Automatic Scoring Method for Open Answer Task in the SJ-CAT Speaking Test Considering Utterance Difficulty Level

Hao Lu*, Takeshi Yamada*, Shingo Imai*, Takahiro Shinozaki†, Ryuichi Nisimura‡,
Kenkichi Ishizuka§, Shoji Makino*, and Nobuhiko Kitawaki*
*University of Tsukuba, Ibaraki, Japan
†Tokyo Institute of Technology, Tokyo, Japan
‡Wakayama University, Wakayama, Japan
§DWANGO Co.,Ltd, Tokyo, Japan

***Abstract*—In this paper, we propose an automatic scoring method for the open answer task of the Japanese speaking test SJ-CAT. The proposed method first extracts a set of features from an input answer utterance and then estimates a vocabulary richness score by human raters, which ranges from 0 to 4, by employing SVR (support vector regression). We devised a novel set of features, namely text statistics weighted by word reliability, to assess the abundance of vocabulary and expression, and degree of word relevance based on the hierarchical distance in a thesaurus to evaluate the suitability of vocabulary. We confirmed experimentally that the proposed method provides good estimates of the human richness score, with a correlation coefficient of 0.92 and an RMSE (root mean square error) of 0.56. We also showed that the proposed method is relatively robust to differences among examinees and among questions used for training and testing.**

I. Introduction

There are currently more than 130,000 international students in Japan as well as around 4 million foreigners studying the Japanese language [1]. The demand for the assessment of Japanese language proficiency is therefore increasing rapidly. The Japanese Computerized Adaptive Test (J-CAT) has been developed as a free online proficiency test for Japanese language learners and is widely used around the world [2]. The J-CAT consists of four test sections designed to assess listening, vocabulary, grammar and reading ability, respectively. J-CAT does not yet include an assessment of speaking ability. Since speaking ability is an important part of any comprehensive assessment of Japanese language proficiency, we are now developing a test for assessing this ability called the Speaking Japanese Computerized Adaptive Test (SJ-CAT) [3], [4], [5].

The SJ-CAT consists of four test tasks, namely reading a sentence aloud, multiple-choice, sentence generation, and open answer. As its name suggests, the reading a sentence aloud task involves reading a given sentence aloud. The multiple-choice task involves selecting a correct answer from given candidates and reading it aloud. The sentence generation task consists of constructing a short sentence in response to a question and speaking it. The open answer task consists of arguing on a given topic, or speaking about material shown on a computer screen (e.g. an advertising pamphlet or a graph) for about 30 seconds. The flexibility of the response to each of the four tasks increases in the order given above. The automatic scoring of the open answer task is considered the most difficult, since it requires an assessment of both the acoustic characteristics and the content of the answer utterance. In this paper, we focus on the automatic scoring of the open answer task.

TABLE I
Definition of the assessment measures in the open answer task.

Fluency	Smoothness in speaking.
Richness	Abundance of vocabulary and expression.
Accuracy	Correctness of syntax and relevance of wording.
Content	Suitableness as an answer.

In an open answer task, it is assumed that the answer utterance is graded as an average of scores given by well-trained human raters (Japanese language teachers). The assessment measures used are *fluency*, *richness*, *accuracy*, and *content*, with a five-level score scale from 0 to 4. The definition of the four measures is shown in Table I. To realize an automatic scoring method that estimates the average score given by human raters, it is essential to extract appropriate features from the answer utterance and to build a robust estimator that estimates the average score from the features.

Previously Nisimura *et al.* proposed an automatic scoring method for the open answer task of the SJ-CAT [4]. Their method first conducts phoneme segmentation using two different sets of acoustic models and obtains two different phoneme segmentation results. Focusing on the onset time and the continuation time in the phoneme segments derived from each of the two segmentation results, the method then extracts a feature set of 14 dimensions, which includes the number of phoneme segments with the same onset time and the maximum difference between onset times in each phoneme segment. Finally it estimates the average score for each of the four assessment measures by using a measure-specific estimator trained using SVR (support vector regression) [6]. Ono *et al.* also proposed an automatic scoring method for the SJ-CAT

978-616-361-823-8

open answer task [5] taking both the acoustical characteristics and the content of the answer utterance comprehensively into consideration. It uses an acoustic feature set consisting of 383 dimensions extracted using the audio feature extractor openSMILE [7] and a content feature set consisting of 7 dimensions extracted from the outputs of the two speech recognizers Julius [8] and T3 [9]. The acoustic feature set includes the signal frame energy and the zero-crossing rate, and the content feature set includes the lexical diversity and the number of important predefined keywords. Finally it estimates the average score in the manner mentioned above.

Although the conventional methods mentioned above attained a relatively high estimation accuracy, there is one drawback in that they use a common feature set to estimate the average score for each of the four different assessment measures. The estimation accuracy would be improved by using a feature set specified for each assessment measure. While the *fluency* is generally assessed by focusing mainly on the acoustic characteristics of the answer utterance, the *richness*, *accuracy* and *content* must to be assessed taking the content of the answer utterance sufficiently into account. However, it is difficult for an automatic scoring method based on speech recognition technology to assess the *accuracy* and *content* since recognition errors are inevitable. On the other hand, we can expect to be able to assess the *richness* even if there are some recognition errors, since *richness* focuses mainly on the abundance of vocabulary and expressions. In this paper, we propose an automatic scoring method for assessing *richness*.

The organization of the rest of this paper is as follows. Section II introduces the proposed method and the feature set. Sections III and IV describe the experimental setup and results. Section V provides our conclusions.

II. Proposed Method

The proposed method first extracts a set of features from an input answer utterance and then estimates the human *richness* score by employing an estimator trained using the SVR, as with the conventional methods. This section provides a detailed description of the features used in the proposed method.

A. Text Statistics

Text statistics are a set of statistics concerning the components of a text, such as sentences, phrases, words, and characters. The effectiveness of using text statistics was recently confirmed in relation to the problem of estimating the difficulty level of a Japanese text, which was determined by well-trained human raters [10].

We can assume that human raters would sufficiently consider the abundance of vocabulary and expression in determining the difficulty level of a Japanese text. Since this is similar to assessing the *richness*, we introduce text statistics to the problem of estimating the human *richness* score. We first investigated the correlation between each of the text statistics and the human richness score. We used 70 answer utterances given by examinees (international students) which were recorded with our prototype SJ-CAT system. The human *richness* score for each answer utterance was obtained by averaging the *richness* scores graded by eight well-trained Japanese teachers. The text statistics were extracted from the hand transcribed text of each answer utterance. This was designed to eliminate the influence of any errors that occurred in speech recognition. To extract the text statistics, we used the Japanese morphological analysis tool MeCab [11] and the syntax-analysis tool CaboCha [12].

TABLE II
Correlation coefficient between each of the text statistics calculated from the hand transcribed text and the human richness score.

Text statistics	Correlation coefficient
Total # of morae (Japanese Hiragana)	0.73
Total # of words	0.76
Total # of unique words	0.75
Total # of phrases	0.72
Total # of sentences	0.11
# of phrases per sentence	0.58
# of morae per sentence	0.52
# of words per sentence	0.57
Maximum # of syntax tree nodes	0.64
Average # of syntax tree nodes	0.57
Maximum value of syntax tree depth	0.62
Average value of syntax tree depth	0.53

The correlation coefficient between each of the text statistics and the human *richness* score is shown in Table II. From Table II, we can see that the correlation on and after the total number of sentences is weak compared with the former 4 statistics. With respect towards the general characteristics of spoken language as observed in the open answer task, the sentences tend to be short and the grammar tends to be simple. We therefore consider that these text statistics do not contribute greatly to our estimation of the human *richness* score. Based on the result, the following four text statistics were selected as features for use in the proposed method.

- **Total number of morae (Japanese Hiragana)**:
  The total number of morae contained in the answer utterance, when all the characters are converted to Japanese hiragana characters.
- **Total number of words**:
  The total number of words in the answer utterance.
- **Total number of unique words**:
  The total number of unique words in the answer utterance.
- **Total number of phrases**:
  The total number of phrases in the answer utterance.

To evaluate the vocabulary variation in the answer utterance, lexical diversity[13] was also adopted as a feature, and it has a correlation coefficient of 0.67 with the human *richness* score under the condition mentioned above.

- **Lexical diversity**:
  A measure of vocabulary variation represented by

TABLE III
CORRELATION COEFFICIENT BETWEEN EACH OF THE TEXT STATISTICS CALCULATED FROM THE SPEECH RECOGNITION OUTPUT AND THE HUMAN RICHNESS SCORE.

	Text statistics	Correlation coefficient
From hand transcribed text	Total # of morae	0.73
	Total # of words	0.76
	Total # of unique words	0.75
From speech recognition output w/o word reliability	Total # of morae	0.65
	Total # of words	0.68
	Total # of unique words	0.66
From speech recognition output w/ word reliablity	Total # of morae	0.69
	Total # of words	0.71
	Total # of unique words	0.69

$\frac{W_{unq}}{\sqrt{2W_{tot}}}$, where W_{unq} and W_{tot} are the total number of unique words and the total number of words, respectively.

However, there is a problem in that the text statistics must be calculated not from the hand transcribed text but from the output text of the speech recognizer. The computation of the text statistics is seriously influenced by recognition errors. Since it is difficult to recognize an answer utterance accurately when it incorporates unnatural pronunciation, filler, and hesitation, recognition errors are unavoidable. To reduce the influence of recognition errors, we introduce word reliability[8] to our text statistics calculation. With the speech recognizer Julius, the reliability for each word is outputted along with the recognition result. The value of the word reliability ranges from 0.0 to 1.0. When the value is close to 1.0, it means that there is no competing candidate word. When calculating the text statistics other than the total number of phrases, the word reliability value was weighted. For example, the number of words becomes 0.9 when there is one word with a word reliability of 0.9. This corresponds to the *richness* being assessed by using the clear part of the answer utterance. From Table III, we confirmed that the correlation coefficient between each of the text statistics and the human *richness* score became higher when we introduced word reliability.

B. Degree of Word Relevance

Although the text statistics mentioned above can assess the quantitative characteristics of the answer utterance, the content of the answer utterance is not taken into consideration. The *richness* tends to be assessed at a higher value when using both the abundant vocabulary and the vocabulary suitable for the intention of a question. To take account of this fact, in the conventional method some important keywords are predefined manually for each question, and the number of important keywords contained in the answer utterance is adopted as a feature that represents the suitability of the vocabulary [5]. However, there are the problems related to the fact that finding all the possible keywords is difficult and no vocabularies other than the predefined keywords are evaluated.

In this paper, we propose a novel method that uses a thesaurus to evaluate the suitability of vocabulary by measuring the word relevance of all the words contained in the answer utterance. A thesaurus is a dictionary that groups words together according to their hierarchical relationship and similarity of meaning. The Japanese Large Thesaurus is a well-known thesaurus of the Japanese language that is widely used in the natural language processing field [14]. It has a hierarchical structure as shown in Fig. 1, so we can evaluate the word relevance of two words by calculating the hierarchical distance between them in the thesaurus. The thesaurus hierarchical distance of two words is calculated by (1).

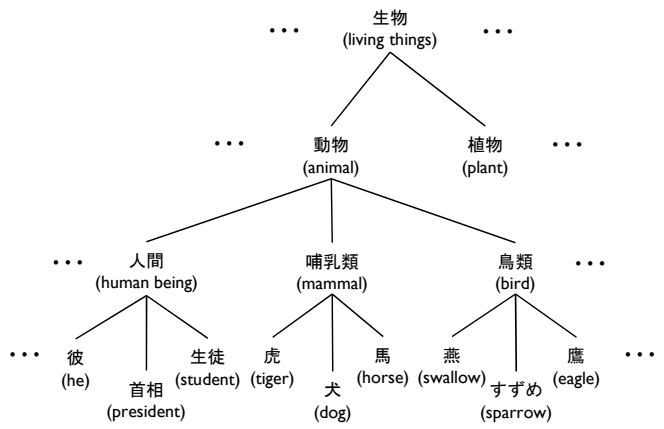

Fig. 1. Example of thesaurus hierarchical structure.

$$D_{i,j} = \frac{d_c \times 2}{d_i + d_j}. \quad (1)$$

In the equation, d_i, d_j, and d_c are the depth of the word i, the word j, and the common upper node c. The $D_{i,j}$ value is between 0.0 and 1.0. This means that the relevance of the two words becomes higher as the $D_{i,j}$ value increases.

The calculation procedure of the proposed degree of word relevance is described as below. We first conduct morphological analysis to the answer utterance and then take out all the nouns appeared. We remove pronouns, numerals, prefixes and suffixes to allow us to focus only on the words that contribute to the expression. The thesaurus hierarchical distance is calculated for every word pair, and the degree of word relevance of the answer utterance is calculated by (2).

$$Rel = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i}(D_{i,j} \cdot w_i \cdot w_j). \quad (2)$$

N is the total number of words. w_i and w_j, respectively, are the reliability of the words i and j outputted by the speech recognizer Julius. To reduce the influence of recognition errors, the word reliability is weighted in the manner mentioned above.

III. EXPERIMENTAL SETUP

To verify the effectiveness of the proposed method, we conducted an experiment according to conditions described in previous research. The experimental conditions are shown in Table IV. The answer utterances from a total of 101 examinees

TABLE IV
EXPERIMENTAL CONDITIONS.

# of answer utterances	1010
# of examinees	101
# of questions	10
Speech recognizer	Julius [8]
Acoustic model	Triphone HMMs
Language model	Tri-gram models
Morphological analysis	MeCab [11]
Syntax analysis	CaboCha [12]
Score estimator	SVR with RBF kernel [15]

(international students), which were collected by the prototype SJ-CAT system, were used in the experiment. Since there are 10 questions in the open answer task, the total number of answer utterances was 1,010. The average *richness* score obtained from eight Japanese teachers was used to train the score estimator.

To calculate the features mentioned above, we used the speech recognizer Julius [8], the morphological-analysis tool MeCab, and the syntax-analysis tool CaboCha. The acoustic model was a speaker-independent triphone model trained with the CSJ native Japanese speech corpus [16] and adapted to non-native speakers. The language model was a trigram model trained with hand transcribed texts of the training data and texts collected from the web and news articles. The word correctness of this recognizer is 60.6% for a subset of the answer utterances used in our experiment [5]. The score estimator was built by using the SVR with the LIBSVM toolkit [15]. We used the RBF kernel and the default parameters during the training and estimation.

IV. EXPERIMENT AND RESULTS

We first conducted a comparative experiment under conditions described in previous studies [4], [5]. We then performed a cross-validation test to investigate whether the estimation accuracy is influenced by differences among examinees and among questions used for training and evaluation.

A. Comparative results

The answer utterances were divided into two sets for training and evaluating the score estimator. The 810 answer utterances from 81 examinees and 10 questions were used as a training set and the remaining 200 answer utterances from 20 examinees and 10 questions were used as an evaluation set. The relationship between the human *richness* score and the score estimated with the proposed method is shown in Fig. 2. A correlation coefficient of 0.92 and RMSE (a root mean square error) of 0.56 were obtained with our proposed method. In contrast, the correlation coefficient with the conventional methods described in [5] and [4] were 0.87 and 0.74, respectively, although the experimental conditions differed slightly. These results confirmed that the proposed method provides good estimates of the human *richness* score. However, we can see that there is a tendency for the answer utterances with a low human *richness* score to be overestimated as shown in Fig. 2.

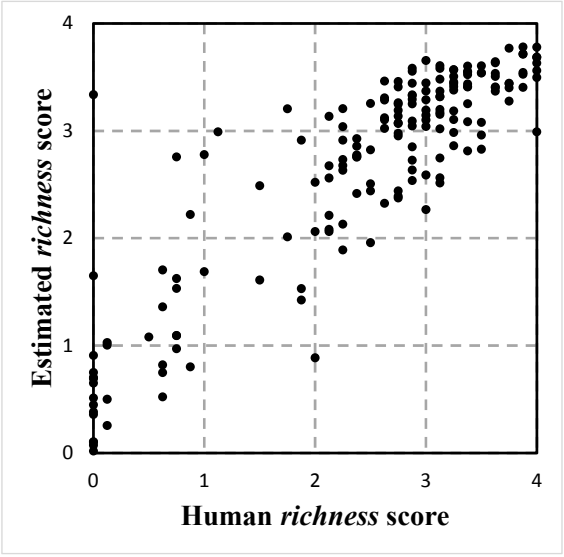

Fig. 2. Relationship between human *richness* score and estimated score with proposed method.

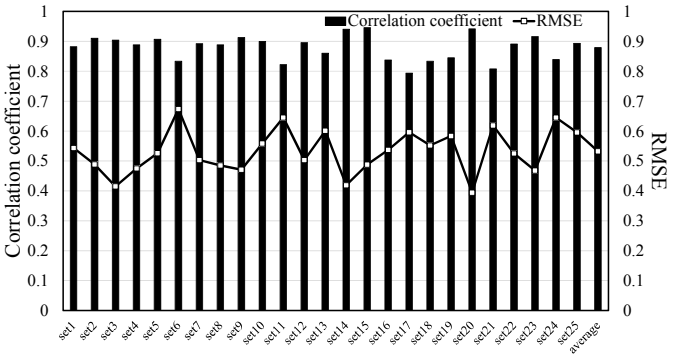

Fig. 3. Correlation coefficient and RMSE in each test set.

B. Cross-validation results

To conduct a cross-validation test, we randomly divided the 10 questions into 5 groups (2 questions per group) and the 80 examinees into 5 groups (20 examinees per group). The 640 answer utterances from 4 groups with a total of 8 questions and 4 groups with a total of 80 examinees were used for training, and the 40 answer utterances from 1 group with 2 questions and 1 group with 20 examinees remaining for testing. By the cross-validation test under this condition, a total of 25 (5 question groups × 5 examinee groups) question-examinee-open test results were obtained. The correlation coefficient and the RMSE in each test set are shown in Fig. 3. The figure shows that the proposed method is relatively robust against differences among examinees and among questions used for training and evaluation. We further conducted an additional cross-validation test to compare the effectiveness of the text statistics and the degree of word relevance. The result confirms that the extent of contribution of the text statistics is high compared with the degree of word relevance.

V. Conclusions

We proposed an automatic scoring method for estimating the human *richness* score in the open answer task of the SJ-CAT. The proposed method uses text statistics weighted by word reliability to assess the abundance of vocabulary and expressions, and the degree of word relevance based on the hierarchical distance in a thesaurus to evaluate the suitability of the vocabulary. The experiments confirmed that the proposed method provides good estimates of the human score, with a correlation coefficient of 0.92 and an RMSE of 0.56. We also showed that the proposed method is relatively robust against differences among examinees and among questions used for training and evaluation.

Acknowledgment

This research was supported by KAKENHI (22242014). We are sincerely grateful to the members of the SJ-CAT project who supported this research.

References

[1] Survey Report on Japanese-Language Education Abroad 2012 Excerpt, http://www.jpf.go.jp/j/japanese/survey/result/dl/survey_2012/2012_s_excerpt_e.pdf.

[2] J-CAT Japanese Language Test, http://www.j-cat.org/en/.

[3] N. Okubo, Y. Yamahata, T. Yamada, S. Imai, K. Ishizuka, T. Shinozaki, R. Nisimura, S. Makino, and N. Kitawaki, "Automatic scoring method considering quality and content of speech for SCAT Japanese speaking test," Proc. OCOCOSDA 2012, pp. 72–77, 2012.

[4] R. Nisimura, R. Kurihara, T. Shinozaki, K. Ishizuka, T. Yamada, S. Imai, H. Kawahara, and T. Irino, "A study on automatic scoring method based on phoneme segmentation using parallel decoding approach for S-CAT speaking test system," Autumn Spring Meeting of the Acoustical Society of Japan, 3-Q-17, pp. 397–399, 2012. (in Japanese)

[5] Y. Ono, M. Otake, T. Shinozaki, R. Nisimura, T. Yamada, K. Ishizuka, Y. Horiuchi, S. Kuroiwa, and S. Imai, "Open answer scoring for S-CAT automated speaking test system using support vector regression," Proc. APSIPA ASC 2012, pp. 1–4, 2012.

[6] D. Basak, S. Pal, and D. Chandra Patranabis, "Support vector regression," Neural Information Processing-Letters and Reviews., Vol. 11, No. 10, pp. 203-224, 2007.

[7] F. Eyben, M. Wollmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," Proc. ACM Multimedia 2010, pp. 1459–1462, 2010.

[8] A. Lee, T. Kawahara, and K. Shikano. "Julius — an open source real-time large vocabulary recognition engine," Proc. EUROSPEECH 2001, pp. 1691–1694, 2001.

[9] P. Dixon, D. Caseiro, T. Oonishi, and S. Furui, "The Titech large vocabulary WFST speech recognition system," Proc. IEEE ASRU 2007, pp. 443–448, 2007.

[10] T. Yamamura, "Effectiveness of text statistics in estimating the difficulty level of Japanese text," IEICE Trans. Information and Systems, Vol. J96-D, No. 8, pp. 1952–1955, 2013.(in Japanese)

[11] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," Proc. EMNLP 2004, pp. 230–237, 2004.

[12] T. Kudo and Y. Matsumoto, "Japanese dependency analyisis using cascaded chunking," Proc. CoNLL-2002, pp. 63–69, 2002.

[13] Carroll. J. B, "On sampling from a lognormal model of word-frequency distribution," Computational analysis of present-day American English, pp. 406-424, 1967.

[14] T. Yamaguchi, "Japanese Large Thesaurus," Taishukan Shoten, ISBN-10:4469790672.(in Japanese)

[15] C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines," ACM Trans. Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[16] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," Proc. SSPR 2003, pp. 135-138, 2003.