Predicting Low vs. High Disparity between Peer and Expert Ratings in Peer Reviews of Physics Lab Reports

Huy V. Nguyen[1] and Diane J. Litman[1,2]

[1] Department of Computer Science | [2] Learning Research and Development Center, University of Pittsburgh

Introduction

OBSERVATIONS:

- Peer reviews address instructor/TA workload and help student develop writing and evaluation skills
- However, a disparity between expert (instructor/TA) and peer grades is unavoidable

OUR GOALS:

- Better understand the validity [3] of peer assessment
- Identify peer outliers in terms of rating disparity with experts

METHODS:

- *Classify peers into groups of low and high rating disparity with experts using only* ***features derived from peer reviews***

Peer Review Data

- Peer and expert reviews of the same report assignment, Physics Lab classes 2010-2011
- Student reports were organized into sections: *abstract*, *introduction*, *experiment*, *analysis*, and *conclusion*
- SWoRD [2] was used to assign reports to reviewers for grading and commenting via rubric
- All classes had 1 or 2 experts review and rate reports
- Number of peers per report varied from 1 to 7
- Rating scaled from 1 (poor) to 7 (excellent)

An Example Instance of Reviews

Fig. 1 Reviews of student and expert of a Introduction section for a student report. Left to right: reviewer, rating, comment

R1	7	[…] everything is explained clearly. Experiment 3 and 4 were perfect.
R2	7	Really nice job! […] I understood everything you were saying.
R3	7	A lot of equations you could probably get rid of some of the basic ones, other than that it was very good.
R4	1	[…] There was little to no theory in this section. […] Try to explain more of the symbols […] as many of them are unclear.
Expert	6	You provide most of the critical equations […]. You are also good at balancing the equation and the description of the theory.

Binary Classification Task

- For each student report section (**instance**), calculate absolute difference (**rating disparity**) between means of peer and expert ratings
- For each dataset, split instances into **Low group** and **High group** according to median of rating disparity
- Predict whether rating disparity of an instance is Low or High

Table 1. Number of instances of each section

Section	Abstract	Intro.	Exper.	Analysis	Concl.
# inst.	362	361	362	280	362

Table 2. Means of rating disparity in the low and high groups ($p < 0.01$) of 5 datasets

Section	Abstract	Intro.	Exper.	Analysis	Concl.
Low	0.37	0.30	0.38	0.40	0.30
High	1.51	1.39	1.53	1.65	1.61

Machine Learning Features

RATING FEATURES:

- **#Peers**: number of peer reviewers per instance
- **Mn** and **Std**: mean and STDEV of peer ratings

COMMENT FEATURES:

- For each dataset, a standard LDA [3] run over all peer comments
- Topic diversity is measured as distance between topic distribution using Euclidean distance (**Euc**) and Kullback–Leibler divergence (**KL**)
- For each instance, inter-comment topic diversity is quantified by the average distance of all comment pairs in the set

Acknowledgement. This work is supported by LRDC Internal Grants Program, University of Pittsburgh. We thank C. Schunn for providing us with the data and feedback.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022.
2. K. Cho and C. D. Schunn (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. Computers and Education, 48(3), 409-426.
3. K. Cho, C. D. Schunn, and R. W. Wilson (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. Journal of Educational Psychology, 98(4), 891-901.

Experimental Results

- Rating features yield significantly higher accuracies than majority baseline (Tab. 3, Col. 2)
- Comment features outperform baseline for 3 of 5 sections (Tab. 3, Col. 3)
- Adding topic features do not further improve the use of rating features (Tab. 3, Col. 4)

Table 3. Prediction accuracies with 10-fold cross validation. * denotes $p < 0.05$ compared to majority baseline

Section	Majority	#Peers + Mn + Std	#Peers + Euc + KL	All
Abstract	54.98	61.66 *	56.27	61.06 *
Intro.	50.69	60.40 *	61.62 *	59.91 *
Exper.	51.10	63.15 *	58.61 *	62.82 *
Analysis	51.07	62.43 *	51.07	62.07 *
Concl.	54.42	67.02 *	59.17 *	66.86 *

Discussion and Future Work

Table 4. Correlation coefficients between Mn and Rating Disparity ($p < 0.01$)

Section	Abstract	Intro.	Exper.	Analysis	Concl.
Corr.	-0.21	-0.37	-0.38	-0.4	-0.35

Table 5. Correlation coefficients between topic diversity and Rating Std ($p < 0.01$). Similar results are for KL metric

Section	Abstract	Intro.	Exper.	Analysis	Concl.
Euc	0.38	0.38	0.45	0.39	0.45

- Peers and experts agree more (lower rating disparity) when peers give high grades (Tab. 4)
- The two topic diversity metrics both positively correlate to the Std peer ratings (Tab. 5)
- No correlation between peer rating reliability, in terms of Std, Euc or KL, and rating validity in terms of disparity with experts
- Figure 1 shows such a case: peer ratings are of low reliability (Std=3) but high validity (Mn=5.5 vs. Expert-rate=6)
- In future, improve predictive accuracy by adding features extracted from student papers
- Study different rating validity measurements