Quality Assessment of Diagnostic Questions Based on Multiple Features

Yuto Shinahara
Aidemy Inc.
Tokyo, Japan
shinahara-y@aidemy.co.jp

Daichi Takehara
Aidemy Inc.
Tokyo, Japan
takehara-d@aidemy.co.jp

Abstract

This paper presents our solution for the *Global Question Quality Assessment*, which is the third task *Diagnostic Questions: The NeurIPS 2020 Education Challenge*. The objective of this challenge is to devise a metric to measure the quality of the diagnostic questions. We extracted several features that are maybe useful for predicting question quality, such as a balance between the choice of answers, level of difficulty, and readability. We finally achieved the accurate prediction of question quality by integrating these features. The code for this solution is available at `https://github.com/haradai1262/NeurIPS-Education-Challenge-2020`.

1 Introduction

To improve the quality of education, we need to receive the feedback from the students' answers and continuously reflected it in an educational system. However, the quality of the insights derived from the feedback is dependent on the quality of the questions in the assessments. Thus, to guarantee the quality of the insights, we must improve the quality of *diagnostic question* [1, 2]. A diagnostic question is a multiple-choice question with four answers, where one answer is correct, and each of the three incorrect answers is chosen to highlight a common misconception. In other words, an excellent diagnostic question enables teachers to know from the answer to the question of whether or not a student correctly understands the relevant knowledge.

Diagnostic Questions: The NeurIPS 2020 Education Challenge [3, 4] aims to develop novel methodologies to understand and improve students' learning and measure the quality of diagnostic questions. The challenge consists of four tasks, and engagement with each task has individual real-world impact. One of them, the objective of the third task *Global Question Quality Assessment*, is to devise a metric to measure the quality of the diagnostic questions. In this paper, we present our solution for the third task. Specifically, we extracted several features that are maybe useful for predicting question quality, such as a balance between the choice of answers, level of difficulty, and readability.

2 Challenge

In this paper, we describe the third task of the competition: *Global Question Quality Assessment*.

2.1 Dataset

In this task, we can utilize a dataset of students' answers to mathematics questions from Eedi[1], a leading educational platform with which millions of students from around the globe interact daily.

[1] `https://eedi.com/`

1st NeurIPS Education Challenge (NeurIPS 2020), Vancouver, Canada.

Table 1: Example of the training data. The table is referenced from [3]. The values of *QuestionId*, *UserId*, and *AnswerId* are unique indices assigned randomly.

QuestionId	UserId	AnswerId	AnswerValue	CorrectAnswer	IsCorrect
10322	452	8466	4	4	1
2955	11235	1592	3	2	0
3287	18545	1411	1	0	0
10322	13898	6950	2	1	0

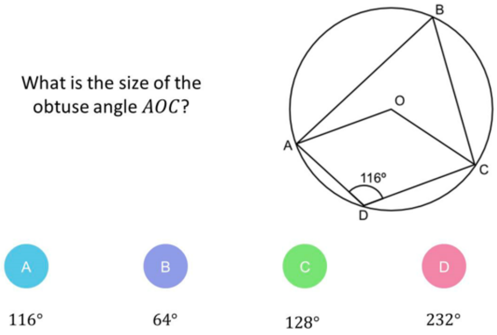


Figure 1: An example question image. The image is referenced from [3].

We were provided with two kinds of data[2]: one was the training data, as shown in Table 1, which consisted of 4,918 students' answers to 948 questions (the total amount of data was more than 1.3 million approximately), while the other was the images presented to the students in the questions, as shown in Figure 1. Each question consists of four choices, and there is only one correct answer.

2.2 Task Description

Goal The goal of this task is to devise a metric to measure the quality of the diagnostic questions. Precisely, we predict each question's quality, as described in [3], and create the ranking of questions' quality based on the dataset introduced in 2.1.

Evaluation In this task, the similarity between experts' judgment and question quality ranking that we created becomes the evaluation metric. Assuming there are $T(T \in \mathbb{N})$ pairs of questions: Q_{A_t} and $Q_{B_t}(t = 1, ..., T)$ and there are $U(U \in \mathbb{N})$ experts. When uth $(u = 1, ..., U)$ expert judges that the quality of Q_{A_t} is higher than the quality of Q_{B_t} and the rank of Q_{A_t} is also higher (i.e., has minimum integer value) than the rank of Q_{B_t} in our ranking, the value s_t is set to 1; otherwise, it is set to 0. Here, uth expert's score S_u is computed as,

$$S_u = \frac{\sum s_t}{T}, \tag{1}$$

and the final score S is defined as,

$$S = \max_u S_u. \tag{2}$$

3 Method

3.1 Overview

For devising suitable metrics to measure the quality of the diagnostic questions, we formed a hypothesis that an appropriate diagnostic question strikes (1) a balance between the choice of answers,

[2] Actually, metadata of the training data was also provided, however, we do not utilize it in this paper.

(2) an appropriate level of difficulty, and (3) readability. Based on this hypothesis, we computed features explained in 3.2 and created a quality ranking of each question by the below process.

1. Compute a feature of each question
2. Sort each question by descending order based on the feature computed in process 1
3. Create an individual question quality ranking by assigning rank 1, 2, ... to sorted questions sequentially
4. Create the other individual question quality rankings by adapting processes through 1 to 3 to the other features
5. Create the final question quality ranking by computing the mean of rankings in the individual question quality rankings

3.2 Features

Feature 1: Selection entropy An appropriate diagnostic question could have a variety of answers selected by students. To quantify this feature, we utilized the variation of *AnswerValue* in the training data. As each question consists of four choices, regarding the number of times kth $(k = 1, 2, 3, 4)$ choice of question whose *QuestionId* value is n (n is non-negative integer) is selected as probability $p_k(n)$ and assuming $p_k(n)$ follows multinomial distribution $P(n)$, we are able to define selection entropy of the question as,

$$H(n) = -\sum_{n} P(n) \log P(n). \tag{3}$$

We expect that a question with a high entropy value would be a balanced question.

Feature 2: Correct/Wrong entropy For quantification of students' variety of answers, we also utilized *IsCorrect* feature. The same formula 3 computes the entropy of this feature.

Feature 3: Difficulty An appropriate diagnostic question should be neither too easy nor too difficult; in other words, students who understand the question can solve it and vice versa. Hence, we tried to estimate each question's difficulty by computing the difference between the mean correctness rate of a student who answered a question and whether the student's answer to the question is correct or wrong. We expect an appropriate question is solved by students with a high mean rate of correctness but is not solved by students with a low mean rate of correctness.

We define the mean correctness rate of each user as,

$$R(j) = \overline{X_j}, \tag{4}$$

where j is equivalent to *UserId* value (non-negative integer) of each user and X_j is set defined as,

$$X_j = \{x_{i,IsCorrect} | x_{i,UserId} = j\}, \tag{5}$$

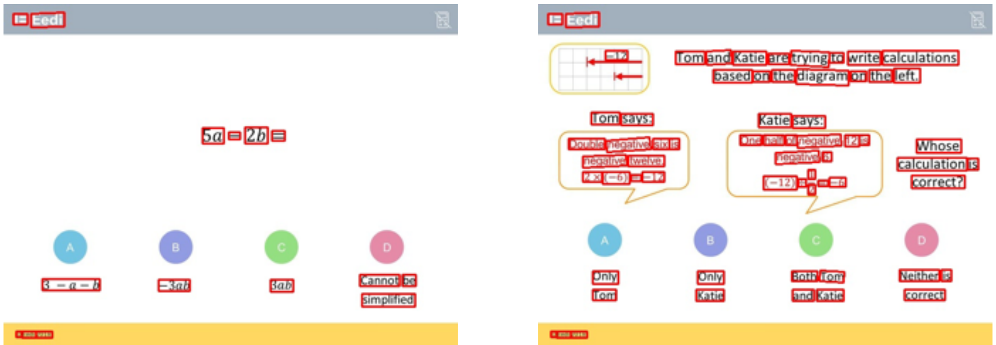
where $x_{i,IsCorrect}$ and $x_{i,UserId}$ are *IsCorrect* and *UserId* value of ith (i is a non-negative integer) training data, respectively. Here, we can compute nth question's difficulty $D(n)$ as,

$$D(n) = \overline{Y_n}, \tag{6}$$

where set Y_n is defined as,

$$Y_n = \{|x_{i,IsCorrect} - R(x_{i,UserId})|\, | x_{i,QuestionId} = n\}. \tag{7}$$

Feature 4: Readability We experimentally utilized text complexity of question images for quantification of question readability. Specifically, we extracted text regions from a question image and then calculated the proportion of the text area to the image's whole area. We adapted CRAFT [5] for text localization. CRAFT is a convolutional neural network-based framework, and it computes the character *region score*, for localization of individual characters in the image, and *affinity score*, for grouping each character into a single instance. It outperformed state-of-the-art text detectors and showed high flexibility with variously shaped texts. Figure 2 shows example images adapting CRAFT to the question images.



$5a - 2b =$

A $3 - a - b$ B $-3ab$ C $3ab$ D Cannot be simplified

Tom and Katie are trying to write calculations based on the diagram on the left.

Tom says: Double negative six is negative twelve $2 \times (-6) = -12$

Katie says: One half of negative 12 is negative 6 $(-12) \div 2 = -6$

Whose calculation is correct?

A Only Tom B Only Katie C Both Tom and Katie D Neither is correct

Figure 2: Examples of text localization to question images. Each red box shows text areas detected by CRAFT. Each example shows a question image including less text area (left) and a question image including more text area (right), respectively.

Table 2: Submission results. The formula 1 computes each score S_u, and the underlined score is equivalent to the final score S, computed by the formula 2, of the submission. The bold and underlined values are the maximum score through all submissions.

	Score (Public evaluation)					Score (Private evaluation)				
Utilized features	S_1	S_2	S_3	S_4	S_5	S_1	S_2	S_3	S_4	S_5
1	0.72	0.76	0.76	0.64	0.64	0.36	0.48	0.48	0.36	0.36
2	0.72	0.84	0.68	0.56	0.64	0.64	0.52	0.52	0.60	0.64
3	0.64	0.60	0.76	0.48	0.80	0.56	0.52	0.68	0.68	0.56
4	0.56	0.52	0.60	0.64	0.56	0.44	0.48	0.64	0.64	0.60
1, 2	0.68	0.80	0.72	0.60	0.60	0.56	0.52	0.44	0.48	0.64
1, 2, 3	0.72	0.76	0.84	0.64	0.72	0.64	0.60	0.60	0.60	**0.80**
1, 2, 3, 4	0.68	0.72	**0.88**	0.60	0.68	0.64	0.60	0.60	0.60	**0.80**

4 Result

Table 2 shows our results. There are two types of evaluation: public (the score and ranking of each participant are always open to the public) and private (the score and ranking of each participant are closed to the public for the duration of the competition). We compete with other participants based on the score in the private evaluation. The formula 1 computes each score S_u, and the underlined score is equivalent to the final score S, computed by the formula 2, of the submission. The bold and underlined values are the maximum score through all submissions.

From the table 2, we marked the highest submission score when utilizing all features both in the public evaluation and in the private evaluation. Besides, feature 3, difficulty, seems to be the most contributory metric for question quality assessment. The submission utilizing this feature marked the highest score compared with other submissions utilizing other single features individually.

Finally, we achieved accurate question quality assessment, and our maximum score in the private evaluation (0.80) marked one of the top scores in the world[3].

5 Conclusion

We have proposed our solution for the third task *Global Question Quality Assessment* in the *Diagnostic Questions: The NeurIPS 2020 Education Challenge*. Experimentally, we found that the difficulty of questions is a contributory metric for question quality assessment. As a result, we achieved the accurate question quality assessment and one of the top scores in the world for this task.

[3] Our user name is *myaunraitau*. Whole ranking table is in `https://competitions.codalab.org/competitions/25449#results`

References

[1] E. C. Wylie and D. Wiliam. Diagnostic questions: is there value in just one. In *Annual Meeting of the National Council on Measurement in Education*, 2006.

[2] J. L. Little. The role of multiple-choice tests in increasing access to difficult-to-retrieve information. *Journal of Cognitive Psychology*, 30(5-6):520–531, 2018.

[3] Z. Wang, A. Lamb, E. Saveliev, P. Cameron, Y. Zaykov, J. M. Hernández-Lobato, R. E. Turner, R. G. Baraniuk, C. Barton, S. P. Jones, S. Woodhead, and C. Zhang. Diagnostic questions: the neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*, 2020.

[4] Diagnostic questions - the neurips 2020 education challenge. `https://competitions.codalab.org/competitions/25449`.

[5] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.