Probabilistic Modeling of Peer Correction and Peer Assessment

Takeru Sunahase
Kyoto University
stakeru@ml.ist.i.kyoto-u.ac.jp

Yukino Baba
University of Tsukuba
baba@cs.tsukuba.ac.jp

Hisashi Kashima
Kyoto University;
RIKEN Center for AIP
kashima@i.kyoto-u.ac.jp

ABSTRACT

Peer assessment is a promising solution for scaling up the grading of a large number of submissions. The reliability of evaluations is one of the critical issues in peer assessment; several probabilistic models have been proposed for obtaining reliable grades from peers. *Peer correction* is a similar framework, in which students are instructed to correct the errors in submissions from other students. Peer correction is typically performed simultaneously with peer assessment; a reviewer is instructed to correct the errors in a submission and to provide a grade to it. We observe the occasional inconsistency between a grade and the correction; for example, a reviewer provides a high grade for a submission but she corrects many errors in it. Such inconsistencies can point to unreliable reviewers. In this paper, we propose probabilistic models for peer correction, and the combination of the peer correction models and the existing peer assessment models for capturing the inconsistency to accurately estimate the reviewer reliability and the student ability. We conduct experiments using the dataset of an actual peer correction platform for language translation, and the results demonstrate that the combination of peer correction models and peer assessment models improves the accuracy of the student ability estimation.

Keywords

Peer correction, peer assessment, statistical models

1. INTRODUCTION

MOOCs have changed education by offering open access to university course materials; however, not everything performed in offline classes is effectively introduced in MOOCs. An example is the ability assessment; in offline classes, teachers evaluate the student abilities by examining their submitted assignments and decide how to improve the educational efficiency. In contrast, assessing the abilities of tens of thousands of students in MOOCs is not feasible for teachers.

Takeru Sunahase, Yukino Baba and Hisashi Kashima "Probabilistic Modeling of Peer Correction and Peer Assessment" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 426 - 431

A promising solution for large-scale ability assessment is to allow students themselves to be involved in the evaluation; instead of teachers, students grade the submissions from other students. Such *peer assessment* approach is beneficial for scaling-up the ability assessment and it has been applied to several MOOCs courses [7]. However, the reliability of evaluations is one of the critical issues in peer assessment because some students may provide unreliable evaluations owing to laziness or lack of evaluation skills. Several probabilistic models have been proposed for estimating the reliabilities of the reviewers in order to accurately assess the student abilities in peer assessment [7, 4, 11, 8, 13, 6]. These models are based on the assumption that students with high ability are likely to provide reliable grades. The models are used to estimate the ability of a student as a test taker and the reliability as a reviewer.

In a similar framework of peer assessment, called *peer correction*, students correct the errors in the submissions from other students. Peer correction is helpful for teachers to reduce their efforts for providing feedback to the students. Typically, peer correction is performed simultaneously with peer assessment; a student is instructed to grade a submission and to correct its errors.

Although the outcomes of peer correction are naturally assumed to be informative for estimating the student abilities, probabilistic models for peer correction have not yet been investigated. Based on a natural assumption that a student who receives fewer corrections are likely to have a higher ability, we propose probabilistic models for peer correction that capture the relationship between the student abilities and the correction outcomes.

Additionally, we noticed an inconsistency between the outcomes of peer correction and those of peer assessment. In one case, a reviewer provides a high grade to a submission but she corrects many errors in it; in another case, a reviewer assigns a low grade but she does not make any corrections. Our idea is that such inconsistencies are beneficial in determining unreliable reviewers; thus, we propose to combine peer assessment models with our peer correction models. This combination allows us to capture the inconsistency and to incorporate it into the estimation of the reviewer reliability and the student ability.

We conduct experiments using a peer correction dataset about language translation. The results of the experiments

show that our probabilistic models for peer correction are capable of estimating the student abilities, and the combined models of peer correction and peer assessment demonstrate a better performance in determining high-ability students than the peer assessment models.

The contributions of this paper are twofold: (i) we propose novel probabilistic models for peer correction that enable us to estimate the student abilities from the received corrections (Section 4), and (ii) we propose to combine our peer correction models and peer assessment models to exploit the inconsistencies among the outcomes of corrections and assessments (Section 6); the results of the experiments show that the combined models are efficient in accurately estimating the student abilities.

2. PROBLEM DEFINITION

We begin with the formulation of peer assessment and peer correction. We assume there is a set of students $\mathcal{S}$. When a student creates a submission for an assignment, other students (that we call *reviewers*) evaluate it and assign grades. The grade for the student $u \in \mathcal{S}$ assigned by the reviewer $v \in \mathcal{S}$ is denoted by $z_{uv} \in \mathbb{R}$. Each reviewer is additionally instructed to correct the errors in a submission. A correction result is denoted by y_{uv}. If a reviewer does not provide any correction for a submission, such information is also embedded in y_{uv}. The representation of y_{uv} is discussed in the next section.

Given a set of peer assessment and peer correction outcomes, $\mathcal{D}$, each of which is represented by a tuple (u, v, z_{uv}, y_{uv}), our goal is to estimate the true abilities of the students $\{s_u\}_{u \in \mathcal{S}}$, where $s_u \in \mathbb{R}$.

3. DATASET

In this work, we use a peer assessment and peer correction dataset collected from Conyac[1], which is a crowdsourcing language translation platform. This platform employs peer correction and peer assessment between translators for collaboratively improving their skills; thus, a translator on this platform can be considered as a student. When a student submits a translation, other students evaluate its quality on a five-point scale (zero (low) to four (high)) and correct the errors in it. Students are invited to high-reward jobs if they have reviewed several submissions.

Students on Conyac can take a qualification test to demonstrate their skills. On this test, a student is instructed to translate the given sentences and then the translations are evaluated by experts employed by the service provider. According to the score, a student is assigned one of five expertise levels (D, C, B, A, and A+). This level is used for the job assignment and the default level is set to one. We consider the assigned levels as the ground truth of the student abilities, that we aim to estimate from the outcomes of peer assessment and peer correction.

We target the peer assessment and peer correction for Japanese to English translations on Conyac. Our dataset contains 5,008 reviews for 413 students, and 135 students provide at least one review. Figure 1(a) shows the distribution of the grades assigned to translations and Figure 1(b) illustrates the distribution of the students' true expertise levels.

We conduct exploratory data analysis to investigate how the outcomes of peer correction can be used for estimating student expertise levels. A natural expectation is that a student whose submissions are likely to be corrected would have lower ability. We calculate the correction ratio of each student, which is the number of corrected submissions divided by the number of submissions. Figure 1(c) shows the average correction ratio of the students in each expertise level. We observe that students with the highest level are likely to have lower correction ratios than the others.

Additionally, we consider that students who have more errors in their submissions would be have lower ability. We calculate the number of corrected parts in each submission by applying the Gestalt pattern matching [10]. We first obtain the matched patterns in pre-correction and post-correction submissions, and then count the number of unmatched patterns in the post-correction submissions. The examples of the calculated numbers are shown in Table 1 and Figure 1(d) shows the distribution of the number of corrected parts in each submission. We calculated the average number of corrected parts of each student and Figure 1(e) presents the average of the values at each level. We found that the students with higher levels are likely to have a lower number of corrected parts.

From these observations, we decide to use the following binary and numerical variables to represent a correction outcome: (1) $y_{uv}^{(\mathrm{b})} \in \{0, 1\}$, which indicates whether the corresponding submission is corrected by the grader ($y_{uv}^{(\mathrm{b})} = 0$) or not ($y_{uv}^{(\mathrm{b})} = 1$), (2) $y_{uv}^{(\mathrm{n})} \in \{0, 1, 2, \dots, \}$, which indicates the number of parts corrected by the grader.

4. PEER CORRECTION MODELS

We propose two peer correction models, $\mathrm{PC_b}$ and $\mathrm{PC_n}$, for estimating the student true abilities. The models are illustrated in Figure 2(a).

4.1 $\mathrm{PC_b}$ model

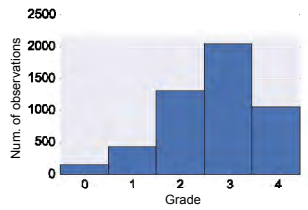
We first present a generative model for $y_{uv}^{(\mathrm{b})} \in \{0, 1\}$, which is a binary indicator whether the submission has been corrected by the reviewer or not. We have two latent parameters into our model, that is, student true ability and reviewer bias; each student is associated with the latent true ability, $s_u \in \mathbb{R}$, which we aim to estimate, and each reviewer has a different bias parameter, $b_v \in \mathbb{R}$, presuming that a reviewer with a lower bias tends to review a submission negatively.

Following the observations, we assume that a submission from a student is likely to be not corrected by a reviewer if the student has high ability. In addition, a reviewer is not likely to correct a submission if he/she has a higher bias. These assumptions are represented as the following generative model:

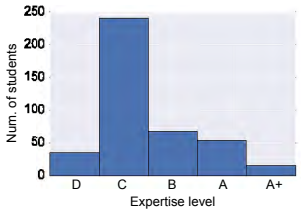
$$y_{uv}^{(\mathrm{b})} \sim \mathrm{Bern}\left(y_{uv}^{(\mathrm{b})} \middle| \sigma\left(s_u + b_v + r\right)\right), \tag{1}$$

where $\sigma(x) = 1/\left(1 + \exp(-x)\right)$, $\mathrm{Bern}(\cdot)$ is the Bernoulli dis-

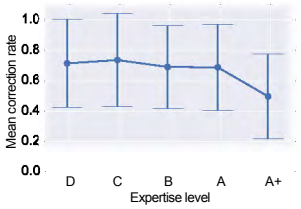
[1] https://conyac.cc/



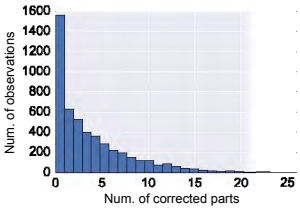
(a) Distribution of submission grades



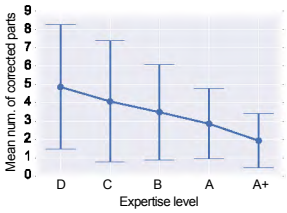
(b) Distribution of expertise levels



(c) Average and standard deviation of correction ratios



(d) Distribution of the number of corrected parts



(e) Average and standard deviation of the mean numbers of corrected parts

Figure 1: Statistics of our dataset

Table 1: Examples of the calculated number of corrected parts. The corrected parts are highlighted (modified parts are highlighted in yellow, added parts are highlighted in pink, and removed parts are highlighted in blue). There were seven corrected parts in the last example because there were three modified parts ("Current members," "are" and "was working on the"), two added parts ("female" and ","), and two removed parts ("girls" and "the").

Pre-correction	Post-correction	Num. of corrected parts
Please enter the title. Please enter the application conditions.	Please enter the title. Please enter the eligibility requirements.	1
41 people from major travel agencies of Japan and land operators participated and had the business meetings with suppliers about latest Thailand MICE circumstances.	41 people from major travel agencies of Japan and land operators participated and had the business meetings with suppliers about the latest Thailand MICE circumstances.	1
Kalafina is the vocal girls band produced by Yuki Kajiura. Currently the member of Kalafina is WAKANA, KEIKO and HIKARU. The group was formed in order to produce the main song when the composer Yuki Kajimura produced music for the film "Boundary of Emptiness".	Kalafina is the female vocal band produced by Yuki Kajiura. Current members of Kalafina are WAKANA, KEIKO, and HIKARU. The group was formed in order to produce the main song when composer Yuki Kajimura was working on the music for the film "Boundary of Emptiness".	7

tribution, and r is a noise. Note that $y^{(\mathrm{b})}_{uv} = 1$ indicates that the corresponding submission is *not* corrected by the reviewer. We denote this generative model by $\mathrm{PC_b}$ model. We can interpret $s_u + b_v + r$ as an apparent ability of the student u for the reviewer v at the time. The model indicates that a submission is likely to be not corrected when the apparent ability is high.

In the same way as the existing peer assessment models that will be reviewed in the next section, we use normal distributions as priors for s_u, b_v, and r:

$$\text{(Student ability)} \quad s_u \sim \mathcal{N}(s_u|\mu_0, 1/\gamma_0) \tag{2}$$
$$\text{(Reviewer bias)} \quad b_v \sim \mathcal{N}(b_v|0, 1/\eta_0) \tag{3}$$
$$\text{(Noise)} \quad r \sim \mathcal{N}(r|0, 1/\kappa_0), \tag{4}$$

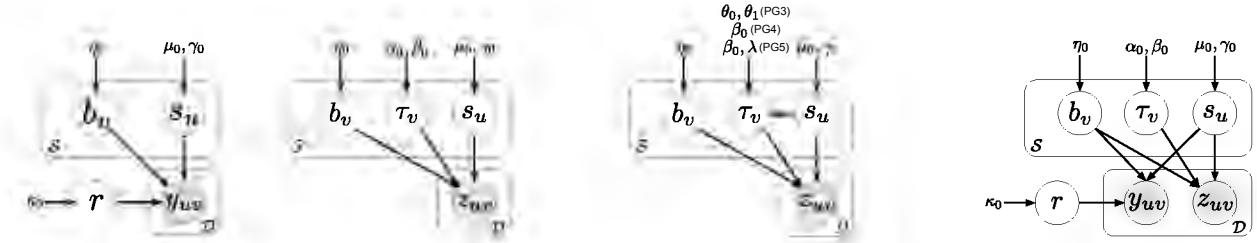
where μ_0, γ_0, η_0, and κ_0 are hyperparameters.

4.2 $\mathrm{PC_n}$ model

Our second model targets the number of corrected parts in each correction, $y^{(\mathrm{n})}_{uv} \in \{0, 1, 2, \dots, \}$. Following the observations from the actual dataset, we assume that a reviewer corrects more parts of a submission when the student has a lower ability. We use the Poisson distribution to represent this assumption:

$$y^{(\mathrm{n})}_{uv} \sim \mathrm{Poisson}\left(y^{(\mathrm{n})}_{uv} \middle| \frac{1}{\exp(s_u + b_v + r)}\right).$$

Similar to the $\mathrm{PC_b}$ model, $s_u + b_v + r$ is considered as the apparent ability of the student u to the reviewer v, and this model indicates that more parts of the submission is likely to be corrected by the reviewer if the apparent ability of the student is lower. We call this model $\mathrm{PC_n}$. The priors given in Eqs. (2), (3), and (4) are incorporated into the $\mathrm{PC_n}$ model as well.



(a) PC_b and PC_n (b) PG_1 (c) PG_3, PG_4, and PG_5 (d) PG_1+PC_b

Figure 2: Peer correction and peer assessment models, and combined models

5. PEER ASSESSMENT MODELS

We next review the existing peer assessment models, which are combined with our peer correction models in the next section. In particular, we summarize the PG_1 [7], PG_3 [7], PG_4 [6], and PG_5 [6] models[2]. These are the generative models for grades, $z_{uv} \in \mathbb{R}$. The peer assessment models are illustrated in Figures 2(b) and 2(c).

The student ability and the reviewer bias parameters are also incorporated in the peer assessment models. All the models use the same priors given in Eqs. (2) and (3).

5.1 PG_1 and PG_3 models

In addition to the latent parameters incorporated in the peer correction models (s_u and b_v), the peer assessment models contain the reviewer reliability $\tau_v \in \mathbb{R}^+$. This parameter indicates how likely a grade given by the reviewer contains a noise. PG_1 is defined as follows:

$$\begin{aligned} \text{(Reviewer reliability)} \quad & \tau_v \sim \text{Gamma}(\tau_v|\alpha_0, \beta_0) \\ \text{(Outcome)} \quad & z_{uv} \sim \mathcal{N}\left(z_{uv}|s_u + b_v, 1/\tau_v\right), \end{aligned}$$

where α_0 and β_0 are hyper parameters. PG_3 is an extension of PG_1, which incorporates the relationship between the reviewer reliability and the ability of the reviewer (as a student). PG_3 is given as follows:

$$\begin{aligned} \text{(Reviewer reliability)} \quad & \tau_v = \theta_1 s_v + \theta_0 \\ \text{(Outcome)} \quad & z_{uv} \sim \mathcal{N}\left(z_{uv}|s_u + b_v, 1/\tau_v\right), \end{aligned}$$

where θ_0 and θ_1 are hyperparameters.

5.2 PG_4 and PG_5 models

PG_4 and PG_5 are variations of PG_3 and they incorporate the relationship between the reviewer reliability and the reviewer ability into the priors of the reliability parameter. The generative models of the reviewer reliability and outcome in PG_4 are given as follows:

$$\begin{aligned} \text{(Reviewer reliability)} \quad & \tau_v \sim \text{Gamma}(\tau_v|s_v, \beta_0) \\ \text{(Outcome)} \quad & z_{uv} \sim \mathcal{N}\left(z_{uv}|s_u + b_v, 1/\tau_v\right), \end{aligned}$$

and those in PG_5 are given as follows:

$$\begin{aligned} \text{(Reviewer reliability)} \quad & \tau_v \sim \mathcal{N}(\tau_v|s_v, 1/\beta_0) \\ \text{(Outcome)} \quad & z_{uv} \sim \mathcal{N}\left(z_{uv}|s_u + b_v, \lambda/\tau_v\right), \end{aligned}$$

where β_0 and λ are hyperparameters.

[2]We do not include PG_2 [7], which is almost similar to PG_1 except it incorporates time-series factors.

6. COMBINED MODELS FOR PEER CORRECTION AND PEER ASSESSMENT

We finally combine our peer correction models and the existing peer assessment models. By combining these two types of models, we expect to capture an inconsistency between the outcome of peer correction and that of peer assessment; the inconsistency can be informative for estimating the reviewer reliabilities.

We use PG_1 and PC_b to explain the model combining and we term the combined model as PG_1+PC_b. We simply consider that s_u and b_v are shared between these two models; namely, the generative model for PG_1+PC_b is given as:

$$\begin{aligned} \text{(Student ability)} \quad & s_u \sim \mathcal{N}\left(s_u|\mu_0, 1/\gamma_0\right) \\ \text{(Reviewer reliability)} \quad & \tau_v \sim \text{Gamma}(\tau_v|\alpha_0, \beta_0) \\ \text{(Reviewer bias)} \quad & b_v \sim \mathcal{N}\left(b_v|0, 1/\eta_0\right) \\ \text{(Noise)} \quad & r \sim \mathcal{N}(r|0, 1/\kappa_0) \\ \text{(Outcomes)} \quad & z_{uv} \sim \mathcal{N}\left(z_{uv}|s_u + b_v, 1/\tau_v\right), \text{ and} \\ & y_{uv}^{(\text{b})} \sim \text{Bern}\left(y_{uv}^{(\text{b})} \middle| \sigma\left(s_u + b_v + r\right)\right). \end{aligned}$$

The PG_1+PC_b model is illustrated in Figure 2(d). Other combined models are defined similarly as PG_1+PC_b.

When an inconsistency occurs between corrections and grades, i.e., a reviewer provides a high grade to a submission but makes many corrections in it, we consider that a large noise occurs on the grade (z_{uv}) and thus the reliability of the reviewer (τ_v) is estimated as low. The combination of peer assessment models and peer correction models allows us to leverage such inconsistencies to estimate the reviewer reliabilities and the student abilities.

7. EXPERIMENTS

We conduct experiments using the actual peer assessment and peer correction dataset about language translation. We investigate the effectiveness of the proposed methods to estimate the student abilities.

7.1 Baselines

We compare the proposed models (PC_b, PC_n, and $PG_{\{1,3,4,5\}}$+$PC_{\{b,n\}}$) with the following baselines: **(a) Correction ratio ($\mathbf{PC_b^\sharp}$)**: this is a naïve version of PC_b and considers the correction ratio of each student as the ability. Specifically, the correction ratio is defined as $-\sum_{y_{uv}^{(\text{b})} \in \mathcal{Y}_u^{(\text{b})}} \delta\left(y_{uv}^{(\text{b})} = 0\right) / |\mathcal{Y}_u^{(\text{b})}|$, where $\mathcal{Y}_u^{(\text{b})}$ is the set of correction outcomes for the student u, and $\delta(\cdot)$ is the in-

Table 2: Average and standard deviation of AUC scores of each method on various classification boundaries. Each column indicates the results for each classification boundary; for example, (D, C, B, A | A+) represents the results for classifying the students at A+ and the others. The winner for each boundary is bold-faced. The cases where a combined model outperforms the corresponding peer assessment model (PG_1, PG_3, PG_4, or PG_5) are underlined.

	AUC			
	(D, C, B, A | A+)	(D, C, B | A, A+)	(D, C | B, A, A+)	(D | C, B, A, A+)
$PC_b^\sharp$	0.713 ± 0.020	0.584 ± 0.015	0.580 ± 0.013	0.498 ± 0.025
PC_b	0.805 ± 0.037	0.628 ± 0.008	0.604 ± 0.013	0.483 ± 0.034
$PC_n^\sharp$	0.714 ± 0.045	0.627 ± 0.013	0.598 ± 0.010	0.611 ± 0.041
PC_n	0.832 ± 0.050	0.710 ± 0.012	0.690 ± 0.008	0.602 ± 0.046
$PG^\sharp$	0.794 ± 0.023	0.723 ± 0.015	0.697 ± 0.013	$\mathbf{0.810} \pm 0.011$
PG_1	0.845 ± 0.016	0.739 ± 0.025	0.742 ± 0.013	0.801 ± 0.015
PG_3	0.751 ± 0.193	0.756 ± 0.033	0.725 ± 0.020	0.786 ± 0.020
PG_4	0.862 ± 0.069	0.774 ± 0.012	$\mathbf{0.743} \pm 0.015$	0.791 ± 0.011
PG_5	0.842 ± 0.115	0.775 ± 0.029	0.731 ± 0.015	0.759 ± 0.043
PG_1+PC_b	0.821 ± 0.020	$\underline{0.755} \pm 0.016$	0.736 ± 0.010	0.792 ± 0.011
PG_3+PC_b	$\underline{0.841} \pm 0.137$	$\underline{0.779} \pm 0.026$	$\underline{0.736} \pm 0.011$	$\underline{0.789} \pm 0.030$
PG_4+PC_b	$\underline{0.870} \pm 0.019$	0.764 ± 0.008	0.730 ± 0.021	$\underline{0.800} \pm 0.016$
PG_5+PC_b	$\underline{\mathbf{0.914}} \pm 0.018$	$\underline{\mathbf{0.782}} \pm 0.016$	0.719 ± 0.026	$\underline{0.782} \pm 0.032$
PG_1+PC_n	$\underline{0.846} \pm 0.019$	0.737 ± 0.017	0.726 ± 0.009	0.661 ± 0.047
PG_3+PC_n	0.788 ± 0.153	0.705 ± 0.044	0.704 ± 0.022	0.710 ± 0.060
PG_4+PC_n	0.844 ± 0.038	0.753 ± 0.017	0.724 ± 0.018	0.646 ± 0.054
PG_5+PC_n	$\underline{0.888} \pm 0.024$	0.746 ± 0.033	0.731 ± 0.012	0.686 ± 0.066

dicator function. For assigning a higher ability for a student with less corrections, we multiply the value with -1. **(b) Mean number of corrected parts ($PC_n^\sharp$)**: this is a naïve version of PC_n and considers the mean number of the corrected parts of each student as the ability. Specifically, the mean number of the corrected parts of the student u is defined as $-\sum_{y_{uv}^{(n)} \in \mathcal{Y}_u^{(n)}} y_{uv}^{(n)} / |\mathcal{Y}_u^{(n)}|$, where $\mathcal{Y}_u^{(n)}$ is the set of correction outcomes for the student u. For assigning a higher ability for a student with less corrected parts, we multiply with -1. **(c) Mean grades ($PG^\sharp$)**: this is a naïve version of PG_1 and considers the mean assigned grades of each student as the ability. The mean grade is defined as $\sum_{z_{uv} \in \mathcal{Z}_u} z_{uv} / |\mathcal{Z}_u|$, where $\mathcal{Z}_u$ is the set of grades assigned to the student u. **(d) PG_1, PG_3, PG_4,** and **PG_5**: existing peer assessment models.

7.2 Experimental setup

We implemented the models using the No-U-Turn Sampler (NUTS) [3], which is a variation of the Hamiltonian Monte Carlo. We executed four chains and they produce 5,000 samples in total. The initial 500 samples were ignored and the average of the rest samples were used as the estimated parameters.

We randomly generated 150 sets of candidate hyperparameters for each method. A method with a set of candidate hyperparameters produces the estimated student abilities. Their performance was evaluated using the groundtruth of 20% of the students. We then decided the best set of hyperparameters for the method and the final result for each method was evaluated by the remaining students. We performed this procedure five times and calculated the average.

Each method outputs the estimated ability of each student. We use the expertise levels assessed by the experts as the ground truth, and investigate how accurately each method classifies the students with high expertise and those with low expertise. We specifically use the area under the ROC curve (AUC) as an evaluation metric.

7.3 Results

Table 2 shows the AUC scores of each method on different classification boundaries. Our peer correction models (PC_b and PC_n) demonstrate better or comparative performance to the existing peer grading models in detecting the students at the highest level; this supports the effectiveness of the peer correction results for estimating student abilities. We see that the "no-correction" cases only occur for high-ability students and the correction information is helpful for distinguishing between the "perfect students" and "almost perfect students", both are likely to obtain the highest grades from the reviewers and the correction outcomes are required to classify them.

In contrast, the performance of peer correction models becomes inferior for detecting the students at lower levels, and $PG^\sharp$ achieves the best performance for detecting the students at the lowest level; the average of the obtained grades is sufficiently informative for detecting low-ability students. Our methods would be beneficial for a situation where teachers aim to detect students who require advanced course materials or assignments.

The combined models of $PG_{\{1,3,4,5\}}$+PC_b outperform the corresponding $PG_{\{1,3,4,5\}}$ in most cases; the outcomes of peer correction are useful for improving the student ability estimation. It is noteworthy that PG_5+PC_b achieves an AUC of 0.914 for classifying the students at A+ and the others. This result is brought by the capability of the combined models for capturing the inconsistencies between the outcomes of assessments and those of corrections.

The number of corrected parts can be more informative than simply considering whether a submission is corrected; in fact, $\mathrm{PC_n}$ is better than $\mathrm{PC_b}$ and $\mathrm{PC}_n^\sharp$ performs better than $\mathrm{PC}_b^\sharp$; however, $\mathrm{PG}_{\{1,3,4,5\}}$+$\mathrm{PC_b}$ outperforms $\mathrm{PG}_{\{1,3,4,5\}}$+$\mathrm{PC_n}$ in our experiments. Because there are more model variations in $\mathrm{PC_n}$ than $\mathrm{PC_b}$, a more meticulous modeling for combining the PG models and the $\mathrm{PC_n}$ model would be required.

8. RELATED WORK

Peer assessment models are categorized into two groups: models for cardinal peer assessment and models for ordinal peer assessment. The former models target a situation where the outcomes are assigned in explicit numerical scores, such as five-point scores. In addition to the probabilistic models reviewed in Section 5, Walsh proposed PeerRank [13], an extension of PageRank for peer assessment. In ordinal peer assessment, each grader is shown multiple submissions and instructed to rank them. The Bradley–Terry model [2] has been applied for ordinal peer assessment [11, 8] and Mi et al. proposed to use the cardinal peer assessment models for ordinal peer assessment [6]. Although several probabilistic models for peer assessment have been studied, peer correction has not yet been investigated.

The design of peer assessment frameworks has been attempted to improve the reliability of evaluation. Kulkarni et al. ([4]) reported that the feedback about the grading bias to graders was beneficial for improving the reliability. Another work proposed to design peer assessment as a multiple choice task where a grader is instructed to choice the best submission [5]. Peer assessment mechanisms based on game theory have been introduced to derive accurate evaluations from peers [14].

Our peer correction models are very related to the models studied in the item response theory, which are for quantifying student abilities and item characteristics in educational tests. One of the simple item response theory model is the Rasch model [9] and our $\mathrm{PC_b}$ (given in Eq.(1)) model has a similar formulation to the Rasch model.

Besides peer assessment, probabilistic models for estimating grader reliability have been studied in crowdsourcing as well. Specifically, a two-stage framework was proposed where crowdsourcing workers in the first stage produce outputs, such as translations or logo designs, and another set of workers in the second stage evaluates the outputs [1]. Probabilistic models for estimating the reliability of each grader and the quality of each output in this two-stage framework have been proposed [1, 12]. Unlike peer assessment, the overlap between students (i.e., creators of outputs) and graders is not assumed in crowdsourcing.

9. CONCLUSIONS

We presented probabilistic models for peer correction, which are used for estimating the student abilities. We proposed two models: one considering whether a grader has corrected a submission, and the other utilizing the number of corrected parts in each submission. We also combined the peer correction models with the peer assessment models; this combination allows us to estimate the reliability of graders from the outcomes of peer corrections and those of peer assessment by considering the consistency between the corrections and assessments. The experiments using the actual dataset of peer correction showed that the combination of peer correction models and peer assessment models was particularly effective in detecting high ability students.

In our models, we did not consider the importance of each corrected part; however, the importance levels differ among corrected parts in which minor corrections (e.g., adding a punctuation mark) and major corrections (e.g., paraphrasing) exist. A major correction would indicate the low quality of a submission and considering such factors is a promising direction to improve the ability estimation accuracy.

10. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 15H01704 and 18K18105.

11. REFERENCES

[1] Y. Baba and H. Kashima. Statistical quality estimation for general crowdsourcing tasks. In *ACM SIGKDD*, pages 554–562, 2013.

[2] R. A. Bradley and M. Terry. The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3):324–345, 1952.

[3] M. D. Hoffman and A. Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[4] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33:1–33:31, 2013.

[5] I. Labutov and C. Studer. JAG: a crowdsourcing framework for joint assessment and peer grading. In *AAAI*, pages 1010–1016, 2017.

[6] F. Mi and D.-Y. Yeung. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In *AAAI*, pages 454–460, 2015.

[7] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. In *EDM*, pages 153–160, 2013.

[8] K. Raman and T. Joachims. Methods for ordinal peer grading. In *ACM SIGKDD*, pages 1037–1046, 2014.

[9] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* 1960.

[10] J. W. Ratcliff and D. E. Metzener. Pattern matching: the Gestalt approach. *DDJ*, 13(7):46, 1988.

[11] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran. A case for ordinal peer-evaluation in MOOCs. In *NIPS-DDE*, 2013.

[12] T. Sunahase, Y. Baba, and H. Kashima. Pairwise HITS: quality estimation from pairwise comparisons in creator-evaluator crowdsourcing process. In *AAAI*, pages 977–984, 2017.

[13] T. Walsh. The PeerRank method for peer assessment. In *ECAI*, pages 909–914, 2014.

[14] W. Wu, C. Daskalakis, N. Kaashoek, C. Tzamos, and M. Weinberg. Game theory based peer grading mechanisms for MOOCs. In *L@S*, pages 281–286, 2015.