Modelling Argument Quality in Technology-Mediated Peer-Instruction

Sameer Bhatnagar
Antoine Lefebvre-Brossard
Michel C. Desmarais
Amal Zouaq
{sameer.bhatnagar,antoine.lefebvre-brossard,michel.desmarais,amal.zouaq}@polymtl.ca
Ecole Polytechnique Montreal
Canada

ABSTRACT

TO DO

CCS CONCEPTS

• **Applied computing → Computer-assisted instruction**; • **Computing methodologies → Natural language processing**.

KEYWORDS

Peer Instruction, Learnersourcing

ACM Reference Format:
Sameer Bhatnagar, Antoine Lefebvre-Brossard, Michel C. Desmarais, and Amal Zouaq. 2021. Modelling Argument Quality in Technology-Mediated Peer-Instruction. In *Proceedings of LAK '21: Learning Analytics & Knowledge (LAK '21).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.000/0000.0000

1 INTRODUCTION

Technology-mediated peer instruction *(TMPI)* platforms [1][11] expand multiple choice items into a two step process. On the first step, students must not only choose an answer choice, but also provide an explanation that justifies their reasoning. On the second step, students are prompted to revise their answer choice, by taking into consideration a subset of explanations written by their peers for another answer choice. In the case that the student wants to keep their original answer choice, but may be unsure of their own explanation, they are also shown peer-explanations for their original answer choice. The student now has three options:

(1) Change their answer choice, by indicating which of their peer's explanations was most convincing
(2) keep their answer choice, but *change explanations* by choosing one that is for the same answer as their own
(3) choose "I stick to my own", indicating that their own explanation is best from amongst those that are shown.

Whenever the student goes with either of the first two scenarios above, we frame this as "casting a vote" for the chosen peer explanation.

The design and growing popularity of TMPI is inspired by three schools of thought: firstly, prompting students to explain their reasoning is beneficial to their learning [4]. Second, classroom based *Peer Instruction*[5], often mediated by automated response systems (e.g. clickers), has become a prevalent, and often effective component in the teaching practice of instructors looking to drive student engagement as part of an active learning experience [2]. In discussing with peers *after* they have formulated their own reasoning, students are engaged in a higher order thinking task from Bloom's taxonomy, as they evaluate what is the strongest argument, before answering again. Thirdly, by capturing data on which explanations students find most convincing, TMPI affords teachers the opportunity to mitigate the "expert blind spot" [8], addressing student misconceptions they might not otherwise have thought of.

In many teaching contexts, however, teachers do not have the time to provide feedback to every student explanation for every question item. The feedback students receive is primarily based on the correctness of their first and second answer choices, not the *explanations* they write and choose.

Moreover, activities from online learning environments are often used for formative assessment, and carry little weight in terms of course credit. Framed as a low-stakes test, this can lead to low student motivation [14]. The expectancy-value model [9], which describes factors that influence the effort students will direct towards a task, includes "how important they perceive the test to be", and the "affective reaction to how mentally taxing the task appears to be" [15].

This makes providing feedback to students on the quality of their explanations a desirable goal, in order to emphasize the importance of the writing activity, as well as promote engagement. The data at hand in TMPI environments enable scaling up how much feedback that can given. The "vote" data represent a proxy for argument quality along the dimension of *convincingness*, as judged by peer learners. These votes can be aggregated into a *convincingness* score to students, as a measure of how effective their explanations are in persuading their peers to change their own answer.

We set out to examine the different measures of argument quality, along the dimension of *convincingness*, and model their role in the TMPI process. Our specific research questions are:

LAK '21, April 2021, UCI,CA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.000/0000.0000

RQ1 What factors influence whether a student will choose a peer's explanation over their own in TMPI?
RQ2 What measures of argument *convincingness* are most useful in aggregating the "vote" data from TMPI?

2 RELATED WORK

2.1 Learnersourcing student explanations

This modality is a specific case of *learnersourcing*[12], wherein students first generate content as part of their own learning process, that is ultimately used to help their peers learn as well.

One of the earliest efforts to leverage peer judgments of peer-written explanations is from the AXIS system[13], wherein students solved a problem, provided an explanation for their answer, and evaluated explanations written by their peers. Using a reinforcement-learning approach known as "multi-armed bandits", the system was able to select peer-written explanations that were rated as helpful as those written by an expert.

2.2 Ranking Arguments for Quality

Rank aggregation is the task of combining the preferences of multiple agents into a single representative ranked list. It has long been understood that obtaining pairwise preference data may be less prone to error on the part of the annotator, as it is a simpler task than rating on scales with more gradations. (This is relevant in TMPI, since each student is choosing one explanation as the most convincing in relation to the subset of others that are shown.)

A classical approach for rank aggregation from pairwise preference data is using the Bradley-Terry model, which has been extended to incorporate the quality of contributions of different annotators in a crowdsourced setting when evaluating relative reading level in a pair passages [3].

When evaluating argument convincingness, one of the first approaches proposed is based on constructing an "argument graph", where a weighted edge is drawn from node A to node B for every pair where argument A is labelled as more convincing than argument B. After filtering example pairs that lead to cycles in the graph, PageRank scores are derived from this directed acyclic graph, and the PageRank scores of each argument are used as the gold-standard to rank for convincingness [7].

More recently, a relatively simpler heuristic Win-Rate score has been shown to be competitive alternative, wherein the rank score of an argument is simply the (normalized) number of times that argument has been chosen as more convincing in a pair, divided by the number of pairs it appears in [10].

Finally, a neural approach based on RankNet has recently yielded state of the art results. By joining two Bidirectional Long-Short-Term Memory Networks in a Siamese architecture, and appending a softmax layer to the output, [6] show that we can jointly model pairwise preferences and overall ranks publicly available datasets.

We will explore two of these options as part of our methodology in our rank aggregation step: the probabilistic Bradley-Terry model, and the simple heuristic scoring model. (We leave the neural approach for future work, as the additional work required to address make the models interpretable enough for the educational context is out of the scope of this study)

3 METHODOLOGY

We borrow our methodological approach from research in argument mining (AM), specifically related to modelling argument quality along the dimension of *convincingness.* A common approach is to curate pairs of arguments made in defence of the same stance on the same topic. These pairs are then presented to crowd-workers who label which of the two is more convincing. These pairwise comparisons can then be aggregated using rank-aggregation methods so as to produce a overall ranked list of arguments.

We extend this work to the domain of TMPI, and change the prediction task: using only vote-data on student explanations, can we effectively predict when a student will choose a student explanation as more convincing than their own?

We build interpretable predictive models for this task, and inspect the relative feature importances in order to address our research questions from above.

Our feature sets include

- *Surface level* features, including: the word count of explanation that the current student wrote; word counts of explanations that were shown on the review step; the number of explanations shown that were much shorter, or much longer than than the student's own explanation; whether the student's first answer is correct;
- *Question and Student level* features, including: the difficulty of the question; the strength of the student (Both of these are defined by overall success rate on getting the correct answer on first attempt);
- *Convincingness* features, including: **win-rate**, defined as the ratio of times an explanation is chosen to the number of times it was shown; **Bradle-Terry** score, which is the argument "quality" parameter estimated for each explanation, according to the Bradley-Terry model, where the probability of argument A being chosen over argument B is given by

$$P(a > b) = \frac{\beta_a}{\beta_a + \beta_b}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{m} [w_{ij} ln \beta_i - w_{ij} ln(\beta_i + \beta_j)]$$

subject to $\sum_i \beta_i = 0$.
The parameters are estimated by minimizing the log-likelihood of the explanations chosen in the pairwise comparison data

4 DATA

5 RESULTS

6 DISCUSSION

7 LIMITATIONS AND FUTURE WORK

(1) Students are not explicitly directed on how to evaluate their peers' explanations. This may have an impact https://link.springer.com/articl 017-0220-3

ACKNOWLEDGMENTS

TO DO

REFERENCES

[1] Elizabeth S. Charles, Nathaniel Lasry, Sameer Bhatnagar, Rhys Adams, Kevin Lenton, Yann Brouillette, Michael Dugdale, Chris Whittaker, and Phoebe Jackson. 2019. Harnessing peer instruction in- and out- of class with myDALITE. In *Fifteenth Conference on Education and Training in Optics and Photonics: ETOP 2019*. Optical Society of America, 11143_89. http://www.osapublishing.org/abstract.cfm?URI=ETOP-2019-11143_89

[2] Elizabeth S. Charles, Nathaniel Lasry, Chris Whittaker, Michael Dugdale, Kevin Lenton, Sameer Bhatnagar, and Jonathan Guillemette. 2015. Beyond and Within Classroom Walls: Designing Principled Pedagogical Tools for Student and Faculty Uptake. International Society of the Learning Sciences, Inc.[ISLS].

[3] 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*. 193–202.

[4] Michelene TH Chi, Nicholas Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science* 18, 3 (1994), 439–477.

[5] Catherine H Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics* 69, 9 (2001), 970–977.

[6] 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. *arXiv preprint arXiv:1907.08971* (2019). https://www.aclweb.org/anthology/P19-1093/

[7] Ivan Habernal and Iryna Gurevych. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, 1589–1599. https://doi.org/10.18653/v1/P16-1150

[8] Mitchell J Nathan, Kenneth R Koedinger, Martha W Alibali, and others. 2001. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the third international conference on cognitive science*, Vol. 644648.

[9] Paul R Pintrich. 1989. The dynamic interplay of student motivation and cognition in the college classroom. *Advances in motivation and achievement* 6 (1989), 117–160.

[10] Peter Potash, Adam Ferguson, and Timothy J Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*. 146–155.

[11] Teaching & Learning Technologies Univeristy of British Columbia. 2019. ubc/ubcpi. https://github.com/ubc/ubcpi original-date: 2015-02-17T21:37:02Z.

[12] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 405–416.

[13] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. ACM Press, Edinburgh, Scotland, UK, 379–388. https://doi.org/10.1145/2876034.2876042

[14] Steven L Wise and Christine E DeMars. 2005. Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment* 10, 1 (2005), 1–17. Publisher: Taylor & Francis.

[15] Lisa F Wolf, Jeffrey K Smith, and Marilyn E Birnbaum. 1995. Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education* 8, 4 (1995), 341–351. Publisher: Taylor & Francis.