Rubric-based Automated Japanese Short-answer Scoring and Support System Applied to QALab-3

Tsunenori Ishioka
The Center for University Entrance Examinations
tunenori@rd.dnc.ac.jp

Kohei Yamaguchi
Kyushu University
k.yamaguchi.795@s.kyushu-u.ac.jp

Tsuneori Mine
Kyushu University
mine@ait.kyushu-u.ac.jp

ABSTRACT

We have been developing an automated Japanese short-answer scoring and support machine for the new National Center written test exams. Our approach is based on the fact that accurately recognizing textual entailment and/or synonymy has been almost impossible. The system generates automated scores on the basis of evaluation criteria or rubrics, and human raters revise them. The system determines semantic similarity between the model answers and the actual written answers as well as a certain degree of semantic identity and implication. An experimental prototype operates as a web system on a Linux computer. To evaluate the performance, we applied the method to the second round of entrance examinations given by the University of Tokyo. We compared human scores with the automated scores for a case in which 20 allotment points were placed in five test issues of a world-history test as part of a trial examination. The differences between the scores were within 3 points for 16 of 20 data provided by the NTCIR QALab-3 task office.

Team Name

Tmkff

Subtask

Evaluation method task

Keywords

writing test, automated scoring, machine learning, random forests, recognizing textual entailment, question answering, university entrance examinations, open-ended question

1. INTRODUCTION

An educational advisory body to the Japanese government has decided that writing tests will be introduced into the new national center test for university entrance examinations, as announced in a final report [MEXT 2016] at the high school and university articulation meeting by the Ministry of Education, Culture, Sports, Science and Technology. The use of AI-based computers was proposed to stabilize the test scores efficiently. The required type of writing test is a short-answer test, where a correct answer is expected to exist. Therefore, the test is scored by judging agreement on the meaning with the correct answer.

Another type of unrequired writing test is essay writing, where a correct answer does not exist. The written answers are evaluated based on the rhetoric, the connection expressions, and the content. Many systems for evaluating essays have been developed and offered in the United States [Shermis and Burstein 2013]. The authors' group also developed the first and most well-known Japanese automated essay scoring system named Jess [Ishioka and Kameda 2006], and it is in practical use now.

While short-answer scoring involves technical difficulty, the number of characters is restricted to 120 at most from dozens of characters. Two characters in Japanese are usually equivalent to one word in English. A short-answer test is widely considered to be more authentic and reliable for measuring ability compared with a multiple-choice test. If technical problems related to the short-answer test are solved, the potential demand for its use –as well as that for the national center test– will be enormous.

A short-answer scoring system has also been developed because of its importance, though various technical problems remain unsolved. New York University (NYU) and the Educational Testing Service (ETS) developed the first automated scoring tools in this field; they evaluated the NYU online program [Vigilante 1999]. Leacock[Leacock and Chodorow 2003] reported the latest specifications of the c-rater developed by ETS. Pulman[Pulman and Sukkarieh 2005] tried to generate several sentences having the same meaning as the correct answer sentence using the natural language technique of information extraction. However, the concordance rate with human examiners was found to be small and impractical.

In 2012, a Kaggle competition for short answer scoring had been completed [Foundation 2012]. Each answer is approximately 50 words in length. The winner, Luis Tandalla [Tandalla 2012], made the best score of 0.77166 evaluated with the quadratic weighted kappa error metric [Hamner 2015], which measures the agreement between two raters (system and human). The real number of 1 shows complete agreement between raters, whereas a human benchmark produced a score of 0.90013. Automated assessment is not yet in the stage of practical application.

Therefore, we conceived of a support system for short written tests where a human rater can correct the automated score by referring to the original scores [Ishioka and Kameda 2017]. When the human rater agrees with the result of the automated score, he/she can just approve the score indicated by default and can produce the corresponding mark. We chose to leave room for human raters to overwrite it without making it a perfect automated scoring system.

Of course, some degree of quality is required for automatic scoring given as an initial value. In order to evaluate the performance of our system as a scoring engine, we decided to participate in the NTCIR-QALab 3 task [Shibuki et al. 2017], this time. A part of Tokyo University's second round of the world-history written test requires essay answers of 450–600 words containing 8 specified terms. This test may not be called a short answer test because of the quantity of writing required, but written answers need to be semantically consistent with the model solution for judgment. By putting the lexical condition on the designation, the short-answer written test could be expanded into about 500 characters. Thus, we attended this conference.

In what follows, Section 2 indicates the test issues and the model answers used in a trial examination for Tokyo University's entrance examinations. Section 3 shows the specifications of our proposed system. Section 4 presents our evaluation of the performance on five tests of social studies. Section 5 concludes with a summary.

2. TEST ISSUES USED IN A TRIAL EXAMINATION

We are assigned five issues in the subject of world history for Tokyo University's second round examinations in the past. The world history test set includes several types of written tests, and we evaluated the test issues required for the most voluminous test of 450–600 characters.

Table 1 shows the "content" asked and the "mandatory words/phrases," which are given by test writers to the examinees.

Besides these, the following are given: (1) three model answers per issue, (2) partial sentences generated from the model answers, and (3) its importance as evaluated by professional raters. However, these are omitted due to space limitations.

The allocated number of points to every test issue is 20. If mandatory words or phrases are missing, 5 points are deducted per omission. Also, if the amount of words exceeds the limit, the score is halved. These are based on our speculation about the actual scoring standards of Tokyo University's entrance examinations.

3. SPECIFICATIONS OF THE SCORING SUPPORT SYSTEM

3.1 Outline

Our system is for automated scoring and for supporting human raters. The approach functions as follows.

1. A system automatically judges each answer posed on whether or not its prepared key phrases agree with those of the model answer using the "scoring criteria" from a surface-like point of view.

2. The system gives not only a temporary score based on the criterion-based judgment but also a prediction score offered by machine learning based on the understanding of other human raters or supervised data. A certain degree of semantic meaning is also used.

3. A human rater can certify the prediction score by which a system presents this information as reference. He or she can correct this and overwrite based on his/her judgment.

To reduce the time and effort, the system precision should possess a certain degree of fitness with human ratings; more than 80% of the precision is desirable for tentative targets. At this conference, step 3 was omitted; we did not use this procedure.

The flowchart of our system is as shown in Figure 1.

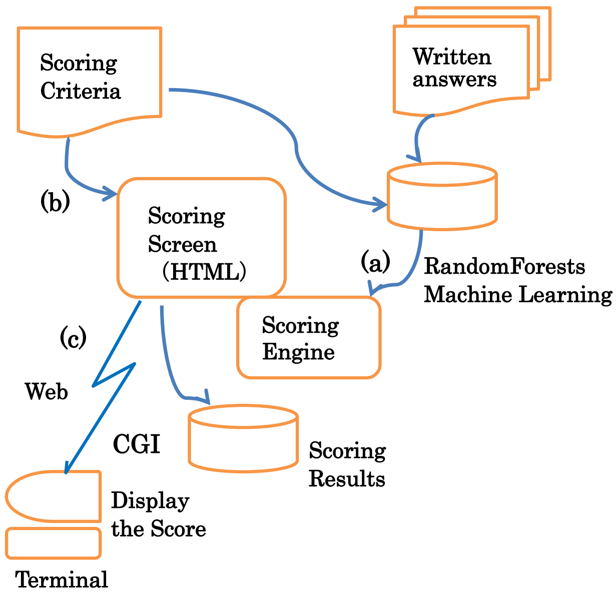

Figure 1: Flowchart of the system

(a) Before scoring, we collected a lot of score data from various human raters and performed a machine learning of "Random Forests" [Breiman 2001]. The degree of fitness with the scoring guideline is also necessary. On the basis of these learning results, we set up a scoring engine to return the scores for new answers.

(b) The system generates a scoring screen written in the Hyper Text Markup Language.

(c) A user or human rater opens a scoring screen of (b) using a web browser on his/her terminal machine. Then, a CGI program is activated. The recommended value as a result of the scoring engine of (a) is indicated here. The scoring result is stocked in a file or a database. The user repeats this mark operation.

3.2 Scoring Screen

Figure 2 shows a screen shot of our prototype system. "The answer sentence that should be scored" (in red ink) is located in the upper part of the system; the middle part has some scoring criteria such as "synonyms and permitted different transcriptions," "model or correct answers that warrant a full mark," "partial phrases that warrant partial scores," and "mandatory phrases." For the "model answer" and "partially correct phrases," the system judges the degree of fitness with the answer sentence to be scored; the

Table 1: Content and mandatory words/phrases of written test issues

Test issue # (Allotment)	Content and the mandatory words/phrases
B792W10-1, The University of Tokyo, 2001 (20 pt.)	**[content]** エジプトが文明の発祥以来，いかなる歴史的展開をとげてきたかを概観せよ。540 字 / Provide an overview of the development of Egypt since the birth of its civilization, taking into consideration both 1. the interests of those arriving in Egypt and the reasons for their advances into Egypt, and 2. the policies and actions taken by Egypt in response to these advances. Limit your answer to 270 English words or less. (540 characters in Japanese) **[mandatory words]** アクティウムの海戦, Battle of Actium/ イスラム教, Islam/ オスマン帝国, Ottoman Empire/ サラディン, Saladin/ ナイル川, Nile River/ ナセル, Nasser/ ナポレオン, Napoleon/ ムハンマド・アリー, Muhammad Ali/
C792W10-1, The University of Tokyo, 2002 (20 pt.)	**[content]** 19 世紀から 20 世紀はじめに中国からの移民が南北アメリカや東南アジアで急増した背景には，どのような事情があったと考えられるか，また海外に移住した人々が中国本国の政治的な動きにどのような影響を与えたか，これらの点について，450 字以内で述べよ。/ Explain, in 225 English words or less, what led to the sudden rise of emigration from China to North and South America and south-eastern Asia from the 19th to the early 20th centuries, described above, and what impact those who emigrated from China had on political movements within China. (450 chars in Japanese) **[mandatory words]** 植民地奴隷制の廃止, Abolition of the colonial slave system/ サトウキビ・プランテーション, Sugar cane plantation/ ゴールド・ラッシュ, Gold rush/ 海禁, Haijin/ アヘン戦争, Opium Wars/ 海峡植民地, Straits Settlements/ 利権回収運動, Rights recovery movement/ 孫文, Sun Yat-sen/
G792W10-1, The University of Tokyo, 2006 (20pt.)	**[content]** 戦争を助長したり，あるいは戦争を抑制したりする傾向が，三十年戦争，フランス革命戦争，第一次世界大戦という 3 つの時期にどのように現れたのかについて， 510 字以内で説明しなさい。/ Explain, in 255 English words or less, how these trends of supporting or suppressing war have appeared in the Thirty Years' War, the French Revolution, and World War I. (510 chars in Japanese) **[mandatory words]** ウェストファリア条約, Treaty of Westphalia/ 国際連盟, League of Nations/ 十四カ条 (の平和原則), Fourteen Points/ 『戦争と平和の法』, "On the Law of War and Peace"/ 総力戦, Total warfare/ 徴兵制, Draft system/ ナショナリズム, Nationalism/ 平和に関する布告, Decree On Peace/
L792W10-1, The University of Tokyo, 2010 (20 pt.)	**[content]** オランダおよびオランダ系の人びとの世界史における役割について，中世末から，国家をこえた統合の進みつつある現在までの展望のなかで，論述しなさい。600 字 / Describe the role of the Netherlands and the Dutch people in world history, from the late Middle Ages to the modern day, when integration is extending beyond national lines. Limit your answer to 300 English words or less. (600 chars in Japanese) **[mandatory words]**) グロティウス, Grotius/ コーヒー, coffee/ 太平洋戦争, Pacific War/ 長崎, Nagasaki/ ニューヨーク, New York/ ハプスブルク家, Habsburgs/ マーストリヒト条約, Treaty of Maastricht/ 南アフリカ戦争, South African War/
P792W10-1, The University of Tokyo, 2014 (20 pt.)	**[content]** ウィーン会議から 19 世紀末までの時期，ロシアの対外政策がユーラシア各地の国際情勢にもたらした変化について，西欧列強の対応にも注意しながら，論じなさい。/ Discuss the changes that Russian foreign policy had on the international situation throughout Eurasia from the Congress of Vienna to the end of the 19th century, noting how the western powers responded. Limit your answer to 300 English words or less. (600 chars in Japanese) **[mandatory words]** アフガニスタン, Afghanistan/ イリ地方, Ili region/ 沿海州, Primorye/ クリミア戦争, Crimean War/ トルコマンチャーイ条約, Treaty of Turkmenchay/ ベルリン会議 (1878 年), Berlin Conference (1878)/ ポーランド, Poland/ 旅順, Port Arthur/

system also judges whether or not the answer sentence includes "mandatory phrases," whether or not it is meaningfully composed, and whether or not it exceeds the character limit; if the answer must be written as a noun or noun phrase, the system judges whether or not it matches the specified "type" format. These judgments are given as either yes or no, and toggle buttons are used. A human rater reviews these judgments and revises them if necessary.

Tentative scores located in the lower part are based on the aforementioned alternative judgment. The right-hand window is to determine the final score. The initial mark is settled by which predictive probability based on the past learned results gives the maximum. The probability values are also indicated. We used only tentative scores in this conference.

When no learning data exist, that is to say, when no prescored data on the relevant test issue exist, the message to that effect is shown in the top windows: no probability and no initial mark are naturally determined. Unfortunately, we or human raters cannot revise the machine scores; we only refer to these.

3.3 Automatic screen creation from a scoring criterion file

Our system is a Web application. Thus, the screen indicated in Figure 2 is generated by HyperText Markup Language. We built the mechanism to make this HTML file automatically from a plain scoring criterion file that a computer beginner can handle.

Figure 3 is a plain original file that makes a screen like the one in Figure 2. Two or three elements are set for criteria. In order, the label, allotment of points, and correspondence are located. The tab is the delimiter.

Synonyms and different transcriptions are recorded in "syno," which appeared in "gold" as a model answer and in "part" as a partially correct phrase. "Syno" is not always limited to a definite lexical meaning. When the text has the same meaning semantically, it is also permitted. "Part" includes two types: one is possible to add to a partial point, and the other is for which a maximum is taken. If multiple same labels are found (for example, part1), we use the maximum of the points; different labels (for example, part1 and part2) can add the allotted points. "Lack" is a mandatory phrase; if no phrases exist, a point is deducted. A comma can be used for the meaning of "both." "Vol" shows the number of characters available. "None" shows a nonsense sentence, and "goji" shows a wrong word such as a kanji that does not exist. Minus points indicate the points to be deducted. At this conference, we did not use "none" and "goji" because the scoring criterion file does not include these.

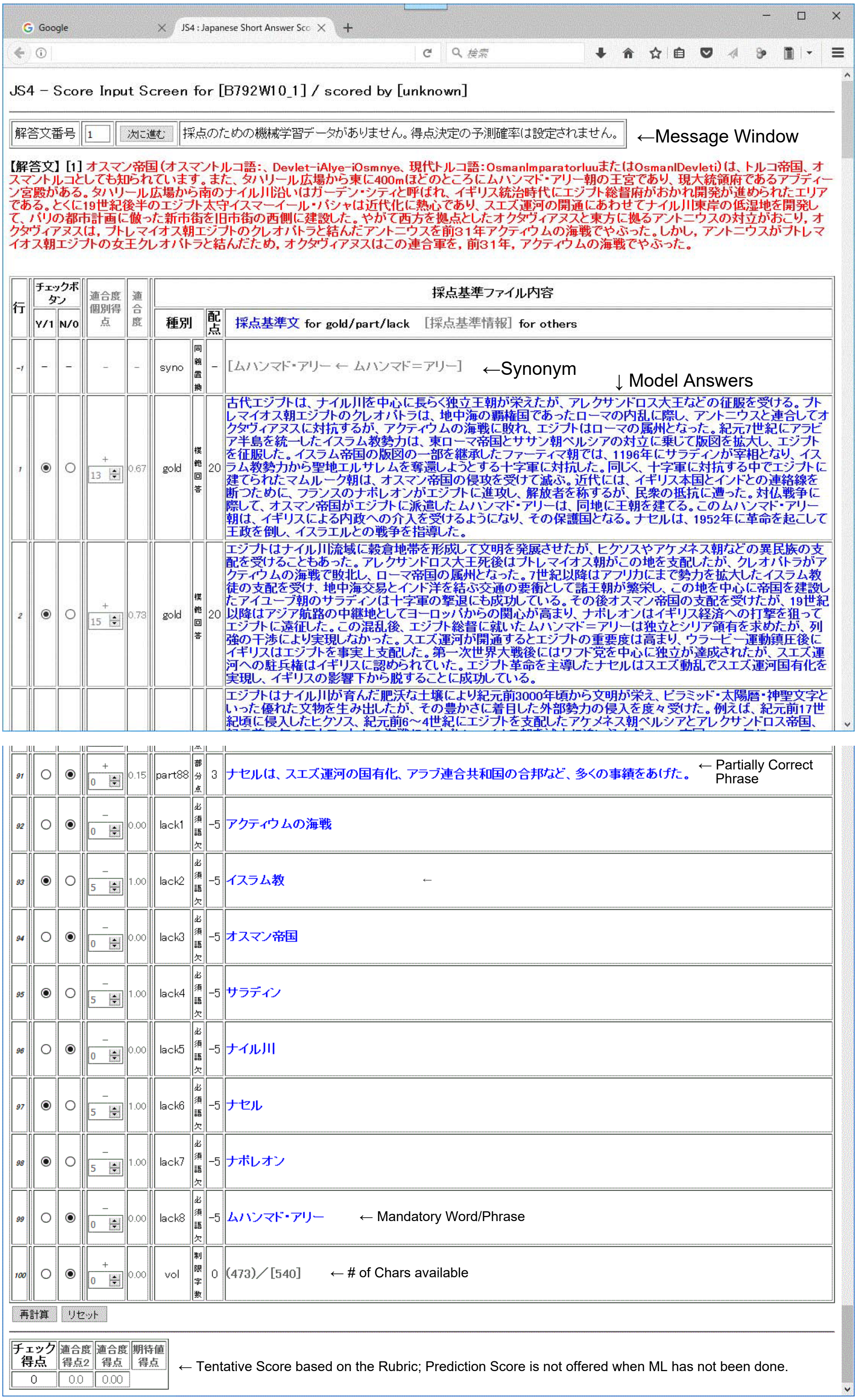

Figure 2: Short-answer scoring and support system screen (In case of world history B792W10_1)

```
syno	ムハンマド・アリー	ムハンマド＝アリー
gold	20	古代エジプトは、ナイル川を中心に長らく独立王朝が栄えたが、アレクサンドロス大王などの征服を受ける。プトレマイオス朝エジプトのクレオパトラは、地中海の覇権国であったローマの内乱に際し、アントニウスと連合してオクタヴィアヌスに対抗するが、アクティウムの海戦に敗れ、エジプトはローマの属州となった。紀元 7 世紀にアラビア半島を統一したイスラム教勢力は、東ローマ帝国とササン朝ペルシアの対立に乗じて版図を拡大し、エジプトを征服した。イスラム帝国の版図の一部を継承したファーティマ朝では、1196 年にサラディンが宰相となり、イスラム教勢力から聖地エルサレムを奪還しようとする十字軍に対抗した。同じく、十字軍に対抗する中でエジプトに建てられたマムルーク朝は、オスマン帝国の侵攻を受けて滅ぶ。近代には、イギリス本国とインドとの連絡線を断つために、フランスのナポレオンがエジプトに進攻し、解放者を称するが、民衆の抵抗に遭った。対仏戦争に際して、オスマン帝国がエジプトに派遣したムハンマド・アリーは、同地に王朝を建てる。このムハンマド・アリー朝は、イギリスによる内政への介入を受けるようになり、その保護国となる。ナセルは、1952 年に革命を起こして王政を倒し、イスラエルとの戦争を指導した。
gold	20	エジプトはナイル川流域に穀倉地帯を形成して文明を発展させたが、ヒクソスやアケメネス朝などの異民族の支配を受けることもあった。アレクサンドロス大王死後はプトレマイオス朝がこの地を支配したが、クレオパトラがアクティウムの海戦で敗北し、ローマ帝国の属州となった。7 世紀以降はアフリカにまで勢力を拡大したイスラム教徒の支配を受け、地中海交易とインド洋を結ぶ交通の要衝として諸王朝が繁栄し、この地を中心に帝国を建設したアイユーブ朝のサラディンは十字軍の撃退にも成功している。その後オスマン帝国の支配を受けたが、19 世紀以降はアジア航路の中継地としてヨーロッパからの関心が高まり、ナポレオンはイギリス経済への打撃を狙ってエジプトに遠征した。この混乱後、エジプト総督に就いたムハンマド＝アリーは独立とシリア領有を求めたが、列強の干渉により実現しなかった。スエズ運河が開通するとエジプトの重要度は高まり、ウラービー運動鎮圧後にイギリスはエジプトを事実上支配した。第一次世界大戦後にはワフド党を中心に独立が達成されたが、スエズ運河への駐兵権はイギリスに認められていた。エジプト革命を主導したナセルはスエズ動乱でスエズ運河国有化を実現し、イギリスの影響下から脱することに成功している。
gold	20	エジプトはナイル川が育んだ肥沃な土壌により紀元前 3000 年頃から文明が栄え、ピラミッド・太陽暦・神聖文字といった優れた文物を生み出したが、その豊かさに着目した外部勢力の侵入を度々受けた。例えば、紀元前 17 世紀頃に侵入したヒクソス、紀元前 6～4 世紀にエジプトを支配したアケメネス朝ペルシアとアレクサンドロス帝国、紀元前 31 年のアクティウムの海戦によりプトレマイオス朝を滅亡に追い込んだローマ帝国、1171 年にファーティマ朝を滅ぼしたトルコ系アイユーブ朝のサラディン、1517 年にマムルーク朝を倒し、エジプトを占領したオスマン帝国がその代表的事例である。しかし、エジプト側も一方的に外部の支配に屈してきたわけではない。現在のエジプト民族は 7～12 世紀までのアラブ系イスラム教勢力の統治期にアイデンティティを形成したものである。1798 年のナポレオンによるエジプト遠征の撃退後、ムハンマド・アリーはエジプトの太守となり、1831 年と 1839 年にエジプト―トルコ戦争を起こした。その後、イギリスの一時支配下に入るが、ナセルが、1952 年のエジプト革命を指導し 1954 年から大統領に就任して、アラブ民族主義の指導者として、スエズ運河の国有化、アラブ連合共和国の合邦など、多くの事績をあげた。
part1	2	古代エジプトは、ナイル川を中心に、古王国から新王国まで、長らく独立王朝が栄えた。
part2	2	古代エジプトは、アケメネス朝やアレクサンドロス大王の征服を受けた。
part3	1	プトレマイオス朝エジプトの女王クレオパトラは、ローマの内乱に際し、アントニウスと連合した。
part4	1	ローマは地中海の覇権国だった。
part5	1	クレオパトラは、ローマの内乱に際し、オクタヴィアヌスに対抗した。
part6	3	クレオパトラは、ローマの内乱に際し、アクティウムの海戦に敗れた。

～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～
～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～～
part87	1	ナセルは、アラブ民族主義の指導者だ。
part88	3	ナセルは、スエズ運河の国有化、アラブ連合共和国の合邦など、多くの事績をあげた。
lack1	-5	アクティウムの海戦
lack2	-5	イスラム教
lack3	-5	オスマン帝国
lack4	-5	サラディン
lack5	-5	ナイル川
lack6	-5	ナセル
lack7	-5	ナポレオン
lack8	-5	ムハンマド・アリー
vol	/2	540
```

Figure 3: Scoring criterion file (labels, allotment of points, and correspondences are tab delimited.)

We use "fitness" as the degree of the relationship between the written answer and "model answer" designated in "gold" or "partially correct phrases" in "part." We define this as the harmonic mean of two kinds of relationships: one is the degree of the reference during the sentence keywords from the viewpoint of a written answer; the other is that from a model answer. These relationships are just like precision and recall often used in information retrieval, e.g., a Google search. This harmonic mean or "fitness" is called an F-measure taking a float number from 0 to 1. Our system rounds this to either 0 or 1 as a toggle button occurrence, and it shows a non-rounded value as a reference for the user.

If the scores by professional human raters are given, a mechanical learning score is presented. Unfortunately, we did not obtain human ratings in advance.

3.4 How to make partially correct phrases

The task of NTCIR provides partial phases, which are created automatically from the correct answers, and gives scores ranging from 1 to 3 by professional reviewers. We call them nugget sentences.

The partially correct answers are given in advance at actual scoring, but they are not given to us. Thus, we substitute the nugget sentences as the partially correct answers.

The allotted points should be the median of three professional evaluations. The total of the partial points may exceed the full score of 20 points, but it ends with the maximum limit.

3.5 Deducted points due to exceeding of character limit

For short answers limited to 30–50 characters, the scores of the answers exceeding the limit number is usually zero. However, in response to about 500 characters like this task, a zero is not appropriate.

Therefore, in the case of exceeding the limit, a specification that halves the score was implemented. We used "`vol /2 500`" instead of "`vol -20 500`" on a scoring criterion file, which shows that system should halve the score instead of the full score of 20 points.

3.6 Japanese sentence processing

Unlike Western languages, Japanese is a sticky language that leaves no blank space between words. Therefore, the performance of the morphological analyzer is more important than that of Western languages. Adequate dictionaries are also indispensable. Wikipedia's entry word dictionary includes a textbook that is suitable for social studies examinations. Our approach is applicable to Western languages as long as we can handle grammatical processing according to the language.

4. PERFORMANCE EVALUATION

4.1 Evaluation Criterion

The task office gave experts' evaluation of each of the four answers prepared by participants on five issues. The experts scored according to the grading criteria they created. This scoring standard was not disclosed to participants in advance.

This task measures the degree of agreement between the participant's evaluation and a professional's. The task office uses two rank correlation coefficients: Spearman's ρ and Kendall's τ.

Table 2: Predicted value, the mean of differences from professional scores, and the mean of squared differences

Issue	predicted values $x; n = 4$	$\Sigma x/n$	$\Sigma x^2/n$	all predicted values
B	0,0,0,2	0.50	1.00	0×11, 2, 8, 14, 15×4
C	0,0,0,0	0.00	0.00	0×13, 3, 9, 12×2, 18×2
G	0,0,0,3	0.75	2.25	0×10, 2, 3, 7, 8, 9, 19×4
L	5,0,0,4	2.25	10.3	0×9, 4, 5, 8, 9, 11, 12, 14×2, 19×2
P	0,4,0,4.5	2.13	9.06	0×8, 4, 4.5×6, 5×2, 7.5, 9

4.2 Evaluation Results

Surprisingly, among the professional evaluations for the 20 answers, four responses for each of five issues were all zero. For this reason, because the standard deviation of professional evaluation was zero, both of the indicators prepared by the task office could not be calculated.

The purpose of this task is how to predict professional evaluations well. Thus, we thought a good solution is evaluating how close the scores presented are to zero.

Table 2 shows our predicted values, the mean of differences from professional evaluation, and the mean of squared differences. For reference, all our predicted values are added including the remaining data that the professionals did not score.

Each response was scored with 20 points as a full mark, and the range was as follows. 0–15 (for B), 0–18 (for C), 0–19 (for G and L), 0–9 (for P). Some answers are given high scores, and the range of the score is wide. Under this situation, the answers evaluated by professional raters produced sufficiently close to zero ratings. Our method produces reasonable scores. The differences between the professionals and ours are within 3 points for 16 of 20 data.

The evaluation criteria based on the residuals with the correct score are the most appropriate, but the evaluation is not an index prepared by the task office. Therefore, we do not explicitly show the other teams' results using this index, but we nevertheless determined that our method is the best.

4.3 Some comments on evaluation indicators

Because the professional evaluations all became zero, the task office presented the values of two correlation coefficients based on the new index that applied the deduction point not depending on the missing words from the part with the additional point.

This is inappropriate for the following three reasons.

1. They did not measure the degree of agreement with the professional rater. The original purpose of the task has not been achieved.

   Our team correctly answered all of the four responses for issue C. Nevertheless, the two indices of the task office gave it NAs. The indicator prepared by the task office is certainly important. However, it is one of the factors that affects the score prediction.

2. Calculating correlation with only 4 data has almost no

Table 3: Predicted values by participants

Issue	Forest1	Forest2	tmkff
B	31,39,62,44	2,3,4,3	0,0,0,2
C	35,39,47,42	0,0,1,1	0,0,0,0
G	18,24,31,41	2,1,1,4	0,0,0,3
L	38,43,59,48	1,4,2,1	5,0,0,4
P	45,53,67,60	5,4,1,7	0,4,0,4.5

meaning.

The bivariate correlation coefficient between x and y is calculated based on the deviation from the average of each of the two variables. In the rank correlation coefficient, the two deviations from the average ranks are taken into account. Therefore, the degree of freedom of distribution associated with the test statistics in this case is only 2, which is 4 minus 2. Statistics based on these few data have little meaning and may lead to a wrong conclusion.

Table 3 shows three participants' predicted values, which are the raw data for five issues. Forest1 seems to have adopted 100 allotment points scoring. Forest2 might suppose 20 points as full marks like we did (tmkff).

Even without using difficult indicators, we can see that our team's (tmkff's) estimates are closest to zero of the correct answer. This is evidence that an index using a correlation coefficient is inappropriate.

3. Evaluations were made based on the indices created after the task execution.

   The new index presented by the task office cannot be calculated from the numerical values associated with XML tag names, e.g., `ans_limits, ans_len, total_0, minus_total, plus_total`. From the points of fairness and accountability, this practice is not appropriate for a competition.

5. CONCLUSION

Evaluating the performance of our system was difficult because of the surprising results that showed all professional evaluations had issues with scoring zero. However, we are convinced that our system can show a certain degree of validity because it returned a score close to zero as being professionally evaluated, while a sufficiently wide range of scores were presented for other answers. Hereafter, we will endeavor to improve the system performance by evaluating unscored answers.

Our system can provide another predicted score by applying machine learning of random forests if sufficient professional scores are given. In such a case, this system can reveal some factors influencing the final forecast score. We can also take into consideration similarities to essay prompted sentences. If you are interested in the scoring, please refer to [Ishioka and Kameda 2017].

Acknowledgments

This project was supported by JSPS KAKENHI Grant Number 26350357 and 17H01843.

6. REFERENCES

[Breiman 2001] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[Foundation 2012] The Hewlett Foundation. 2012. *Short Answer Scoring*. Automated Student Assessment Prize, Phase Two. https://www.kaggle.com

[Hamner 2015] Ben Hamner. 2015. *Package 'Metrics'*. Evaluation metrics for machine learning. https://github.com/benhamner/Metrics/tree/master/R

[Ishioka and Kameda 2006] Tsunenori Ishioka and Msayuki Kameda. 2006. Automated Japanese essay scoring system based on articles written by experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (Coling-ACL 2006)*. Association for Computational Linguistics, 233–240. https://doi.org/10.3115/1220175.1220205

[Ishioka and Kameda 2017] Tsunenori Ishioka and Masayuki Kameda. 2017. Overwritable automated Japanese short-answer scoring and support system. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*. 50–56. https://doi.org/10.1145/3106426.3106513

[Leacock and Chodorow 2003] Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities* 37, 4 (2003), 389–405. https://doi.org/10.1023/A:1025779619903

[MEXT 2016] MEXT. 2016. *Publishing the final report of high school and university articulation meeting.* Ministry of Education, Culture, Sports, Science and Technology in Japan. http://www.mext.go.jp

[Pulman and Sukkarieh 2005] Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic short answer marking. In *EdAppsNLP 05 Proceedings of the second workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 9–16. http://dl.acm.org/citation.cfm?id=1609831

[Shermis and Burstein 2013] Mark D. Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation*. Routledge.

[Shibuki et al. 2017] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Yoshinobu Kano, Teruko Mitamura, and Tatsunori Mori andNoriko Kando. 2017. Overview of the NTCIR-13 QA Lab-3 Task. In *Proceedings of NTCIR-13*.

[Tandalla 2012] Luis Tandalla. 2012. *Scoring Short Answer Essays*. The Hewlett Foundation: Short Answer Scoring. https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/TechnicalMethodsPaper.pdf

[Vigilante 1999] Richard Vigilante. 1999. Online Computer Scoring of Constructed-Response Questions. *Journal of Information Technology* 1, 2 (1999), 57–62.