Modelling Argument Quality in Technology Mediated Peer Instruction

Sameer Bhatnagar
Polytechnique Montreal

Michel C. Desmarais
Polytechnique Montreal

Amal Zouaq
Polytechnique Montreal

TO DO

Keywords:

---

1. INTRODUCTION

Technology-mediated peer instruction *(TMPI)* platforms (Charles et al., 2019)(Univeristy of British Columbia, 2019) 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.

The student now has three options:

1. Change their answer choice, by indicating which of their peer's explanations for a *different* answer choice was most convincing;

2. keep the *same* answer choice, but indicate which the peer's explanations the student found more convincing than their own;

3. choose "I stick to my own", which indicates that they are keeping to the same answer choice, and that their own explanation is best from among 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 (Chi et al., 1994). Second, classroom based *Peer Instruction* (Crouch and Mazur, 2001), 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 (Charles et al., 2015). 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" (Nathan et al., 2001), addressing student misconceptions they might not otherwise have thought of.

We suggest that the "vote" data collected on each explanation, is a proxy for argument quality, along the dimension of *convincingness*, as judged by peer learners. These votes can be aggregated into a *convincingness* score, as a measure of how effective that explanation is in persuading peers to change their own answer. Student explanations can then be ranked along such a score, allowing for instructors to gain insights on the thinking of their students with respect to specific content, and potentially even help for students improve how they communicate ideas with their discipline.

The problem of aggregating the results of evaluative peer-judgments extends beyond TMPI. For example, in response to the difficulty students can have providing a holistic score to their peers' work, there is a growing number of peer-review platforms built on *comparative* judgments. Notable examples include ComPAIR(Potter et al., 2017) and JuxtaPeer(Cambre et al., 2018), both of which present students with a just a pair of their peers' submissions, and prompt the learner to evaluate them with respect to one another. As in TMPI, students apply a comparative judgment to only the subset of peer content that they are shown during the review step. There is a need for a principled approach to aggregating this learnersourced data in a pedagogically relevant manner.

This sets the stage for our central research questions:

- since each student's "vote" in this context represents an incomplete evaluative judgement, which rank aggregation methods are best suited for the *measuring* the quality of student explanations in TMPI?

- once we establish an appropriate measure of explanation quality, along the dimension of *convinciness*, can we *model* this property, and identify the linguistic features of the most effective student explanations, as judged by their peers?

To our knowledge, we are among the first to examine unsupervised rank aggregation methods as applied to these student "votes" in TMPI, in order to reliable measurements of *convincingness*.

We suggest that the results of our work can inform the design of TMPI platforms. However, in a broader context, we aim to contribute to the growing body of research surrounding technology-mediated peer-review, specifically where learners do not provide holistic scores, but generate their evaluative judgments in a comparative setting. Our research questions generalize to these broader settings: whenever a learner engages in comparative peer assessment, the supporting technology must design some principled approach to constructing the subsets of peer-submissions that will be shown (e.g. random sampling, based on a learner model, showing the *popular* items more often, etc). Instructors need support in parsing through the overwhelming mounts of data that can be generated when students not only create, but also curate content as part of their learning activity.

This paper begins with an overview of research related to learnersourcing (section 2). We then describe our TMPI dataset, as well as publicly available reference datasets of argument quality, which we use to evaluate our methodology (section 3). Our most important contribution is in proposing a methodology for evaluating the quality of student explanations, along the dimension of *convincingness*, in TMPI environments; we demonstrate this methodology in

section 4. We then present our results on choosing the appropriate *measure* of explanation convincingness (section 5), and finally, we describe how we *model* these convincingness "scores" so as to identify the linguistic features of explanations most often associated with high-quality explanations (section 6).

2. Related Work

2.1. Learnersourcing student explanations

This modality is a specific case of *learnersourcing*(Weir et al., 2015), wherein students first generate content as part of their own learning process, that is ultimately used to help their peers learn as well. Notable examples include PeerWise (Denny et al., 2008) and RiPPLE (Khosravi et al., 2019), both of which have student generate learning resources, which are subsequently used and evaluated by peers as part of formative assessment activities.

One of the earliest efforts to leverage peer judgments of peer-written explanations specifically is from the AXIS system(Williams et al., 2016), 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. Our research follows from these studies in scaling to multiple domains, and focusing on how the vote data can be used more directly to model argument quality as judged by peers.

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 (Chen et al., 2013).

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 (Habernal and Gurevych, 2016).

More recently, a relatively simpler heuristic WinRate 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 (Potash et al., 2019).

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, (Gleize et al., 2019) show that we can jointly model pairwise preferences and overall ranks publicly available datasets.

transition	N	N_{pairs}	$wc_{med}(IQR)$
Right -¿ Right	79816	308509	21 (12)
Right -¿ Wrong	1340	9587	16 (14)
Wrong -¿ Right	8373	59541	17 (10)
Wrong -¿ Wrong	16606	65826	18 (10)

Table 1: Summary statistics of data, aggregated by transition type. N is the number of student answers, N_{pairs} is the number of pairs generated from those answers, and $wc_{med}(IQR)$ is the median word count for student explanations, with the inter-quartile range as a measure of dispersion.

We will explore two of these options as part of our methodology in our rank aggregation step, via several related methods: the probabilistic Bradley-Terry model, as well as two of its variants (CrowdBT and the Elo rating system), 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. Data

The data for this study come from myDALITE.org, which is a hosted instance of an open-source project, `dalite`[1], maintained by a Canadian researcher-practitioner partnership focused on supporting teachers develpping active learning pedagogy SALTISE.

Table 1 gives an overview of the dataset included in this study. The data is from introductory level university science courses, and generally spans different teachers at different colleges and universities in Canada.

4. 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, whose task it is to 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 define prediction tasks that not only aim to validate this methodology, but help answer our specific research questions.

4.1. Rank Aggregation

The raw data emerging from a TMPI platform is tabular, in the form of student-item observations. The fields include the item prompt, the student's *first* answer choice, their accompanying explanation, the peer explanations shown on the review step, the student's *second* answer choice,

[1] https://github.com/SALTISES4/dalite-ng

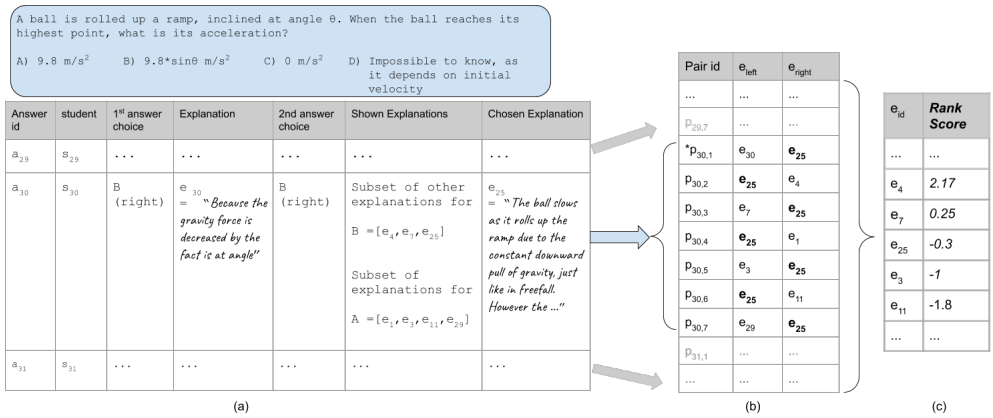

Figure 1: Example of student-item observations from a TMPI environment. (a) Student s_{30} chose the correct **B** as the answer on their first attempt, and provided the explanation e_{30} in the dataset for this question. Without providing feedback on whether this is the correct answer, the student is shown a subset of explanations from previous students for **B**, as well as for **A** (the most popular incorrect answer). The student decides to keep the same answer choice **B**, and indcates that the explanation e_25 is the most convincing. This is referred to as a *Right− >Right* transition. (b) This observation is transformed into 7 explanation pairs. The first pair is for the choice of e_{25} over what the student themself, and the other six are for the choice of e_{25} over the other shown explanations. The pairs are labelled as either having the left or right exlanation being more *convincing*. (c) This pairwise preference data is aggregated global ranked list, where each explanation is assigned a Rank Score.

and the peer explanation they chose as most convincing (`None` if they choose to "stick to their own").

It should be noted that there is no credit associated with which explanation is chosen in this TMPI platform (all points are attributed based on the correctness of the answer choice on the first and second steps). After carefully looking at timestamp data, we observe that a large fraction of students who choose to "stick to their own", spend as little as 5 seconds on the review step. For this reason, we exclude these students' data, and build all rank scores only based on students who explicitly choose a peer's explanation over their own.

After this first filtering step, we take the TMPI observations for each question, and construct explanation pairs, as in figure 1.

1. **WinRate**, defined as the ratio of times an explanation is chosen to the number of times it was shown.

2. **BT** 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

$$P(a > b) = \frac{1}{1 + e^{\beta_b - \beta_a}}$$

where β_i is the latent strength parameter of argument i.

We decompose each student-item observation into argument pairs, where the chosen explanation is paired with each of the other shown ones, and the pair is labelled with

$y = -/+1$, depending on whether the chosen explanation is first/second in the pair. Assuming there are N explanations, labelled by K students, and S_K labelled pairs,the latent strength parameters are estimated by maximizing the log-likelihood given by:

$$\ell(\boldsymbol{\beta}) = \sum_{K} \sum_{(i,j)\epsilon S_K} log \frac{1}{1 + e^{\beta_i - \beta_j}}$$

subject to $\sum_i \beta_i = 0$.

3. The **Elo** rating system(**?**), which was originally proposed for ranking chess players, has been successfully used in adaptive learning environments (see (**?**) for a review). This rating method can be seen as a heuristic re-parametrization of the **BT** method above, where the probability of argument A being chosen over argument B is given by

$$P(a > b) = P_{ab} = \frac{1}{1 + 10^{(\beta_b - \beta_a)/400}}$$

All arguments are initialized with an initial value of 1500 points, an the rating of any argument is only updated after it appears in a pairwise comparison with another. The rating update rule transfers points from the winner, to the loser, in proportion to the difference in strength:

$$\beta_a := \beta_a + K(P_{ab} - \beta_a)$$

While the **BT** model can be thought of a *consensus* approach, **Elo** ratings are dynamic and implicitly give more weight to recent data(**?**).

4. **Crowd-BT** (Chen et al., 2013) is an extension of the **BT** model tailored to settings where different annotators may have assigned opposite labels to the same pairs, and the reliability of each annotator may vary significantly. A reliability parameter is estimated for each student,

$$\eta_k \equiv P(a >_k b | a > b)$$

where $\eta_k \approx 1$ if the k^{th} student agrees with most other students, and $\eta_k \approx 0$ if the student is in opposition to their peers. This changes the model of argument a being chosen over b by student k to

$$P(a >_k b) = \eta_k \frac{1}{1 + e^{\beta_b - \beta_a}} + (1 - \eta_k) \frac{1}{1 + e^{\beta_b - \beta_a}}$$

and the log-likelilood maximized for estimation to

$$\ell(\boldsymbol{\eta}, \boldsymbol{\beta}) = \sum_{K} \sum_{(i,j)\epsilon S_K} log(\eta_k \frac{1}{1 + e^{\beta_i - \beta_j}} + (1 - \eta_k) \frac{1}{1 + e^{\beta_i - \beta_j}})$$

5. **Length**, a method used purely as a baseline, where for each pair, we simply predict that the explanation with more words is the more convincing. This is a commonly used baseline for the pairwise classification task of predicting argument quality (Toledo et al., 2019) has

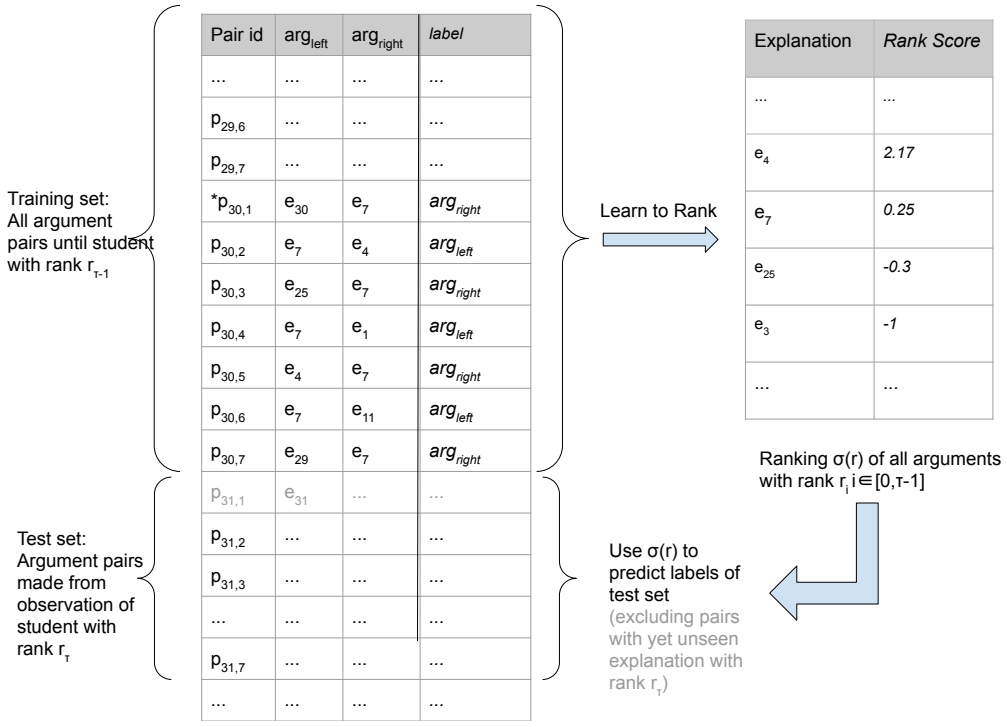

Figure 2: All of the pairs constructed for answers by students $i <= 30$ are used to learn rank scores σ, which are evaluated in their ability to predict the label in the held out test set: argument pairs constructed from the answer of student $i = 31$. We must exclude those pairs which include the 31st student, as this explanation is unseen in the training set, and does not have a rank score

been shown to be competitive for data from learning environments (**?**). (Since we only use a basic white-space tokenizer, we round the token-counts of each explanation down to the nearest multiple of five, as it is unlikely that a student could discern which is longer if the difference in lengths is less than this.)

In order to evaluate these rank aggregation different scores, and address our research question, we employ a time-series based cross-validation scheme: at each timestep, we calculate the aggregated argument *convincingness* scores from past students, and set out to predict: which arguments will be chosen as more convincing from the pairs constructed for the current student?

5. Measuring Argument Quality

TO DO

6. Modelling Argument Quality Scores

The goal **RQ1** is establish which rank aggregation methods are best suited for the context of TMPI, such that one can take the comparative preference data from many students who each see different subsets of peer explanations. We build on the results from the previous section to now predict these aggregate scores for each explanation, using linguistic properties of those explanations

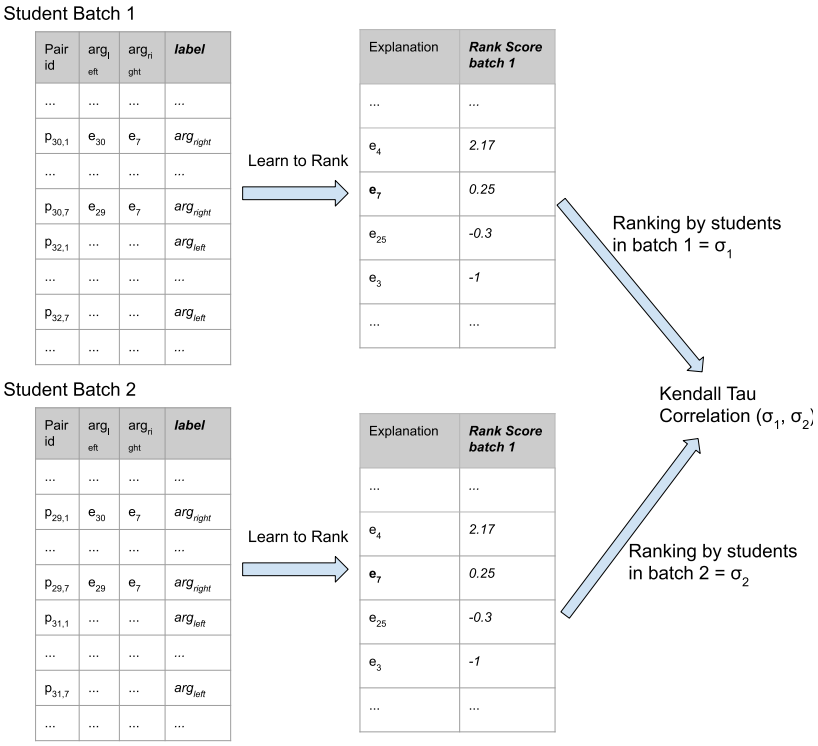

Figure 3: Validation scheme for evaluating reliability of rankings. At each time step τ we take all students $i <= \tau$, and split student into two batches (chosen at alternating time steps) We learn rankings for each of these batches, and evaluate the Kendall Tau rank correlation as estimate of reliability of the rankings.

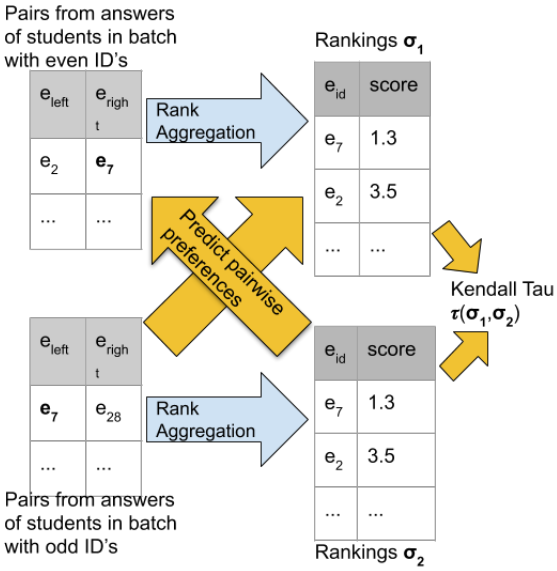

Figure 4: Methodology for evaluation of rank scores

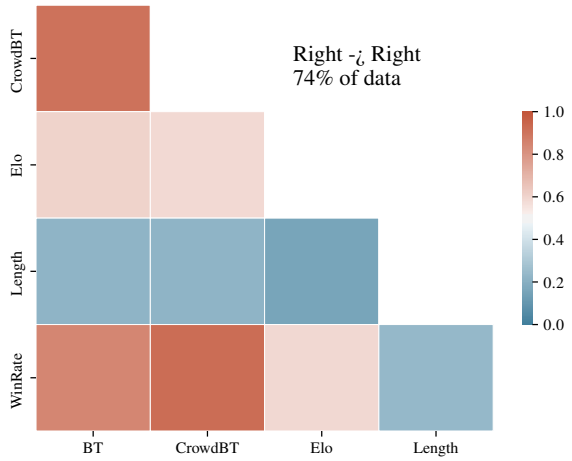

Figure 5: Correlation between different Ranking Scores for each explanation, disaggregated by transition type

Figure 6: Comparing the classification accuracy of different rank aggregation scores in predicting which argument is more convincing from a pair. Rank scores are calculated with the vote data of half the students, and tested on the pairs generated by the other half. Data is averaged across all questions, dis-aggregated by different TMPI transition types.

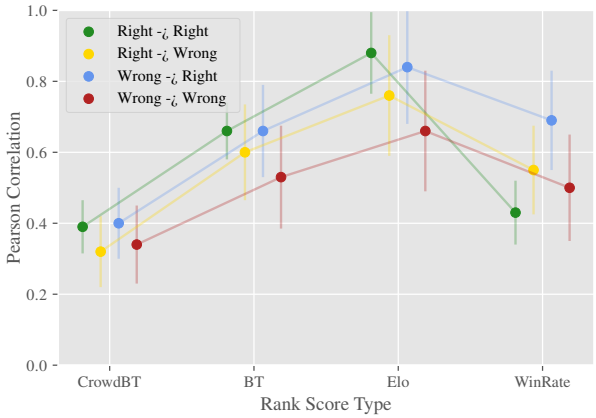

Figure 7: Pearson correlation coefficient between different rank score types, derived from two independent groups of students, averaged over all questions, dis-aggregated by different TMPI transition types.

We address **RQ2** with a regression task of predicting the argument *convincingness* scores using two different approaches to representing the student text: as an embedding inside a vector space models, or via a feature-rich document vector.

We experiment with vector space models with different document representations:

1. LSA vectors (10,50,100 components) (Deerwester et al., 1990)
2. Glove embeddings (Pennington et al., 2014)
3. BERT embeddings (Devlin et al., 2018), out-of-the-box, and fine-tuned for the current classification task

The advantage of a feature-rich approach lies in the interpretability for teachers in their reporting tools, as well as generalizability to new items before vote data can be collected. The list of features included here are derived from related work in argument mining (Habernal and Gurevych, 2016)(Persing and Ng, 2016)on student essays, automatic short answer scoring (Mohler and Mihalcea, 2009)

- Surface Features: word count, sentence count, max/mean word length, max/mean sentence length;
- Lexical: uni-grams & bigrams, type-token ratio, number of keywords (defined by open-source discipline specific text-book), number of equations;
- Syntactic: POS n-grams (e.g. *nouns, prepositions, verbs,conjunctions,negation, adjectives, adverbs, punctuation*), modal verbs (e.g. *must, should, can, might*), contextuality/formality measure (Heylighen and Dewaele, 2002), dependency tree depth;
- Semantic: using LSA vectors trained on domain specific corpora, in this case an open-source textbook in the discipline, we calculate similarity to all other explanations in LSA space;

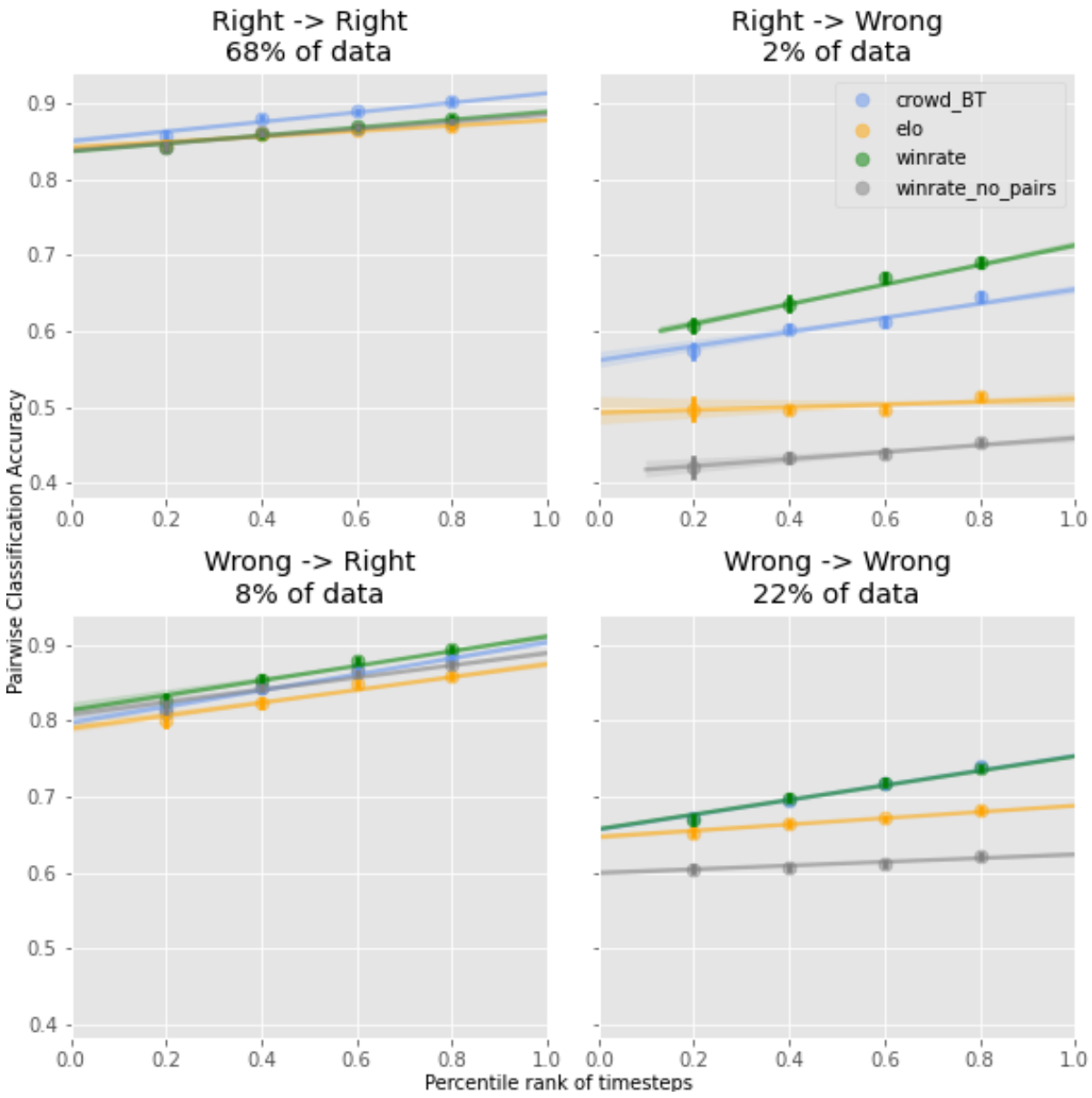

Figure 8: to do

- Co-reference (Persing and Ng, 2016): fraction of entities from the prompt mentioned in each sentence, averaged over all sentences (using neural Co-reference resolution) vector cosine similarity between student explanation and prompt, and answer choices;
- Readability: Fleish-Kincaid, Coleman-Liau, spelling errors

Features typical to NLP analyses in Learning Analytics that are not included here are cohesion, sentiment, and psycholinguistic features.

REFERENCES

CAMBRE, J., KLEMMER, S., AND KULKARNI, C. 2018. 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*. ACM Press, Montreal QC, Canada, 1–13.

CHARLES, E. S., LASRY, N., BHATNAGAR, S., ADAMS, R., LENTON, K., BROUILLETTE, Y., DUGDALE, M., WHITTAKER, C., AND JACKSON, P. 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.

CHARLES, E. S., LASRY, N., WHITTAKER, C., DUGDALE, M., LENTON, K., BHATNAGAR, S., AND GUILLEMETTE, J. 2015. Beyond and Within Classroom Walls: Designing Principled Pedagogical Tools for Student and Faculty Uptake. International Society of the Learning Sciences, Inc.[ISLS].

CHEN, X., BENNETT, P. N., COLLINS-THOMPSON, K., AND HORVITZ, E. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202.

CHI, M. T., LEEUW, N., CHIU, M.-H., AND LAVANCHER, C. 1994. Eliciting self-explanations improves understanding. *Cognitive science 18,* 3, 439–477.

CROUCH, C. H. AND MAZUR, E. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics 69,* 9, 970–977.

DEERWESTER, S. C., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W., AND HARSHMAN, R. A. 1990. Indexing by latent semantic analysis. *JASIS 41,* 6, 391–407.

DENNY, P., HAMER, J., LUXTON-REILLY, A., AND PURCHASE, H. 2008. PeerWise: Students Sharing Their Multiple Choice Questions. In *Proceedings of the Fourth International Workshop on Computing Education Research*. ICER '08. ACM, New York, NY, USA, 51–58. event-place: Sydney, Australia.

DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

GLEIZE, M., SHNARCH, E., CHOSHEN, L., DANKIN, L., MOSHKOWICH, G., AHARONOV, R., AND SLONIM, N. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. *arXiv preprint arXiv:1907.08971*.

HABERNAL, I. AND GUREVYCH, I. 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.

HEYLIGHEN, F. AND DEWAELE, J.-M. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of science 7,* 3, 293–340. Publisher: Springer.

KHOSRAVI, H., KITTO, K., AND WILLIAMS, J. J. 2019. Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *arXiv preprint arXiv:1910.05522*.

MOHLER, M. AND MIHALCEA, R. 2009. Text-to-text Semantic Similarity for Automatic Short Answer Grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. Association for Computational Linguistics, Stroudsburg, PA, USA, 567–575. event-place: Athens, Greece.

NATHAN, M. J., KOEDINGER, K. R., ALIBALI, M. W., 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.

PENNINGTON, J., SOCHER, R., AND MANNING, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. Vol. 14. 1532–1543.

PERSING, I. AND NG, V. 2016. End-to-End Argumentation Mining in Student Essays. In *HLT-NAACL*. 1384–1394.

POTASH, P., FERGUSON, A., AND HAZEN, T. J. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*. 146–155.

POTTER, T., ENGLUND, L., CHARBONNEAU, J., MACLEAN, M. T., NEWELL, J., ROLL, I., AND OTHERS. 2017. ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry 5,* 2, 89–113.

TOLEDO, A., GRETZ, S., COHEN-KARLIK, E., FRIEDMAN, R., VENEZIAN, E., LAHAV, D., JACOVI, M., AHARONOV, R., AND SLONIM, N. 2019. 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)*. 5629–5639.

UNIVERISTY OF BRITISH COLUMBIA, T. . L. T. 2019. ubc/ubcpi. original-date: 2015-02-17T21:37:02Z.

WEIR, S., KIM, J., GAJOS, K. Z., AND MILLER, R. C. 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.

WILLIAMS, J. J., KIM, J., RAFFERTY, A., MALDONADO, S., GAJOS, K. Z., LASECKI, W. S., AND HEFFERNAN, N. 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.