International Journal of Learning, Teaching and Educational Research
Vol. 11, No. 1, pp. 1-11, April 2015.

Using Natural Language Processing Technology to Analyze Teachers' Written Feedback on Chinese Students' English Essays

Ming Liu, Weiwei Xu, Qiuxia Ran and Yawen Li
Southwest University
Beibei District, Chongqing, China

Abstract. Writing an essay is a very important skill for students to master, but a difficult task for them to overcome. It is particularly true for English as Second Language (ESL) students in China. It would be very useful if students can receive timely and effective feedback about their writing. In order to build an automatic feedback system, we need to understand the relationship between textual features and human teacher feedback, and how well those features were used for predicting feedback rating. In this study, we analyzed 105 Chinese English majors' essays with teachers' feedback and used Coh-Metrix, a computational linguistic tool, to extract features from their writing. The study results showed some feedback was moderately correlated to some textual features (e.g. text easability cohesion and lexical diversity were related to coherence feedback) and those feedback are more predictable, such as spelling, grammar, supporting ideas and coherence. This finding has important implications for building automated writing feedback tool.

Keywords: Writing Feedback, Text Analysis, Natural Language Processing.

1. Introduction

With the coming of the 21st century and the globalization of English, English essay writing, as one of the four basic skills of language learning, has become a more and more important skill. It not only requires some basic writing skill, such as spelling and grammar, but also asks some high competency of writing, such as coherence, structure and reasoning. Thus, it is also a difficult task to overcome. It is particularly so in China. Statistics show that the number of college students in China has soared to twenty-six million in 2013 (Bureau of Statistis of China, 2013), accounting for the largest proportion of ESL learners worldwide. Since 1987, the writing test has become one important aspect of the College English testing in China. As for college students in China, college English has been an obligatory course to take. In a typical English course,

students have to do 2-3 essay writing assignments and take 1 essay writing test in order to pass national English tests, such as College English Test (CET) 4 or Test for English-Major (TEM) 4. Essay writing is the last part of these tests. Novice writers need feedback to develop their writing skills; however, providing timely and meaningful feedback is time-consuming and expensive.

Since the early 1980s, researchers have investigated feedback on students' writing (Brannon & Knoblauch, 1982). These study results showed that written feedback provided a potential value in motivating students to revise their draft and improving their writing (Leki, 1991). As a result, written feedback is the most popular method among various feedback delivery modes (oral feedback, audiotaped and writing conference) that teachers use to interact and communicate with students. Straub (Straub, 2000) suggested that the effective teacher feedback should be written in complete sentences, avoid abstract, technical language and abbreviations, relate their comments back to specific words and paragraphs from the students' text, by viewing students' writing seriously, as part of a real exchange. In addition, an increasing number of studies have also been conducted to see whether certain types of feedback are more likely than others to help ESL students improve the accuracy of their writing, such as direct and indirect feedback (Lee, 2004). Direct or explicit feedback occurs when the teacher identifies an error and provides the correct form, while indirect strategies refer to situations when the teacher indicates that an error has been made but does not provide a correction, thereby leaving the student to diagnose and correct it.

With the advanced development of information technology and natural language processing techniques, various numbers of automatic essay scoring (AES) systems have been proposed. Haswell (Haswell, 2006) reviewed systems for automated feedback tracing back to the 1950s. These systems focused more on assessment of end products, and less on providing formative feedback (Shermis & Burstein, 2003; Williams & Dreher, 2004) The Writer Workshop (Anderson, 2005) and Editor (Thiesmeyer & Theismeyer, 1990) both focus on grammar and style. Sourcer's Apprentice Intelligent Feedback system (SAIF) (Britt, Wiemer-Hastings, Larson, & Perfetti, 2004) is a computer assisted essay writing tool used to detect plagiarism, uncited quotations, lack of citations, and limited content integration problems. The Glosser system (Villalon, Kearney, Calvo, & Reimann, 2008) aims to support reflection in writing through trigger questions. It uses text mining algorithms to help learners think about issues such as coherence, topics, and concept visualization. However, Glosser only provides generic trigger questions. Liu et al. (Liu, Calvo, & Rus, 2014; Liu, Calvo, & Rus, 2010) investigated an automatic trigger question generation system which could support critical review writing.

The aim of this study is to investigate the frequent type of feedback used by human teachers and the relationship between the feedback and the textual features extracted by using the natural language processing techniques.

The rest of this paper is constructed as follows: Section 2 presents related work on feedback classification. Section 3 describes the study and discusses the results. Finally, Section 4 concludes this paper.

Table 1: Criterion Category

Criterion Category	Examples
Grammar	Fragments, Run-on Sentences Subject-verb agreement, Ill-formed verbs Pronoun Error, Missing Possessive Error
Usage	Wrong article, Missing article Confusing words, Wrong form of word Preposition Error
Mechanics	Spelling, Capitalize Proper Nouns Missing Question mark, Missing final punctuation Missing Apostophe, Missing Comma
Style	Repetition of words, Inappropriate words or phrases Too many short sentences, Too many long sentences
Organization	Background, Thesis, Main-point Supporting ideas, Conclusion

2. Related Work

Recent development in natural language processing techniques has made it possible for researchers to develop a wide range of sophisticated techniques that facilitate text analysis. Some tools, such as Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), LIWC (Pennebaker & Francis, 1999) and Gramulator (Rufenacht, McCarthy, & Lamkin, 2011), are useful in this respect, and have certainly contributed to ESL knowledge (S. A. Crossley & McNamara, 2012). Coh-Metrix is a powerful computational tool that provides over 100 indices of cohesion, syntactical complexity, connectives and other descriptive information about content (Graesser et al., 2004). Coh-Metrix has extensively been used to analyze the overall quality of writing (S. A. Crossley & McNamara, 2012) and one important aspect of writing quality, such as coherence (Scott a. Crossley & McNamara, 2011a). For example, Crossley and McNamara found that computational indices related to text structure, semantic coherence, lexical sophistication, and grammatical complexity best explain human judgments of text coherence. This study focused on using Coh-Metrix to analyze more aspects of writing quality including, Supporting Ideas, Conclusion and Sentence Diversity.

The AES systems, such as Criterion (Burstein, Chodorow, & Leacock, 2004), can provide feedback on some aspects of writing including grammar, usage, mechanics, style, organization, development, lexical complexity and prompt-

specific vocabulary usage (See Table 1). The Criterion categories are more relevant to our case since we aim to generate corrective feedback on different aspects of ESL student writing.

3. Study

We conducted an empirical study in analyzing Chinese ESL college student essays with teachers' comments and the relationship between the teacher feedback and textual features. Section 3.1 describes the annotation process, where each essay is scored in different aspect, such as *Grammar, Spelling, Coherence, Organization and Supporting Ideas*. Section 3.2 shows the textual feature extraction process. Section 3.3 illustrates the relationship between the textual features and each feedback category, while section 3.4 examines the predictive strength of the features in explaining the score variance in the each feedback score.

3.1 Proposed Feedback Taxonomy

Table 2: Feedback Frequency and Pearson Correlations between Raters

Feedback Category	Frequency	r
Grammar	48	.824
Spelling	12	.504
Word Count	24	.707
Sentence Diversity	40	.454
Conclusion	44	.747
Supporting Ideas	98	.632
Coherence	40	.716
Chinglish Expression	24	.352
Organization	89	.534

Our dataset containing 105 English majors' essays with teachers' feedback was collected from a large university in China. Two experienced English teachers volunteered to rate the quality of the essays. They had at least five years of teaching composition course for English majors. Their first task was to identify the most frequent feedback type adapted from the standardized rubric used for grading college English. 9 frequent feedback categories were found, including *Grammar, Spelling, Word Count, Sentence Diversity, Conclusion, Supporting Ideas, Organization, Coherence and Chinglish* (See Appendix I). Table 2 shows that Supporting Ideas and Organization categories were more frequent than others, while Spelling and Chinglish Expression and word count were less frequent. We observed some feedback categories were similar to the Criterion categories, such as Grammar, Spelling and Supporting Ideas. But, the Chinglish Expression and Conclusion categories only appeared in our dataset.

The teachers' second task was to give a score to each feedback category regarding to the rubric (See Appendix I) on a scale of 3. 1 means negative

feedback on the category while 3 means positive feedback on the category. The Correlations between the raters are located in Table 2. The raters had the highest correlations for judgments of Grammar, Word Count, Conclusion and Supporting Ideas and the lowest correlations for Chinglish and Sentence Diversity.

For further analysis, the dataset was randomly divided into training set (n=70) and testing set (n=35). A training set was used to identify which of the textual features most highly correlated with each feedback score. Moreover, the training set was used to train a multiple regression model to examine the amount of variance explained by each writing feature. The model was then applied to a test set to calculate the accuracy of the analysis.

3.2 Textual Feature Extraction

We used Coh-Metrix 3.0, which could retrieve 108 scores of textual features. More information can be found on the website (http://cohmetrix.Memphisedu/cohmetrixpr/index.html).

Descriptive indices: It includes the number of paragraphs, number of sentences, number of words, number of syllables in words, mean length of paragraphs etc.

Cohesion: Cohesion is a key aspect of understanding language discourse structure and how connections within a text influence cohesion and text comprehension(Kintsch & van Dijk, 1978). Coh-Metrix employs referential cohesion including noun overlap, argument overlap, stem overlap, and LSA-based semantic overlap.

Sentence Complexity: The grammatical structure of a text is also an important indicator of human evaluations of text quality. Difficult syntactic constructions (syntactic complexity) include the use of embedded constituents, and are often dense, ambiguous, or Ungrammatical (Graesser et al., 2004). Syntactic complexity is also informed by the density of particular syntactic patterns, word types and phrase types.

Lexical sophistication: Lexical sophistication refers to the writer's use of advanced vocabulary and word choice to convey ideas. Lexical sophistication is captured by assessing the type and amount of information provided by the words in a text. Words are assessed in terms of rarity (frequency), abstractness (concreteness), evocation of sensory images (imagability), salience (familiarity), and number of associations (meaningfulness). Words can also vary in the number of senses they contain (polysemy) or levels they have in a conceptual hierarchy (hypernymy).

Moreover, we propose and extract 8 new features that are not available in Coh-Metrix. These features refer to characteristics of ESL learners' writing style and reflect on the importance of the introduction section, conclusion section and mechanics in errors including spelling errors and grammatical errors. In the database, each essay is stored as a plain text, where each line is a paragraph. We

use Java API to extract the first line and last line text, as introduction and conclusion section respectively. For checking spelling errors, an open source spelling error checker, called LanguageTool (http://www.languagetool.org/), is employed to scan each word. For checking grammatical errors, the Link Grammar Parser (Lafferty, Sleator, & Temperley, 1992) is used to check the grammar of a sentence based on natural language processing technology. If the link grammar could not generate links (relations between pairs of words) after parsing a sentence, this sentence would be considered as ungrammatical.

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

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

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

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

Spelling errors: the number of spelling errors. We employ an open source spelling error checker called LanguageTool (http://www.languagetool.org/), which is part of the OpenOffice suite.

Grammatical errors: the number of sentences with grammatical errors. We use the Link Grammar Parser (Lafferty et al., 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 word spelling errors to the total number of words in the document.

Percentage of grammatical errors: the ratios of the number of sentence with grammatical errors to the total number of sentences in the document.

Therefore, there are totally 116 features extracted from each essay.

3.3 Pearson Correlation

Table 3: Correlations between Textural Features Scores and Raters' feedback scores

Feedback Category	Features	R	P value
Chinglish	Gerund incidence	.415	<0.05
	paragraph length	.459	<0.05
	first person singular pronoun incidence	.493	<0.01
Coherence	Text Easability Cohesion	.433	<0.05
	Lexical diversity	.402	<0.05
Conclusion	Conclusion Portion	.477	<0.05
	Lexical diversity	.394	<0.05
Supporting Ideas	Intentional verbs incidence	.496	<0.05
	Adjective incidence	.503	<0.05
	CELEX Log minimum frequency for content words	.541	<0.01
Grammar	Grammar errors	-.606	<0.01
Sentence Variety	Hypernymy for verbs	.506	<0.01
	Standard deviation of Sentence length	.413	<0.05
Spelling	Spelling Errors	-.617	<0.05
Organization	Number of paragraphs	.507	<0.01
Word Count	Word count	.666	<0.01

Based on the system producing feature scores and the human annotators' score on each category, we used IBM SPSS for evaluating the Pearson correlation between textual features and each category. Over 30 textual features demonstrated significant correlations with the human ratings of each feedback category. Table 3 shows the *Chinglish* was more related to the number of Gerund used, the paragraph length and the first person singular pronoun incidence. The *Coherence* was correlated to Text Easability PC Deep cohesion, consistent with Crossley and McNamara's study result (S. Crossley & McNamara, 2010). As expected, the *Conclusion* was more related to the features of Conclusion Portion and Lexical Diversity. We have not defined specific features which can detect the *Supporting Ideas*. However, some features, such as Intentional verbs and Adjective incidence, have shown their moderate correlations with the category of *Supporting Ideas*. As we had expected, the *Grammar* and *Spelling* were negatively related to the features of grammar error and spelling error. The *Word Count* was correlated to the number of words in an essay. *Organization* was correlated to the number of paragraphs since the essays with only 1 or 2 paragraphs were given lower scores by human annotators since they did not have a clear essay structure, *introduction, body and conclusion*. Crossley and MacNamara (Scott a. Crossley & McNamara, 2011b) got the similar study results, where six features including *the total number of paragraphs* were significant predictors in the regression to the raters' organization evaluations.

3.3 Test Set Model

We used the training set to train a regression model for each feedback category and evaluated the model in testing set. Table 4 shows the performance of each regression model for predicting essay feedback ratings. It has been found that Grammar (r^2=.881) and Spelling feedback (r^2=.886) were easier for prediction, since some textual features were highly related to those feedbacks. It also demonstrated that the combination of the textual features accounted for 88.1% of the variance in the grammar evaluation of the 35 essays comprising the test set. On the other hand, organization and conclusion were difficult to predict since r^2=.223 and r^2=.380 respectively since the textual features were not correlated to those feedback ratings.

Table 4: Linear Regression Analysis to Predict Essay Feedback Ratings in Testing Set

Feedback	R	R^2	S.E.
Chinglish Expression	.764	.584	.349
Coherence	.790	.624	.472
Conclusion	.616	.380	.486
Supporting Ideas	.745	.555	.407
Grammar	.939	.881	.260
Sentence Variety	.735	.540	.423
Spelling	.941	.886	.242
Organization	.475	.223	.473
Word Count	.756	.572	.535

Notes: S.E. is standard error

4. Conclusion

Human teachers' written feedback is very useful for students to revise their draft and improve writing. A great number of researches has been conducted to investigate the theoretical foundation of feedback in terms of feedback mode, feedback strategies and feedback classification. With the development of information technologies, automated essay scoring tools have been proposed, which can extract textual features and generate corrective feedback on the traits of writing including grammar, usage, style, mechanics and organization. However, these AES systems are mainly designed for international ESL students, who take TOFEL test. Those students can only represent a small portion of ESL students, because they obviously possess a higher English competency. Thus, we conducted an empirical study to investigate the frequent feedback types and examine the feasibility of using existing natural language processing tools to automatically measure the feedback.

In the study, we collected 105 essays written by English majors and some teachers' comments at a large university in China. Two English teachers first found 9 frequent feedback categories based on the teachers' comments. Some feedback categories are consistent with the Criterion category. Then, they gave a

score on a scale of 1 to 3 to each feedback category of each student essay. The study results showed that the feedback had moderate correlations with some features extracted by using Coh-Metrix, a computational writing analysis tool, and some proposed new features. For example, coherence feedback was highly related to Text Easability Cohesion and Lexical diversity, while Supporting Ideas was related to Intentional verbs incidence and Adjective incidence. Moreover, it has been found that some feedback, such as supporting ideas, coherence, grammar and spelling, were more predictable. It indicated the feasibility of using existing NLP tools to measure the quality of feedback.

Our future work will examine teachers' comments in detail and collect non-English major student essays for analysis. In addition, we will focus on building an automatic essay feedback generation system. Specifically, we will investigate the feedback generation mechanism by using association rule mining algorithms. In addition, we will look at how to incorporate effective feedback strategies, such as formative feedback theory, into feedback generation templates.

Acknowledgment

The authors would like to thank those teachers and student participants. This work is partially supported by Chongqing Social Science Planning Fund Program under grant No. 2014BS123, Fundamental Research Funds for the Central Universities under grant No. SWU114005, No. XDJK2014A002 and No. XDJK2014C141 in China.

Appendix A

Table 5: Nine Traits Rubric for Essay Writing

Category	Scoring
Organization	1 Rudiment of organization apparent, but may be illogical, ineffective or different to understand the sequencing of ideas 2 Satisfactory organization of sections, but the sequencing of paragraphs within sections may be problematic. 3 Effective method of organization for both section and for paragraphs within sections.
Supporting Ideas	1 Minimal use of examples and facts to support the writer's idea. 2 using some examples and facts to discuss strengths/weakness of some opinions, but may have difficulties (1) choosing appropriate facts; (2) sufficiently explaining those facts; (3) connecting them to present thing. 3 Effective supports the strengths and weakness of one's opinion; Generally effective use of choice of examples and facts, although some material may be extraneous or not adequately explained
Grammar	1 Uses simple sentence constructions, but there are still numerous errors (greater than 7).

	2 Uses simple sentence with minor errors (between 5-7). 3 Uses complex sentence with minor errors (less than 5).
Sentence Variety	1 Little complex sentences or longer sentences (less than 2) are used 2 Moderate number of complex sentences or longer sentences (between 2 and 4) are used 3 A Effective use of complex sentence construction or longer sentence (greater than 4)
Coherence	1 Some apparent sequencing of sentences within paragraphs, relying primarily on a limited set of cohesive devices (e.g. first, second, third) and basic connection words (e.g. however, also, because). However, there may be frequent points in which the reader has difficulties understanding sequencing of ideas. 2 Writer sequences ideas, relying primarily on a limited set of cohesive devices; some errors or unclear transitions, but they do not significantly impair understanding of the text. 3 Coherent and logical sequencing of ideas, using a wider range of cohesive devices (e.g. pronominalization, passive, etc;) only minor and occasional errors.
Word Count	1 Less than 50 words 2 Between 50 and 150 words 3 Around 200 words
Conclusion	1 No conclusion key words found; Conclusion is inappropriate; No conclusion 2 briefly summarized some points 3 It stresses the importance of the thesis statement, gives the essay a sense of completeness.
Spelling	1 greater than 3 2 within 1 and 3 3 no spelling error
Chinglish Expression	1 greater than 5 2 within 3 and 5 3 less than 2

References

Anderson, J. (2005). Mechanically Inclined:Building Grammar, Usage, and Style into Writer's Workshop.

Brannon, L., & Knoblauch, C. H. (1982). On students' rights to their own texts: A model of teacher response. *College Composition and Communication, 33*, 157-166.

Britt, M. A., Wiemer-Hastings, P., Larson, A. A., & Perfetti, C. A. (2004). Using Intelligent Feedback to Improve Sourcing and Integration in Students' Essays. *Int. J. Artif. Intell. Ed., 14*, 359-374.

Bureau of Statistis of China, N. (2013). China Statistical YearBook.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25*, 27. doi: 10.1002/rcm.5057

Crossley, S., & McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *The 32nd Annual Conference of the Cognitive Science Society*. Austin: TX.

Crossley, S. a., & McNamara, D. S. (2011a). Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence. *The 33rd Annual Conference of the Cognitive Science Society*.

Crossley, S. a., & McNamara, D. S. (2011b). 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*, 170. doi: 10.1504/IJCEELL.2011.040197

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency： The role of cohesion， readability， and lexical difficulty. *Journal of Research in Reading, 35*, 115-135.

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*, 193-202.

Haswell, R. (2006). The complexities of responding to student writing; or, looking for shortcuts via the road of excess. *Across the Disciplines, 3*.

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*. Paper presented at the Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*, 285-312. doi: 10.1016/j.jslw.2004.08.001

Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals, 24*, 203-218.

Liu, M., Calvo, R., & Rus, V. (2014). Automatic Generation and Ranking of Questions for Critical Review. *Educational Technology & Society, 17*, 333-346.

Liu, M., Calvo, R. A., & Rus, V. (2010). Automatic Question Generation for Literature Review Writing Support. Carnegie Mellon University, USA: Springer's Lecture Notes in Computer Science

Pennebaker, J. W., & Francis, M. E. (1999). Linguistic inquiry and word count (LIWC).

Rufenacht, R. M., McCarthy, P. M., & Lamkin, T. A. (2011). Fairy Tales and ESL Texts: An Analysis of Linguistic Features Using the Gramulator. *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*.

Shermis, M. D., & Burstein, J. (2003). Automated essay scoring: A cross-disciplinary perspective. *16*.

The student, the text, and the classroom context: A case study of teacher response, 7 23-55 (2000).

Thiesmeyer, E. C., & Theismeyer, J. E. (1990). Editor:A System for Checking Usage, Mechanics, Vocabulary, and Structure.

Villalon, J., Kearney, P., Calvo, R. A., & Reimann, P. (2008). Glosser: Enhanced Feedback for Student Writing Tasks.

Williams, R., & Dreher, H. (2004). Automatically Grading Essays with Markit©. *Issues in Informing Science and Information Technology, 1*, 693-700.