Peer Reviewing Short Answers using Comparative Judgement

Pushkar Kolhe
Georgia Institute of Technology
801 Atlantic Drive,
Atlanta, GA 30332, USA
pushkar@cc.gatech.edu

Dr. Michael L. Littman
Brown University
115 Waterman St.
Providence, RI 02912, USA
mlittman@cs.brown.edu

Dr. Charles L. Isbell
Georgia Institute of Technology,
801 Atlantic Drive,
Atlanta, GA 30332, USA
isbell@cc.gatech.edu

L@S 2016, April 25–26, 2016, Edinburgh, UK.
ACM 978-1-4503-3726-7/16/04.
http://dx.doi.org/10.1145/2876034.2893424

Abstract

We propose a comparative judgement scheme for grading short answer questions in an online class. The scheme works by asking students to answer short answer questions. Then a multiple choice question is created whose choices are the answers given by students. We show that we can formulate a probabilistic graphical model for this scheme which lets us infer each students proficiency for answering and grading questions.

Introduction

Asking questions in a class is an important element in keeping students engaged. In an online class this is usually done with multiple choice questions or answers that can be easily graded. However for some topics it is essential to ask open ended short answer questions. Online classes don't often have these type of questions because it is hard to give timely feedback on them. We propose a new scheme that uses comparative judgment to give feedback for such questions.

Our method can be used to completely automate grading of short answer questions. An adaptive comparative judgement scheme can generally be used to rank short answers. Our method allows this scheme to grade these answers as correct or wrong. Just like adaptive comparative judgement schemes our method does not suffer from grader bias. Our method also allows us to recognize ambiguous answers and refer them to an expert.

Related Work

There has been significant work in the peer grading literature. For example, [3, 1, 4] have shown various approaches for peer grading short essay questions. Our work extends their work by using the new approach explained below.

Our work is related to the Completely Automated Public Turing test that is used to tell Computers and Humans Apart (reCAPTCHA) system [5]. It is a challenge response system used online to determine whether a user is a human or a computer. In it, users are asked to translate images into words. The reCAPTCHA system specifically challenges the user with two images. In its most simplest instantiation the system knows the translation of an image but does not know the translation of the other image. When the user answers the challenge, the system verifies if the user was indeed a human or not from the known translation. The other answer is then directly used as a label to translate the other image. The reliability of the translation increases when several users translate this image with the same label.

In the same vein, our questions are True/False questions (but they could be simple short answer questions too). Students provide a label and an explanation when they answer this question. These answers are then converted to a multiple choice questions by combining answers of atleast two students. Then we challenge the students to choose correct answers. We use this grading to determine two things: proficiency of the student at grading and correctness of the answer.

Problem Formulation

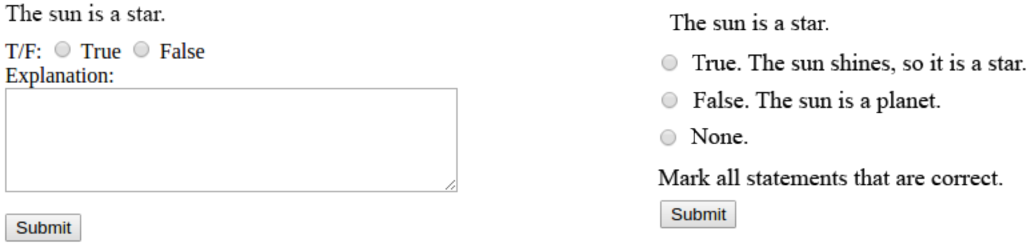
Every student u has some proficiency and every question v has a hardness parameter associated with it. Simply put, if the student's proficiency is higher than the question's hardness, there is a good chance that the student answers the question correctly. In any given exam the student is asked a series of True/False questions and a series of Multiple Choice questions. In the multiple choice questions, the choices are answers given by other students. Our models assume the existence of the following parameters which are either observed or latent variable which we wish to estimate.

- Grader Proficiency: Every student who is participating in the exam has a reliability score associated with them. There are two types of reliabilities - one for **generating** a correct answer and the other for **recognizing** a correct answer.
- Question Hardness: Every question has a hardness parameter associated with it.
- Answer Correctness: Every answer is labeled by the student. This parameter shows the probability that the grade was correct.
- Observed Answer Grades: Finally this is the observed grade for an answer.

Computing the posterior is non-trivial since all the variables are correlated with each other. That is why we use an approximate inference technique like Gibbs sampling.

Creating Multiple Choice Questions

As we have shown in Figure 1 student's answers are converted to multiple choice questions. When we run or Gibbs



The sun is a star.

T/F: ○ True ○ False

Explanation:

Submit

The sun is a star.

○ True. The sun shines, so it is a star.

○ False. The sun is a planet.

○ None.

Mark all statements that are correct.

Submit

Figure 1: The figure shows an example question as it can appear to two different students. In the first case the student is going to provide an answer to the question. In the second case, the student is shown two answers and asked to choose correct answers.

sampler we get confidences associated with the correctness of every answer's grading. If our confidence is low on a particular answer, we can get more information by posing it for grading to a highly proficient student. A systematic approach is used here so that we can improve confidence in our grading with the minimum number of times we pose a multiple choice question.

We can pair up an answer of high confidence with an answer of low confidence. If the grader grades the high confidence answer correctly, we can be assured that will also grade the other answer correctly.

If we know that the grader is highly proficient in grading, we can pose answers with low confidence values to them. Since we have confidence over their grading, we can be assured that they will grade the answers correctly.

Grading through multiple choice questions is used to improve our estimates on the answer correctness. We run our Gibbs sampler again after a student solves the multiple choice questions posed to them.

Conclusion

We showed a new approach to model the peer grading problem. In this model we have two reliability parameters for the students: one for their proficiency in answering questions and the second for grading. Also our formulation does not ask the student grader to grade with a numerical value, so it removes grader bias.

Our experiments show that with Gibbs Sampling we can estimate these parameters very reliably. Also as the number of students in a class increase, the reliability improves.

Discussion

Polar Grading

In linguistics, a question with a binary choice is known as a polar question. We call our grading *polar grading* because we ask the students to tell us if an answer is correct or wrong. We don't ask them to give a numerical grade. We assume that students are better at making a polar choice than making a qualitative judgment to determine a numerical score. There has been some study in the linguistics area towards this question. Some studies show how children respond to a yes-no question as opposed to an open ended wh-question [2]. We would like to perform similar experiments in the context of peer grading.

Student Ranking

As teachers we have to give grades to our students. A grade is usually a measure of how much the student proficiency in the class material. Usually these are determined by the score the students scores in the exams, that is, the sum of the scores of the *true grade* in this exam.

However if a grade is a measure of the student proficiency, then in our scheme two parameters that could relate to this are the reliability parameters - *reliability of generation* and *reliability of recognition*. In fact, one can argue that good students can answer hard questions and good students can grade hard questions. In a way Teaching Assistants (TAs) are the best students of their class.

If we had to rank students, traditionally we have used true grades. But there is a case to consider the reliability parameters for ranking too. Which of these parameters best represents the actual student proficiency is an open question. Answering it will be a part of our future work in this subject.

REFERENCES

1. Edward F Gehringer. 2001. Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bulletin* 33, 1 (2001), 139–143.
2. Carole Peterson, Craig Dowden, and Jennifer Tobin. 1999. Interviewing preschoolers: Comparisons of yes/no and wh-questions. *Law and Human Behavior* 23, 5 (1999), 539.
3. Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579* (2013).
4. Karthik Raman and Thorsten Joachims. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1037–1046.
5. Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science* 321, 5895 (2008), 1465–1468.