

An Unsupervised Automated Essay-Scoring System

Yen-Yu Chen, *Industrial Technology Research Institute*

Chien-Liang Liu and Chia-Hoang Lee, *National Chiao Tung University*

Tao-Hsing Chang, *National Kaohsiung University of Applied Sciences*

Automated essay scoring (AES) is the ability of computer technology to evaluate and score written prose. Proposed in 1966, AES has since been used successfully on large-scale essay exams. The goal is not to replace human raters. In current large exams, each essay is scored by two or more human

raters, and the final scores are averaged over these scores. For example, in the Graduate Record Examination (GRE) analytical writing section, two trained readers score each essay. If there is more than a one-point difference between the two readers' scores, then a third reader grades the essay, and the score for that essay will be the average of the two highest scores. In general, the whole essay-scoring process is time consuming and requires considerable manpower. Therefore, instead of having two people score the essays, each essay could be scored by AES and a human rater, with the final then determined by both. The combined approach would still require the AES system and the human rater to assign a score within one scale point of each other. Otherwise, a third human rater would resolve the discrepancy.

Companies such as Vantage Learning and ETS Technologies have published research results that demonstrate strong correlations and nonsignificant differences between AES and human scoring.¹ In essence, the human

raters grade the essays according to some criteria. For example, the GRE analytical writing score is based on a strong focus on the topic, good evidence to support arguments, and proper use of grammar. If an essay includes all of these factors, it could earn a top score. Therefore, the aim of AES systems is to simulate a human rater's grading process, and a system is usable only if it can perform the grading as accurately as human raters.

In this article, we propose an unsupervised AES system that requires only a small number of essays within the same topic without any scoring information. (See the "Related Research in Automated Essay Scoring" sidebar for details on other approaches.) The scoring scheme is based on feature information and the similarities between essays. We use a voting algorithm based on the initial scores and similarities between essays to iteratively train the system to score the essays. Our experiments yield an adjacent agreement rate of approximately 94 percent and

The proposed automated essay-scoring system uses an unsupervised learning approach based on a voting algorithm. Experiments show that this approach works well compared to supervised learning approaches.

Related Research in Automated Essay Scoring

Automated Essay Scoring (AES) has been a real and viable alternative and complement to human scoring for many years. In 1996, Ellis Page designed the Project Essay Grader (PEG) computer grading program.¹ Page looked for the kind of textual features that computers could extract from the texts and then applied multiple linear regressions to determine an optimal combination of weighted features that best predicted the teachers' grades. The features Page identified as having predictive power included word length and the number of words, commas, prepositions, and uncommon words in the essay. Page called these features proxies for some intrinsic qualities of writing competence. He had to use indirect measures because of the computational difficulty of implementing more direct measures.²

Because it only uses indirect features, however, this type of system is vulnerable to cheating. Therefore, it is a significant research challenge to identify and extract more direct measures of writing quality. For example, later research used machine learning to identify discourse elements based on an essay-annotation protocol.³ Meanwhile, many researchers used natural language processing (NLP) and information retrieval (IR) techniques to extract linguistic features that might more directly measure essay qualities.

During the late 1990s, more systems were developed, including the Intelligent Essay Assessor (IEA), e-rater, and IntelliMetric. IntelliMetric successfully scored more than 370,000 essays in 2006 for the Analytical Writing Assessment (AWA) portion of the Graduate Management Admission Test (GMAT).

Intelligent Essay Assessor (IEA) uses latent semantic analysis (LSA) to analyze essay semantics.⁴ The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. LSA captures transitivity relations and collocation effects among vocabulary terms, thereby letting it accurately judge the semantic relatedness of two documents regardless of their vocabulary overlap.⁵

IEA measures the content, style, and mechanics components separately and, whenever possible, computes each component in the same way so that score interpretation is comparable across applications. The system must be trained on a set of domain-representative texts to measure an essay's overall quality. For example, a biology textbook could be used when scoring biology essays. LSA characterizes student essays by representing their meaning and compares them with highly similar texts of known quality. It adds corpus-statistical writing-style and mechanics measures to

help determine overall scoring, validate an essay as appropriate English (or other language), detect plagiarism or attempts to fool the system, and provide tutorial feedback.⁶

E-rater employs a corpus-based approach to model building, using actual essay data to examine sample essays. The features of e-rater include syntactic, discourse, and topical-analysis modules. The origin of the syntactic module is parsing. In discourse analysis, it assumes the essay can be segmented into sequences of discourse elements, which include introductory material, a thesis statement, main ideas, supporting ideas, and a conclusion.⁷ To identify the various discourse elements, the system was trained on a large corpus of human-annotated essays. Finally, the topical-analysis module identifies vocabulary usage and topical content. In practice, a good essay must be relevant to the topic assigned. Moreover, the variety and type of vocabulary used in good essays differ from that of poor essays. The assumptions behind this module are that good essays resemble other good essays.

In recent years, many supervised-learning approaches on essay-scoring systems have been proposed. Lawrence M. Rudner and Tahung Liang used a Bayesian approach to perform AES, showing the effectiveness of the supervised-learning approach for essays.⁸ Essentially, the supervised-learning model needs enough labeled data to construct the classification model. Our experiments indicate that such approaches require at least 200 scored essays, which make them inappropriate for environments where there are not enough scored essays.

References

1. E.B. Page, "The Imminence of Grading Essays by Computer," *Phi Delta Kappan*, vol. 47, 1966, pp. 238–243.
2. K. Kukich, "Beyond Automated Essay Scoring," *IEEE Intelligent Systems*, vol. 15, no. 5, 2000, pp. 22–27.
3. J. Burstein, D. Marcu, and K. Knight, "Finding the Write Stuff: Automatic Identification of Discourse Structure in Student Essays," *IEEE Intelligent Systems*, vol. 18, no. 1, 2003, pp. 32–39.
4. T. Landauer and S. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Rev.*, vol. 104, no. 2, 1997, pp. 211–240.
5. M.A. Hearst, "The Debate on Automated Essay Grading," *IEEE Intelligent Systems*, vol. 15, no. 5, 2000, pp. 22–37.
6. T.K. Landauer, D. Laham, and P.W. Foltz, "The Intelligent Essay Assessor," *IEEE Intelligent Systems*, vol. 15, no. 5, 2000, pp. 27–31.
7. Y. Attali and J. Burstein, "Automated Essay Scoring with E-Rater v.2," *J. Technology, Learning and Assessment*, vol. 4, no. 3, 2006, <http://escholarship.bc.edu/jtla/vol4/3>.
8. L.M. Rudner and T. Liang, "Automated Essay Scoring Using Bayes' Theorem," *J. Technology, Learning and Assessment*, vol. 1, no. 2, 2002, <http://escholarship.bc.edu/jtla/vol1/2>.

an exact agreement rate of approximately 52 percent.

Overview of Unsupervised Learning

In supervised learning, we can regard the AES as a classification learner and

the scores of the training essays as the training data categories. New essays will be classified into an appropriate category based on the features and the classification model. On the other hand, the training data in an unsupervised-learning classifier does not

contain label information, so the classifier must determine how the data is organized from unlabeled examples. We propose a novel unsupervised-learning method and apply it to an essay-scoring application without scored essays as the training data.

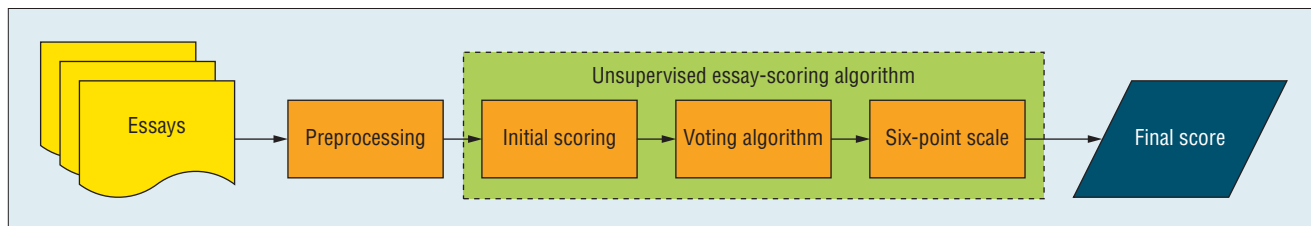


Figure 1. Automated essay-scoring system flow. Our approach includes preprocessing and an unsupervised essay-scoring algorithm.

A traditional clustering algorithm such as the k -means clustering algorithm needs users to determine the number of clusters before performing the algorithm. In this work, the number of clusters is equal to the number of grading levels. We divide the essay scores into six levels. Therefore, there are six clusters, and the essays with the same scores could be regarded as being in the same cluster.

The unsupervised learning we propose is based on a voting algorithm. We give each essay an initial score, which is transformed into the standard score. We obtain this Z -score from

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

where X is the score from the original normal distribution, μ is the mean of the original normal distribution, and σ is the standard deviation of original normal distribution. In statistics, the Z -score reflects how many standard deviations an observation is above or below the mean.

Each essay's Z -score will be changed iteratively. In each iteration, the score of an essay, which we call a *target essay*, is determined by the rest of the essays, which we call *voting essays*, based on two factors. One of the factors is the similarity between the target and voting essays. The other factor is the voting essay's Z -score. In other words, the voting essays are prone to attract the essays that are similar to them and that will lead to essay clustering.

The frequency histogram of the essay scores in a group could be

approximated as a normal distribution, and the normal distribution could be transformed into the standard normal distribution, where the area under the curve over the event description could be obtained based on a Z table. Thus, when the voting algorithm finishes, the essay's Z -score will be obtained and the Z -score could be used to determine the essay's score location.

Unsupervised-Learning Approach on Essay Scoring

Essentially, clustering is the assignment of objects into groups so that objects from the same cluster are more similar to each other than objects from different clusters. In practice, the *distance measurement*—which determines the similarity of two elements—plays an important role in clustering applications. Common distance functions include Euclidean distance, Manhattan distance, and hamming distance. Because the data set we use is text content, the terms in the essays will become the basis of the distance function in this article.

Our data set contains the essays written by junior high school students on the topic “Recess at School.” Our similarity function should be able to measure the similarities between two essays. Essentially, the similarities between essays should include the sharing terms, content organization, the semantics of the terms, and so forth. Traditionally, techniques for detecting similarity between long texts (documents) have centered on analyzing shared words.²

Therefore, we use the bag-of-words model to represent the essay content and then we represent the text as an unordered collection of words, disregarding grammar and even word order. As a result, the similarity between essay i and essay j is based on the shared terms between two essays.

Figure 1 shows the system flow, which includes preprocessing and an unsupervised essay-scoring algorithm. The unsupervised essay-scoring algorithm includes initial scoring, the voting algorithm, and six-point-scale stages.

Preprocessing

Different languages might have different preprocessing processes. In general, the preprocessing stage includes words stemming, stop-words filtering, phrase identification, and so on. Unlike English, the Chinese language does not use spaces as a boundary to separate words in a sentence, so Chinese word segmentation is required at this stage. In this article, we use a maximum matching method to perform Chinese word segmentation. This algorithm extracts the longest substring that makes sense in the current string.

Performing the maximum matching method for segmenting Chinese texts is not the best-known Chinese word-segmentation algorithm, but the difference is negligible. We base the similarity of two essays on the number of shared terms rather than essays' semantics, so the maximum matching method does not influence our result.

```

1: for all  $i$  such that  $1 \leq i \leq N$  do
2:   Assign initial score to  $S_{i,0}$ 
3: end for
4: for all  $i$  such that  $1 \leq i \leq N$  do
5:   Calculate  $Z_{i,0}$  (Based on Equation 3)
6: end for
7: repeat
8:   for all  $j$  such that  $1 \leq j \leq N$  do
9:     Calculate  $S_{j,t}$  (Based on Equation 2)
10:   end for
11:   for all  $i$  such that  $1 \leq i \leq N$  do
12:     Calculate  $Z_{i,t}$  (Based on Equation 3)
13:   end for
14: until The change rate of the Z-score between consecutive iterations is less than  $\epsilon$ 
15: for all  $i$  such that  $1 \leq i \leq N$  do
16:   Assign essay  $[i]$ 's score based on its Z-score and historical distribution information
17: end for

```

Figure 2. Unsupervised essay-scoring algorithm. Lines 1 through 3 show the initial scoring stage, lines 7 through 14 give the voting algorithm stage, and lines 15 through 17 are the six-point-scale stage.

Unsupervised Essay-Scoring Algorithm

Based on the similarity function, we propose an unsupervised-learning algorithm for essay scoring to cluster the essays. Figure 2 shows the unsupervised essay-scoring algorithm, which includes initial scoring, the voting algorithm, and a six-point scale.

Initial Scoring. In our unsupervised essay-scoring algorithm, the scores of essays will change iteratively and the initial scores should be given by the system based on the features. The initial score is based on the feature that could reflect the quality of the essays in the beginning. First, we employed SPSS (www.spss.com), which is a computer software used for statistical analysis, to analyze the scatter plot between scores and the mean number of unique terms of essays in each score interval. The plot shows that the mean number of unique terms is related to the scores of essays. Second, the Pearson correlation value, which is used to measure the degree of the linear relationship between two variables (obtained using SPSS software) is 0.661, which means a positive relationship exists between these two variables. In other words,

it is reasonable to employ the number of unique terms as the initial scores of the essays.

Voting Algorithm. The core of the algorithm is a voting algorithm, where the other essays vote on an essay's score. Because the distribution of essay scores could be viewed as a normal distribution, we use the Z-score to normalize the essay score. The algorithm uses the following equations:

$$S_{j,t} = \sum_{i \neq j} Sim_{i,j} * Z_{i(t-1)} \quad (2)$$

$$Z_{i,t} = \frac{\left(S_{i,t} - \left(\sum_{k \neq i} S_{k,t} \right) / (N-1) \right)}{\sigma_t} \quad (3)$$

where $S_{j,t}$ is the score of essay j at time t , $Sim_{i,j}$ is the similarity between essays i and j , $Z_{i,t}$ is the Z-score of essay i at time t , N is the total number of essays, and σ_t is the standard deviation of the scores at time t .

In Equations 2 and 3, the *target essay* is the essay that needs to be computed, and the other essays are called *voting essays*. In Equation 2, essay j is a target essay and its score is determined by voting essays. The score

contributed by the voting essay i is determined by $Sim_{i,j}$, which represents the similarity between the essays j and i , and its previous Z-score $Z_{i,(t-1)}$. We use Equation 3 to compute the Z-score of essay i at time t based on the definition of Z-score.

If a voting essay's score is higher than the average at the previous iteration, the target essay would get a positive score; otherwise the target essay gets a negative score. The higher the score of a voting essay, the higher the score the target essay would get. In other words, every voting essay has a tendency to attract each target essay to its cluster. The similarity factor will give a voting essay more weight if it is similar to the target essay. The target essay's score will be closer to its similar essays' scores at each iteration.

As a result, the essays' initial scores could be obtained based on the number of unique terms and the similarity function is based on the number of sharing terms between two essays. We could apply the voting algorithm to all the essays, and the result will start to converge if the number of iterations is sufficient. According to our experiments, the system could converge within 50 iterations.

Six-Point Scoring Scale. When the voting algorithm reaches a stable state, an essay's final Z-score will be obtained and this Z-score could be mapped to the essay's grading scores. In this work, we evaluated each essay on a six-point scoring scale, with a six being the highest score. If historical distribution information is available, the cumulative percentage of essays graded from one to five could be transformed to corresponding Z-score intervals. On the other hand, if historical information is unavailable, we could use normal distribution to represent the score frequency distribution. With the

Table 1. Voting algorithm grading result along with the essays' shared term similarities.

Data sets	1	2	3	4	5	6	Adjacent agreement rate (%)	Exact agreement rate (%)
1 (23 essays)	17	5	1	0	0	0	95.7	73.9
2 (64 essays)	9	32	17	6	0	0	90.6	50.0
3 (105 essays)	2	15	46	36	6	0	92.4	43.8
4 (104 essays)	0	2	27	62	13	0	98.1	59.6
5 (46 essays)	0	0	0	23	23	0	100.0	50.0
6 (4 essays)	0	0	0	2	2	0	50.0	0
Total (346 essays)	28	54	91	129	44	0	94.5	52.0

historical Z-score interval information and an essay's Z-score information, we could determine the essay's score by referencing the historical Z-score interval where the essay's Z-score is located.

In this article, the historical information shows that the cumulative percentage of essays graded from one to five are 6.5, 25.1, 55.6, 85.8, and 99.0 percent, respectively, which we translated into -1.512, -0.671, 0.141, 1.070, and 2.320 for their Z-score representations.

The essays with a final Z-score less than -1.512 will be graded with a score of one, the essays with a final Z-score between -1.512 and -0.671 will be graded with a score of two, and so on.

Evaluation

We used two data sets to evaluate our approach. These essays, written by junior high school students, were graded by two or three teachers. The first set included 689 Chinese essays on the topic "Recess at School." We applied the voting algorithm and supervised-learning algorithms to this data set to compare the performance between different algorithms. Supervised-learning systems in the experiment used 346 essays as test

Table 2. The grading result of essays with random terms.

Number	Number of random unique terms	Score
1	30	0
2	30	0
3	30	0
4	100	0
5	100	0
6	100	0
7	200	0
8	200	0
9	200	0
10	500	0
11	500	0
12	500	0

data and the other 343 essays as training data. Meanwhile, the second data set included 342 Chinese essays on the topic "Dining Hours." We also applied the voting algorithm to this data set for comparison.

Table 1 shows the adjacent and exact agreement rates when applying the voting algorithm to 346 essays. The *exact agreement* occurs when two or more raters give an essay the exact same score. On the other hand, *adjacent agreement* requires two or more raters to assign a score within one scale point of each other.

Attack Experiments and Discussion

Because the features that the system uses come from indirect features, one

of the possible attacks is to employ unmeaningful terms to fool the system. Thus, we performed two kinds of attacks to simulate the scenarios that people might use to fool the system. For the first attack experiment, we used a different number of random terms to attack the system. As Table 2 shows, the number of random unique terms ranges from 30 to 500 and the score result shows that the system could detect these essays.

For the second attack experiment, we simulated a scenario where a student might use the content appearing in other literature to fool the system. The experiment's data set included literature written by famous scholars and writers, novels, Internet documents, news, and lyrics. This experiment also included an essay that consists of the sentences coming from six-point essays. The result shows that if the terms employed by these essays are not related to the essays' topic, the system could identify them and give them a zero score. If the students used sentences coming from the six-point essays, however, their essay scores would be five.

Because the system's feature space comes from a bag-of-words model, essays will get a better score if they

Table 3. The grading result for the “Recess at School” data set.

Approach	Adjacent agreement rate (%)	Exact agreement rate (%)
Voting system (sharing term)	94.5	52.0
Random rater	64.1	23.9
Three-rater	78.9	30.3
Support vector machine (SVM)	93.6	49.4
Bayesian	93.4	50.3

Table 4. The grading result for the “Dining Hours” data set.

System	Adjacent agreement rate (%)	Exact agreement rate (%)
Voting system (sharing term)	92.7	50.0
Random rater	68.7	26.7
Three rater	72.2	16.2
SVM	91.8	55.4
Bayesian	89.7	46.6

use the terms coming from the bag-of-words collection drawn from high-scoring essays. Meanwhile, if the users use the terms that are not associated with the topic, the system can detect these essays. As a result, in the first attack experiment, the essays listed in Table 2 could be given high initial scores in the beginning, but they will be identified as off-topic by the system in the end. The main reason is that the similarities between these essays and the other essays are low, so they could not earn high scores.

The second attack experiment shows that even if an essay is well written but is not within the bounds of the topic of the other essays, it will be identified as off-topic as well. However, if the students use the terms from other high-scoring essays, the system will give them a high score. However, if students could use the terms that are frequently used by high-scoring essays, it means that they know how to express their ideas about the subject to a certain extent.

Our experiments show that it is not easy to fool the system, and if students try to do so, the

human rater will easily identify their essays.

Comparison

Tables 3 and 4 list the exact and adjacent agreement rates of the experiments we describe in this article. The random-rater approach randomly grades the essays based on the essays’ score frequency distribution, while the three-rater approach means that all the essays are given three grades.

In addition, we implemented two supervised Chinese AES systems for comparison purposes. All the supervised-learning approaches take 343 essays as training data and use the remaining essays as testing data. The features employed by supervised-learning approaches include the number of terms, paragraphs, phrases, complete sentences, and sememes extracted from HowNet (see www.keenage.com), which is an online common-sense knowledge base unveiling interconceptual relations and interattribute relations of concepts in Chinese and English.

Furthermore, we applied the voting algorithm to the “Dining Hours” data set. Table 4 shows the experiment’s

results. The exact and adjacent agreement rates are similar to the first data set. In this experiment, the voting algorithm’s performance was not affected by the different data sets.

Our experiments show that our unsupervised-learning approach works well in the essay-scoring domain. As we mentioned earlier, the similarity function is based on a bag-of-words model. The underlying idea behind the voting algorithm is that good essays resemble other good essays, and it conforms to e-rater’s assumption (see the “Related Research in Automated Essay Scoring” sidebar). Because the essays are on the same topic, the students might describe the same activities using different terms. The low-scoring essays tend to describe these activities using common terms. On the other hand, the high-scoring essays tend to use graceful terms to describe these activities and use simile expressions. For example, “as white as snow” could be used to describe the color white, and such similes rarely appear in the low-scoring essays.

The advantage of this approach is that we could apply it to any language with a little modification because it does not use any specific language feature. The disadvantage is that it does not consider organization, style, and grammar features. In general, high-quality essays might involve a selection of creative expressions to reflect the writer’s point of view. The similarity function we use is not good at recognizing high-quality essays, and the experiment shows the result. However, our proposed design could be extended to integrate other scoring modules. For example, a two-phase essay-scoring system could be constructed to include other specific linguistic features.

THE AUTHORS

In the first phase, the voting algorithm could be applied to the essays to determine the essays' initial scores. The second phase could include other natural language processing (NLP) or information retrieval (IR) techniques to adjust the scores.

The attack experiments show that it is not easy to fool the system unless the users use the terms appearing in high-scoring essays. Currently, the limitation of this approach is that the essays must be on the same topic. In addition, the bag-of-words model makes it inapplicable to creative writing essays. ■

Yen-Yu Chen is an associate engineer in the Information and Communications Research Laboratories at the Industrial Technology Research Institute, Taiwan. His research interests include artificial intelligence, natural language processing, and automated essay scoring. Chen has an MS in computer science from National Chiao Tung University. Contact him at chenyy@itri.org.tw.

Chien-Liang Liu is a postdoc in the Department of Computer Science at National Chiao Tung University, Taiwan. His research interests include machine learning, natural language processing, and data mining. Liu has a PhD in computer science from National Chiao Tung University. Contact him at cliu@mail.nctu.edu.tw.

Chia-Hoang Lee is a professor in the Department of Computer Science and a senior vice president at National Chiao Tung University, Taiwan. His research interests include artificial intelligence, human-machine interface systems, and natural language processing. Lee has a PhD in computer science from the University of Maryland, College Park. Contact him at chl@cs.nctu.edu.tw.

Tao-Hsing Chang is an assistant professor in the Department of Computer Science and Information Engineering at National Kaohsiung University of Applied Sciences, Taiwan. His research interests include artificial intelligence in education, natural language processing, and automated essay scoring. Chang has a PhD in computer science from National Chiao Tung University. Contact him at changth@cc.kuas.edu.tw.


Acknowledgments

The data we analyzed here were collected by the Research Center for Psychological and Educational Testing at National Taiwan Normal University. This work was supported in part by the National Science Council under grants NSC-98-2221-E-009-141 and NSC-98-2811-E-009-038.

References

1. J. Wang and M.S. Brown, "Automated Essay Scoring versus Human Scoring: A Comparative Study," *J. Technology, Learning and Assessment*, vol. 6, no. 2, 2007, <http://escholarship.bc.edu/jtla/vol6/2>.

2. C.T. Meadow, B.R. Boyce, and D.H. Kraft, *Text Information Retrieval Systems*, 2nd ed., Academic Press, 2000.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.



 IEEE  computer society

Limited Time Offer
Half Year Subscriptions Available Now!

See how the Computer Society publications are leading the way to new discoveries.

Go to: www.computer.org/promos/HY100MNI
Orders must be received by 10 August 2010.

There's still time to subscribe to your favorite Computer Society magazines and journals. You can also select books, ReadyNotes and Essential Sets on your specific topics of interest.

