Automated Japanese Essay Scoring System : Jess

Tsunenori ISHIOKA

Research Division
National Center for Univ. Entrance Exam.
Tokyo, Japan 153-8501

Masayuki KAMEDA

Software Research Center
Ricoh Co., Ltd.
Tokyo, Japan 112-0002

Abstract

We have developed an automated Japanese essay scoring system named jess. The system evaluates an essay from three features: (1) Rhetoric — ease of reading, diversity of vocabulary, percentage of big words (long, difficult words), and percentage of passive sentences, (2) Organization — characteristics associated with the orderly presentation of ideas, such as rhetorical features and linguistic cues, (3) Contents — vocabulary related to the topic, such as relevant information and precise or specialized vocabulary. The final evaluated 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. Our system does not need any essays graded by human experts.

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. Most of the factors are present in giving tests, and the human "rater," in particular, is a major error factor in the scoring of essays. In fact, there are many other factors that influence the scoring of essay tests as listed below, and much research has been devoted to them.

- Handwriting skill (handwriting quality, spelling)
- Serial effects of rating (the order in which essay answers are rated)
- Topic selection (how should essays written on different topics be rated?)
- Other error factors (writer's gender, ethnic group, etc.)

In recent years, with the aim of removing these error factors and to establish fairness, considerable research has been performed on computer-based automated essay scoring systems [1, 3, 11, 12]. The most famous of these is probably e-rater [1] developed by the Educational Testing Service (ETS) in the United States and currently managed and extended by ETS Technologies, a subsidiary organization. E-rater is presently being used to score essays in the Graduate Management Admission Test (GMAT), an entrance examination for business graduate schools. E-rater evaluates essays from the following three points of view.

Structure: syntactic variety, i.e., use of diverse structures in arrangement of phrases, clauses, and sentences.

Organization: logical presentation of ideas using rhetorical expressions, logical connectors between clauses and sentences, etc.

Contents: use of vocabulary related to the topic.

The e-rater system features a database of hundreds of essays scored by expert readers. Performing linear regression against those expert scores and computer-based scores makes it possible to determine regression coefficients for multiplying the matrices used in scoring. In Japan, however, there is no collection of such authorized scores, and after careful consideration, it was decided that the same kind of approach would not be practical for implementing a Japanese version of e-rater. It is possible, though, to obtain complete articles from the Mainichi Daily News newspaper up to 2002 from Nichigai Associates, Inc. and complete articles from the Nihon Keizai newspaper up to 2001 from Nikkei Books and Software, Inc. for purposes of linguistic studies. 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. 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 developed by the Language Media Laboratory of Kyoto University; ChaSen (http://chasen.aist-nara.ac.jp/; used by the authors

Proceedings of the 15th International Workshop on Database and Expert Systems Applications (DEXA'04)
1529-4188/04 $ 20.00 IEEE

in this study) from the Matsumoto Laboratory of the Nara Institute of Science and Technology. Likewise, for syntactic analysis, there are free resources such as KNP from Kyoto University.

With resources such as these, we can prepare 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 in Ishioka [5] and elsewhere. It is the authors' belief that this learning approach to published essays and columns as models makes it possible to develop a system essentially the same as e-rater, that is, an automated scoring system for essays written in Japanese, but using technically superior methods.

We have named this automated Japanese essay scoring system "jess." This system evaluates essays based on the three essay features of (1) rhetoric, (2) organization, and (3) contents, which are basically the same as structure, organization, and contents 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 point allotment, default weights are 5, 2, and 3 for structure, organization, and contents, respectively, for a total of 10 points. This default point allotment of 5, 2, and 3 in which "rhetoric" is weighted higher than "organization" and "contents" is based on the work of Watanabe et al. [13]. Users can change the point allotment.

The following sections describe the scoring criteria of jess in detail. Sections 2, 3, and 4 examine rhetoric, organization, and contents, respectively. Section 5 presents an application example and associated operation times.

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 [8] and Nagao [9]. 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 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 as the authors consider it an index contributing to not only "rhetoric" but to "contents" as well.

2.1 Ease of reading

The following items are considered as indexes of "ease of reading." These indexes do not agree with usual reading complexity [4].

1. Median and maximum sentence length:

   It is generally assumed that shorter sentences make for easier reading [7]. 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 why the median value is used as opposed to the average value 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.

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, the authors found it to have a median value 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.

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

   The declined or conjugated words 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 [14] 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, and increases as vocabulary becomes more concentrated, and conversely decreases as vocabulary becomes more diversified. The median value of K for editorials and columns in the Mainichi Daily News was found to be 87.3 and 101.3, respectively.

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 since simply measuring the length of a word may lead to erroneous conclusions. While a "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. It is therefore necessary to count the number of characters in a word after converting it to kana characters (i.e., to its "reading") to judge whether that word is big or small.

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 an example of poor writing [7]. It is for this reason that percentage of passive sentences is also used as an index of rhetoric.

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. We therefore attempt to determine the logical structure of a document by detecting the occurrence of conjunctive expressions. 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 [10]. 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."

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 that 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 [6] for the appearance patterns of forward and reverse connections. 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. For example, the probability of occurrence p of the pattern $\{a, b, a, a\}$ turns out to be $0.44 \times 0.52 \times 0.55 \times 0.28 = 0.035$. 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.

4 Contents

A technique called latent semantic indexing (LSI) [2] can be used to check whether the contents 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. The process extracts diagonal elements from singular value matrix up to the kth element to form a new matrix S. Likewise, it extracts left and right hand singular value decomposition matrices up to the kth column to form new matrices T and D. Matrix $\widehat{X}$ can be expressed as follows.

$$\widehat{X} = TSD'$$

Here, $\widehat{X}$ is an approximation of X with T and S being $t \times k$ and $k \times k$ square diagonal matrices, respectively, and D' a $k \times d$ matrix. The $'$ symbol denotes transposition. According to Deerwester [2], a k of from 50 to 100 is sufficient for linguistic data based on empirical results.

Essay e to be scored can be expressed by t-dimension word vector x_e based on morphological analysis, and using this, $1 \times k$ document vector d_e corresponding to a row in document space D can be derived as follows.

$$d_e = x_e' T S^{-1}$$

Similarly, k-dimension vector d_q corresponding to essay prompt q can be obtained. Similarity between these documents is denoted by $r(d_e, d_q)$, which can be given by the cosine of the angle formed between the two document vectors. Note that the normalization of sizes of two document vectors is not necessary. Theoretically speaking, r can take on negative values, but setting its lower limit to zero appears to be appropriate here.

5 Application Example

An e-rater demonstration can be viewed at http://www.etctechnologies.com/html/eraterdemo.html. In this demonstration, seven response patterns (seven essays) are evaluated. We translated essays A-to-G on that Web site into Japanese and then scored them using jess as shown in Table 1.

Table 1: 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

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

in jess is 10 with 5 points allocated to rhetoric, 2 to organization, and 3 to contents as standard. For purposes of comparison, the jess score converted to e-rater's 6-point system is shown in parentheses. It can be seen here that essays given good scores by e-rater are also given good scores by jess and that 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. It is here where a difference between e-rater and jess, which uses the point-deduction system for scoring, appears. 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 1 shows jess processing time (CPU time). Further research by using 590 essays proves that jess has same degree of the performance of human experts. 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. Jess can be executed on the Web at http://zaza.rd.dnc.ac.jp/jess/.

6 Conclusion

An automated Japanese essay scoring system called "jess" has been created for use in scoring essays in college-entrance exams. This system has been shown to be valid for essays in the range of 800 to 1600 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. It was consequently found that jess could return a low evaluation of "contents" even for an essay that responds well to the essay prompt. When analyzing contents, a mechanism is needed for automatically selecting a term-document cooccurrence matrix in accordance with the essay targeted for evaluation.

Acknowledgement

The authors 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

[1] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder & M. D. Harris, Automated Scoring Using A Hybrid Feature Identification Technique. In the Proceedings of *the Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, 1998.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer & R. Harshman, Indexing by latent semantic analysis. *Journal of the American Society for Information Science* Vol. 41 No. 7, pp. 391–407, 1990.

[3] P. W. Foltz, D. Laham & T. K. Landauer, Automated Essay Scoring: Applications to Educational Technology. In proceedings of *EdMedia '99.*, 1999.

[4] R. Gunning, The Technique of Clear Writing, New York, McGraw Hill, 1968.

[5] T. Ishioka & M. Kameda, Document retrieval based on Words' cooccurrences — the algorithm and its applications" (in Japanese), Japanese Journal of Applied Statistics, Vol. 28, No. 2, pp. 107–121, 1999.

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

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

[8] M. Maekawa, *Scientific Analysis of Writing* (in Japanese), Iwanami Shotten, 1995.

[9] M. Nagao (ed.), *Natural Language Processing* (in Japanese), The Iwanami Software Science Series **15**, Iwanami Shotten, 1996.

[10] S. Noya *Logical Training* (in Japanese), Sangyo Tosho, 1997.

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

[12] L. M. Rudner & L. Liang, Automated essay scoring using Bayes' theorem, *National Council on Measurement in Education*, New Orleans, LA., 2002.

[13] H. Watanabe, Y. Taira & T. Inoue, An Analysis of Essay Examination Data (in Japanese), *Research bulletin, Faculty of Education, University of Tokyo*, Vol. 28, pp. 143–164, 1988.

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