International Journal of Learning, Teaching and Educational Research
Vol. 15, No. 5, pp. 16-26, April 2016

Using Coh-Metrix to Analyze Chinese ESL Learners' Writing

Weiwei Xu
College of International Studies,
Southwest University,
Chongqing, China

Ming Liu
School of Computer and Information Science,
Southwest University,
Chongqing, China

Abstract. Scoring essays is costly, laborious and time-consuming. Automated scoring of essays is a promising approach to face this challenge. Coh-Metrix is a computer tool that reports on cohesion, sentence complexity, lexical sophistication and other descriptive features at sentence- and paragraph-level. It has been widely used to analyze native English speakers' essay writing. However, few studies have used Coh-Metrix to analyze essays written by English as a Second Language (ESL) students. In this study, we analyzed the correlation between several Coh-Metrix features combined with a set of newly proposed features and the quality of essays, written by Chinese university students, both English and non-English majors. This study shows that each group of students tends to write essays that have their own signature features. The quality of essays written by English majors highly correlate to the importance of introduction, conclusion and cohesion at the sentence level, while the quality of essays written by chemistry majors are highly related to mechanics errors, sentence complexity and cohesion at the paragraph level.

Keywords: ESL essay writing, Textual feature analysis, Automatic Essay Scoring, Computer in education

Introduction

Important constructs, central to ESL writing and proposed by several researchers, are grammatical and spelling errors. Cohesion is also important, although it is a much more difficult aspect of writing to account for due to its deeper nature (Rus & Niraula, 2012). This study focuses on grammatical and spelling errors and cohesion which are directly observed through the explicit presence or absence of specific tokens. Errors may be caused by inappropriate transfer of first language patterns and/or incomplete knowledge of the target

language, in this case, English. Researchers (Q. Liang, 2004; Liu, 2008) have pointed out that Chinese college students, especially those with low proficiency in English, often make errors at the surface level, such as spelling and grammatical errors (e. g. run-on sentences); errors at high level, such as using Chinglish (ungrammatical English expressions used in Chinese context, having deprecating connotation); and low cohesion or incorrect use of connectives. Even for students with high proficiency, like English majors, writing high quality essays with high cohesion, well-established introduction and conclusion, remains a challenge. Thus, a marking tool, specifically developed to analyze ESL learners' errors, is very much needed. It should be noted that errors are categorized as word-level (spelling errors) and sentence-level (grammatical errors) and, as mentioned above, are consequences of incomplete knowledge of the target language or of the transfer difficulty due to major dissimilarities between the foreign language and students' native language. On the other hand, cohesion is a discourse-level aspect of writing and lack of cohesion in an essay may reflect lack of composition training and practice. This distinction is important to make, because one can argue that the only net advantage of native speakers of English over ESL speakers is their knowledge of English vocabulary and grammar. Discourse-level aspects, on the other hand, are governed by general cross-language principles of cohesion and coherence, and are equally impacting for both native and EFL speakers. As is shown in this study, English majors who presumably have mastered the mechanics of the language (vocabulary and grammar) struggle mainly with the compositional aspect, which is in contrast with non-English majors who struggle with both the mechanics and composition aspect of essay writing.

Researches in computer-based essay scoring, referred to as Automatic Essay Scoring (AES), have been going on for more than 40 years. The first known AES system, called Project Essay Grader (Page, 2003) based on a regression model, was developed by Ellis Page in 1966. With the advancement of Natural Language Processing (NLP) and Information Retrieval (IR) techniques, four more advanced AES systems were developed during the late 1990s (M. Shermis & Burstein, 2003). In recent years, different approaches to AES were proposed (McNamara, Crossley, Roscoe, Allen, & Dai, 2015; Mark D. Shermis, 2014). AES systems in China is still at an early stage (Ge & Chen, 2007; Han, 2009; Li, 2009; M. Liang & Wen, 2007; M. Liang, 2011). Most of researchers focus on the reviews of existing AES systems and their potential applications to Chinese ESL context (Ge & Chen, 2007; Han, 2009; M. Liang & Wen, 2007). Few researchers (Li, 2009; M. Liang, 2011) have attempted to develop AES systems in Chinese ESL context by using latent semantic analysis technique (Landauer, Foltz, & Laham, 1998).

This paper aims to explore what textual features are good predictors for writing quality and investigate its implication for developing AES system in Chinese ESL context. Textual features such as syntactic patterns, cohesions and connectives were extracted by using the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix is used to analyze essays written by Chinese ESL students, and this study analyzed the correlations between features and the quality of essays written by both English and non-English majors.

Coh-Mextrix

Coh-Metrix is a computational tool that provides over 100 indices of cohesion, syntactical complexity, connectives and other descriptive information about content (Graesser et al., 2004). Due to space restriction, only a summary of Coh-Metrix's key features is presented here. The current public version available is Coh-Metrix 3.0, which can retrieve 108 scores of textual features. More information can be found on the website http://cohmetrix.Memphisedu/cohmetrixpr/index.html. A wide-range overview is provided in (Graesser et al., 2004):

Descriptive Indices: It includes the number of paragraphs, sentences, words, syllables in words, etc.

Cohesion: It is a key aspect for understanding the discourse structure of a language and how connectives used in a text have an impact on cohesion (Kintsch & van Dijk, 1978).

Sentence Complexity: It indicates human graders' evaluations of the quality of the text.

Lexical Sophistication: It refers to the writer's use of advanced vocabulary and word choice to express his or her thought.

New Features

This study proposes and extracts 8 new features that are not available in Coh-Metrix. These features refer to characteristics of ESL learners' writing styles and reflect on the importance of the introduction section, conclusion section and mechanics in errors including spelling and grammatical errors. Students often make the mistake of jumping straight to answering the essay question in the first paragraph without following a background statement, essay statement or outline statement. In addition, students rush to finish up in conclusion. The conclusion section should restate the author's stance with respect to the essay question, make a brief summary of evidences and finish with some sort of judgment about the topic. Moreover, spelling and grammatical errors are always good indicators of essay quality.

Number of Words in Introduction: the total number of words in the first paragraph considered as introduction.

Number of Words in Conclusion: the total number of words in the last paragraph considered as conclusion.

Introduction Portion: the ratios of number of words in introduction to the total number of words in the essay.

Conclusion Portion: the ratios of number of words in the conclusion to the total number of words in the essay.

Spelling Errors: the number of spelling errors. This study employs an open source spelling error checker called Language Tool (http://www.languagetool.org/), which is a part of the Open Office suite.

Grammatical Errors: the number of sentences with grammatical errors. This study uses the Link Grammar Parser (Lafferty, Sleator, & Temperley, 1992) to check the grammar of a sentence, which is also widely used in ESL context.

Percentage of Spelling Errors: the ratios of the number of spelling errors to the total number of words in the essay.

Percentage of Grammatical Errors: the ratios of the number of sentences with grammatical errors to the total number of sentences in the essay.

Methodology

Participants

Essays were collected from 90 freshmen at one of China's key universities. Among them, 41 students were English majors at the College of International Studies, while 49 students were chemistry majors at the School of Chemistry. English majors are considered to have the higher English proficiency. For the English majors, their average score in English as a testing subject in the National Higher Education Entrance Examination (also called Gaokao) was 131.30, and the standard deviation was 7.37. For the chemistry majors, their average score was 110, and the standard deviation was 10.14. Three experienced English teachers at the College of International Studies at the university volunteered to rate the quality of essays. All of them have at least five years of experience in teaching a composition course for both English and non-English majors.

Task and Instruments

The writing task was timed and considered as an assignment in English class. Students were required to finish it within 30 minutes. The writing task was to write a persuasive essay following the standard of college English essay writing set by the Ministry of Education in China.

The essays were rated by the three experienced English teachers mentioned above. They evaluated students' essays based on the standardized rubric commonly used to grade college English essay tests on the scale of 1 to 100. They first evaluated 18 essays. If the correlations between the teachers did not exceed r=.50 on each item, the evaluation process were rechecked until correlation was greater an equal to 0.5. After they reached a moderate agreement, each teacher then evaluated the 72 essays that comprised the whole sample used in this study.

It was found out that their inter-rater reliability was high with r=.756, r=.745, r=.607, respectively, p<.001. The scoring rubric included organization (e. g. clear organization of subtopics), content (e. g. clearly expressing ideas, text coherence, interesting and balanced introduction and conclusion) and mechanics (e. g. errors in punctuation and grammar).

These essays were chosen because the types they represented better reflected the conditions under which students usually completed prompt-based essays, such as CET or TEM. In addition, these two student groups can be representatives of most of the university students including English majors and non-English majors. Hence, the results of the selected features and algorithms are more likely to be accurate in the context of Chinese ESL writing. Indeed, the English majors' essays exhibit more discourse-level issues, while the non-English majors' essays exhibit both basic-level issues (spelling- and grammar-level) and discourse-level issues. This is the case due to English majors' more knowledge about the basics of the target language, English.

Results and Discussion

Descriptive statistics for the English majors and chemistry majors as well as the hybrid group (the combination of both essays) are reported in TABLE I.

TABLE I. DESCRIPTIVE AND ANOVA STATISTICS FOR ENGLISH MAJORS' AND CHEMISTRY MAJORS' ESSAYS IN THE DATASET

Features	English Majors	Chemistry Majors	F(1,71)	Hybrid
Raters' Essay Evaluations	70.45(9.95)	73.30(7.64)	1.362	72.10(8.72)
Number of Words	274(46.28)	136.47(4.00)	203.95*	194.65(76.56)
Number of Sentences	17.23(3.87)	9.30(0.52)	72.23*	12.65(5.14)
Number of Paragraphs	4.41(0.79)	3.03(1.03)	27.12*	3.62(1.16)
Number of Syllables per word	1.41(0.06)	1.61(0.10)	73.03*	1.53(1.31)
Number of Spelling Errors per Document of Words	2.9(1.94)	3(2.91)	8.77*	2.94(2.37)
Number of Grammar Errors per Document of Sentences	4.23(2.28)	6(2.23)	6.05*	4.98(2.40)

The average scores of the English majors' and chemistry majors' essays were not significantly different. The English majors and chemistry majors' essays were significantly different when number of words, sentences, paragraphs and syllables per word are involved. It indicates that the essays written by the English-major students contain more words, sentences and paragraphs, but less complicated words (less syllables), compared with the essays written by the chemistry majors. In addition, the English majors made less grammatical and spelling errors than the chemistry majors did.

Key Features for English Majors' Essays

Top six features were selected by using the same feature selection method used above, but this time applied on the training set (21 essays) written by the English majors. The linear regression yielded a significant model, $F(6,14)=10.982$, $p<0.001$, $r=.944$, $r2=.892$. Table II shows the six features that correlate with the essay scores. The conclusion portion was positively related to essay quality. But, the feature of the introduction portion was negatively related to the scores. It indicates the importance of the summarization of arguments in the final section of essays, as found in previous study (Freeman & Freeman, 1998). Cohesion as measured by content word overlap and Wordnet overlap were positively related to essay quality, which was similar to the results reported in a previous study (SA Crossley & McNamara, 2010). However, the argument overlap was

negatively correlated to essay quality. Argument overlap occurred when there were matching personal pronouns between sentences. It is observed that unskillful writers like to use a person's experience as an example to support the arguments in an illogical way. These essays contain many pronouns such as "he" and "his". The following text segment is extracted from one poor quality essay from the dataset. Although this example essay has high argument overlap, it lacks logic between the following two adjacent sentences: "My friend Bob, he often helped his parents do household jobs and got reward when he was young. So up to now, he always the best person I think, his experience makes him learn how to independent."

TABLE II. CORRELATIONS BETWEEN FEATURES AND RATERS' SCORES IN ENGLISH MAJOR GROUP IN THE TRAINING SET

Feature	Type	R	P
Introduction Portion	New feature	-.635	<.050
Conclusion Portion	New feature	.576	<.050
Argument Overlap	Cohesion	-.551	<.050
Content Word Overlap	Cohesion	.521	<.050
Temporal Connectives	Cohesion	-.803	<.001
WordNet Overlap between Verbs	Cohesion	.714	<.050

"*Temporal Connectives*" was negatively related to essay quality, because some poor-skilled writers incorrectly used some temporal connectives, such as "when", "since" and "as". As expected, English majors' essays showed issues at the discourse level such as temporal connectives and argument overlap which negatively correlated with the quality of the essays.

Regression Model Performance in the English Major Group

In order to validate the regression model consisting of six features, the model in these test sets (11 essays) were evaluated. It yielded r=.784, r2=.615. Therefore, this result demonstrates that the combination of six features account for 61.5% of the variance in the test set.

Categorical scores, including "distinction" (80-100), "credit" (70-79), "pass" (60-69) and "fail" (0-59), are also one of the common credit systems used at China's universities, such as Southwest University (University, 2007). These categorical scores are also used in many writing tests (Lawrence M. Rudner & Liang, 2002). The scores derived from the test set were used to assess categorical accuracy of the regression scores, compared with the human-graded scores. The regression model produced categorical matches for 7 of the 11 essays (64 % accuracy). The

reported, weighted Cohen's kappa for the categorical matches was 0.516, demonstrating a moderate agreement. A confusion matrix for this analysis is provided in TABLE III.

TABLE III. HUMAN CATEGORICAL SCORE PREDICTION IN THE ENGLISH MAJOR GROUP IN THE TEST SET

System Predicted Scores	Actual Human Scores			
	Distinction	Credit	Pass	Fail
Distinction	2	0	0	0
Credit	0	2	0	0
Pass	0	1	1	1
Fail	0	1	1	2

Key Features for the Chemistry Majors' Essays

The top seven features were selected by using the same feature selection algorithm as before, but this time applied on the training set (27 essays) written by the students majoring in chemistry. The linear regression yielded a significant model, $F(7,19)=3.186$, $P<0.05$, $r=.709$, $r2=.503$. TABLE IV presents the correlations between these features and scores. Among these features, it is observed that results are similar to those reported in other studies (Scott a. Crossley & McNamara, 2011; McNamara, Crossley, & Roscoe, 2013; Mcnamara, Crossley, & Mccarthy, 2010). Essay quality is positively related with essay length (number of words) and cohesion (semantic similarity between adjacent paragraphs). As expected, the new features "*Percentage of Spelling Errors*" and "*Percentage of Grammatical Errors*" were negatively related to the essay quality. Surprisingly, the syntactic complexity (incidence score of verbal phrases) was negatively related to essay quality, which was different to the results found in the previous study (Mcnamara et al., 2010). This may be one characteristic of ESL writers, since they are more likely to make grammatical mistakes if they try to write complex sentences. Another cohesion feature, "*Standard Deviation of the Semantic Similarity between Sentences*", showed negative correlations with essay quality. It indicates that the semantic inconsistency between sentences was negatively correlated to essay quality. Unlike studies in the past, there is a negative correlation between "*Logical Connectivity*" and essay scores. It is found out that many essays with poor marks had many "and" as a logical connective. It was used almost always for connecting two nouns or adjectives, such as "more and more popular", and "China and the West". As expected, the non-English majors show more problems at basic-levels of writing, such as spelling and grammatical errors.

TABLE IV. CORRELATION OF THE FEATURES AND RATERS' SCORES IN THE CHEMISTRY MAJOR GROUP IN THE TRAINING SET

Feature	Type	R	P
Number of Words	Descriptive	.676	<.050
Percentage of Spelling Errors	New Feature	-.486	<.050
Percentage of Grammatical Errors	New Feature	-.460	<.050
Logical Connectivity	Cohesion	-.450	<.050
Standard Deviation of the Semantic Similarity between Sentences	Cohesion	-.531	<.050
Semantic Similarity between Adjacent Paragraphs	Cohesion	+.528	<.050
Incidence Score of Verbal Phrases	Syntactic Pattern	-.641	<.050

Regression Model Performance in the Chemistry Major Group

In order to validate the regression model consisting of seven features, this model in the test set (13 essays) written by the chemistry majors was evaluated. It yielded r=.569. The scores derived from the test set were used to assess the categorical accuracy of the regression scores, compared with the human-graded scores. The regression model produced categorical matches for 8 of the 13 essays (54 % accuracy). The reported, weighted Cohen's kappa for the categorical matches was 0.404, demonstrating a moderate agreement. A confusion matrix for this analysis is provided in TABLE V.

TABLE V. HUMAN CATEGORICAL SCORE PREDICTION IN THE CHEMISTRY MAJOR GROUP IN THE TEST SET

System Predicted Scores	Actual Human Scores			
	Distinction	Credit	Pass	Fail
Distinction	2	0	0	0
Credit	0	2	0	0
Pass	0	1	1	1
Fail	0	1	1	2

This matrix reflects a decrease in the categorical agreement using the model tested in the dataset of the chemistry majors' essays. The predicted scores tend to be in the "credit" category since 8 of the 13 essays have been predicted in the "credit" category. This level of performance is partially due to the frequent credit scores and small variations of actual human scores (SD: 7.64), which renders the prediction task more difficult.

Conclusion and Future work

This study has used a set of Coh-Metrix indices combined with a set of newly proposed features to analyze ESL essays written by the English majors and non-English majors at a university in China. It showed the predictive values of several features extracted using Coh-Metrix; some of the newly proposed features significantly correlated to essay quality as well. These features include cohesion at the sentence- and paragraph-level, introduction and conclusion portion, syntactical complexity and surface errors. The results indicate the usefulness of Coh-Metrix and the newly proposed new features. Interestingly, different features are more significant for different groups of essays. The English majors emphasize cohesion between sentences, writing a good summarization, whereas the non-English majors focus on making less surface errors, such as spelling and grammatical errors, and cohesion between adjacent paragraphs.

This study has some limitations. For example, the sample size is not big enough, since 72 essays and two groups of ESL writers were analyzed. However, these essays were written by university students in a real scenario and the data analysis process was sound. In the future, improving the performance of the prediction model will be the focus. At the present, most of the studies use a linear regression model for essay score prediction. Non-linear regression models, such as SVM Regression (Shevade, Keerthi, Bhattacharyya, & Murthy, 1999) and other machine learning techniques (Hongbo Chen, Ben He, Tiejian Luo, 2012; Larkey, 1998) will be investigated. Moreover, more ESL essays written by university students from different disciplines will be collected and analyzed.

Acknowledgement:

This article was supported by Chongqing Social Science Planning Fund Program [2014BS123], Fundamental Research Funds for the Central Universities [XDJK2014A002], [XDJK2014C141] and [SWU114005] in China.

References

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21*(2/3), 170. http://doi.org/10.1504/IJCEELL.2011.040197

Crossley, S., & McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In *The 32nd Annual Conference of the Cognitive Science Society* (pp. 984–989). Austin: TX. Retrieved from http://csjarchive.cogsci.rpi.edu/Proceedings/2010/papers/0310/paper0310.pdf

Freeman, Y. S., & Freeman, D. E. (1998). *ESL/EFL Teaching: Principles for Success*. Heinemann.

Ge, S., & Chen, X. (2007). Automated Essay Scoring for Chinese EFL learners. *Foreign Language World*, *122*(5), 43–50.

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, & Computers*, *36*(2), 193–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15354684

Han, N. (2009). The theory and practice of Automated English Essay Scoring Systems. *China Test*, *3*, 38–44.

Hongbo Chen, Ben He, Tiejian Luo, B. L. (2012). A ranked-based learning approach to automated essay scoring. In *2012 Second International Conference on Cloud and Green Computing (CGC)* (pp. 448–455).

Kintsch, W., & van Dijk, T. (1978). Towards a model of text comprehension and production. *Psychological Review*, *85*, 363–394.

Lafferty, J., Sleator, D., & Temperley, D. (1992). Grammatical Trigrams: A Probabilistic Model of Link Grammar. In *Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language*.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '98*, (1), 90–95. http://doi.org/10.1145/290941.290965

Lawrence M. Rudner, & Liang, T. (2002). Automated essay scoring using Bayes' Theorem. *The Journal of Technology, Learning and Assessment*, *1*(2).

Li, J. (2009). *Using latent semantic analysis for automated essay scoring in the Chinese EFL context*. Guangdong University of Foreign Studies.

Liang, M. (2011). *Construting a model for the computer assisted scoring of Chinese EFL learners' argumentative essays*. Foreigh Language Teaching and Research Press.

Liang, M., & Wen, Q. (2007). A critical review and implications of some automated essay scoring systems. *Computer Assisted Foreign Language Education*, 18-24.

Liang, Q. (2004). Contrastive study on the objectivity of English and Chinese argumentative writing: A survey on the objectivity of Chinese college students' English argumentative writing. *Journal of China West Normal University (Philosophy & Social Sciences)*, (5).

Liu, X. (2008). A case study of a non-English Major's Writing: Error analysis. *Foreign Language Research*, *2*, 140–142.

McNamara, D. S., Crossley, S. a, & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515. http://doi.org/10.3758/s13428-012-0258-1

Mcnamara, D. S., Crossley, S. A., & Mccarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1). http://doi.org/10.1177/0741088309351547

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. http://doi.org/10.1016/j.asw.2014.09.002

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Rus, V., & Niraula, N. (2012). Automated detection of local coherence in short argumentative essays based on centering theory. In *13th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 450–461). New Delhi, India.

Shermis, M., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah NJ: Lawrence Erlbaum Associates.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. http://doi.org/10.1016/j.asw.2013.04.001

Shevade, S., Keerthi, S., Bhattacharyya, C., & Murthy, K. (1999). Improvements to the SMO Algorithm for SVM Regression. *IEEE Transactions on Neural Networks*.

University, S. (2007). *Implementation guidances on the reform of credit system in Southwest University*. Chongqing. Retrieved from agronomy.swu.edu.cn/up/87.doc