Constructing Open Essay-Writing Data and an Automatic Essay-Scoring System in Japanese

Masayuki Ohno*, Koichi Takeuchi†, Kota Motojin‡,
Masahiro Taguchi§, Yoshihiko Inada¶, Masaya Iizuka‖, Tatsuhiko Abo**and Hitoshi Ueda**
*Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan
Email:pw2z9792@s.okayama-u.ac.jp
†Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan
Email:koichi@cl.cs.okayama-u.ac.jp
‡Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan
Email: pm9n6cei@s.okayama-u.ac.jp
§Graduate School of Humanities and Social Science, Okayama University, Okayama, Japan
¶Graduate School of Education, Okayama University, Okayama, Japan
‖Institute for Education and Student Services, Okayama University, Okayama, Japan
**Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan

***Abstract*—In this paper, we describe an on-going study of developing an automatic essay-scoring system in Japanese. Several Japanese essay-scoring systems have been developed; however, they are evaluated based on only closed essay data; thus, it is difficult to compare their performances. Thus, we first propose methods of constructing baseline data of essay-writing tests in Japanese, which can be used for evaluating essay-scoring systems, and describe the framework of essay-writing tests and manual evaluation procedure. We are currently developing an essay-scoring system using word-based similarity and plan to evaluate the system using the baseline data constructed with our proposed methods. Experimental results of the evaluation indicate that the proposed methods work well with a maximal correlation coefficient of 0.629.**

***Keywords*-Automatic scoring of answers of essay-writing tests; baseline data of essay-writing tests; word-based similarity;**

I. Introduction

In this paper, we propose methods of constructing baseline essay data that can be used to evaluate automatic essay-scoring systems. The aim of this on-going research is to reduce the burden of manual scoring and variation in different human scorers by developing an automatic essay-scoring system.

There are two main types of essay-writing-tests. The first type includes essay-writing tests that are assumed to have certain correct sentences as a gold-standard answer text. The second includes essay-writing tests that cannot be assumed to have correct sentences as answers. The former is called *Short-Answer Type*, and the latter is called *Essay Type* [6].

Machine-learning-based approaches have been applied for Short-Answer-Type tests [8][6]; however, in Essay-Type tests, it is difficult to prepare model answers for students to write their own ideas. Even if a model answer is prepared, since the answers contain 200 to 800 of characters, textual entailment technology should be used to measure the degree of agreement between a model answer and students' essays. Many English-essay-scoring systems have been developed [9][5] and several systems (i.e., e-rator [1] and IntelliMetric [2]) have been used in scoring actual essay-writing tests. Ishioka developed [10], a Japanese essay-scoring system (Jess) that uses a statistical regression model and linguistic patterns for scoring essays in Essay-Type tests.

One of the difficulties in developing an automatic scoring system for Essay-Type tests is that there is no common essay data that can be used. Thus, we first developed methods of constructing essay data by carrying out practice essay-writing tests while obtaining copyright transfer agreements from examinees.

We are constructing four evaluation modules for Essay-Type tests we are currently developing. The modules evaluate essays based on 1) comprehensiveness, 2) logical consistency, 3) validity, and 4) spelling and grammar. We constructed the comprehensiveness module using word-based similarity and applied the constructed essay data to this module. Experimental results of the evaluation show that the word-based similarity approach is promising.

II. Constructing Japanese essay data for essay-writing tests

We had students take essay-writing practice tests. In the following subsections, we describe the framework of the data collection by carrying out these essay-writing tests, manual scoring framework, and content of the current data set[1].

[1]In this on-going research project, we are planning to carry out several types of essay-writing tests and collect the answers over the next three years.

A. Framework of collecting essay data by carrying out essay-writing test

The students took two lectures and wrote essays for three questions for each lecture. To evaluate the students' spelling (and kanji (Chinese-based characters used in Japanese writing)), the students wrote their essays on paper then manually input the essays as electronic text data.

The two lectures given were entitled "Light and shadow in globalization" (Lecture 1) and "Structure of natural science and scientific education" (Lecture 2). In Lecture 1, the lecture slides and three essay questions were printed and distributed to the students. In Lecture 2, however, only the questions were distributed. This was to obtain a wide variety of essays by changing the difficulty of the task.

B. Framework of manual evaluation score

The collected essays require manual evaluation scores. As previous studies [9][4][5] have been pointed out, there are no standard evaluation criteria; thus, the criteria should be defined according to what aspect of student performance should be evaluated. Thus, we defined the following four criteria from the discussions in our research team: (1) Comprehensiveness of questions, (2) logical consistency (evaluate whether the essay is logically written), (3) validity (whether the content of the statement is reasonable and persuasive), and (4) spelling and grammar. A score of 1 to 5 is given to these four criteria (the larger the number, the better), and the sum of these scores express the final score of the essay. The automatic essay-scoring system we are developing will be designed to evaluate essays based on these four evaluation criteria.

C. Current status of essay data

Essay data consist of two parts, i.e., base reference data and student essays. The base reference data consists of the lecture titles, reference texts of the lectures (less than 2000 characters), and question texts. These data are used in essay-writing systems. There are two types of student-essay data: text and image. Image data are the original answer sheets written by students, and image data will be used in the study of OCR error correction in the future. The details of the student essays collected from the practice essay-writing tests are given in Table I.

Table I
DETAILS OF ESSAY-WRITING TESTS AND COLLECTED ESSAYS

	Lecture1	Lecture 2
Title	Light and shadow in globalization	Structure of natural science and scientific education
Max. number of characters for questions	(1) 300 (2) 250 (3) 300	(1) 100 (2) 400 (3) 500 to 800
# of Students # of Essays	328 984	327 981

The total number of students that took the practice tests was 328. The 328 students answered the three questions in Lecture 1, but in Lecture 2, 327 students answered the questions (due to one person being absent). Thus, we collected 1965 essays. Each essay contained an execution date and student ID given in the practice tests. For manually scoring the student essays, we constructed a rubric that defines how we score the essays based on the established four evaluation criteria for each question. Using the rubric, we have finished scoring the essays of 160 students for all questions for Lecture 1. In the next section, we discuss the evaluation of several essay-scoring methods on the basis of the essay data.

III. Our Automatic essay-scoring system under development

There are two scoring approaches for automatic essay-scoring systems: machine-learning, which uses student essays with human-annotated scores [8], and statistical scoring without human-annotated scores [5]. Since the former requires human-annotated essays for the target essay-writing tests, we take the latter approach for evaluating target essays with some reference text data.

As mentioned in the previous section, we defined four evaluation criteria for essay scoring. Thus, for our automatic essay-scoring system we constructed four evaluation modules, comprehensiveness, logical consistency, validity, spelling and grammar. Figure 1 shows the system configuration diagram of our automatic essay-scoring system with the four evaluation modules.

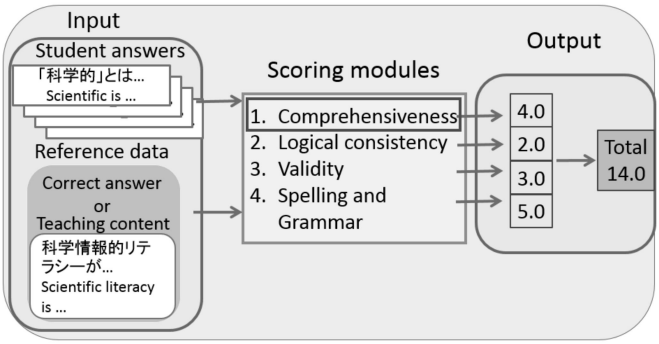

Figure 1. Overview of automatic essay-scoring system

Each module outputs an independent score from 0 to 5.0, and our system outputs the sum of the all module scores as the final score. As reference data, the modules use lecture content text, questions, and examples of correct answer essays if available. We are constructing the comprehension evaluation module and validity evaluation module. The policy and method of comprehensiveness evaluation module are described below.

IV. Comprehensiveness evaluation module

The comprehensiveness module is used to evaluate how students understand the lectures and questions in writing their essays. We assume that this module gives a high score to answer essays with content related to the reference texts, i.e., lecture content and questions.

Thus, we developed several similarity-evaluation methods to compare the lecture content and answer essays. We first apply our morpheme-based content-word-matching evaluation method then our morpheme-based N-gram similarity evaluation methods which are based on BLEU [7]. In a question-type essay-writing test, important keywords that should exist in answer essays can be assumed. Thus, we apply our important-keyword-enhancement method in scoring essays and to above two methods.

A. Content morpheme matching

Our morpheme-based content-word-matching evaluation method is used to count the number of content words that match the lecture text and answer essays. The content words are the nouns, verbs, and adjectives that are outputted using the Japanese morphological analyzer MeCab[2].

In the following formulas, function $P(a, b)$ returns 1 if the input morphemes a and b are the same and both are content words (Equation (1)). The final score is the total matching number of the content words between the lecture text (A) and an answer text (B) (Equation (2)).

$$P(a,b) = \begin{cases} 1 & (a = b) \\ 0 & (a \neq b) \end{cases} \tag{1}$$

$$score\_match(A,B) = \sum_{a \in A, b \in B} P(a,b) \tag{2}$$

Figure 2 shows examples of content morphemes matching the lecture text and an answer essay. The symbol "/" indicates morpheme boundaries. The matching is done on the morphemes of the base form for inflectional morphemes. In both sentences, the content morphemes "グローバリゼーション (globalization)", "格差 (gap)" exist; thus, the score of this comprehensiveness module is 2.

In the experimental results section, we discuss the evaluation of these matching scores by comparing them to the human-annotated scores for the comprehensiveness evaluation module using the correlation coefficients. Since the above scores are simple matching numbers of content morphemes, they are not limited in 5.0. To evaluate the performance of our morpheme-based content-word-matching evaluation method, we do not need to normalize the scores, but we do normalize them to 5.0 by dividing the maximal scores in the essays when we apply this method for comprehensiveness module.

[2]http://taku910.github.io/mecab/.

lect. グローバリゼーション/に/伴い/世界/的/格差/は/...
Globalization/Dat/with/world/wide/gap/is/...
essay グローバリゼーション/によって/、/先進/国/と/
Globalization/by/,/developed/country/and/
発展/途上/国/の/所得/格差/は/...
developed/country/'s/income/gap/is/...

Figure 2. Example of content morphemes matching between lecture text and answer essay

B. N-gram similarity based on BLEU

We apply our morpheme N-gram methods to evaluate essays with more precise matching between texts. Several approaches, e.g., BLEU [7] and RIBES [3], have been proposed regarding evaluating automatic machine-translation systems. Both approaches are based on N-gram similarities; but RIBES measures the morphological N-gram similarity for each sentence, while BLEU measures the morpheme N-gram similarity between texts. In this essay evaluation task, we cannot assume a similarity that measures texts sentence by sentence; thus, we apply the BLEU-based similarity to the essay evaluation task.

With the original BLEU, a penalty is assumed when the evaluation sentence is too short with respect to the given correct sentence; thus, this assumption is not suitable for the essay evaluation task. We simply apply morpheme N-gram similarity to evaluate text similarity by varying the length N from 1 to 4.

Each N-gram similarity score $score\_n\_gram(A, B)$ between texts A and B is expressed using Equation (3).

$$score\_n\_gram(A,B) = \begin{cases} calbleu(1) \\ calbleu(2) \\ calbleu(3) \\ calbleu(4) \end{cases} \tag{3}$$

Function $calbleu(n)$ denotes an N-gram score outputted by BLEU. When summing all the N-gram scores, we give manually defined weights to enhance the long N-gram matching scores because a longer N-gram matching can be regarded as the comparing texts would be similar.

In the following formulas, $score\_sumN(A, B)$ denotes the sum of the N-grams in Equation (4) and $score\_sumN_w(A, B)$ denotes the weighted sum in Equation (5).

$$score\_sumN(A,B) = \sum_{i=1}^{4} calbleu(n) \tag{4}$$

$$score\_sumN_w(A,B) = \sum_{n=1}^{4} calbleu(n) \times w_n \tag{5}$$

In these formulas, we defined the coefficients w_1, w_2, w_3, w_4 as 1.0, 2.0, 3.0, 4.0 in the experiments. With our mor-

pheme N-gram matching methods, we did not assume to skip functional morphemes such as particles and auxiliary verbs; thus, our N-grams contain some functional morphemes. For example, in the sentences shown in Figure 2, the morphemes of 1-gram matching would be "グローバリゼーション (globalization)", "格差 (gap)", and "は (is)", the morphemes of 2-gram matching would be "格差/は (gap/is)", and both 3-gram and 4-gram do not match the sentences.

C. Enhancing important keywords

A previous study [6] revealed that designating important keywords is effective in scoring essays in the rubric-based scoring approach.

In Equation (6), the $score\_keyword(A, B)$ denotes the similarity score by taking into account important keywords in evaluating two texts, A and B. The $score(A, B)$ indicates the other similarity evaluation scores between A and B defined in Sections IV-A and IV-B.

$$score\_keyword(A, B) = \begin{cases} R \times score(A, B) & (b \in K) \\ score(A, B) & (b \notin K) \end{cases} \quad (6)$$

Where K denotes a set of keywords, and a and b are morphemes in texts A and B, respectively. We assume that text B denotes answer essays; thus, $b \in K$ indicates that one of the morphemes in text B is a keyword. The symbol R denotes the enhancing weight value to multiply the base score. In Section V, we set R to 2.0.

We also propose combining our methods, i.e., content morpheme matching, N-gram similarity, and important-keyword-enhancement methods.

V. Evaluation and Experimental Results

As described in Section II-C, we had 160 essays with manually evaluated scores for questions 1, 2 and 3 in Lecture 1. In this section, we discuss the evaluation of the proposed methods for the comprehensiveness module; thus, we apply the correlation coefficient between the human-annotated scores evaluating for comprehensiveness and the output scores of the proposed methods.

We now describe the details of the questions in Lecture 1.

Lecture: Light and shadow of globalization

- Q1: How has globalization changed the income disparity throughout the world or each country? Also, why do you think the phenomenon of income disparity expansion or reduction has occurred? Answer this question within 300 characters.
- Q2: What role did multinational companies play in the progress of globalization? Answer this question with an example of a multinational company within 250 characters.
- Q3: How has globalization of culture influenced our lives? Also, how do you evaluate it? Answer this question with a concrete example within 300 characters.

In the above three questions, we defined an important keyword to Question 1 because the questioner designated an important keyword in the rubric. For questions 2 and 3, however, we did not set important keywords because the questions contain free discussions, and we could not find any keyword in the rubric.

A. Results of content morpheme matching

Our morpheme-based content-word-matching evaluation method outputs the scores by counting the content morphemes in both lecture texts and answer essays. Table II lists the evaluation results using the correlation coefficients between the outputs of content-morpheme-matching and comprehensiveness scores.

Table II
Correlation coefficients between content-morpheme-matching and manual comprehensiveness scores

Question	Morpheme similarity
1	0.363
2	0.441
3	**0.629**

The results in Table II show that the correlation coefficient for Question 1 is the lowest, while that of Question 3 is the highest. This does not seem to fit the type of questions because Question 1 should be a type of clear inquiry that asks about partial content of the lecture. Thus, well-written essays might be long enough to have some of the same phrases and expressions in the lecture text. In Question 3, however, the questioner expected the students to describe a wide range of discussions on globalization; thus, it can be difficult to assume a correct answer essay.

The reason of the low correlation coefficient in Question 1 can be considered due to the simple content morpheme matching method; this indicates that this simple method cannot be used to correctly evaluate paraphrasing or entailments between lecture texts and answer essays.

The cause for the high correlation coefficient in Question 3 is related to student performance. Most of the students did not give an original discussion but gave an example explained in the lecture presentation. Also, the human annotator gave a high score in evaluating the comprehensiveness of the answer essays if the essays contained a correct example of the culture of the globalization, i.e., the example was the same as that explained in the lecture texts. The reason these essays were given high scores was because they were not the best answers but were not wrong. Thus, most of these essays obtained a comprehensiveness score of 4 (maximum is 5). Our morpheme-based content-word-matching evaluation method worked well for Question 3 because proper nouns

and characteristic nouns in the examples in the answer essays matched those in the examples in the lecture texts.

B. Results of morpheme N-gram similarity

We applied our morpheme N-gram similarity evaluation methods to compare the lecture content and answer essays, and the output scores were evaluated with correlation coefficients and compared with the manually annotated comprehensiveness scores. Table III lists the experimental results of using various N-gram similarities, i.e., from 1-gram to 4-gram, the sum of the N-grams, and sum of weighted N-grams.

Table III
CORRELATION COEFFICIENTS OF MORPHEME N-GRAM SIMILARITIES

Q.	1-gram	2-gram	3-gram	4-gram	sumN	sumNw
1	0.072	0.055	0.053	0.086	**0.319**	0.070
2	0.246	0.239	−0.147	0.123	**0.249**	−0.190
3	0.094	−0.009	0.108	0.038	**0.594**	−0.060

The experimental results show that the sum of N-grams performs the best among the single N-gram similarities and weighted sum of N-grams in correlation coefficients for all questions. The correlation coefficients of each N-gram similarity were quite low; this can be considered as the functional morphemes contained in N-grams mot matching between the lecture texts and answer essays. We also found that the weighted N-gram was much worse than the sum of N-grams in correlation coefficients. This can be considered as most matched long N-grams might be wrong because of certain functional morphemes that do not relate to the similarity of the content.

C. Results of enhancing important keywords

Question 1 can be assumed to have the important keyword "Gini" in the Gini coefficient; thus, we separately apply our important-keyword-enhancing method to our morpheme-based content-word-matching evaluation method and morpheme-based N-gram similarity evaluation methods. These correlation-coefficients results are listed in Tables IV and V.

Table IV
CORRELATION COEFFICIENTS OF CONTENT MORPHEME MATCHING BY APPLYING OUR IMPORTANT-KEYWORD-ENHANCING METHOD

Question	Content matching with enhancing keywords
1	0.339

Table V
CORRELATION WITH MORPHEME N-GRAM SIMILARITIES WHEN IMPORTANT KEYWORDS ARE SET

Q.	1-gram	2-gram	3-gram	4-gram	sumN	sumNw
1	0.211	0.207	0.175	0.167	0.350	0.210

Table IV shows that our important-keyword-enhancing method does not work well in content morpheme matching compared to the results in Table II. This can be considered as the morpheme-based content-word-matching evaluation method having already evaluated the keywords; thus, the weighted scores did not clearly affect the performance. The experimental results of enhancing keywords with morpheme N-gram similarity were improved for all N-gram methods. This indicates that content words are effective in evaluating the similarity between two texts.

D. Combinations of proposed methods

In this section we discuss applying combinations of the proposed methods to evaluate the comprehensiveness of essays. First, we combined the morpheme-based content-word-matching evaluation method and morpheme-based N-gram similarity evaluation methods. The method of the combination is the sum of the scores of these two methods. Table VI shows the results of their correlation coefficients.

Table VI
CORRELATION COEFFICIENTS OF COMBINED MORPHEME-BASED CONTENT-WORD-MATCHING EVALUATION METHOD AND MORPHEME-BASED N-GRAM SIMILARITY EVALUATION METHODS

Question	Cont + sumN	Cont + sumNw
1	0.340	0.314
2	0.361	0.271
3	0.615	0.595

Table VI shows that the morpheme-based content-word-matching evaluation method and morpheme-based N-gram similarity evaluation methods performed better than only the weighted N-gram method. Compared with Table V, we found that this combination method performed better than the individual N-gram and weighted N-gram methods. Both methods, however, did not outperform the morpheme-based content-word-matching evaluation method, as shown in Table II.

We then combined the morpheme-based content-word-matching evaluation, important-keyword-enhancement, and morpheme-based N-gram similarity evaluation methods. Table VII shows the correlation coefficients for Question 1.

Table VII
CORRELATION COEFFICIENTS OF COMBINATION OF ALL METHODS

Question	important keyword	Cont + sumN
1	Matching of content words	0.339
1	Morpheme N-gram similarity	**0.386**
1	Both	0.327

Table VII shows that the combination all methods performed better than any of the individual methods. Comparing the results from Table II with Table VI, this combination is the best for Question 1.

VI. Discussion

We applied our morpheme-based content-word-matching evaluation method, morpheme-based N-gram similarity evaluation methods, and important-keyword-enhancement methods and combinations of these methods to essay evaluation. According to the correlation-coefficient results discussed in Section V, the morpheme-based content-word-matching evaluation outperformed the other single methods for the three questions. Thus, the morpheme-based content-word-matching evaluation method can be considered effective in evaluating the similarity between lecture texts and answer essays.

Regarding the combinations of these methods, the combination of the morpheme-based content-word-matching evaluation method, morpheme-based N-gram similarity evaluation methods, and combined with important-keyword-enhancement method performed the best for Question 1, while the simple addition of the important-keyword-enhancement method to the morpheme-based content-word-matching evaluation method decreased the correlation-coefficient score. This indicates that we can improve the performance of the comprehensiveness module if we use the appropriate combination of the proposed methods.

We used Jess [10] to evaluate an essay with the scores of three evaluation criteria, i.e., rhetoric, logical, and content. Thus, we applied Jess to evaluate the answer essays of Question 1 for Lecture 1. The correlation-coefficient results regarding comprehensiveness and the total scores were negative scores due to Jess does not taking into account any reference data. Thus, using reference data, is quite effective for evaluating essays.

Since all the above results are from Lecture 1, we are currently adding human annotated scores from Lecture 2, which is different from Lecture 1. Thus, we will evaluate the proposed methods in answer essays of different types of questions for future work.

VII. Conclusion and future work

We described the construction of essay data that can be used for automatic scoring and experimental results of the proposed methods. Essay data contain reference data, i.e., lecture texts, question texts, answer essays, manually annotated scores, and rubrics. The scores are given based on the four evaluation criteria we defined. We have been constructing two types of essay data with about 320 essays. We plan to construct more essay data by conducting trial essay-writing tests over the next two years. We also plan to distribute the essay data for research purposes after data collection is completed.

We implemented a comprehensiveness evaluation module for essays, which is one of the evaluation modules of our essay-scoring system under development. We proposed morpheme-based content-word-matching evaluation, morpheme-based N-gram similarity evaluation, and important-keyword-enhancing methods and combinations of these methods for essay evaluation. The experimental results using correlation coefficients indicate that the morpheme-based content-word-matching evaluation method is effective in evaluating the comprehensiveness of answer essays.

We are currently adding human-annotated scores from Lecture 2, which is different from Lecture 1. Thus, we will evaluate the proposed methods regarding answer essays of different types of questions for future work.

VIII. Acknowledgments

We appreciate Prof. Tsunenori Ishioka for providing many useful comments for this research.

References

[1] Y. Attali and J. Burstein. *Automated essay scoring with e-rater v. 2.0 (ETS RR-04-45)*. Educational Testing Service, 2005. Princeton, NJ.

[2] S. Elliot. IntelliMetric from here to validity. In M. Shermis and J. Bursten, editors, *Automated essay scoring: A cros disciplinary perspective*, pp. 71–86. Lawrence Erlbaum Associates, Hillsdale, NJ, 2003.

[3] T. Hirao, H. Isozaki, Kevin Duh, K. Sudo, H. Tsukamoto, M. Nagata. RIBES: Automatic Evaluation of Machine Translation Based on Rank Correlation. In *the Seveneenth Annual Meeting of the Association for Natural Language Processing*, pp. 1115–1118, 2011. (in Japanese).

[4] T. Ishikawa. *What is the "good document": thoughts of essay for entrance examination*. Chilumashobo, 2010. (in Japanese).

[5] T. Ishioka. Computer-Based Writing Tests. *IEICE*, Vol. 99, No. 10, pp. 1005–1011, 2016. (in Japanese).

[6] T. Ishioka, M. Kameda, and D. Liu. AI-based Japanese Short-answer Scoring and Support System. In *IEICE Technocal Report, NLC*, pp. 87–92, 2016. (in Japanese).

[7] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[8] R. Terada and K. Kubo and T. Shibata and S. Kurohashi and T. Ohkubo. Automatic Essay Scoring Using Neural Network. In *the Twenty-second Annual Meeting of the Association for Natural Language Processing*, pp. 370–373, 2016. (in Japanese).

[9] E. V. Steedman, M. Tillema, G. Rijlaarsdam, and H. van den Bergh, editors. *Measuring Writing Recent Insights into Theory, Methodology and Practices (Studies in Writing)*. Brill Academic Pub, 2012.

[10] T. Ishioka nad M. Kameda. Automated Japanese essay scoring system based on articles written by experts. In *the 21st International Conference on Computational Linguistics*, pp. 233–240, 2006.