Utilizing Latent Semantic Word Representations for Automated Essay Scoring

Cancan Jin
School of Computer and Control Engineering
University of Chinese Academy of Sciences
Beijing, China
jincancan9338@gmail.com

Ben He
School of Computer and Control Engineering
University of Chinese Academy of Sciences
Beijing, China
benhe@ucas.ac.cn

***Abstract*—Automated essay scoring (AES) utilizes a set of features to measure the writing quality of essays. However, due to the limits of the existing natural language processing techniques, current AES systems are only capable of making use of shallow text features such as the essay length and the number of the clause. In this paper, we argue that the current AES systems can be further improved by taking into account the latent semantic features. To this end, on top of the commonly used shallow features, we propose three deep semanitc features based on Continuous Bag-of-Words Model (CBOW) and Recursive Autoencoder Model. We use Support Vector Machine for Ranking (SVM^{rank}) to learn a rating model and test the performance of three new features. Experiments on the publicly available English essay dataset, Automated Student Assessment Prize (ASAP), show that our proposed features are beneficial to automated essay scoring.**

***Keywords*-Automated essay scoring; Deep neural network; Semantic word representations**

I. Introduction

Automated essay scoring (AES) is seen as a machine learning problem [1], which automatically rates essays written for given prompts, namely, essay topics, in an educational setting. The typical workflow of an AES system is illustrated in Figure 1. It firstly extracts a number of features which can reflect the quality of the essay, such as lexical and syntax. Next, it learns a rating model. Finally, the learned model is used to the evaluate the score the given new essays.

Nowadays, AES systems have been put into practical use in large-scale English tests and play the role of one human rater. For example, before AES systems enter the picture, essays in the writing assessment of Graduate Record Examination (GRE) are rated by two human raters. A third human rater is needed when the difference of the scores given by the two human raters is larger than one in the 6-point scale. Currently, GRE essays are rated by one human rater and one AES system. A second human rater is required only when there exists a non-negligible disagreement between the first human rater and the machine rater. With the help of an AES system that highly agrees with human raters, the human workload can be reduced by half at most. Therefore, the agreement between the AES system and the human rater is an important indicator of an AES system's effectiveness.

There have been efforts in developing AES methods since the 1960s. Various kinds of algorithms and models based on NLP and machine learning techniques have been proposed to implement the AES systems. However, due to the limits of the existing natural language processing techniques, current AES systems are only capable of making use of shallow text features such as the essay length and the number of clause, they are not able to represent the deep semantic content of essays, resulting in limited robustness and effectiveness [2]. To this end, this paper aims to investigate the relationship between various features and the writing quality. Based on our prior studies, we propose three novel features based on the latent semantic representation of words, which are extracted through deep natural networks (DNN).

To the best of our knowledge, this work is the first effort in utilizing latent semantic features for training the AES algorithm, which aims at the optimization of the agreement between the human and machine raters.

Experimental results on the publicly available dataset ASAP indicate that our proposed features achieve higher agreement with human raters than the original statistical features, measured by root mean squared error [3], Pearsons correlation coefficient [4] and Spearman correlation coefficient [5].

The rest of this paper is organized as follows. In section II, we introduce the research background of automated essay scoring and give a brief introduction to deep semantic features. In section III, a detailed description of latent semantic word representations and statistical features are presented. Section IV explains the experimental setup and section V presents the experimental results. Finally, in section VI we conclude this research.

II. Related Work and Background

We give a brief description of four representative automated essay scoring systems (AES) in section II-A and discuss the Continuous Bag-of-Words Model used in AES systems in section II-B. Then, a brief introduction to Recursive Autoencoder is given in section II-C.

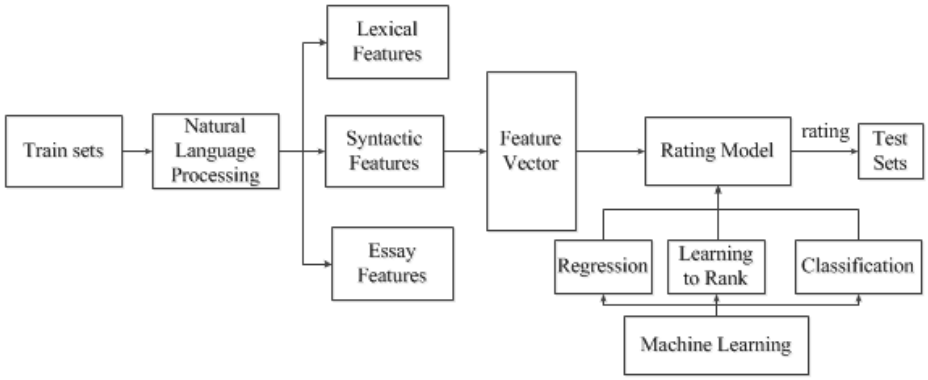

Figure 1. The typical workflow of an automated essay scoring system.

A. Existing AES Systems

In general, existing solutions consider AES as a learning problem. Various learning techniques are applied based on a large number of predefined objectively measurable features.

In 1966, the first AES system, Project Essay Grading (PEG) , was developed by Ellis Page upon the request of the American College Board. The PEG system defines a large set of surface text features from essays, e.g. fourth root of essay length, and uses regression-based approach to predict the score that human raters will give. The PEG considers only surface text features while ignoring the content of the essays. Therefore, it is easy for the students to cheat for high scores.

E-rater, developed by Educational Testing Services (ETS) in America, in late 1990s, is a commercial AES system which has been put into practical use in the Graduate Record Examination (GRE) and the Test of English as a Foreign Language (TOEFL). The E-rater system uses natural language processing techniques to extract various kinds of linguistic features of essays, such as lexical, syntactic, etc. Then it predicts the final score by the stepwise regression method [6].

Intelligent Essay Assessor (IEA) [7], developed also in late 1990s, evaluates essay by measuring semantic features. Each ungraded essay, represented by a semantic vector generated by Latent Semantic Analysis (LSA) [8], is rated according to the similarity degree with semantic vectors of graded essays.

Bayesian Essay Test Scoring System, developed by Larkey in 2003, is based on naive Bayesian model. It is the only open-source AES system, but has not been put into practical use yet.

There are many kinds of linguistic features commonly used in automated essay scoring, such as lexical complexity, grammar errors, syntactic complexity, organization and development, coherence, etc. Lexical complexity is often measured by cosine similarity of term vectors and the change in term length. Grammer errors can be estimated in two ways. One way is to design a grammar error detector for each kind of grammar error. Syntactic complexity is evaluated by features about sentences, clauses and T-units, such as mean length, number of verb phrase, complex nominals, dependent clause ratio, etc.

B. Learning Vector Representation of Words

This section introduces the concept of distributed vector representation of words. A well known framework to learn the word vectors is Continuous Bag-of-Words Model (CBOW). Compared to the different versions of word vectors, Continuous Bag-of-Words Model (CBOW) works better than others in efficiency and accuracy [9]. We therefore choose CBOW to obtain the word vectors.

In this model, every word is mapped to a unique vector, which can be found in the matrix *W*. Assuming a sentence consists of $(w_1, w_2, ..., w_n)$, where w_i is the ith word in the sentence, n is the number of words in the sentence. The objective of this model is to maximize the average probability of the current word based on the surrounding text. k is the adjustable parameter for the number of surrounding words to be considered.

$$\frac{1}{n}\sum_{i=k}^{n-k} lnp(w_i|w_{i-k}, ..., w_{i+k}) \tag{1}$$

The model is trained using stochastic gradient descent and the parameters are obtained through back-propagation. In the end, word vectors will be the by-product.

C. Learning Vector Representation of Sentences

Sentence understanding is useful in many natural language processing applications such as automated sentiment extraction. Recursive Autoencoder (RAE) [10] is a popular choice for this task, by which meaning of a sentence is represented using recursive structures and a high dimentional space, where points represent certain notions of meaning.

Recursive neural network aims to learn vector representations of phrases or sentences through a hierarchical structure. Autoencoder also is a neural network which is mainly used to learn a compressed representation for a set of data, typically for the purpose of dimensionality reduction. Autoencoder has three layers: input layer, hidden layer and output layer. The output layer has the same node numbers as the input layer, and is trained to reconstruct its own inputs. Generally, the node numbers in hidden layer is smaller than the input layer. What we need usually is the node datas in the hidden layer.

In order to obtain the vector representations which can reflect the structure and syntactic of the phrases and sentences, RAE, the combination of recursive nueral network and autoencoder, is applied. In addition, the network structure of RAE is the same as the parse tree that obtained by NLP tools, so we can get vector representations for each node of a parser tree. RAE is an unsupvised learning algorithm, often trained by backpropagation method and optimized by gradient descent.

III. Utilizing Latent Semantic Word Representations for Automated Essay Scoring

The main work-flow of our proposed approach is as follows. Firstly, a set of essays rated by professional human raters are gathered for the training. We analyze this set of human rated essays, and extract a set of pre-defined features. We get two rating models through SVM^{rank} based on the human rated essays represented by vectors of the features. Then the learned rating models output a score for each essay separately, including both rated and unrated essays. Finally, the model score is mapped to a predefined scale of valid ratings, such as an integer from 1 to 6 in a 6-point scale.

In this section, we give a detailed description of the basic statistical features used in III-A, and the proposed latent semantic features in III-B.

A. Statistical Features

We define four types of features as the indicators of the essay quality, including lexical, syntactical, grammar and fluency, content and prompt-specific features. A brief description of these four classes of features is given below.

Lexical features: We define two subsets of lexical features. Each subset of features consists of one or several sub features.

– *Statistics of word length*: The mean and variance of word length in characters. The variety of the word length reflect a essay wording condition since the unusual words generally more longer.

– *Unique words*: The number of unique words appeared in each essay, normalized by the essay length in words.

Syntactical features: There are three subsets of syntactical features.

– *Statistics of sentence length*: The mean and variance of sentence length in words. The variety of the length of sentences potentially reflect the complexity of syntactic.

– *clauses*: The mean number of clauses in each sentence, normalized by sentence numbers in a essay. The mean length of sentences contains clause. Clauses are labeled as *SBAR* in the parser tree generated by a commonly used NLP tool, Stanford Core NLP [11], which is an integrated suite of natural language processing tools for English in Java[1], including part-of-speech tagging, parsing, co-reference, etc..

– *Sentence level*: The height of the parser tree is also incorporated into the feature set. Commonly, the more complex sentences, the height of the parser tree will be more larger.

Grammar and fluency features: There are two subsets of grammar and fluency features.

– *Word bigram and trigram*: We evaluate the grammar and fluency of an essay by calculating mean tf/TF of word bigrams and trigrams [12] (tf is the term frequency in a single essay and TF is the term frequency in the whole essay collection). We assume a bigram or trigram with high tf/TF as a grammar error because high tf/TF means that this kind of bigram or trigram is not commonly used in the whole essay collection but appears in the specific essay.

– *POS bigram and trigram*: Mean tf/TF of POS bigrams and trigrams. The reason is the same with word bigrams and trigrams.

Content and prompt-specific features: We define two subsets of content and prompt-specific features.

– *Essay length*: Essay length in characters and words, respectively. The fourth root of essay length in words is proved to be highly correlated with the essay score [13].

– *Word vector similarity*: Mean cosine similarity of word vectors, in which the element is the term frequency multiplied by inverse document frequency (tf-idf) [14] of each word. It is calculated as the weighted mean of all cosine similarities and the weight is set as the corresponding essay score.

B. Latent Semantic Features

In this section, we introduce three latent semantic features, the detailed explanation about On-topic degree and CBOW essay coherence is in III-B1, and RAE essay coherence is presented in III-B2.

1) Features based on word vectors: To get the On-topic degree for each essay in a specific essay set, all we need are the topic vector and the essay vector. Firstly, we compute the weight value for each word in each essay as follows:

$$weight_{ij} = \frac{tf_i}{n_j} \ln \frac{tf_i N}{TF_i n_j} \tag{2}$$

where $weight_{ij}$ is the weight value of the ith word in the essay set j, N is the numbers of word in the eight essay sets, n_j is the number of words in essay set j, tf_i is the

numbers of the ith word in essay set j, TF_i is the numbers of the ith word in the eight essay sets. We choose ten words with the largest weights in each essay set, which are made as the essay set keywords. Table I presents two example essays and the extracted keywords of the essay set 2. We can see that the extracted keywords adequately reflect what the essay topic is about, showing that our approach for the topic words extraction produces reasonable results.

Then, we exact every word vector in a commom semantic vector through Continuous Bag-of-Words Model (CBOW). In the training process, we choose Hierarchical Softmax algorithm and Negative Sampling algorithm to optimize word vectors. In addition, we choose the sum of word vectors rather than the concatenation to predict the next word in sentence. The model is trained using stochastic gradient descent where the gradient is obtained by back propagation [15]. After that, every word is mapped to a unque vector, represented by a column in a matrix *W*.

Finally, because of every word can find the corresponding word vector in matrix W, it is easy to get the keywords vector and the essay vector. Assuming $w_1, w_2, ..., w_10$ are the ten key word vector of the essay set j, the essay set j keywords vector is:

$$key_j = \frac{\sum_{i=1}^{10} w_i}{10} \tag{3}$$

, assuming $d_1, d_2, ..., d_n$ are the words in an essay k, n is the numbers of the word in essay k, the essay vector is:

$$d_k = \frac{\sum_{i=1}^{n} d_i}{n} \tag{4}$$

The On-topic degree of essay k in essay set j is defined as the cosine of the essay set j keywords vector and the essay vector d_k. Assuming $w_1, w_2, ..., w_m$ are the word vectors in a sentence s, m is the number of the words in sentence s, the sentence vector is :

$$s_i = \frac{\sum_{i=1}^{m} w_i}{m} \tag{5}$$

Assuming $s_1, s_2, ..., s_n$ are the sentence vectors in the essay i, the CBOW essay coherence of essay i is defined as follows:

$$semco_i = \sum_{i=1}^{n} \sum_{j=1, j!=i}^{n} \frac{cosine(S_i, S_j)}{\binom{n}{2}} \tag{6}$$

where n is the number of sentences in the essay i.

2) Feature based on sentence vectors: Our choice of the Recursive Autoencoder to get the sentence vectors is due to the fact that it can get not only the fixed-length vector but also the representative sentence vector. To further improve the sentence representation, we first obtain the parser tree for each sentence through the Stanford parser algorithm, then construct the neural network on the basis of the parser tree.

Assume a sentence is made up of a list of word vectors $s = (w1, w2, w3)$. In a binary tree, we define a branching triplets of parents with children, which is in the form of $(p - c_1c_2)$. We assume that we can get the following triplets in encoder processing: $((y_1 - x_2x_3), (y_2 - x_1y_1))$, also the following triplets can be obtained in the decoding process: $((y_2 - x_1^{'}y_1^{'}), (y_1^{'} - x_2^{'}x_3^{'}))$. We define the loss function as follows:

$$|[x_1; y_1] - [x_1^{'}; y_1^{'}]|^2 + |[x_2; x_3] - [x_2^{'}; x_3^{'}]|^2 \tag{7}$$

Through the trained model, we can exact semantic vectors for each sentence in same dimentional. Next, we can compute the RAE essay coherence for each essay through formula (6).

IV. Experimental Setup

This section presents our experimental design, including the test dataset used, configuration of testing algorithms, and the evaluation methodology.

A. Test Dataset

The dataset used in our experiments comes from the Automated Student Assessment Prize (ASAP), which is sponsored by the William and Flora Hewlett Foundation. Dataset in this competition[2] consists of eight essay sets. Each essay set was generated from a single prompt. The number of essays associated with each prompt ranges from 900 to 1800 and the average length of essays in word in each essay set ranges from 150 to 650. All essays were written by students in different grades and received a resolved score, namely the actual rating, from professional human raters. Moreover, ASAP comes with a validation set that can be used for parameter training. There is no overlap between this validation set and the test set used in our evaluation.

We use root mean squared error, Pearsons correlation coefficient and Spearman correlation coefficient to evaluate the agreement between the ratings given by the AES algorithm and the actual ratings. They are widely accepted as reasonable evaluation measures for AES systems [16].

Normalized root-mean-squared error (nRMSE) [3] is a evaluation criterion which aims to measure the error between predicted ratings and true ones. The essay scores of a given essay topic are normalized to be within [0, 1] between computing the mean squared error. nRMSE is given by:

$$e = \sqrt{\frac{\sum d_i^2}{n}} \tag{8}$$

where d_i is the deviation between predicted ratings and true ones, n is the times of the measurement.

Pearson correlation coefficient [4] is used to measure the strength of a linear association between two variables. It is computed as follow:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}} \tag{9}$$

where $\overline{x}$ is the mean of the $x_1, x_2, ..., x_n$, $\overline{y}$ is the mean of the sequence $y_1, y_2, ..., y_n$.

Table I
EXAMPLE: THE GIVEN PROMPT AND THE EXTRACTED KEYWORDS OF ESSAY SET 2 IN THE ASAP DATASET.

Promt of essay set2	"All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf – that work I abhor – then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us." –Katherine Paterson, Author Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience,observations, and/or reading.
Keywords	book library people read offensive movie can like think just
Example essay 1	Do you think that libraries should remove certain materials off the shelves? People have different oppions, of whats good and whats bad. I have read and seen a lot of books in my life time. I hear people telling me, 'oh dont read that book its a bad book.'But I ask myself, @CAPS2 do I know it's a bad book when I haven't even given it a chance?' @CAPS1 are some books, music, movies, and magazines out @CAPS1 that are offensive. Yet we still want to read, listen, watch, and look at them. If we tried to remove all the offensive books, from the libraries we wouldn't have anything left on the shelves. Katherine Paterson said, 'If I have the right to remove that book from the shelf that work i abhor- then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us.' Katherine Paterson makes a great point out of her quote. Why should we have to remove a book if just some people think its offensive? Ask yourself the question again, '@CAPS2 do you know it's a bad book when you haven't even given it a chance?' @CAPS3't judge a book by what you hear. Find out what your own oppion is
Example essay 2	A book represents a person's beliefs and feelngs about a topic. Therefore censoring books is wrong because it goes against freedom of speech. I do not believe books, or other media should be removed because while it @MONTH1 be offensive, if it is the truth about something then we need to read about it to prevent tragedies from happening again. A person who finds a topic offensive isn't forced to read about it. If we do not accept other people beliefs, we are just as prejudice as the books we are trying to censor. Many books that are threatened by censorship are books about historical facts, such as the holocaust. I think we have the right to know about important events. For one reason, if we forget about such things, than they are more likely to be repeated. Do we really want the murder of millions of people to be repeated? Just becaused we are ashamed of something dosen't mean we should sweep it under the carpet where it can lay hidden waiting to stike again like a hungry alligator. Instead books about prejudice events should be kept out to show an example of how we shouldn't act. While some books are censored because of shame, some are censored because a certain group of people @MONTH1 find it offensive. There is a much easier solution that censoring those books. The offended groups should just ignore the book. If it is offensive in the first place then no one should be forced to read it. The book shouldn't be forced in any other way like posters, and television broadcast. It should be kept in a certain section of a library so parties wishing to avoid it can. The main reason for a book to be censored in the first place is because someone finds in offensive. However, isn't keeping someone from saying something just as offensive? To @CAPS1 A @CAPS2 is a big target for censorship, but it is about treating everyone fairly. If that book like that was censored, it would be like saying some people are better than others. Another big target for censorship in some countries is the @CAPS3. Censoring that book would basically tell the @CAPS4 population that they are wrong. Censoring something is wrong. Not only could it cause sad events to repeat themselves, offended parties can just simply avoid the book. While some books offend people, censoring a book can be just as offensive to other parties. The simplest thing to do is to leave it up to the reader to decid if they should read it, or not.

Spearman correlation coefficient [5] assesses how well the relationship between two variables can be described using a monotonic function. It is given by:

$$\rho = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)} \tag{10}$$

where d_i is the difference between two variables, n is the number of the variables.

B. Configuration of Testing Algorithms

We use support vector machine (SVM) to rate the essays. The linear kernel is used in the experiments. The parameter C, which controls the trade-off between empirical loss and regularizer, is set by grid search on the ASAP validation set.

[1]http://nlp.stanford.edu/software/corenlp.shtml
[2]http://www.kaggle.com/c/asap-sas/data

C. Evaluation Methodology

We conduct two sets of experiments to evaluate the effectiveness of our proposed new semantic features and other shallow features for automated essay scoring.

The first set of experiments evaluate our proposed three semantic features through SVM^{rank}. This experiment consist of five parts. First, we choose thirteen shallow txt features which introduced in section III-A as the input of the rating algorithm, and call this part as *baseline*; Then, in the next three parts, we seperately add one feature on the basis of the *baseline*, for instance, *On-topic degree* represent add the *On-topic feature*on the basis of *baseline*; In the last part, we add our proposed three features on the basis of the first part.

The second set of experiments conduct an ablation test on the importance of each individual feature. It examines the effectiveness and redundancy of each feature by substracting the feature from the entire feature set.

The AES system is evaluated by three measures, namely

the normalized root mean squared error, Pearson's correlation coefficient, and Spearman correlation coefficient with the *baseline*. We conduct 10-fold cross-validation, the essays of each prompt are randomly partitioned into 10 subsets. In each fold, 9 subsets are used for training, and one is used for testing.

V. Experimental Results

Table II, III and IV all present the first experimental results obtained on the ASAP dataset, measured by root mean squared error, Pearson's correlation coefficient and Spearman correlation coefficient. In all tables, *baseline* stands for that each essay is represented as a thirteen dimentional feature vector and the features are introduced in the III-A, *On-topic degree* represent that the input of the rating algorithm add a feature *On-topic degree* on the base of the *baseline*, as same as the *CBOW essay coherence* and *RAE essay coherence*. *Total* stands each essay is represented as a sixteen dimentional vector, namely, the rating input contains predefined features introduced in section III-A and three semantic features which introduced in section III-B. *overall* stands for the average evaluated value of the eight evaluated values on the current feature sets.

At the same time, the result of *Total* suggests that each of these three factors acts to magnify the impact of the other. From table 1, compared to the baseline, our proposed three features can reduce the root mean squared error between human and machine raters; For generic rating model, one can conclude from Table 2 and 3 that our proposed semantic features all improve the agreement between human and machine raters to some extent compare to the baseline. Also, we can find from the table 2 and 3 that accuracy of the rating system is 0.7156 and 0.7248, it represent our machine raters have strongly correlation with human raters.

Table V presents the second experimental results obtained on the ASAP dataset, also measured by root mean squared error, pearson correlation coefficient and spearman correlation coefficient. In the table, *baseline* stands for that each essay is represented as a sixteen dimentional feature vector and the features are introduced in the III. *word variance* represents the dimentional of the input features vector is fifteen, namely, lacking word variance compared to the baseline.

In Table IV, it is obvious that the lack of every feature discribed in section III-A causes the decrease of performance of rating system. This shows that each feature in the feature set of the baseline is effiective for the essay scoring.

Table V presents the results obtained by an ablation test on the effect of removing each features on the essay scoring accuracy. From the results, apart from the essay length which is a naive but straightforward evidence of essay quality, it is not obvious which features contribute the most to the essay scoring accuracy. Indeed, it is necessary to combine the features to construct a robust essay scoring system.

VI. Conclusion

We have proposed three latent semantic features to improve the quality of the essay features in automated essay scoring (AES). Experiments on the public English dataset ASAP show that the semantic features all improve the the agreement between human and machine raters to some extent. Also, we find that every feature extracted in this essay is efficient through ablation test.

Most existing research on AES focus on extracting prompt-specific features. While such features have the advantage of providing a satisfactory rating standard for essays written for a specific topic, they also suffer from validity and feasibility problem when be used in unrelated subject model. In our future work, we plan to continue the research on dining more deep semantic and syntactic features. Because of the diversification of writing features of essays associated with different prompts, a viable approach is to explore more generic writing features that can well reflect the writing quality. In our future work, we plan to continue the research on dining more deep semantic features. Because of more generic writing features of essays can well relect the writing quality.

Acknowledgments

This work is supported in part by the National Science Foundation of China (6147239/61103131), Beijing Natural Science Foundation (4142050) and SRF for ROCS, SEM.

References

[1] D. S, "An overview of automated scoring of essays," *The Journal of Technology, Learning and Assessment*, vol. 5, no. 1, 2006.

[2] P. J. J. D. B. Y. Yang, C. W. Buckendahl, "A review of strategies for calidating computer-automated scoring [j]," *Applied Measurement in Education*, vol. 15, no. 4, pp. 391–412, 2002.

[3] G. Z. Ferl, R. E. Port, and G. Ferl, "Root mean square errors."

[4] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[5] C. Croux and C. Dehon, "Influence functions of the spearman and kendall correlation measures," *Statistical methods & applications*, vol. 19, no. 4, pp. 497–515, 2010.

[6] Y. Attali and J. Burstein, "Automated essay scoring with e-rater," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, 2006.

[7] P. W. Foltz, D. Laham, and T. K. Landauer, "Automated essay scoring: Applications to educational technology," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, vol. 1999, no. 1, 1999, pp. 939–944.

[8] S. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188–230, 2005.

Table II
EVALUATION RESULTS OF THE THREE NEW FEATURES IN ADDITION TO THE BASELINE FEATURES BY THE NORMALIZED ROOT MEAN SQUARED ERROR. SVM^{rank} IS USED FOR LEARNING THE RATING MODEL. A LOWER VALUE IS BETTER.

Eassy set No.	1	2	3	4	5	6	7	8	overall
baseline	0.0928	0.1300	0.1959	0.2064	0.1535	0.1732	0.1043	0.0785	0.1421
On-topic degree	0.0927	0.1295	0.1957	0.1994	0.1504	0.1667	0.1067	0.0763	0.1399(1.5%)
sentenceBased coherence	0.0934	0.1292	0.1949	0.2048	0.1514	0.1714	0.1043	0.0774	0.1411(0.7%)
wordBased coherence	0.0932	0.1296	0.1944	0.2040	0.1529	0.1687	0.1038	0.0784	0.1409(0.8%)
Total	0.0933	0.1288	0.1941	0.2044	0.1512	0.1638	0.1051	0.0763	0.1399(1.5%)

Table III
EVALUATION RESULTS OF THE THREE NEW FEATURES IN ADDITION TO THE BASELINE FEATURES BY PEARSON'S CORRELATION COEFFICIENT. SVM^{rank} IS USED FOR LEARNING THE RATING MODEL. A HIGHER VALUE IS BETTER.

Eassy set No.	1	2	3	4	5	6	7	8	overall
baseline	0.7790	0.6298	0.7536	0.7096	0.7543	0.6966	0.7142	0.5698	0.7063
On-topic degree	0.7816	0.6566	0.7554	0.7347	0.7354	0.7241	0.7010	0.6008	0.7151(1.3%)
sentenceBased coherence	0.7793	0.6372	0.7582	0.7094	0.7412	0.7041	0.7156	0.5868	0.7086(0.3%)
wordBased coherence	0.7783	0.6259	0.7596	0.7100	0.7467	0.7151	0.7179	0.5726	0.7087(0.4%)
Total	0.7787	0.6706	0.7633	0.7109	0.7403	0.7226	0.7088	0.5999	0.7156(1.3%)

Table IV
EVALUATION RESULTS OF THE THREE NEW FEATURES IN ADDITION TO THE BASELINE FEATURES BY SPEARMAN CORRELATION COEFFICIENT. SVM^{rank} IS USED FOR LEARNING THE RATING MODEL. A HIGHER VALUE IS BETTER.

Eassy set No.	1	2	3	4	5	6	7	8	overall
baseline	0.7948	0.6727	0.6861	0.7360	0.8033	0.6987	0.7134	0.5469	0.7140
On-topic degree	0.7942	0.6711	0.6890	0.7605	0.8092	0.7353	0.6993	0.5783	0.7241(1.4%)
sentenceBased coherence	0.7972	0.6808	0.6864	0.7442	0.8044	0.7117	0.7147	0.5656	0.7200(0.8%)
wordBased coherence	0.7912	0.6704	0.6899	0.7482	0.8037	0.7430	0.7183	0.5520	0.7220(0.8%)
Total	0.7864	0.6819	0.6938	0.7431	0.8040	0.7353	0.7111	0.6001	0.7248(1.5%)

Table V
EVALUATION RESULTS OF THE ABLATION TEST IN ADDITION TO THE BASELINE FEATURES.

Eassy set No.	root mean squared error	Pearson's correlation coefficient	Spearman correlation coefficient
baseline	0.1399	0.7156	0.7248
word expectation	0.1403	0.7147	0.7213
word variance	0.1401	0.7102	0.7242
sentence expectation	0.1400	0.7153	0.7248
sentence variance	0.1404	0.7078	0.7224
word bigram	0.1407	0.7098	0.7158
word trigram	0.1402	0.7114	0.7215
POS bigram	0.1404	0.7072	0.7208
POS trigram	0.1401	0.7124	0.7231
similarity	0.1409	0.7077	0.7185
clause length	0.1590	0.6004	0.5841
clause numbers	0.1407	0.7101	0.7198
sentence depth	0.1402	0.7116	0.7216
essay length	0.1407	0.7127	0.7214
On-topic degree	0.1416	0.7056	0.7157
sentenceBased coherence	0.1403	0.7087	0.7247
wordBased coherence	0.1405	0.7109	0.7240

[9] G. C. J. D. Tomas Mikolov, Kai Chen, "Efficient estimation of word representations in vector space," *In Proceedings of Workshop at ICLR*, 2013.

[10] P. J. e. a. Socher R, Huang E H, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection [c]," *NIPS*, pp. 801–809, 2011.

[11] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2003, pp. 423–430.

[12] T. Briscoe, B. Medlock, and Ø. Andersen, "Automated assessment of esol free text examinations," University of Cambridge Computer Laboratory Technical Reports,UCAM-CL-TR-790, Tech. Rep., 2010.

[13] M. Shermis and J. Burstein, *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum, 2002.

[14] G. Salton, *The SMART Retrieval System-Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

[15] R. J. David E. Rumelhart, Geoffrey E. Hinton, *Learning representations by backpropagating errors*. Neurocomputing: foundations of research, 1988.

[16] D. Williamson, "A framework for implementing automated scoring," in *Annual Meeting of the American Educational Research Association andthe National Council on Measurement in Education, San Diego, CA*, 2009.