Automated Japanese Essay Scoring System based on Articles Written by Experts

Tsunenori Ishioka
Research Division
The National Center for
University Entrance Examinations
Tokyo 153-8501, Japan
tunenori@rd.dnc.ac.jp

Masayuki Kameda
Software Research Center
Ricoh Co., Ltd.
Tokyo 112-0002, Japan
masayuki.kameda@nts.ricoh.co.jp

Abstract

We have developed an automated Japanese essay scoring system called Jess. The system needs expert writings rather than expert raters to build the evaluation model. By detecting statistical outliers of predetermined aimed essay features compared with many professional writings for each prompt, our system can evaluate essays. The following three features are examined: (1) rhetoric — syntactic variety, or the use of various structures in the arrangement of phases, clauses, and sentences, (2) organization — characteristics associated with the orderly presentation of ideas, such as rhetorical features and linguistic cues, and (3) content — vocabulary related to the topic, such as relevant information and precise or specialized vocabulary. The final evaluation score is calculated by deducting from a perfect score assigned by a learning process using editorials and columns from the Mainichi Daily News newspaper. A diagnosis for the essay is also given.

1 Introduction

When giving an essay test, the examiner expects a written essay to reflect the writing ability of the examinee. A variety of factors, however, can affect scores in a complicated manner. Cooper (1984) states that "various factors including the writer, topic, mode, time limit, examination situation, and rater can introduce error into the scoring of essays used to measure writing ability." Most of these factors are present in giving tests, and the human rater, in particular, is a major error factor in the scoring of essays.

In fact, many other factors influence the scoring of essay tests, as listed below, and much research has been devoted.

- Handwriting skill (handwriting quality, spelling) (Chase, 1979; Marshall and Powers, 1969)
- Serial effects of rating (the order in which essay answers are rated) (Hughes et al., 1983)
- Topic selection (how should essays written on different topics be rated?) (Meyer, 1939)
- Other error factors (writer's gender, ethnic group, etc.) (Chase, 1986)

In recent years, with the aim of removing these error factors and establishing fairness, considerable research has been performed on computer-based automated essay scoring (AES) systems (Burstein et al., 1998; Foltz et al., 1999; Page et al., 1997; Powers et al., 2000; Rudner and Liang, 2002).

The AES systems provide the users with prompt feedback to improve their writings. Therefore, many practical AES systems have been used. E-rater (Burstein et al., 1998), developed by the Educational Testing Service, began being used for operational scoring of the Analytical Writing Assessment in the Graduate Management Admission Test (GMAT), an entrance examination for business graduate schools, in February 1999, and it has scored approximately 360,000 essays per year. The system includes several independent NLP-based modules for identifying features relevant to the scoring guide from three categories: syntax, discourse, and topic. Each of the feature-recognition modules correlate the essay scores with assigned by human readers. E-rater uses a model-building module to select and weight predictive features for essay scoring. Project Essay

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 233–240,
Sydney, July 2006.

Grade (PEG), which was the first automated essay scorer, uses a regression model like e-rater (Page et al., 1997). IntelliMetric (Elliot, 2003) was first commercially released by Vantage Learning in January 1998 and was the first AI-based essay-scoring tool available to educational agencies. The system internalizes the pooled wisdom of many expert scorers. The Intelligent Essay Assessor (IEA) is a set of software tools for scoring the quality of the conceptual content of essays based on latent semantic analysis (Foltz et al., 1999). The Bayesian Essay Test Scoring sYstem (BETSY) is a windows-based program that classifies text based on trained material. The features include multi-nomial and Bernoulli Naive Bayes models (Rudner and Liang, 2002).

Note that all above-mentioned systems are based on the assumption that the true quality of essays must be defined by human judges. However, Bennet and Bejar (1998) have criticized the overreliance on human ratings as the sole criterion for evaluating computer performance because ratings are typically based as a constructed rubric that may ultimately achieve acceptable reliability at the cost of validity. In addition, Friedman, in research during the 1980s, found that holistic ratings by human raters did not award particularly high marks to professionally written essays mixed in with student productions. This is a kind of negative halo effect: create a bad impression, and you will be scored low on everything. Thus, Bereiter (2003) insists that another approach to doing better than ordinary human raters would be to use expert writers rather than expert raters. Reputable professional writers produce sophisticated and easy-to-read essays. The use of professional writings as the criterion, whether the evaluation is based on holistic or trait rating, has an advantage, described below.

The methods based on expert rater evaluations require much effort to set up the model for each prompt. For example, e-rater and PEG use some sort of regression approaches in setting up the statistical models. Depending on how many variables are involved, these models may require thousands of cases to derive stable regression weights. BETSY requires the Bayesian rules, and IntelliMetric, the AI-based rules. Thus, the methodology limits the grader's practical utility to large-scale testing operations in which such data collection is feasible. On the other hand, a method based on professional writings can overcome this; i.e., in our system, we need not set up a model simulating a human rater because thousands of articles by professional writers can easily be obtained via various electronic media. By detecting a statistical outlier to predetermined essay features compared with many professional writings for each prompt, our system can evaluate essays.

In Japan, it is possible to obtain complete articles from the Mainichi Daily News newspaper up to 2005 from Nichigai Associates, Inc. and from the Nihon Keizai newspaper up to 2004 from Nikkei Books and Software, Inc. for purposes of linguistic study. In short, it is relatively easy to collect editorials and columns (e.g., "Yoroku") on some form of electronic media for use as essay models. Literary works in the public domain can be accessed from Aozora Bunko (http://www.aozora.gr.jp/). Furthermore, with regard to morphological analysis, the basis of Japanese natural language processing, a number of free Japanese morphological analyzers are available. These include JUMAN (http://www-lab25.kuee.kyoto-u.ac.jp/nlresource/juman.html), developed by the Language Media Laboratory of Kyoto University, and ChaSen (http://chasen.aist-nara.ac.jp/, used in this study) from the Matsumoto Laboratory of the Nara Institute of Science and Technology.

Likewise, for syntactic analysis, free resources are available such as KNP (http://www-lab25.kuee.kyoto-u.ac.jp/nlresource/knp.html) from Kyoto University, SAX and BUP (http://cactus.aist-nara.ac.jp/lab/nlt/{sax,bup}.html) from the Nara Institute of Science and Technology, and the MSLR parser (http://tanaka-www.cs.titech.ac.jp/pub/mslr/index-j.html) from the Tanaka Tokunaga Laboratory of the Tokyo Institute of Technology. With resources such as these, we prepared tools for computer processing of the articles and columns that we collect as essay models.

In addition, for the scoring of essays, where it is essential to evaluate whether content is suitable, i.e., whether a written essay responds appropriately to the essay prompt, it is becoming possible for us to use semantic search technologies not based on pattern matching as used by search engines on the Web. The methods for implementing such technologies are explained in detail by Ishioka and Kameda (1999) and elsewhere. We believe that this statistical outlier detection ap-

proach to using published professional essays and columns as models makes it possible to develop a system essentially superior to other AES systems.

We have named this automated Japanese essay scoring system "Jess." This system evaluates essays based on three features : (1) rhetoric, (2) organization, and (3) content, which are basically the same as the structure, organization, and content used by e-rater. Jess also allows the user to designate weights (allotted points) for each of these essay features. If the user does not explicitly specify the point allotment, the default weights are 5, 2, and 3 for structure, organization, and content, respectively, for a total of 10 points. (Incidentally, a perfect score in e-rater is 6.) This default point allotment in which "rhetoric" is weighted higher than "organization" and "content" is based on the work of Watanabe et al. (1988). In that research, 15 criteria were given for scoring essays: (1) wrong/omitted characters, (2) strong vocabulary, (3) character usage, (4) grammar, (5) style, (6) topic relevance, (7) ideas, (8) sentence structure, (9) power of expression, (10) knowledge, (11) logic/consistency, (12) power of thinking/judgment, (13) complacency, (14) nuance, and (15) affinity. Here, correlation coefficients were given to reflect the evaluation value of each of these criteria. For example, (3) character usage, which is deeply related to "rhetoric," turned out to have the highest correlation coefficient at 0.58, and (1) wrong/omitted characters was relatively high at 0.36. In addition, (8) sentence structure and (11) logic/consistency, which is deeply related to "organization," had correlation coefficients of 0.32 and 0.26, respectively, both lower than that of "rhetoric," and (6) topic relevance and (14) nuance, which are though to be deeply related to "content," had correlation coefficients of 0.27 and 0.32, respectively.

Our system, Jess, is the first automated Japanese essay scorer and has become most famous in Japan, since it was introduced in February 2005 in a headline in the Asahi Daily News, which is well known as the most reliable and most representative newspaper of Japan.

The following sections describe the scoring criteria of Jess in detail. Sections 2, 3, and 4 examine rhetoric, organization, and content, respectively. Section 5 presents an application example and associated operation times, and section 6 concludes the paper.

2 Rhetoric

As metrics to portray rhetoric, Jess uses (1) ease of reading, (2) diversity of vocabulary, (3) percentage of big words (long, difficult words), and (4) percentage of passive sentences, in accordance with Maekawa (1995) and Nagao (1996). These metrics are broken down further into various statistical quantities in the following sections. The distributions of these statistical quantities were obtained from the editorials and columns stored on the Mainichi Daily News CD-ROMs.

Though most of these distributions are asymmetrical (skewed), they are each treated as a distribution of an ideal essay. In the event that a score (obtained statistical quantity) turns out to be an outlier value with respect to such an ideal distribution, that score is judged to be "inappropriate" for that metric. The points originally allotted to the metric are then reduced, and a comment to that effect is output. An "outlier" is an item of data more than 1.5 times the interquartile range. (In a box-and-whisker plot, whiskers are drawn up to the maximum and minimum data points within 1.5 times the interquartile range.) In scoring, the relative weights of the broken-down metrics are equivalent with the exception of "diversity of vocabulary," which is given a weight twice that of the others because we consider it an index contributing to not only "rhetoric" but to "content" as well.

2.1 Ease of reading

The following items are considered indexes of "ease of reading."

1. Median and maximum sentence length

   Shorter sentences are generally assumed to make for easier reading (Knuth et al., 1988). Many books on writing in the Japanese language, moreover, state that a sentence should be no longer than 40 or 50 characters. Median and maximum sentence length can therefore be treated as an index. The reason the median value is used as opposed to the average is that sentence-length distributions are skewed in most cases. The relative weight used in the evaluation of median and maximum sentence length is equivalent to that of the indexes described below. Sentence length is also known to be quite effective for determining style.

2. Median and maximum clause length

In addition to periods (.), commas (,) can also contribute to ease of reading. Here, text between commas is called a "clause." The number of characters in a clause is also an evaluation index.

3. Median and maximum number of phrases in clauses

   A human being cannot understand many things at one time. The limit of human short-term memory is said to be seven things in general, and that is thought to limit the length of clauses. Actually, on surveying the number of phrases in clauses from editorials in the Mainichi Daily News, we found it to have a median of four, which is highly compatible with the short-term memory maximum of seven things.

4. Kanji/kana ratio

   To simplify text and make it easier to read, a writer will generally reduce kanji (Chinese characters) intentionally. In fact, an appropriate range for the kanji/kana ratio in essays is thought to exist, and this range is taken to be an evaluation index. The kanji/kana ratio is also thought to be one aspect of style.

5. Number of attributive declined or conjugated words (embedded sentences)

   The declined or conjugated forms of attributive modifiers indicate the existence of "embedded sentences," and their quantity is thought to affect ease of understanding.

6. Maximum number of consecutive infinitive-form or conjunctive-particle clauses

   Consecutive infinitive-form or conjunctive-particle clauses, if many, are also thought to affect ease of understanding. Note that not this "average size" but "maximum number" of consecutive infinitive-form or conjunctive-particle clauses holds significant meaning as an indicator of the depth of dependency affecting ease of understanding.

2.2 Diversity of vocabulary

Yule (1944) used a variety of statistical quantities in his analysis of writing. The most famous of these is an index of vocabulary concentration called the K characteristic value. The value of K is non-negative, increases as vocabulary becomes more concentrated, and conversely, decreases as vocabulary becomes more diversified. The median values of K for editorials and columns in the Mainichi Daily News were found to be 87.3 and 101.3, respectively. Incidentally, other characteristic values indicating vocabulary concentration have been proposed. See Tweedie et al. (1998), for example.

2.3 Percentage of big words

It is thought that the use of big words, to whatever extent, cannot help but impress the reader. On investigating big words in Japanese, however, care must be taken because simply measuring the length of a word may lead to erroneous conclusions. While "big word" in English is usually synonymous with "long word," a word expressed in kanji becomes longer when expressed in kana characters. That is to say, a "small word" in Japanese may become a big word simply due to notation. The number of characters in a word must therefore be counted after converting it to kana characters (i.e., to its "reading") to judge whether that word is big or small. In editorials from the Mainichi Daily News, the median number of characters in nouns after conversion to kana was found to be 4, with 5 being the 3rd quartile (upper 25%). We therefore assumed for the time being that nouns having readings of 6 or more characters were big words, and with this as a guideline, we again measured the percentage of nouns in a document that were big words. Since the number of characters in a reading is an integer value, this percentage would not necessarily be 25%, but a distribution that takes a value near that percentage on average can be obtained.

2.4 Percentage of passive sentences

It is generally felt that text should be written in active voice as much as possible, and that text with many passive sentences is poor writing (Knuth et al., 1988). For this reason, the percentage of passive sentences is also used as an index of rhetoric. Grammatically speaking, passive voice is distinguished from active voice in Japanese by the auxiliary verbs "reru" and "rareru". In addition to passivity, however, these two auxiliary verbs can also indicate respect, possibility, and spontaneity. In fact, they may be used to indicate respect even in the case of active voice. This distinction, however, while necessary in analysis at the semantic level, is not used in morphological analysis and syntactic analysis. For example, in the case that the object

of respect is "teacher" (sensei) or "your husband" (goshujin), the use of "reru" and "rareru" auxiliary verbs here would certainly indicate respect. This meaning, however, belongs entirely to the world of semantics. We can assume that such an indication of respect would not be found in essays required for tests, and consequently, that the use of "reru" and "rareru" in itself would indicate the passive voice in such an essay.

3 Organization

Comprehending the flow of a discussion is essential to understanding the connection between various assertions. To help the reader to catch this flow, the frequent use of conjunctive expressions is useful. In Japanese writing, however, the use of conjunctive expressions tends to alienate the reader, and such expressions, if used at all, are preferably vague. At times, in fact, presenting multiple descriptions or posing several questions seeped in ambiguity can produce interesting effects and result in a beautiful passage (Noya, 1997). In essays tests, however, examinees are not asked to come up with "beautiful passages." They are required, rather, to write logically while making a conscious effort to use conjunctive expressions. We therefore attempt to determine the logical structure of a document by detecting the occurrence of conjunctive expressions. In this effort, we use a method based on cue words as described in Quirk et al. (1985) for measuring the organization of a document. This method, which is also used in e-rater, the basis of our system, looks for phrases like "in summary" and "in conclusion" that indicate summarization, and words like "perhaps" and "possibly" that indicate conviction or thinking when examining a matter in depth, for example. Now, a conjunctive relationship can be broadly divided into "forward connection" and "reverse connection." "Forward connection" has a rather broad meaning indicating a general conjunctive structure that leaves discussion flow unchanged. In contrast, "reverse connection" corresponds to a conjunctive relationship that changes the flow of discussion. These logical structures can be classified as follows according to Noya (1997). The "forward connection" structure comes in the following types.

Addition: A conjunctive relationship that adds emphasis. A good example is "in addition," while other examples include "moreover" and "rather." Abbreviation of such words is not infrequent.

Explanation: A conjunctive relationship typified by words and phrases such as "namely," "in short," "in other words," and "in summary." It can be broken down further into "summarization" (summarizing and clarifying what was just described), "elaboration" (in contrast to "summarization," begins with an overview followed by a detailed description), and "substitution" (saying the same thing in another way to aid in understanding or to make a greater impression).

Demonstration: A structure indicating a reason-consequence relation. Expressions indicating a reason include "because" and "the reason is," and those indicating a consequence include "as a result," "accordingly," "therefore," and "that is why." Conjunctive particles in Japanese like "node" (since) and "kara" (because) also indicate a reason-consequence relation.

Illustration: A conjunctive relationship most typified by the phrase "for example" having a structure that either explains or demonstrates by example.

The "reverse connection" structure comes in the following types.

Transition: A conjunctive relationship indicating a change in emphasis from A to B expressed by such structures as "A ..., but B..." and "A...; however, B...).

Restriction: A conjunctive relationship indicating a continued emphasis on A. Also referred to as a "proviso" structure typically expressed by "though in fact" and "but then."

Concession: A type of transition that takes on a conversational structure in the case of concession or compromise. Typical expressions indicating this relationship are "certainly" and "of course."

Contrast: A conjunctive relationship typically expressed by "at the same time," "on the other hand," and "in contrast."

We extracted all $(= 125)$ phrases indicating conjunctive relationships from editorials of the Mainichi Daily News, and classified them into the above four categories for forward connection and

those for reverse connection for a total of eight exclusive categories. In Jess, the system attaches labels to conjunctive relationships and tallies them to judge the strength of the discourse in the essay being scored. As in the case of rhetoric, Jess learns what an appropriate number of conjunctive relationships should be from editorials of the Mainichi Daily News, and deducts from the initially allotted points in the event of an outlier value in the model distribution.

In the scoring, we also determined whether the pattern in which these conjunctive relationships appeared in the essay was singular compared to that in the model editorials. This was accomplished by considering a trigram model (Jelinek, 1991) for the appearance patterns of forward and reverse connections. In general, an N-gram model can be represented by a stochastic finite automaton, and in a trigram model, each state of an automaton is labeled by a symbol sequence of length 2. The set of symbols here is $\Sigma = \{a$: forward-connection, b : reverse-connection$\}$. Each state transition is assigned a conditional output probability as shown in Table 1. The symbol $\sqcup$ here indicates no (prior) relationship. The initial state is shown as $\sqcup\sqcup$. For example, the expression $P(a|\sqcup\sqcup)$ signifies the probability that "a : forward connection" will appear as the initial state.

Table 1: State transition probabilities on $\{a$: forward-connection, b : reverse-connection$\}$

$P(a\|\sqcup a) = 0.48$	$P(b\|\sqcup a) = 0.52$	$P(a\|\sqcup b) = 0.36$
$P(b\|\sqcup b) = 0.64$	$P(a\|aa) = 0.35$	$P(b\|aa) = 0.65$
$P(a\|ab) = 0.55$	$P(b\|ab) = 0.44$	$P(a\|ba) = 0.28$
$P(b\|ba) = 0.72$	$P(a\|bb) = 0.35$	$P(b\|bb) = 0.65$
$P(a\|\sqcup\sqcup) = 0.44$	$P(b\|\sqcup\sqcup) = 0.56$	

In this way, the probability of occurrence of certain $\{a$: forward-connection$\}$ and $\{b$: reverse-connection$\}$ patterns can be obtained by taking the product of appropriate conditional probabilities listed in Table 1. For example, the probability of occurrence p of the pattern $\{a, b, a, a\}$ turns out to be $0.44\times0.52\times0.55\times0.28 = 0.035$. Furthermore, given that the probability of $\{a\}$ appearing without prior information is 0.47 and that of $\{b\}$ appearing without prior information is 0.53, the probability q that a forward connection occurs three times and a reverse connection once under the condition of no prior information would be $0.47^3 \times 0.53 = 0.055$. As shown by this example, an occurrence probability that is greater for no prior information would indicate that the forward-connection and reverse-connection appearance pattern is singular, in which case the points initially allocated to conjunctive relationships in a discussion would be reduced. The trigram model may overcome the restrictions that the essay should be written in a pyramid structure or the reversal.

4 Content

A technique called latent semantic indexing can be used to check whether the content of a written essay responds appropriately to the essay prompt. The usefulness of this technique has been stressed at the Text REtrieval Conference (TREC) and elsewhere. Latent semantic indexing begins after performing singular value decomposition on $t \times d$ term-document matrix X (t : number of words; d : number of documents) indicating the frequency of words appearing in a sufficiently large number of documents. Matrix X is generally a huge sparse matrix, and SVDPACK (Berry, 1992) is known to be an effective software package for performing singular value decomposition on a matrix of this type. This package allows the use of eight different algorithms, and Ishioka and Kameda (1999) give a detailed comparison and evaluation of these algorithms in terms of their applicability to Japanese documents. Matrix X must first be converted to the Harwell-Boeing sparse matrix format (Duff et al., 1989) in order to use SVDPACK. This format can store the data of a sparse matrix in an efficient manner, thereby saving disk space and significantly decreasing data read-in time.

5 Application

5.1 An E-rater Demonstration

An e-rater demonstration can be viewed at www.ets.org, where by clicking "Products→e-rater Home→Demo." In this demonstration, seven response patterns (seven essays) are evaluated. The scoring breakdown, given a perfect score of six, was one each for scores of 6, 5, 4, and 2 and three for a score of 3.

We translated essays A-to-G on that Web site into Japanese and then scored them using Jess, as shown in Table 2.

The second and third columns show e-rater and Jess scores, respectively, and the fourth column shows the number of characters in each essay.

Table 2: Comparison of scoring results

Essay	E-rater	Jess	No. of Characters	Time (s)
A	4	6.9 (4.1)	687	1.00
B	3	5.1 (3.0)	431	1.01
C	6	8.3 (5.0)	1,884	1.35
D	2	3.1 (1.9)	297	0.94
E	3	7.9 (4.7)	726	0.99
F	5	8.4 (5.0)	1,478	1.14
G	3	6.0 (3.6)	504	0.95

A perfect score in Jess is 10 with 5 points allocated to rhetoric, 2 to organization, and 3 to content as standard. For purposes of comparison, the Jess score converted to e-rater's 6-point system is shown in parentheses. As can be seen here, essays given good scores by e-rater are also given good scores by Jess, and the two sets of scores show good agreement. However, e-rater (and probably human raters) tends to give more points to longer essays despite similar writing formats. Here, a difference appears between e-rater and Jess, which uses the point-deduction system for scoring. Examining the scores for essay C, for example, we see that e-rater gave a perfect score of 6, while Jess gave only a score of 5 after converting to e-rater's 6-point system. In other words, the length of the essay could not compensate for various weak points in the essay under Jess's point-deduction system. The fifth column in Table 2 shows the processing time (CPU time) for Jess. The computer used was Plat'Home Standard System 801S using an 800-MHz Intel Pentium III running RedHat 7.2. The Jess program is written in C shell script, jgawk, jsed, and C, and comes to just under 10,000 lines. In addition to the ChaSen morphological analysis system, Jess also needs the kakasi kanji/kana converter program (http://kakasi.namagu.org/) to operate. At present, it runs only on UNIX. Jess can be executed on the Web at http://coca.rd.dnc.ac.jp/jess/.

5.2 An Example of using a Web Entry Sheet

Four hundred eighty applicants who were eager to be hired by a certain company entered their essays using a Web form without a time restriction, with the size of the text restricted implicitly by the Web screen, to about 800 characters. The theme of the essay was "What does working mean in your life." Table 3 summarizes the correlation coefficients between the Jess score, average score of expert raters, and score of the linguistic understanding test (LUT), developed by Recruit Management Solutions Co., Ltd. The LUT is designed to measure the ability to grasp the correct meaning of words that are the elements of a sentence, and to understand the composition and the summary of a text. Five expert raters reted the essays, and three of these scored each essay independently.

Table 3: Correlation between Jess score, average of expert raters, and linguistic understanding test

	Jess	Ave. of Experts
Ave. of Experts	0.57	
LUT	0.08	0.13

We found that the correlation between the Jess score and the average of the expert raters' scores is not small (0.57), and is larger than the correlation coefficient between the expert raters' scores of 0.48. That means that Jess is superior to the expert raters on average, and is substitutable for them. Note that the restriction of the text size (800 characters in this case) caused the low correlation owing to the difficulty in evaluating the organization and the development of the arguments; the essay scores even in expert rater tend to be dispersed.

We also found that neither the expert raters nor Jess, had much correlation with LUT, which shows that LUT does not reflect features indicating writing ability. That is, LUT measures quite different laterals from writing ability.

Another experiment using 143 university-students' essays collected at the National Institute for Japanese Language shows a similar result: for the essays on "smoking," the correlation between Jess and the expert raters was 0.83, which is higher than the average correlation of expert raters (0.70); for the essays on "festivals in Japan," the former is 0.84, the latter, 0.73. Three of four raters graded each essay independently.

6 Conclusion

An automated Japanese essay scoring system called Jess has been created for scoring essays in college-entrance exams. This system has been shown to be valid for essays of 800 to 1,600 characters. Jess, however, uses editorials and columns taken from the Mainichi Daily News newspaper as learning models, and such models are not sufficient for learning terms used in scientific and technical fields such as computers. Consequently, we found that Jess could return a low evaluation of "content" even for an essay that responded well to the essay prompt. When analyzing content, a mechanism is needed for automatically selecting

a term-document cooccurrence matrix in accordance with the essay targeted for evaluation. This enable the users to avoid reverse-engineering that poor quality essays would produce perfect scores, because thresholds for detecting the outliers on rhetoric features may be varied.

Acknowledgements

We would like to extend their deep appreciation to Professor Eiji Muraki, currently of Tohoku University, Graduate School of Educational Informatics, Research Division, who, while resident at Educational Testing Service (ETS), was kind enough to arrange a visit for us during our survey of the e-rater system.

References

Bennet, R.E. and Bejar, I.I. 1998. Validity and automated scoring: It's not only the scoring, *Educational Measurement: Issues and Practice*. 17(4):9–17.

Bereiter, C. 2003. Foreword. In Shermis, M. and Burstein, J. eds. *Automated essay scoring: cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berry, M.W. 1992. Large scale singular value computations, *International Journal of Supercomputer Applications*. 6(1):13–49.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris, M.D. 1998. Automated Scoring Using A Hybrid Feature Identification Technique. *the Annual Meeting of the Association of Computational Linguistics*, Available online: www.ets.org/research/erater.html

Chase, C.I. 1986. Essay test scoring : interaction of relevant variables, *Journal of Educational Measurement*, 23(1):33–41.

Chase, C.I. 1979. The impact of achievement expectations and handwriting quality on scoring essay tests, *Journal of Educational Measurement*, 16(1):293–297.

Cooper, P.L. 1984. The assessment of writing ability: a review of research, *GRE Board Research Report*, GREB No.82-15R. Available online: www.gre.org/reswrit.html#TheAssessmentofWriting

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(7):391–407.

Duff, I.S., Grimes, R.G. and Lewis, J.G. 1989. Sparse matrix test problem. *ACM Trans. Math. Software*, 15:1–14.

Elliot, S. 2003. IntelliMetric: From Here to Validity, 71–86. In Shermis, M. and Burstein, J. eds. *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P.W., Laham, D. and Landauer, T.K. 1999. Automated Essay Scoring: Applications to Educational Technology. *EdMedia '99*.

Hughes, D.C., Keeling B. and Tuck, B.F. 1983. The effects of instructions to scorers intended to reduce context effects in essay scoring, *Educational and Psychological Measurement*, 43:1047–1050.

Ishioka, T. and Kameda, M. 1999. Document retrieval based on Words' cooccurrences — the algorithm and its applications (in Japanese), *Japanese Journal of Applied Statistics*, 28(2):107–121.

Jelinek, F. 1991. Up from trigrams! The struggle for improved Language models, *the European Conference on Speech Communication and Technology (EUROSPEECH-91)*, 1037–1040.

Knuth, D.E., Larrabee, T. and Roberts, P.M. 1988. *Mathematical Writing*, Stanford University Computer Science Department, Report Number: STAN-CS-88-1193.

Maekawa, M. 1995. *Scientific Analysis of Writing* (in Japanese), ISBN4-00-007953-0, Iwanami Shotton.

Marshall, J.C. and Powers, J.M. 1969. Writing neatness, composition errors and essay grades, *Journal of Educational Measurement*, 6(2):97–101.

Meyer, G. 1939. The choice of questions on essay examinations, *Journal of Educational Psychology*, 30(3):161–171.

Nagao, M.(ed.) 1996. *Natural Language Processing* (in Japanese), The Iwanami Software Science Series 15, ISBN 4-00-10355-5,

Noya, S.: 1997. *Logical Training* (in Japanese), Sangyo Tosho, ISBN 4-7828-0205-6.

Page, E.B., Poggio, J.P. and Keith, T.Z. 1997. Computer analysis of student essays: Finding trait differences in the student profile. *AERA/NCME Symposium on Grading Essays by Computer*.

Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M.E., and Kukich, K. 2000. Comparing the validity of automated and human essay scoring, GRE No. 98-08a. Princeton, NJ: Educational Testing Service.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*, Longman.

Rudner, L.M. and Liang, L. 2002. *National Council on Measurement in Education*, New Orleans, LA. Available online: http://ericae.net/betsy/papers/n2002e.pdf

Tweedie, F.J. and Baayen, R.H. 1998 How Variable May a Constant Be? Measures of Lexical Richness in Perspective, *Computers and the Humanities*, 32:323–352.

Watanabe, H., Taira, Y. and Inoue, T. 1988 An Analysis of Essay Examination Data (in Japanese), Research bulletin, Fuculty of Education, University of Tokyo, 28:143–164.

Yule, G.U. 1944. *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.