Developing Chinese Automated Essay Scoring Model to Assess College Students’ Essay Quality

Ju-Lu, Yu
Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taiwan
No.140, Minsheng Rd., West Dist., Taichung City 40306, Taiwan (R.O.C.)
ddog5633@yahoo.com.tw

Bor-Chen Kuo
Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taiwan
No.140, Minsheng Rd., West Dist., Taichung City 40306, Taiwan (R.O.C.)
kbc@mail.ntcu.edu.tw

Kai-Chih Pai
Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taiwan
No.140, Minsheng Rd., West Dist., Taichung City 40306, Taiwan (R.O.C.)
minbai0926@gmail.com

ABSTRACT

The present study aimed at proposing a Chinese automated essay scoring model to assess college students writing quality. Thirty-one related Chinese linguistic indicators were developed based on Coh-Metrix indices and characteristics of Chinese texts. Essay collected from 277 college students were analyzed using automated Chinese text analyze tool. A stepwise regression was used to explain the variance in human scores. The number of words, number of low strokes, content words frequency, minimal edit distance (all words) and minimum frequency for content words predicted 55.8% variance in human scores. On the other hand, seven indicators: number of words, content words frequency, concreteness, Measure of Textual Lexical Diversity, minimal edit distance (part of speech), minimal edit distance (all words) and words per sentence were predictive of human essay ratings by using discriminant analysis. The present study further explored the effectiveness of the Chinese automated essay scoring model by using three different methods: stepwise linear regression, discriminant analysis, and Nonparametric Weighted Feature Extraction classification (NWFE). The preliminary results showed that NWFE classification method produced higher exact matches (51.3%) between the predicted essay scores and the human scores than stepwise regression (47.3%) and discriminant analysis (47.3%).

Keywords

Chinese automated essay scoring, writing quality, NWFE classification, Chinese linguistic indicators

1. INTRODUCTION

Essay scoring has traditionally relied on expert raters. These scoring methods need to spend more time and a large amount of human scoring. Based on these limitations, automated essay scoring becomes the important research for essay assessment. According to the results of past studies, automated essay scoring reported perfect agreement (i.e., the exact match of human and computer scores) from 30-60% and adjacent agreement (i.e., within 1 point of the human score) from 85-99% [1]. Moreover, recently the study of analyzing the scored essays using Coh-Metrix has increased noticeably [2, 4, 5, 6, 7, 8, 13, 14, 15]. Coh-Metrix is an automated text analysis tool that provides lots of different linguistic indices [10]. The tool can provide these indices by combining lexicons, a syntactic parser, and several other components that are widely used in computational linguistics.

Chinese language features in the characteristics of different from the English, cannot be directly applied to the Chinese essay writing. Most of the experts will consider the following sections: Number of words, structure organization, vocabulary diversification, typos, and punctuation. Based on the development of Coh-Metrix, automated text analyze tool were developed in Chinese. Totally 66 Chinese related linguistic indicators were used to analyze the characteristics of Chinese texts [12].

Writing the literacy assessment is an important standardized testing to assess college students’ writing skill in Taiwan. The assessment is to detect whether students can express personal comments on specific issues. Students need to read an article, respectively, and express personal comments by writing the essay in two hundred words. These essays were scored by two experts and score from 0-5. However, we need to a lot of experts and spend more time to score. To propose a suitable automated scoring model is important and needed.

2. PROPOSED CONTRIBUTIONS

The purpose of the study is to explore the characteristics of Chinese writing and propose a suitable Chinese automated essay scoring model to assess college students writing quality. Past studies explored the variety of human scoring were predicted by different text features using regression analysis. Moreover, they proposed automated essay scoring model and examined the essay matches by linear regression and discriminant analysis. A Nonparametric Weighted Feature Extraction (NWFE) classification method was also used to examine the essay matches in the present study.

Nonparametric Weighted Feature Extraction (NWFE) is based on a nonparametric extension of scattering matrices. It could reduce parametric dimensional and increase classification accuracy [11]. The present study used linear regression analysis and discriminant analysis of the gradual selection of variables for the NWFE classification method and examine the accuracy of essay matches.

3. Method

3.1 Text Indices Selection Procedure

The present study collected Chinese essay from college students in Taiwan. All essay was analyzed by Chinese automated text analyze tool. The tool provides 62 Chinese linguistic indices, includes basic text measures (e.g., text, sentence length), words information (e.g., word frequency, concreteness), cohesion (semantic and lexical overlap, lexical diversity, along with the incidence of connectives), part of speech and phrase tags (e.g., nouns, verbs, adjectives), and syntactic complexity (e.g., Sentence syntax similarity, Minimal Edit Distance).

The first step, correlation analyses was conducted to examine the strength of relations between the selected indices and the human scores of essay quality. Text indices retained based on a significant correlation with human scores. Multicollinearity was then assessed between the indices (r >.900). The index retained based the strongly with human scores when two or more indices demonstrated multicollinearity. Finally, totally thirty-one indices were used in the study.

3.2 Essay Scoring

277 essays were collected from college students in Taiwan. Each essay in the study was scored independently by two expert raters using a 5-point rating. The rating scale was used to assess the quality of the essays and had a minimum score of 0 and a maximum score of 5. The experts evaluated the essays based on a standardized rubric used in the Chinese writing literacy assessment in Taiwan. The results of correlation between two experts are 0.788. It indicated that consistency of expert scoring.

3.3 Essay Evaluation

Three different methods were used to examine the accuracy of automated essay scoring: linear regression analysis, discriminant analysis, and NWFE classification. Text features were selected by linear regression and discriminant analysis. The leave-one-out method was used to experiment with training essay set and testing the essay set. The present compared the exact matches of the essay by using the three methods.

4. Preliminary Results

4.1 Linear Regression Analysis: Text Features

A stepwise regression analysis was conducted to examine which text indicators were predictive of human essay ratings. 40 Chinese text features were used in the study. The results presented in Table 1. Five indicators were a significant predictor in the regression model: Number of words, the number of low strokes, content word's frequency, minimal edit distance (all words) and the minimum frequency of content words, F = 12.074, p <.001, r =.747, r^2 =.558. The results from the linear regression demonstrate that the five variables account for 55.8% of the variance in the human scoring of writing quality.

Table 1. Stepwise regression results for text features

Indicators	*B*	*SE*	Β
number of words	.011	.001	.529
number of low strokes	.000	.000	-.131
content words frequency	.824	.402	.086
minimal edit distance (all words)	2.334	.618	.238
minimum frequency for content words	-.148	.042	-.154

4.2 Discriminant Analysis: Text Features

The purpose of the discriminant analysis was to examine whether features are predictive of human scoring. The results of the discriminant analysis showed that seven text features could predict human scorning, includes the number of words, content word frequency, concreteness, Measure of Textual Lexical Diversity, minimal edit distance (part of speech), minimal edit distance (all words) and words per sentence.

4.3 Exact and Adjacent Matches

Table 2 and Table 3 presented the results of exact and adjacent matches. The linear regression analysis (stepwise) selected features: The number of words, number of low strokes, content words frequency, minimal edit distance (all words) and minimum frequency for content words. The exact matches (leave-one-out) between the predicted essay scores (rounded to 0-5) and the human scores is 47.3% exact accuracy and 95.3% adjacent accuracy.

The discriminant analysis (stepwise) selected features had the number of words, word frequency of content words, minimal edit distance (local), MTLD, the number of terms, concreteness, and minimal edit distance (part of speech). The exact matches (leave-one-out) between the predicted essay scores and the human scores is 47.3% exact accuracy and 93.9% adjacent accuracy.

The present study conducted NWFE classification method to examine the effectiveness of automated essay scoring. The results showed that 48.7% exact matches between predicted scores and human scoring, which text features selected by linear regression. Moreover, 51.3% exact matches between predicted scores and human scoring, which text features selected by discriminant analysis.

Table 2. Comparison of Exact

Classification method	Text features selected by linear regression	Text features selected by Discriminant
Linear regression	47.3%	46.6%
Discriminant	45.5%	47.3%
NWFE	48.7%	51.3%

Table 3. Comparison of Adjacent

Classification method	Text features selected by linear regression	Text features selected by Discriminant
Linear regression	95.3%	93.9%
Discriminant	94.2%	93.9%
NWFE	89.9%	90.3%

5. Conclusion

Past studies have found that the number of words was an important indicator of human score [4, 15]. The results of the study also presented that the number of words has a high significant correlation with human scores. The number of words, the minimal edit distance (local), and the number of low strokes three indicators belong to Descriptive and Syntactic Complexity categories in Coh-Metrix. MTLD belongs to Lexical Diversity. These indicators are related the scoring guide of writing for college students in Taiwan.

Comparing exact matches between linear regression analysis (stepwise) and discriminant analysis (stepwise). The results of leave-one-out of exact matches linear regression and discriminant analysis showed consistency. Moreover, regardless of method linear regression analysis (stepwise) or discriminant analysis (step-wise) selection indicators, the accuracy of exactly matched of NWFE method is higher than the other two classification methods.

6. Future Works

Past studies have investigated the potential for component scores that are calculated using the linguistic features by Coh-Metrix in assessing text readability [9, 12]. Moreover, one study has explored correlations between human ratings of essay quality and component scores based on similar natural language processing indices and weighted through a principal component analysis [2]. However, this approach has not been extended to computational assessments of essay quality In Chinese. The present study will adapt a similar approach to passing studies [9, 12]. We will conduct a principle component analysis (PCA) or factor analysis to reduce the number of indices selected from Chinese automated text analyze tool into a smaller number of components comprised of related features. The present study will further explore the correlation between component scores and human scoring. A Chinese automated essay scoring model based on text component scores will be developed and explored.

7. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of this study by the Ministry of Education.

8. REFERENCES

[1] Attali, Y., & Burstein, J. 2006. Automated Essay Scoringwith E-rater V.2. *Journal of Technology, Learning and Assessmen*t, 43.

[2] Crossley, S. A., & McNamara, D. S. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66-79.

[3] Crossley, S. A., & McNamara, D. S. 2014. Developing component scores from natural language processing tools to assess human ratings of essay quality. In W. Eberle & C. Boonthum-Denecke (Eds.), *Proceedings of the 27th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 381-386). Palo Alto, CA: AAAI Press.

[4] Crossley, S. A., Dempsey, K., & McNamara, D. S. 2011. Classifying paragraph types using linguistic features: Is paragraph positioning important? *Journal of Writing Research*, 3, 119-143.

[5] Crossley, S. A., Roscoe, R. D., & McNamara, D. S. 2013. Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the 26th International Flordia Artificial Intelligence Research Society (FLAIRS) Conference*, 208-213. Menlo Park, CA: The AAAI Press.

[6] Crossley, S.A. & McNamara, D.S. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 984-989. Austin, TX: Cognitive Science Society.

[7] Crossley, S.A., & McNamara, D.S. 2011. Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1236-1231. Austin, TX: Cognitive Science Society.

[8] Guo, L., Crossley, S. A., & McNamara, D. S. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18, 218-238.

[9] Graesser, A.C., McNamara, D.S., and Kulikowich, J. 2012. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223-234.

[10] Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193-202.

[11] Kuo B.-C., and Landgrebe, D. A. 2004. Nonparametric weighted feature extraction for classification, *IEEE Transactions on Geoscience and Remote Sensing*, 42(5), 1096-1105.

[12] Kuo B.-C., and Liao C.-H. 2014. The Automated text analysis for Chinese text. *2014 Workshop on the Analysis of Linguistic Features (WoALF 2014)*, Taipei, Taiwan.

[13] McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. 2015. A Hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35-59.

[14] McNamara, D.S., Graesser, A.C., McCarthy, P., & Cai, Z. *Automated evaluation of text and discourse with Coh-Metrix.Cambridge*: Cambridge University Press, 2014.

[15] Roscoe, R.D., Crossley, S.A., Weston, J.L., & McNamara, D.S. 2011. Automated assessment of paragraph quality: Introductions, body, and conclusion paragraphs. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 281-286. Menlo Park, CA: AAAI Press.