but also switch answer

Table 4: 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.

	rr	rw	wr	ww
Biology	2459	124	733	603
Chemistry	1151	51	228	236
Physics	2288	66	278	334

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"*)?

4 Methods

This section might be before the data sets and should clearly name and describe the methods.

Choosing which argument is more convincing from a pair, is a binary ordinal regression task, where the objective is to learn a function that, given two feature vectors, can assign the better argument a rank of $+1$, and the other a rank of -1. It has been proven that such a binary ordinal regression problem, can be cast into an equivalent binary classification problem, wherein the model is trained on the *difference* of the feature vectors of each argument in the pair [10]. Referred to as *SVM-rank*, this method of learning pairwise preferences has been used extensively in the context of information retrieval (e.g. ranking search results for a query based on past clickthrough data) [12], but also more recently in evaluating the journalistic quality of newspaper and magazine articles [15], and for predicting which argument is more convincing [9].

We follow-up on this work, building simple vector space models to represent our text, and train Support Vector Machine classifiers on the arithmetic difference of these input vectors. Our first model, *ArgBow*, is topic-specific, and uses a Tf-Idf representation for the input text, trained on a 80% of the data in a five-fold cross-validation scheme (unseen token in the different held-out test sets are mapped to the same value).

As these models suffer from the *cold-start problem*, in that we need training examples to build the term-document matrices before being able to make predictions on unseen data, we propose a model that leverages semantic information learned from large bodies of text. In *ArgGlove*, we encode each token using 300-dimensional GloVe vectors [18], and each argument as the average of its token vectors.

Finally, in order to leverage recent advances in transfer-learning for NLP, the final model we explore is *ArgBert*. We begin with a pre-trained language model built using Bi-directional Encoder Representation from Transformers, known as*BERT*[7], trained on large bodies of text for the task of masked token prediction and sentence-pair inference. As proposed in [27], we take the final 768-dimensional hidden state of the base-uncased BERT model, feed it into a binary classification layer, and fine-tune all of the pre-trained weights using our argument-pair data [3].

Dataset	Acc	AUC	model
UKP	0.83	0.89	ArgBERT[27]
IBM_ArgQ	0.80	0.86	ArgBERT[27]
IBM_Evi	0.73	-	EviConvNet[8]

Table 5: State of the art performance for pairwise argument classification of convincingness for three publicly available datasets, using cross-topic validation scheme

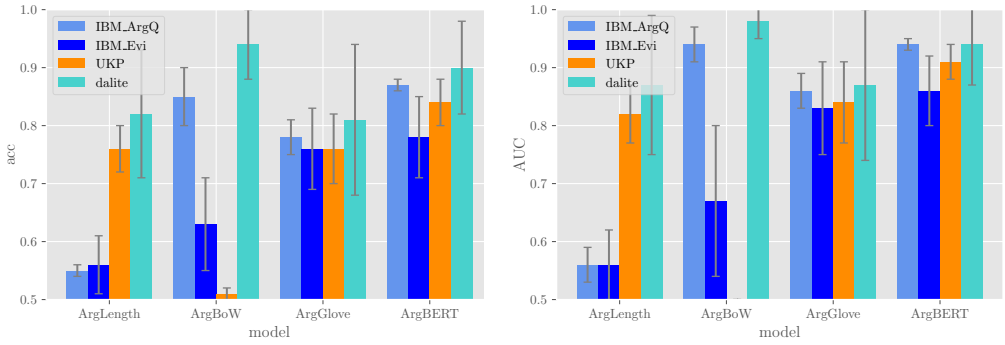
In Table 5, we denote, to the best of our knowledge, the state-of-the-art performance for each dataset on the task of pairwise classification for *convincingness*. These values are meant provide a reference as we present our results in the next section, but cannot be directly compared, as we employ a stratified 5-fold cross-validation scheme, while the reference studies measure cross-topic validation.

5 Results

- ArgLength sets baseline for each dataset. IBM datasets curate most aggressively for pairs with comparable lengths.
- Vector space models performance comparable to BERT.
- GloVe most stable across datasets. UKP and IBMEvi performance on ArgBoW due to wide-ranging vocabulary (lots of unknown tokens in test set), which reveals limitation of approach

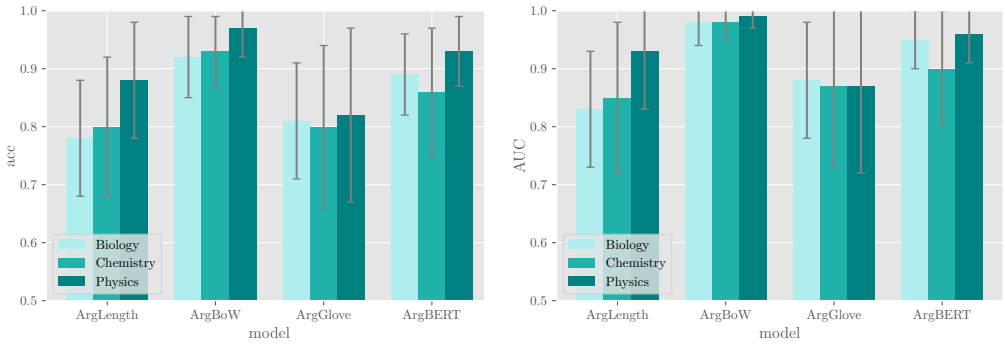
- performance of in chemistry worst, physics best, given their relative proportions in data
- proportionately more errors for observations switch their answer (check if true in IBM Evi, where stances can be different)

[3] modified from the `run_glue.py` script provided by the `tranformers` package, built by company hugging face



(a) Accuracy (b) RoC-AuC

Fig. 1: Pairwise ranking classification accuracy and ROC-AUC for different models across datasets



(a) Accuracy (b) RoC-AuC

Fig. 2: Pairwise ranking classification accuracy and ROC-AUC for different models in myDalite dataset, across disciplines

6 Discussion

In [15], for task of pairwise ranking of newspaper articles based on "quality", the authors achieve a similar result: when comparing the performance of SVMrank models using different input feature sets (e.g. *use of visual language*, *use of named entities*, *affective content*), their top performing models achieve "same-topic" pairwise ranking accuracy of 0.84 using a combination of content and writing features, but also a 0.82 accuracy with the content words as features alone.

7 Future Work

[17] and [15] use combination of writing and content (BoW) features to achieve their best results, and thus this avenue must be explored more thoroughly, especially as this maybe vary across disciplines an teaching contexts.

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 [4]. Work on deriving point wise scores for argument pairs is proposed as a Gaussian Process Preference Learning task by [25]. Seeing the lack of pointwise labels for overall convincingness, [27] 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 their explanations.

References

1. Univeristy of British Columbia, T..L.T.: ubc/ubcpi (Aug 2019), `https://github.com/ubc/ubcpi`, original-date: 2015-02-17T21:37:02Z
2. 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
3. Charles-Woods, E., Whittaker, C., Dugdale, M., Lasry, N., Lenton, K., Bhatnagar, S.: Designing of DALITE: Bringing Peer Instruction on-line. In: Rummel, N., Kapur, M., Nathan, M., Puntambekar, S. (eds.) Computer Supported Collaborative Learning
4. 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)
5. Crouch, C.H., Mazur, E.: Peer instruction: Ten years of experience and results. American Journal of Physics **69**(9), 970–977 (2001)
6. Denny, P., Hamer, J., Luxton-Reilly, A., Purchase, H.: PeerWise: Students Sharing Their Multiple Choice Questions. In: Proceedings of the Fourth International Workshop on Computing Education Research. pp. 51–58. ICER '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1404520.1404526, `http://doi.acm.org/10.1145/1404520.1404526`, event-place: Sydney, Australia

7. 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)
8. 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/
9. 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
10. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression (1999), publisher: IET
11. Jhangiani, R.S.: The impact of participating in a peer assessment activity on subsequent academic performance. Teaching of Psychology **43**(3), 180–186 (2016), publisher: SAGE Publications Sage CA: Los Angeles, CA
12. 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)
13. Khosravi, H., Kitto, K., Williams, J.J.: Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522 (2019)
14. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Transactions on Internet Technology (TOIT) **16**(2), 10 (2016)
15. Louis, A., Nenkova, A.: What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. Transactions of the Association for Computational Linguistics **1**, 341–352 (2013)
16. 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)
17. Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F.: Computational Sociolinguistics: A Survey. arXiv preprint arXiv:1508.07544 (2015)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
19. Persing, I., Ng, V.: End-to-End Argumentation Mining in Student Essays. In: HLT-NAACL. pp. 1384–1394 (2016)
20. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in MOOCs. arXiv preprint arXiv:1307.2579 (2013)
21. Pollitt, A.: The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice **19**(3), 281–300 (2012). https://doi.org/10.1080/0969594X.2012.665354, https://doi.org/10.1080/0969594X.2012.665354, publisher: Routledge _eprint: https://doi.org/10.1080/0969594X.2012.665354
22. Potter, T., Englund, L., Charbonneau, J., MacLean, M.T., Newell, J., Roll, I., others: ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. Teaching & Learning Inquiry **5**(2), 89–113 (2017)
23. SALTISE: SALTISES4/dalite-ng (Jun 2019), https://github.com/SALTISES4/dalite-ng, original-date: 2018-01-11T16:36:55Z

24. Shermis, M.D., Burstein, J.C.: Automated essay scoring: A cross-disciplinary perspective. Routledge (2003)
25. 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
26. 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
27. 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
28. Weir, S., Kim, J., Gajos, K.Z., Miller, R.C.: Learnersourcing subgoal labels for how-to videos. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. pp. 405–416. ACM (2015)
29. 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