PAPER

A Topic-Independent Method for Scoring Student Essay Content

Ryo NAGATA[†a)], **Jun-ichi KAKEGAWA**[††], *Members*, *and* **Yukiko YABUTA**[†††], *Nonmember*

SUMMARY This paper proposes a topic-independent method for automatically scoring essay content. Unlike conventional topic-dependent methods, it predicts the human-assigned score of a given essay without training essays written to the same topic as the target essay. To achieve this, this paper introduces a new measure called MIDF that measures how important and relevant a word is in a given essay. The proposed method predicts the score relying on the distribution of MIDF. Surprisingly, experiments show that the proposed method achieves an accuracy of 0.848 and performs as well as or even better than conventional topic-dependent methods.

key words: *essay scoring, language teaching and learning, student essay, essay topic, learner of English*

1. Introduction

One of the effective ways to improve one's writing skills is to write, receive feedback, revise based on the feedback, and then repeat the whole process. Unfortunately, however, this requires a considerable effort from the classroom teacher; he or she is faced with reading and scoring a number of essays every time an essay topic is assigned. This is especially true when the writer is a non-native speaker (student) of English because their essays normally contain a wide variety of errors and unnatural expressions. This can be easily seen in an actual student essay:

> *I became univercity student, I get up early every morning. I go to the school when I listening to music in train. Stady is very different. Especiary I think that programing and math doesn't know. But, frances is very interesting. Because I think that teacher is interesting.*

For the same reason, essay scoring also becomes problematic in essay tests where students are given a particular essay topic to write about. Again, human raters are faced with reading and scoring a great number of student essays.

In view of this background, researchers including Page [15], Attali and Burstein [1], Burstein et al. [6], and Foltz et al. [11] have done a great deal of work on automated essay scoring to reduce the human effort to score student essays. Automated essay scoring involves grammatical error detection [2], [9], [13], off-topic essay detection [5], evaluation on style, mechanics [1], organization [14], and rhetoric, and essay content evaluation [1], [4], [7], [12].

Automated essay content evaluation, which is the target of this paper, is used to predict the human-assigned score of a given essay in terms of its content. It can be formalized as a text classification problem or a document retrieval problem. Namely, the human-assigned score of a given essay is predicted by classifying it into a score category or retrieving the most highly similar essays which are manually scored in advance (i.e., training essays). For example, Burstein and Chodorow [4] proposed a method for predicting the human-assigned score of a given essay by retrieving the most highly similar essays in training essays where the similarity is calculated based on word frequencies. More precisely, the target essay and training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights. The similarity is measured by the cosine between the target essay vector and training vectors.

One of the major drawbacks to the conventional methods is that they are topic-dependent. They require human-reader scored essays that are written to the same essay topic as the target essay. This means that for every new essay topic, one has to collect essays that are written to the new essay topic and has to score them, which is costly and time-consuming. This also implies that it is almost impossible for classroom teachers to set essay topics by themselves.

To solve the drawback, this paper proposes a topic-independent method for automatically scoring essay content. Here, topic-independent means that once a set of human-reader scored essays written to a topic is obtained, one can apply the proposed method to essays written to any essay topic. In other words, there is no need for making training essays every time a new essay topic is assigned. Surprisingly, it rivals the conventional topic-dependent methods as our experiments show. To achieve this, this paper introduces a new measure called MIDF that measures how important and relevant a word is in a given essay. The proposed method predicts the score relying on the distribution of MIDF.

The rest of this paper is structured as follows. Section 2 describes the proposed method. Section 3 describes experiments conducted to evaluate the proposed method. Section 4 discuses the experimental results.

Manuscript received August 18, 2009.
Manuscript revised October 22, 2009.

[†]The author is with Konan University, Kobe-shi, 658–8501 Japan.
[††]The author is with Tokyo University of Science, Yamaguchi, Sanyoonoda-shi, 756–0884 Japan.
[†††]The author is with Seisen Jogakuin College, Nagano-shi, 381–0085 Japan.
a) E-mail: rnagata@konan-u.ac.jp

DOI: 10.1587/transinf.E93.D.335

2. Proposed Method

The proposed method assumes that a good essay (in terms of its content) contains more important words that are relevant to its essay topic than poor essays do. Under this assumption, it predicts the score of a given essay based on the importance and relevance of words in it. The rest of this section first describes how to calculate the importance and relevance and then describes how to predict the score based on the importance and relevance.

The importance and relevance are calculated based on statistics obtained from a corpus (i.e., a set of documents such as newspaper articles). To formalize the calculation, we use the symbols $w_1, w_2, \ldots, w_M$ to denote M different words in the corpus. We refer to the number of documents in the corpus as N. We refer to the number of documents in which w_i appears as n_i. Similarly, we refer to the number of documents in which w_i and w_j appear as n_{ij}. Also, we refer to the number of occurrences of w_i in the whole corpus as f_i.

The importance of w_i measures how much information it contains. It can be formalized by Inversed Document Frequency (IDF) [18]

$$\mathrm{IDF}(w_i) = \log \frac{N}{n_i}, \tag{1}$$

which can be interpreted as the information measure noting that it has the form of $-\log p$ where p is the estimate for the probability of w_i occurring. IDF gives a high value to important words that appear in a few particular documents. For instance, $\mathrm{IDF}(w_i = \textit{photo}) = 8.4$ and $\mathrm{IDF}(w_i = \textit{I}) = 1.9$ according to statistics obtained from a corpus[†], which agrees with our intuition. IDF is also used in the conventional topic-dependent methods [1], [4], [7], [12].

The relevance of a word to an essay topic is approximated to the relevance of the word to each word in the essay topic (e.g., *summer* and *vacation* in the case of the essay topic *summer vacation*). The relevance of w_i to w_j is measured by Mutual Information (MI) [10]

$$\mathrm{MI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{2}$$

where $p(w_i)$ and $p(w_i, w_j)$ are the probability of w_i occurring and that of w_i and w_j co-occurring, respectively. The probabilities $p(w_i)$ and $p(w_i, w_j)$ are estimated by

$$p(w_i) = \frac{f_i}{\sum_{t=1}^{M} f_t}, \tag{3}$$

and

$$p(w_i, w_j) = \frac{n_{ij}}{N}, \tag{4}$$

respectively. MI measures how relevant a word is to another (for instance, $\mathrm{MI}(\textit{trip}, \textit{photo}) = 3.9$ and $\mathrm{MI}(\textit{trip}, \textit{idea}) = 1.3$ according to the same corpus).

To measure how important and relevant a word is in a given essay, this paper introduces a new measure called MIDF

$$\mathrm{MIDF}(w_i, T) = \sum_{w_j \in T} \frac{\mathrm{MI}(w_i, w_j)}{|T|} IDF(w_i) e_i \tag{5}$$

where e_i and T denote the number of occurrences of w_i in the essay and a set of words in the essay topic (excluding function words), respectively. MIDF defined by Eq. (5) can be regarded as a weight which is assigned to the given word.

Using MIDF just defined, the proposed method predicts the score of a given essay as follows. First, each training essay is divided into words; note that training essays can be written to any topic (i.e., topic-independent). All words are reduced to their morphological stem and converted entirely to lower case. Function words such as determiners are discarded.

Second, each training essay is transformed into a histogram of MIDF distribution as shown in Fig. 1. MIDF is calculated for each word which is obtained in the first step. The resulting pairs (word and its MIDF) are expressed as a histogram. The horizontal and vertical axes correspond to bins of MIDF and the number of words falling in each bin, respectively. The width of the bins is determined based on

$$d = \frac{1}{L}\{\max_{v \in V}(v) - \min_{v \in V}(v)\} \tag{6}$$

where V and L denote a set of the values of MIDF and the parameter that determines the number of bins, respectively; L is set to 10 in this paper.

In practice, Eq. (6) severely suffers from outliers giving a very wide bin. In that case, most words fall in the first few bins and outliers in the last bins, and no words between them. To avoid this, outliers are excluded from the determination of the width. This can be done by

$$d = \frac{1}{L}\{\min(\max_{v \in V}(v), \hat{\mu} + 2\hat{\sigma}) - \max(\min_{v \in V}(v), \hat{\mu} - 2\hat{\sigma})\} \tag{7}$$

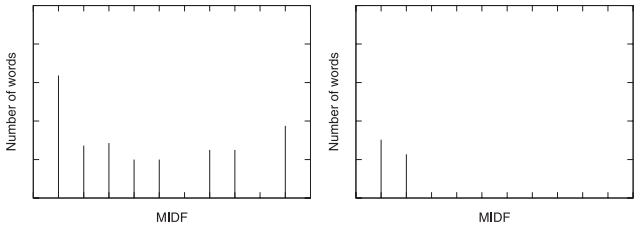
where $\hat{\mu}$ and $\hat{\sigma}$ are the mean and unbiased standard deviation of MIDF. Equation (7) is used to determine the width of the bins in the proposed method.

Intuitively, a good essay contains a number of important and relevant words and its histogram should be a relatively flat one as in Fig. 1 (a) whereas a poor one should look like a histogram that has a peak on the leftmost bins and no or very low bars on the others as in Fig. 1 (b) (the figures are drawn from actual student essays).

Third, the histograms are transformed into vectors whose elements and values are the bins of the histograms and the corresponding frequencies, respectively. The number of words in each training essay is added to its corresponding vector in order to evaluate very short essays as poor.

Finally, the score of the target essay is predicted by classifying it based on the training essay vectors. To do this, the target essay is first transformed into a vector in the same

[†]The details of the corpus is described in Subsect. 3.1.



(a) good essay (b) poor essay

Fig. 1 Student essays transformed into MIDF histograms.

manner. Then, it is classified into a score category by a classifier learned from the training essay vectors. Almost any kind of classifier can be used for the purpose. Support Vector Machines (SVMs), which have been shown to be effective in text classification, are used in the proposed method.

So far, we have limited our discussion to the binary-score category (*good* or *poor*) setting for the purpose of illustration. However, the proposed method is capable for multi-score category settings such as the one to six score category which is often used in English writing tests. In multi-score category settings, training essays are labeled with either one of the multi-scores (one to six, for example) instead of *good* and *poor* by human raters. Each training essay is transformed into a MIDF histogram and then into a training essay vector as in the binary-score category case. The target essay is also transformed into a vector in the same manner except for the score label. The target essay vector is classified into one of the multi-scores by SVMs, which is trained on the training essay vectors. SVMs are originally developed for binary-class problems. However, SVMs are extended to multi-class problems [16]. Thus, SVMs are applicable to multi-score category settings. Also, other kinds of classifier are applicable. For example, k-Nearest Neighbor (k-NN) classifier would be suitable for our purpose when training essays are sufficiently available.

This is how the proposed method predicts the human-assigned score of a given essay. As having been discussed, the proposed method does not require human-reader scored essays written to the same topic as the target essay whereas the conventional topic-dependent methods do. It only requires a set of human-reader scored essays written to a topic and a corpus that has been very easy to obtain lately. In other words, once a set of human-reader scored essays written to a topic is obtained, one can apply the proposed method to essays written to any topic.

3. Experiments

3.1 Experimental Conditions

For evaluation, 455 essays were collected which had been written by Japanese learners of English. Their topics were either *my family*, *future dreams*, or *memories in junior high school*. The writers were third grade junior high students and first grade high school students. Table 1 shows the statistics on the target essays.

Table 1 Statistics on the target essays.

Topic	Writer	# essays	# poor essays
My family	Jr. high	194	20
Future dreams	high	242	70
Memories in Jr. high	high	19	5
TOTAL	Jr. high & high	455	95

Each essay was separately assigned to two teachers of English (out of three). They gave *poor* to essays if they thought the essay had a problem in terms of content and he or she wanted to give some feedback comments to the writer, and otherwise *good*. If both of them gave *poor* to an essay, then its score was determined to be *poor*, otherwise *good*.

We used the binary-score category setting in the experiments because of the following two reasons. The first reason is that the writers of the essays were beginning learners of English writing (third grade junior high students and first grade high school students) whose writing abilities were limited compared to advanced learners. Because of the limitation, the evidences for the writing abilities were not enough to reliably assign appropriate multi-scores to the target essays. Thus, we used the binary-score category setting for obtaining more reliable test data. The second reason is that the proposed method is still useful with the binary-score category setting. As already mentioned, the human raters (teachers of English) gave *poor* if they thought the essay had a problem in terms of content and they wanted to give some feedback comments to the writer, and otherwise good. This means that teachers of English can effectively find learners whose essays have a problem in terms of content and they can give some feedback comments if the proposed method accurately predicts their scores (*good* or *poor*). These are the reasons why we used the binary-score category setting in the experiments.

A set of texts derived from English language learning materials were used as a corpus to calculate IDF and MI. The corpus approximately consisted of 180000 words.

The performance was evaluated by accuracy, recall, precision, and F-measure. Accuracy was defined by

$$A = \frac{\text{Number of essays correctly predicted}}{\text{Number of essays}}. \quad (8)$$

Accuracy measures how accurately the proposed method predicts the human-assigned scores of the target essays. Recall was defined by

$$R = \frac{\text{Number of poor essays correctly predicted}}{\text{Number of poor essays}}. \quad (9)$$

Recall measures how well the proposed method detects all the poor essays in the target essays. Precision was defined by

$$P = \frac{\text{Number of poor essays correctly predicted}}{\text{Number of essays predicted poor}}. \quad (10)$$

Precision measures how well the proposed method detects only the poor essays in the target essays. F-measure was defined by

$$F = \frac{2RP}{R+P}, \tag{11}$$

which measures the performance considering both recall and precision.

3.2 Experimental Procedures

The proposed method was implemented using the collected essays. When it was tested on essays written to a topic, the rest written to the other topics were used as training essays. For instance, when it was tested on essays written to *My family*, essays written to *Future dreams* and *Memories in Junior high school* were used as training essays.

The topic-dependent method based on TFIDF [4] was also implemented for comparison. In the original method, IDF was calculated based on the statistics obtained from training essays. However, in the experiments, it was calculated based on the statistics obtained from the corpus because pre-experiments showed that the topic-dependent method performed much better with IDF obtained from the corpus.

In addition, three other topic-dependent methods were implemented that were derived from the topic-dependent method above. One was the topic-dependent method based on TFIDF where SVMs were used as a classifier instead of the cosine, which was used in the original, to remove the difference between the classifiers used in the proposed method and the topic-dependent method. The others were topic-dependent methods based on MIDF instead of TFIDF where the elements and values of the vectors were words and corresponding MIDFs, respectively. The difference between the two was their classifiers (the cosine or SVMs). In all methods for comparison, the number of words was included as a feature in the vectors in order to evaluate very short essays as *poor* like in the proposed method. The second polynomial kernel was used in all SVMs.

To evaluate the performance of the methods for comparison, leave-one-out cross-validation [17] was used because they were topic-dependent methods and it was impossible to use the same evaluation method as in the proposed method (namely, trained on two of the three essay topics and tested on the rest). Each essay in turn was left out as a test essay from the training essays written to an essay topic, and the topic-depended methods were trained on the remaining training essays written to the same essay topic. All predictions were averaged to calculate the performance measures. For comparison, the proposed method was also evaluated by leave-one-out cross-validation. Note that the proposed method is still topic-independent since it does not directly use the information on the essay topics.

3.3 Experimental Results

Table 2 shows the performance of the proposed method. It shows that the proposed method achieves a very high accuracy of around 80% to 90%. In other words, the predictions of the proposed method agree with the human-assigned scores most of the time. When viewed from another perspective, the proposed method detects about half poor essays with less than 30% of false-positives.

These results imply that the proposed method is useful in helping human raters and teachers score essay content. It can be used to block human raters' false-negatives and false-positives in essay tests; if the prediction of the proposed method disagrees with a score given by a human rater, then the essay should be double-checked by another human rater. The proposed method can also help teachers find students who need feedback comments in terms of essay content.

Table 3 to 6 show the results of comparison between the

Table 2 Performance of proposed method.

Topic	*A*	*R*	*P*	*F*
My family	0.895	0.450	0.900	0.600
Future dreams	0.793	0.529	0.685	0.597
Memories in Jr. high	0.938	0.600	1.00	0.750
TOTAL	0.859	0.516	0.731	0.605

Table 3 Comparison between proposed method and topic-dependent methods (Topic:My Family).

Topic	*A*	*R*	*P*	*F*
Baseline	0.103	**1.00**	0.103	0.187
TFIDF + cosine	0.902	0.050	**1.00**	0.095
TFIDF + SVMs	0.876	0.250	0.357	0.294
MIDF + cosine	0.887	0.100	0.333	0.154
MIDF + SVMs	0.892	0.150	0.429	0.222
Proposed method	**0.948**	0.600	0.857	**0.706**

Table 4 Comparison between proposed method and topic-dependent methods (Topic: Future Dreams).

Topic	*A*	*R*	*P*	*F*
Baseline	0.289	**1.00**	0.289	0.449
TFIDF + cosine	0.731	0.114	**0.727**	0.198
TFIDF + SVMs	0.764	0.729	0.573	**0.642**
MIDF + cosine	0.645	0.329	0.371	0.348
MIDF + SVMs	0.756	0.657	0.568	0.609
Proposed method	**0.769**	0.514	0.621	0.562

Table 5 Comparison between proposed method and topic-dependent methods (Topic: Memories in Jr. High).

Topic	*A*	*R*	*P*	*F*
Baseline	0.263	**1.00**	0.263	0.417
TFIDF + cosine	0.789	0.200	**1.00**	0.333
TFIDF + SVMs	0.579	0.600	0.333	0.429
MIDF + cosine	0.421	0.400	0.200	0.267
MIDF + SVMs	0.684	0.600	0.429	0.500
Proposed method	**0.842**	0.600	0.750	**0.667**

Table 6 Comparison between proposed method and topic-dependent methods (Topic: All Topics).

Topic	*A*	*R*	*P*	*F*
Baseline	0.209	**1.00**	0.209	0.345
TFIDF + cosine	0.807	0.105	**0.714**	0.183
TFIDF + SVMs	0.804	0.621	0.527	0.570
MIDF + cosine	0.738	0.284	0.346	0.312
MIDF + SVMs	0.820	0.547	0. 547	0.547
Proposed method	**0.848**	0.537	0.671	**0.596**

proposed method and the topic-dependent methods. They also show the baseline performance where all essays are scored as *poor*. Note that the performances of the proposed methods in Table 3 to 6 are different from those in Table 2 because the evaluation methods are different between them (although the performances are similar).

The results reveal that the proposed method performs as well as or even better than the topic-dependent methods. Also, they reveal that F-measure of the proposed method is relatively stable whereas that of the topic-dependent methods varies from topic to topic.

4. Discussion

The reason why the proposed method performs as well as or even better than the topic-dependent method despite the fact that it is topic-independent is that training essays are sparse and that the topic-dependent methods suffer from the sparseness. This is exemplified as follows. Suppose that there are two essays written to the essay topic *My family* in training essays; one is about the occupations of family members and its score is *good*, and the other is about a dog in the family and its score is *poor*. Further suppose that the target essay is about a dog. In that case, the predictions of the topic-dependent methods are likely to be *poor* no matter how good the content is because the target essay shares more similar words with the training essay about a dog than with the training essay about occupations. By contrast, the proposed method does not rely on the word similarity between the target essay and training essays. Instead, it predicts the score relying on the similarity between the histograms that express the distribution of MIDF. Using the same example above, the training essay about a dog would be transformed into a histogram that looks like Fig. 1 (b) in Sect. 2 whereas the training essay about occupations would be transformed into a histogram that looks like Fig. 1 (a). The target essay would be transformed into a histogram that looks like Fig. 1 (a) if its true score is *good* and contains a number of important and relevant words. As a result, the proposed method is likely to predict the score of the target essay as *good* even if it is about a dog.

As having been discussed, the proposed method performs as well as or even better than the topic-dependent methods. In addition to the performance, it has an advantage over the topic-dependent methods. That is, it is topic-independent, which makes it much more useful and practical.

At the same time, there are some false-negatives and false-positives even in the proposed method. Especially, the recall of the proposed method is low compared to its precision (meaning that false-negatives are more problematic than false-positives). One of the major causes of false-negatives is repetitions of the same and/or similar phrases. For example, the proposed method would give *good* to the target essay written to the topic *My family*:

> *My father is kind. My mother is kind. My sister is kind. My brother is kind. My grandfather is kind...*

since it contains a number of important and relevant words such as *father*, *mother*, and *sister*. Apparently, human raters would give *poor* to it. It requires a new technique such as detection of repetitions of the same or similar phrases to handle such cases. Fortunately, Burstein and Wolska [8] proposed a method for identifying overly repetitious word use. This method can be applied to avoiding these false-negatives.

False-negatives are also due to grammatical errors. It is a problem of grammar, but too many grammatical errors hinder the reader from understanding the content no matter how many important and relevant words are in the target essay. This kind of false-negatives may be avoided by using grammatical error detection such as [2], [9], [13].

False-positives tend to occur when the content is very specific. For instance, the proposed method mistakenly gave *poor* to an essay (topic: *Future dreams*) about being a calligrapher, which was actually a good essay. The content was so specific that important and relevant words such as *calligrapher* and *calligraphy* did not appear in the corpus used to calculate MIDF. Because of this, their MIDF was not defined and these words were not considered in the histogram. As a result, the histogram tends to be one like Fig. 1 (b). This problem might be reduced by using a bigger corpus such as the British National Corpus [3].

Another cause of false-positives is spelling errors. If important and relevant words are mistakenly spelt in the target essay, the situation is very similar to the specific-content situation above. Namely misspelt words do not appear in the corpus, and thus they are not considered in the histogram whereas human raters seem not to care about a few spelling errors as long as they are able to guess the meanings of the words and understand the content. This causes the discrepancy between the proposed method and human raters.

5. Conclusions

This paper proposed a topic-independent method for automatically scoring essay content. The experiments show that it achieves an accuracy of 0.848 and performs as well as or even better than topic-dependent methods.

For future work, we will investigate how to correctly predict the scores of essays whose content is very specific. We will also investigate how the proposed method can be combined with other systems such as grammatical error detection systems to achieve better performance.

References

[1] Y. Attali and J. Burstein, "Automated essay scoring with e-rater V.2," Journal of Technology, Learning, and Assessment, vol.4, no.3, pp.3–30, Feb. 2006.

[2] C. Brockett, W.B. Dolan, and M. Gamon, "Correcting ESL errors using phrasal SMT techniques," Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp.249–256, Sydney, Australia, July 2006.

[3] L. Burnard, Users Reference Guide for the British National Corpus. version 1.0, Oxford University Computing Services, Oxford, 1995.
[4] J. Burstein and M. Chodorow, "Automated essay scoring for non-native English speakers," Proc. ACL99 Workshop on Computer-mediated Language Assessment and Evaluation of Natural Language Processing, pp.68–75, College Park, USA, June 1999.
[5] J. Burstein and D. Higgins, "Advanced capabilities for evaluating student writing: Detecting off-topic essays without topic-specific training," Proc. 12th international conference on artificial intelligence in Education, pp.112–119, Amsterdam, The Netherlands, July 2005.
[6] J. Burstein, M. Chodorow, and C. Leacock, "Automated essay evaluation: The *Criterion* online writing service," AI Magazine, vol.25, no.3, pp.27–36, Sept. 2004.
[7] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M.D. Harris, "Automated scoring using a hybrid feature identification technique," Proc. 36th Annual Meeting of the Association for Computational Linguistics, pp.206–210, Montreal, Canada, Aug. 1998.
[8] J. Burstein and M. Wolska, "Toward evaluation of writing style: Finding overly repetitive word use in student essays," Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics, pp.35–42, Budapest, Hungary, April 2003.
[9] M. Chodorow and C. Leacock, "An unsupervised method for detecting grammatical errors," Proc. 1st Meeting of the North America Chapter of the Association for Computational Linguistics, pp.140–147, Seattle, USA, April 2000.
[10] K.W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," Computational Linguistics, vol.16, no.1, pp.22–29, March 1990.
[11] P.W. Foltz, D. Laham, and T.K Landauer, "Automated essay scoring: Applications to educational technology," Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp.939–944, Seattle, USA, June 1999.
[12] L.S. Larkey, "Automatic essay grading using text categorization techniques," Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.90–95, Melbourne, Australia, Aug. 1998.
[13] R. Nagata, K. Morihiro, A. Kawai, and N. Isu, "A feedback-augmented method for detecting errors in the writing of learners of English," Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp.241–248, Sydney, Australia, July 2006.
[14] S.T. O'Rourke and R.A. Calvo, "Visualizing paragraph closeness for academic writing support," Proc. 9th IEEE International Conference on Advanced Learning Technologies, pp.688–692, Riga, Latvia, July 2009.
[15] E.B. Page, "Computer grading of student prose, using modern concepts and software," J. Experimental Education, vol.62, no.2, pp.127–142, Jan. 1994.
[16] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," Proc. European Symposium on Artificial Neural Networks, pp.219–224, Bruges, Belgium, April 1999.
[17] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, 2000.
[18] M. Yamamoto and K.W. Church, "Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus," Computational Linguistics, vol.27, no.1, pp.1–30, March 2001.

Ryo Nagata graduated from the Department of Electrical Engineering, Meiji Univ. in 1999 and completed the doctoral program in information engineering at Mie Univ./ in 2005 and became a research associate at Hyogo Univ. of Teacher Education. Since 2008, he has been an associate professor at Konan University. His research interests are language modeling, grammatical error detection and correction, and edu-mining (educational data mining). He is a member of the Association for Natural Language Processing.

Jun-ichi Kakegawa graduated from the Department of Applied Electronics, Tokyo Univ. of Science in 1999 and completed the doctoral program in engineering at Tokyo Univ. of Science in 2004. In the same year, he became a research associate at Hyogo Univ. of Teacher Education. Since 2009, he has been a research associate at Tokyo Univ. of Science, Yamaguchi. His research interests are computer-assisted learning and edu-mining. He is a member of the Japanese Society for Artificial Intelligence, Japanese Society for Information and Systems in Education, the Japan Society for Educational Technology, and the Association for Natural Language Processing.

Yukiko Yabuta obtained BA in Linguistics form Monash University, Melbourne, and MA in Applied Linguistics from The University of Tokyo. Since 2008, she has been an associate professor at Seisen Jogakuin College.