Automated Essay Feedback Generation and Its Impact in the Revision

Ming Liu, Yi Li, Weiwei Xu and Li Liu

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 could receive timely and effective feedback about their writing. Automatic essay feedback generation is a challenging task, which requires understanding the relationship between the text features of the essay and feedback. In this study, we first analyzed 1290 teacher comments on their 327 English-major students and annotated the feedback on seven aspects of writing, including the *grammar*, *spelling*, *sentence diversity*, *structure*, *organization*, *supporting ideas*, *coherence and conclusion*, for each paper. Then, an automatic feedback classification experiment was conducted with the machine learning approach. Finally, we investigated the impact of the system generated-indirect corrective feedback (ICF) and human teachers' direct corrective feedback (DCF) in two English writing classes (N=56 in ICF class; N=54 in DCF class) at a key Chinese university through a web-based assignment management system. The study results indicated the feasibility of this approach that system generated ICF can be as useful as direct comments made by the teachers in terms of improving the quality of the content regarding to the *structure*, *organization*, *supporting ideas*, *coherence and conclusion*, and encouraging students to spend more time on self-correction.

Index Terms—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 for some high competency of writing, such as coherence, structure and reasoning. Thus, it is a difficult task to overcome. It is particularly true for students in China. Statistics show that the number of college students in China has soared to twenty-six million in 2013 [1], accounting for the largest proportion of English as Second Language (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 is an obligatory course to take and a fair score of the College English Test is required of all Chinese students graduating from any university. In a typical English course, students have to do two or three essay writing assignments and take an 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. As to the EFL (English as Foreign Language) teaching practice in China, where big class is the norm with enormous amount of information to be dealt with and learning is largely exam oriented, getting timely feedback for each EFL learner's writing task is thus often difficult.

Since the early 1980s, researchers have investigated the effectiveness of teacher feedback as a way of improving students' writing [2]. A substantial amount of research on teacher written feedback in ESL writing contents has been concerned with the benefits of the corrective feedback in students' writing development [3]. Corrective feedback is a commonly used feedback type in classrooms: the marking of a student's error by the teacher. Fathman and Whalley [4] found positive effects for rewriting from corrective feedback on both grammar and content. However, trying to establish a direct link between corrective feedback and successful second language acquisition is oversimplistic and highly problematic [5].

An increasing number of studies have been conducted to see whether certain types of written feedback are more likely than others to help ESL students improve the accuracy of their writing, such as ICF and DCF [6]–[9]. DCF or explicit feedback occurs when the teacher identifies an error and provides the correct form or explicit suggestions to fix the problem, while ICF or implicit feedback refers 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.

- *M.Liu is with the School of Computer and Information Science, Southwest University, Chongqing 400715, China. E-mail: mingliu@swu.edu.cn*
- *Y.Li is with the Faculty of Education, Southwest University, Chongqing 400715, China, E-mail: liyi1807@swu.edu.cn.*
- *W.W. Xu is with the College of International Studies, Southwest University, Chongqing 400715, China, E-mail: 534514401@qq.com*
- *L. Liu is with School of Software Engineering, Chongqing University, Chongqing 400044, P.R.China, E-mail: dcsliuli@cqu.edu.cn.*

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography(**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

Studies examining the effects of these different types of error feedback on students' second language (L2) writing, have reported positive impacts of ICF on the ability of students to edit their own composition and to improve levels of accuracy in writing because the ICF leads to a reflection on writing and a greater cognitive engagement [10]–[12]. Indeed, Reflection is an important language learning step [13]. ICF encourages students to critically evaluate their own written performance in the target language with the goal of improving not only their linguistic competence and skill, but also their ability to learn [10], [14], [15].

With the advanced development of natural language processing techniques and statistical models, several commercial automated essay scoring systems were developed, such as e-rater developed by Educational Testing Service (ETS) [16], Knowledge Analysis Technologies and IntelliMetric [17] to analyze a wide range of text features at lexical, syntactic, semantic, and discourse levels. Based on these scoring systems, some automated writing evaluation (AWE) tools, such as Criterion[18] and MYAccess, have been developed to provide corrective feedback or scores on various rhetorical (e.g., organization) and language-related dimensions (e.g., grammar and mechanics) [19]. Advantages of automated feedback are its anonymity, instantaneousness, and encouragement for repetitive improvements by giving students more practice for writing essays [20]. Some researchers have argued that AWE might lead to negative effects on students' writing behavior [21] since students focused on improving scores, rather than content development.

Few researchers [22]–[25] attempted to generate ICF on content development. The generated feedbacks or questions were used to scaffold student reflection on the different aspects of the writing content, such as cohesion and organization. For example, the Glosser system [22] used text mining algorithms, such as Latent Semantic Analysis [26], to provide content clues about issues related to coherence and topics to scaffold students reflection with a set of generic trigger questions.

We consider the work of Glosser that points in the direction that we have followed in this project. In our approach, the system first points out the weaknesses of some aspects of the writing, such as organization and structure. Then, it provides trigger questions to support student self-reflection on the weaknesses in the writing. Lastly, students performed self-correction on the writing. The system-generated ICF suggested students to "double-check" their essay with several trigger questions provided. Instead of revising only to correct errors, students try to reconsider and refine the whole text. The aim of this study is to explore the challenges of automated essay feedback generation and the effects of system-generated ICF during the revision in the context of Chinese ESL college students' writing. To fulfill the above-mentioned aims, the following research questions were posed:

1. What are the frequent aspects of the essay commented by college English teachers in the context of Chinese ESL learners?
2. Can these comments or feedback be automatically detected using the machine learning approach?
3. What is the impact of the system generated ICF and human teachers' direct comments on the quality of writing?

The rest of this paper is constructed as follows: Section 2 presents related work on automated essay feedback systems. Section 3 and 4 describe our approach to automatic essay feedback generation and the system evaluation result. Section 5 presents the user study that investigates the impact of two types of feedbacks, ICF and DCF, in the quality of writing and discusses the results. Finally, Section 6 concludes this paper.

2 Survey on Automated Essay Feedback

Automated feedback systems for writing support can be traced back to Automated Essay Scoring (AES) developed in the 1950s. Essay assessment is a time-consuming and costly process. Sometimes, it leads to many inconsistencies in the grades given by different human raters [27]. The early AES systems were used to overcome time, cost and reliability issues in writing assessment [28]. Project Essay Grader (PEG) is one of the first AES systems that used an essay's objective features, such as word count or spelling errors, and gave a score about the quality of each feature. The experimental results indicated that the system-predicted scores were comparable to those of human raters. However, PEG only focused on the surface features and ignored the semantic aspects of essays, such as coherence [29]. With the advance of text mining techniques, these AES systems can provide scores as feedback on semantic aspect of writing, such as topic coverage, discourse structure and coherence [30]. Haswell et al. [19] argued that these AES systems focused primarily on providing holistic grades and less on meaningful feedback on writing[31]. Ericsson and Haswell [32] also criticized these AES systems, as they deemed them focused mainly on providing feedback to improve the grades. The authors claimed that such approach devaluated the role of teachers as well as warped students' notions of good writing. However, according to other researchers [18] these tools could motivate students to write and revise. Gibbs and Simpson [33] defined several characteristics for an effective feedback, stating that it should be timely, specific enough, and focus on learning rather than marks. Despite a variety of initiatives to improve the quality of automatic feedback, the effectiveness of proposed systems remains to be proven and further research is needed.

Because AES systems have been developed for assessment, rather than to assist learning, many researchers [34] have tried to bring the focus back to learning(through automated feedback) instead of scores. They used technologies similar to AES systems to extract document features, trying to translate these features into useful information, typically related to the common writing problem. One of the challenges of these approaches is to make the feedback specific, so that students can understand how to improve their writing.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

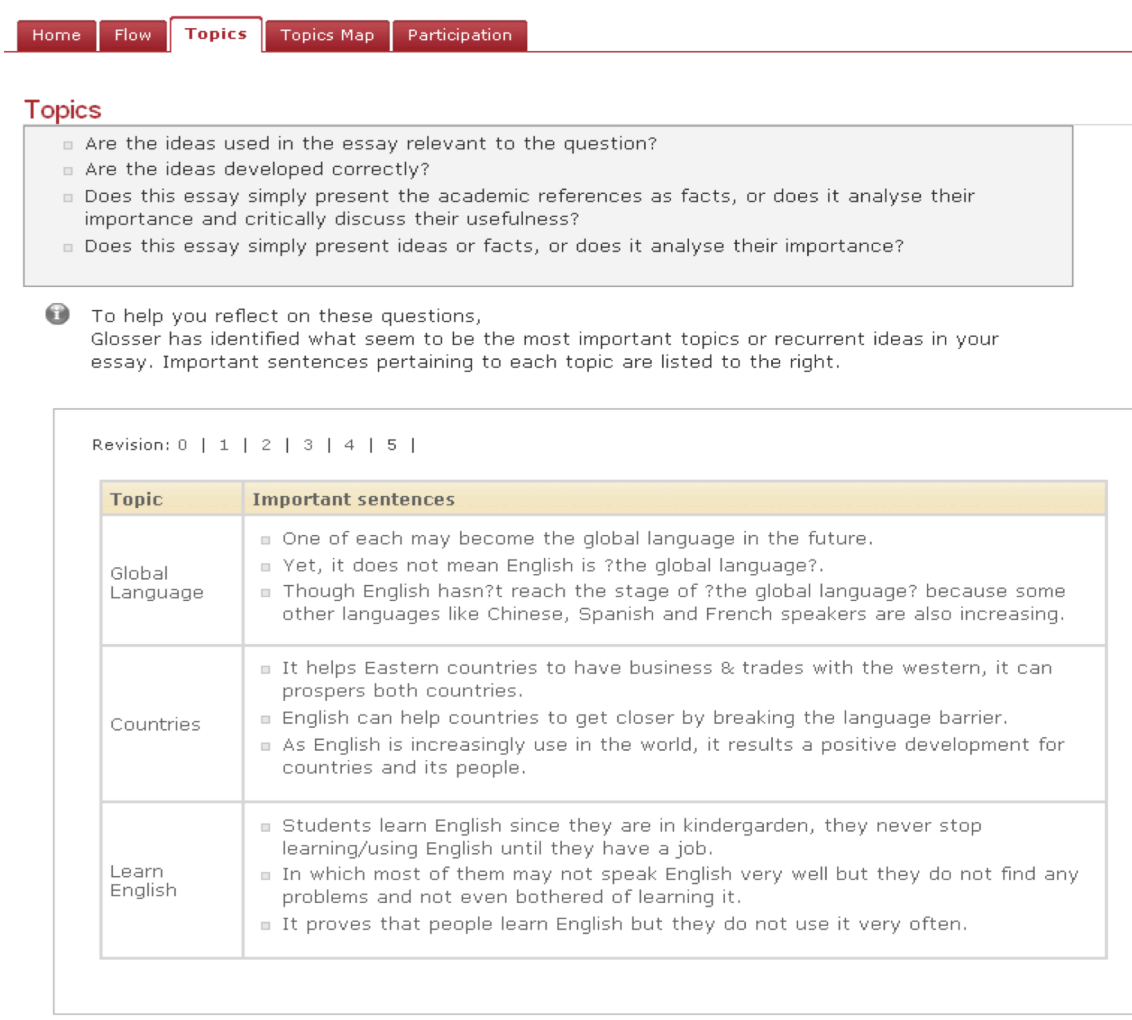


Figure 1: The user interface of the topics tool in Glosser. The figure is reproduced from Calvo et al. [22]

Many automatic writing feedback systems have been designed to address specific writing problems. Some of the early systems including Editor [35], developed at Rochester Institute of Technology and Writers Workshop [36], developed at Bell Laboratories, focused on grammar and style check. Research studies on the impact of Editor [37] concluded that the pedagogical benefits of grammar and style checking were limited. It could also be argued that these systems only focus on the final product.

Recently, many automatic writing feedback systems started using text mining techniques to provide more sophisticated feedback. Sourcer's Apprentice Intelligent Feedback system (SAIF) [38] also used text mining techniques to provide feedback for students to write essays. The system can be used to detect plagiarism, uncited quotations, lack of citations, and limited content integration problems. Once a problem is detected, SAIF can give helpful feedback to the student, such as "Reword plagiarism and model proper format," if the problem is unsourced copied material (plagiarism). SAIF uses Latent Semantic Analysis (LSA) to calculate the average distance between consecutive sentences and provide feedback on the overall coherence of the text. LSA is a technique used to measure the semantic similarity of texts [26]. For finding citations, SAIF uses a Regular Expression Pattern Matching technique to detect the explicit citations by recognizing phrases containing author name (e.g. According to, As stated in, State). Evaluations have showed that SAIF provides feedback that encourages more explicit citations in students' essays. However, SAIF only addresses some basic problems related to sourcing and integration. In addition, it requires a large number of source documents to build the LSA semantic space and a large number of predefined pattern matching rules. Based on this technology, Kakkonen and Sutinen [39] proposed a model for the assessment of free text that combines both computerized and human models of assessment.

The most relevant work to the present study is Glosser, which is an automatic writing feedback system that provides academic essay writing support for college students [22], [34]. It uses text mining algorithms to analyze various features of texts, based on which feedback is provided to student writers. Glosser (1.0) provides feedback on some aspects of the writing, such as flow, topics, and topic map visualization. The feedback is given in the form of generic trigger questions (adapted to each course) and document features that relate to each set of questions. Figure 1 displays the user interface of Topics feedback. The generic trigger questions (e.g. are the ideas used in the essay relevant to the question? Are the ideas devel-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

oped correctly?) are provided at the top of the page to help the writer focus their evaluation of the essay. The extracted document feature called 'gloss' is shown below the questions. In this case, the gloss refers to Topics on the left-hand side of the table shown in the figure, such as Global Language, Countries and Learn English, and important sentences are listed on the right-hand side of the table. As Glosser highlights the 'gloss' in the essay, students are learning during the process of reflection. However, Glosser used a set of generic questions to trigger reflection. Our previous approach used natural language processing techniques to generate content specific trigger questions based on citations provided in the student academic essays for helping reflection [23], [25].

Our current research can be considered as an extension of the existing Glosser system. Like Glosser, we analyzed the student essays and automatically generated ICF addressing some aspects of writing. However, our approach focused on more aspects of the writing, such as *Grammar, Spelling, Sentence Diversity, Supporting Ideas* and *Organization*, since these aspects were frequently addressed in the teachers' feedback in the context of Chinese ESL learners' writing based on our empirical study findings. From the technology point of view, we adapted the supervised machine learning approach to classify the quality of essays regarding to each writing aspect. Specifically, the textual feature model was built by using the latest natural language processing techniques.

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 [40], LIWC [41] and Gramulator [42], are useful in this respect, and have certainly contributed to ESL knowledge [43]. Coh-Metrix is a powerful computational tool that provides over 100 indices of cohesion, syntactical complexity, connectives and other descriptive information about content [40]. Coh-Metrix has been extensively used to analyze the overall quality of writing [43] and important aspects of writing quality, such as coherence [44]. In this study, we used Coh-Metrix to extract features to build the feedback classification model. The major contributions of this paper are the following:

1. Proposed a novel approach to automatically generate essay feedback. Compared with the previous automated essay feedback system [22], our system applies the supervised machine learning approach to classify the quality of essays regarding to each writing aspect and focuses on more aspects of the writing, which are frequently commented by Chinese English teachers.

2. Conducted the quasi-experimental evaluation of automatic feedback technologies for writing, in the context of an English as a second language course in China. There are very few studies in which control and experimental groups are on their use of novel technologies [18]. Particularly, our study examined the impact of the ICF on the content related aspects of the essay.

3 Data Collection and Feedback Annotation

The diagnostic assessment of writing is an important aspect of second language abilities test, which focuses more on specific features rather than global abilities [45]. Rating scales represent the construct on which the performance evaluation is based. North [46] reviewed several rating scales including four skill models and model of communicative language ability. Based on these findings and writing theories, Knoch [47] proposed a more comprehensive and practical model for assessing second language writing tests. He defined eight feature categories, including accuracy, fluency, complexity, mechanics, cohesion, coherence, reader/writer interaction and content. The accuracy category contains grammar feature, while the mechanics category contains spelling feature. The complexity contains sentence diversity feature. The content contains supporting ideas and organization features. In this study, we adapted Knoch model to annotate teachers' comments because it is more relevant to our case and covers a wider range of features than other models.

We investigated 1290 feedback comments written by 10 college English teachers based on 327 essays written by those second-year college students enrolled at the comprehensive English class from 2013 to 2015. All those students were English majors from College of International Studies at Southwest University. The writing task was timed and considered as an assignment in the English class. Students were required to finish it within 40 minutes. The writing task was to write a persuasive essay following the standard of college English essay writing set by Ministry of Education in China. The essay question is about "*Children Should Get Paid By Doing Housework*".

The first task was to group the comments into frequent essay feature categories based on the Knoch model. Two experienced English teachers volunteered to annotate the teachers' comments on these essays. They had at least five years of teaching composition course for English majors. Seven frequent essay feature categories were found, including *Grammar (N=144), Spelling (N=36), Sentence Diversity (N=120), Conclusion (N=132), Supporting Ideas (N=294), Organization (N=267), and Coherence (N=120)*. These findings are supported with existing research evidence that ESL teachers pay a great deal of attention to student writ ten errors [6] (including grammatical and spelling errors), Content (including supporting ideas, coherence and sentence diversity) [48], Organization [48]. In fact, current commercial AES, such as Criterion [18], included some of these features, such as *Grammar, Spelling and Supporting Ideas, Conclusion*.

The second task was to ask the two teacher annotators to score each essay feature based on the rubric defined in the Appendix on a scale of 3. 1 means negative feedback on an essay feature, 2 means neutral while 3 means positive feedback on an essay feature. This analytic rubric is an adapted version of Knoch [47], focusing on seven essay features mentioned above. We constructed it after informal interviews with those 10 English teachers about the criteria they used for evaluating their students written work. As we mentioned before, those English teachers helped us to collect the dataset, including the student es-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

says and teacher feedbacks. The annotators were first trained to use the rubric with 10 essays. A Pearson correlation for each rubric evaluation was conducted between the raters' responses. If the correlations between the raters did not exceed r = .50 on all items, the ratings were reexamined until scores reached the r = .50 threshold. After the raters had reached an inter-rater reliability of at least r = .50, each rater then independently evaluated the 327 essays that comprise the corpus used in this study.

The Correlations between the raters are located in Table 1. The raters had the highest correlations for judgments of *Grammar* (r=. 824), *Spelling* (r=. 704), *Conclusion* (r=. 747) and *Supporting Ideas* (r=. 632) and the lowest correlations for *Sentence Diversity* (r=. 454), *Coherence*(r=. 516) and organization (r=. 504). These results were consistent with Crossley and McNamara 's findings [49], where the deficiency in Grammar, Spelling, Conclusion and Supporting Ideas features was easier to identify than Coherence and Organization.

In addition, it has been found that a relatively larger number of negative feedbacks were given to the content related features, such as *Supporting Ideas* (49.1%), *Coherence* (23.6%), *Sentence Diversity* (33.0%), *Conclusion* (24.5%), and *Organization* (20.8%) features. These results reveal the common writing problems that most of English-major students have were content-related. For example, students have difficulties in providing persuasive evidences to support essay arguments, and using more complex sentences in the writing.

4 Automatic Feedback Generation

This section describes our approach to automatic essay feedback generation using the supervised machine learning approach based on the annotated dataset described in section 3. Subsection 4.1 presents the feature model was built mainly based on Coh-Metrix [40], [51], while subsection 4.2 describes the evaluation of three common classifiers, Naïve Bayes, Support Vector Machine and Decision tree. Subsection 4.3 illustrates the automatic feedback generation process.

4.1 Textual Feature Extraction

We used Coh-Metrix 3.0, retrieving 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[52].

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 [40]. 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 proposed and extracted 8 new features that were 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 used 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/), was employed to scan each word. For checking grammatical errors, the Link Grammar Parser [53] was 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.

TABLE 1: RATERS' INTER-AGREEMENT AND THE DISTRIBUTION OF FEEDBACK VALUE ON EACH CATEGORY

Essay Feature	r	Negative (%)	Neutral (%)	Positive (%)
Grammar	.824	17.0	63.2	19.8
Spelling	.704	19.8	59.4	20.8
Sentence Diversity	.454	33.0	57.6	9.4
Conclusion	.747	24.5	54.7	20.8
Supporting Ideas	.632	49.1	37.7	13.2
Coherence	.516	23.6	58.5	17.9
Organization	.504	20.8	36.7	42.5

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

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

Grammatical errors: the number of sentences with grammatical errors. We used the Link Grammar Parser [53] to check the grammar of a sentence, which was 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.

4.2 Automatic Negative Feedback Classification Result

Since pointing out the weaknesses of the essays is important in the written feedback, our approach emphasizes on automatic negative feedback classification in one of seven essay features (Grammar, Spelling, Sentence Diversity, Conclusion, Supporting Ideas, Coherence and Organization). Teachers annotated one feature of a student essay as positive, neutral or negative feedback, which is mentioned in Section 3. In order to improve the performance of the system classification, we modified each instance, one feature of a student essay, labeled as either negative or non-negative feedback. In other words, previous labeled positive feedback or neutral feedback changed to non-negative feedback. Precision, recall and F-score were used as the evaluation matrix. Precision is the fraction of system-classified instances that are correctly classified as negative feedback, while recall is the fraction of the negative feedback that are detected by the system. The F-score is calculated as follows:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

The ten-fold cross-validation was used to evaluate the classification performance of three classifiers. In addition, we defined a baseline for the performance evaluation of classifiers. The baseline classifier assumes that all the instances are classified as negative feedback, so the precision is the percentage of negative instances to the size of the dataset, while the recall is 1.

Table 2 shows that the J48 outperforms the rest of classifiers across seven essay features. Grammar and Spelling features were easier to be classified (F-score is above 0.820 across three classifiers) since they did not require deep text understanding. Regarding to the rest of features, the sentence diversity and supporting idea features were easier to be detected by using J48, where those f-score was above 0.500. Conclusion, Organization and Coherence category were generally difficult to be classified, where f-score for conclusion was 0.485, f-score for coherence was 0.439 and f-score for organization was 0.405. Some findings were consisted with Crossley and McNamara's study [54]. The cohesive features extracted by the Coh-Metrix were not sufficient predictors to the quality of the coherence rated by the human teachers.

4.3 Automatic Feedback Generation System

Similar to Glosser [22], SAM is a web-based automatic feedback generation system including indirect corrective feedback generation, assignment management and online text editing. The system contains seven negative feedback classifiers for seven different essay features implemented in J48 algorithm described in section 4.2 and one text editor with the revision control ability.

The automatic feedback generation process is described as follows. A student firstly writes an essay in the online text editor. After the student finishes the writing, the feedback classifier then classifies the essay regarding to particular essay feature. If the classification result is negative, the system generates indirect corrective feedback questions. Two English teachers those who annotated the teachers' comments predefined feedback questions for each essay feature based on their teaching experience (See Table 3). This feedback suggests students to "double-check" their essay with several trigger questions provided. It tends to help students reflect on what they have written. Currently, the system selects all the automatically classified negative feedbacks to the students (See Figure 2).

We used the Etherpad (www.etherpad.org) to implement the online text editor embedded in the system. Etherpad is a highly customizable open source online editor. Like Google Docs, the text editor can automatically

TABLE 2: THE CLASSIFICATION RESULT OF THE FOUR CLASSIFIERS IN NEGATIVE FEEDBACK.

	J48			Naïve Bayes			SMO			Baseline		
Essay Feature	P	R	F	P	R	F	P	R	F	P	R	F
Grammar	0.850	0.804	0.820	0.650	0.604	0.620	0.704	0.726	0.711	0.170	1	0.291
Spelling	0.984	0.984	0.984	0.898	0.841	0.869	0.841	0.921	0.879	0.594	1	0.745
Sentence Diversity	0.532	0.547	0.524	0.360	0.257	0.300	0.483	0.491	0.486	0.330	1	0.496
Conclusion	0.490	0.491	0.485	0.243	0.346	0.286	0.424	0.425	0.421	0.245	1	0.394
Supporting Ideas	0.522	0.462	0.490	0.474	0.519	0.495	0.458	0.423	0.440	0.491	1	0.659
Coherence	0.423	0.462	0.439	0.136	0.158	0.146	0.381	0.358	0.369	0.179	1	0.304
Organization	0.404	0.406	0.405	0.130	0.136	0.133	0.471	0.364	0.410	0.208	1	0.344
Note: P refers to the precision, R recall while F is F-score												

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

stores the revision history of an edited document, which allows us to easily access the number of revisions and the number of edited words in a revision. In addition, we installed the reference plugin in the text editor, which provides students an easy access to the concept of a particular essay feature, such as organization and coherence, so that the student can understand the concept well.

In addition, the system allows teachers to give direct feedback on a specific text segment of the essay in the Etherpad text editor using the comment plugin. Like working on the comment function in the Microsoft Word, the student can accept changes based on the comments.

Figure 2 shows an example of the system-generated feedback about the coherence issue. The following predefined feedback questions are used for the self-correction:

1) *Does your topic sentence connect to the previous paragraph?*
2) *Do you understand how each paragraph and sentence follows from the previous one?*
3) *The communication of your ideas needs more strong cohesive cues and devices, such as transitions and connective phrases that link ideas. Do you use them correctly?*

The student checks her essay based on the feedback

TABLE 3: EXAMPLES OF ICF ON SEVEN ASPECTS OF THE WRITING

Essay Feature	Examples
Grammar	*Do you use the grammar check function in Microsoft Word? Are you sure each sentence has a subject and a verb? Do subjects and verbs agree with one another in number in your sentences?*
Spelling	*Do you adopt the spelling that your dictionary gives first in any entry with variant spellings? Do you use the spell check function in Microsoft Word? Make sure you double-check for capitalization errors and misspelled words in the essay*
Sentence Diversity	*Do you use a variety of short and long sentences? Try to use condensed and split sentences in turn. Avoid excessively long sentences (25 words or more). If you find yourself pausing for breath when reading aloud a sentence, it may already be too long.*
Conclusion	*Does your conclusion bring all of your ideas together and make a concise summary of main points? Does your conclusion restate the thesis in different words or phrases? Don't introduce new concepts that were not addressed previously.*
Supporting Ideas	*Do your body paragraphs rely heavily on unsupported hypothetical and opinion-based claims rather than fact-based claims? Do not make too many overgeneralizations. You can always add more elaboration, factual information, persuasive evidence, and specific examples to get your point across.*
Coherence	*Does your topic sentence connect to the previous paragraph? Do you understand how each paragraph and sentence follows from the previous one? The communication of your ideas needs more strong cohesive cues and devices, such as transitions and connective phrases that link ideas. Do you use them correctly?*
Organization	*Planning before an essay will get you a clearer structure and a higher grade. Mare sure your introduction sets out a robust position of your own and previews some of the ideas to be developed in the body. Is your paragraph organization linear and explicit?*

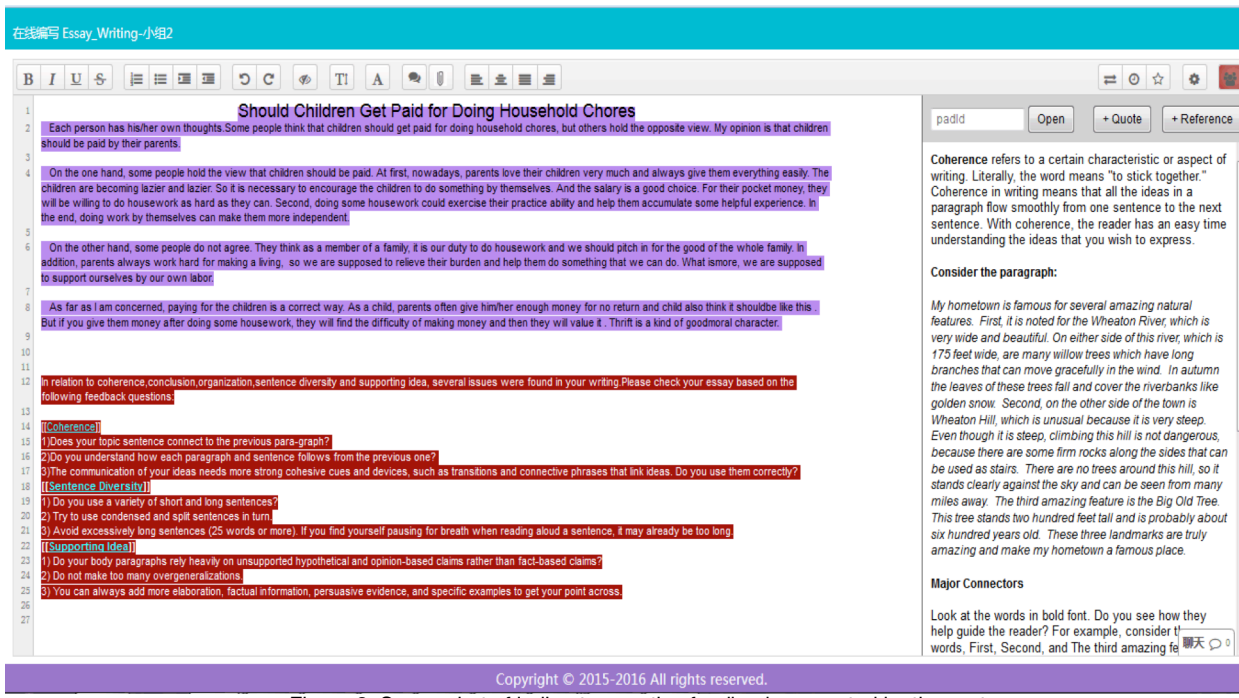


Figure 2: Screenshot of indirect corrective feedback generated by the system

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

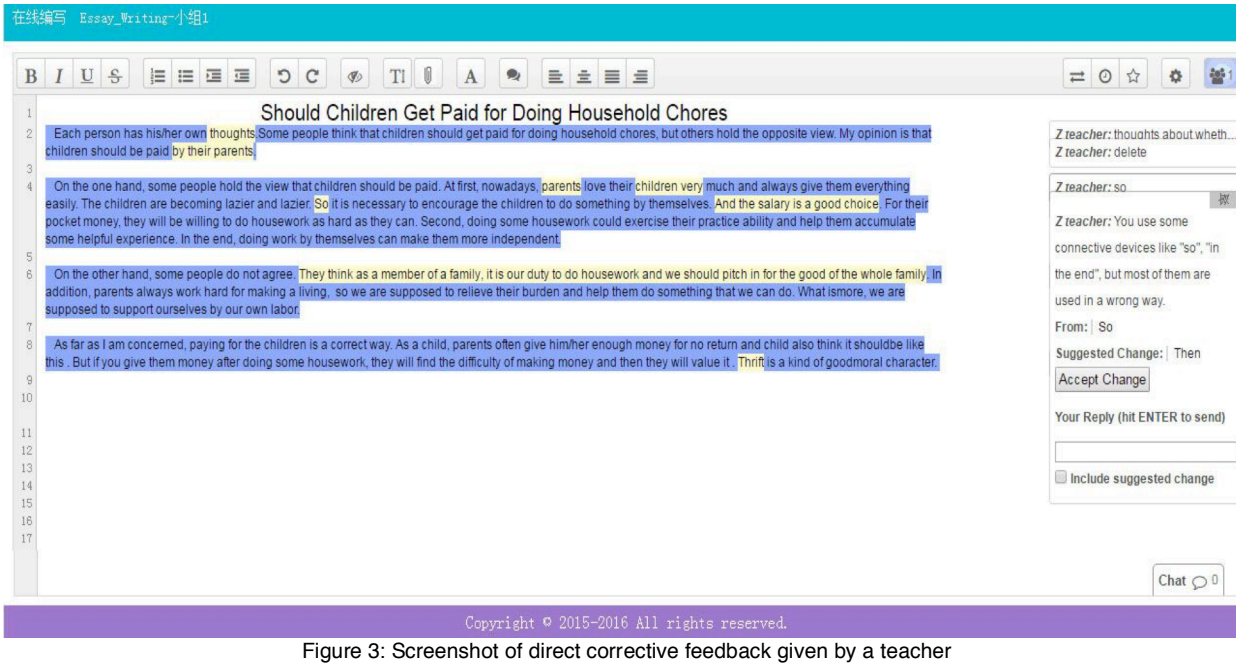


Figure 3: Screenshot of direct corrective feedback given by a teacher

provided and try to improve the coherence aspects of the essay. If the student does not understand the importance of the coherence well, then he or she gets more information about the coherence by clicking the Coherence link (See Figure 2). This approach is similar to the Glosser [22], where the feedback or trigger questions do not contain specific answers. They are intended for students to perform self-reflection with the trigger questions provided.

Moreover, the human teacher can give corrective comments on the essay through the system. Figure 3 shows an example of a teacher's the direct corrective comments on the same student essay. In the example, regarding to the essay coherence, several connectives were misused. In this case, the teacher highlights the 'so' and comments on it for correction shown on the right hand side of the screen.

5 User Study

The aim of this study is to investigate the impact of the system-generated indirect corrective feedback and human-teacher given direct corrective feedback in the revision. Particularly, we have analyzed the improvement of essay quality regarding to seven essay features, the student perception of the feedback and computer recorded revision history in two different feedback classes.

5.1 participants and procedure

Like the first study, participants were second year English major students at Southwest University in China, who enrolled in two English comprehensive classes in 2016. Two English teachers were volunteered to participate this study. Those teachers also participated in teachers' comment annotation experiment. In the traditional DCF class (n=54, 50 females, 4 males), each student received only DCF made by those teachers regarding to the seven features. In the ICF class (n=56, 48 females, 8 males), students received only ICF that were automatically generated by the system. In addition, we examined the participants' prior knowledge about writing by analyzing their scores in the first year writing class. An independent sample t-test was conducted with 95% confidence interval to compare the scores obtained in DCF and ICF class. There was no significant difference in the scores for DCF (M=71.91, SD=11.05) class and ICF (M=73.50, SD=11.60) conditions (t(108)=.737, p=.463).

The writing task includes two main stages. In the first stage, students were asked to write a pervasive essay in 40 minutes during the class. The essay question is the same as the first study described in section 3. Then, those two teachers gave comments for the DCF class (Each teacher gave feedback for 27 essays), while the system automatically generated ICF for the ICF class. Students had one week to revise their drafts based on the feedback.

In the next week, students submitted their revised version in the system, and then those two teachers scored each essay feature of 110 essays, based on their previous annotation experience described in section 3. Finally, all the student participants responded to a survey and some of them accepted an informal interview.

5.2 Results and Discussion

5.2.1 Feedback Generation

Two types of feedback generated by the system and human teachers covered both form (e.g. grammar) and content Feedback Generation. As we expected, teachers generated more feedback than the system did since teachers could give specific feedback in multiple specific places in a student essay regarding to a type of problem, such as Grammar (See Table 4). In addition, similar to the previous findings described in section 3, we observed that more feedbacks were related to the content related issues in *Sentence Diversity, Organization, Supporting Ideas and Coherence features*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

TABLE 4: THE NUMBER OF FEEDBACK GENERATED FROM TWO CLASSES

	Spelling	Grammar	Coherence	Conclusion	Supporting Ideas	Sentence Diversity	Organization
DCF	11	15	21	16	19	42	23
ICF	6	10	18	13	16	38	14

TABLE 5: THE IMPACT OF THE FEEDBACK ON DIFFERENT ESSAY FEATURES IN THE ICF AND DCF CLASS. THE VALUE OF MEAN DIFFERENCE CALCULATED BY THE SECOND DRAFT SCORE MINUS THE FIRST DRAFT SCORE.

Essay Feature	Class	Mean Difference	Standard Deviation	t	N	Sig. (Two Tails)
Spelling	ICF	0.11	0.066	1.627	55	.110
	DCF	0.44	0.094	4.724	53	<.001
Grammar	ICF	0.13	0.075	1.729	55	.089
	DCF	0.29	0.061	4.690	53	<.001
Coherence	ICF	0.54	0.105	5.104	55	<.001
	DCF	0.63	0.085	7.423	53	<.001
Conclusion	ICF	0.57	0.076	7.535	55	<.001
	DCF	0.30	0.098	3.036	53	.004
Supporting Ideas	ICF	0.46	0.071	5.201	55	<.001
	DCF	0.54	0.094	5.698	53	<.001
Sentence Diversity	ICF	0.07	0.088	0.814	55	.419
	DCF	0.06	0.077	0.724	53	.472
Organization	ICF	0.39	0.110	3.567	55	<.001
	DCF	0.87	0.095	9.112	53	<.001

5.2.2 The Improvement of Essay Quality in Seven Aspects

Students first wrote the first draft, then revised it based on the feedback received, and finally produced the second draft. The human teacher scored each aspect of the second draft. Table 5 shows that the quality of the first draft in terms of seven features was improved across both classes. Paired t-tests with 95% confidence interval revealed that the improvement in spelling and grammar were significant in the DCF class, $t(53)=4.724$, $p<0.001$ in spelling, and $t(53)=4.690$, $p<0.001$ in grammar. But, no significant improvement on these two aspects was observed in the ICF class, $t(55)=1.625$, $p>0.05$ in spelling, and $t(55)=1.729$, $p>0.05$ in grammar. Moreover, statistical tests were conducted to examine the differences between ICF and DCF class. Independent t-test showed significant differences, $t(106)=18.300$, $p<0.001$ in grammar and, $t(106)=12.236$, $p<0.001$ in spelling between DCF and ICF class. The study results indicate that students improved their linguistic accuracy better when DCF strategy was applied rather than ICF. This finding is along with Ferris [14] who reported that direct error correction led to better grammatical accuracy than indirect error feedback.

Regarding to the content related errors, we observed that student writers like to use more assertions (e.g. *we must/we should*) and rather than factual evidence to support the essay argument. In addition, the illogical organization is another big issue. The main reason for these problems is that Chinese ESL writers have deductive thinking [55], and the reader's responsibility opinion (It is the reader's responsibility to connect the ideas, understand the context and draw some conclusions) in the writing[56]. The statistical t-test results show that the second draft significantly outperformed the first draft in terms of *organization, structure, coherence, supporting ideas and conclusion* across both classes. ICF points out the way to improve content related aspects of the writing. In fact, the majority of the students from the ICF class agreed that the feedbacks on these aspects were helpful. Instead of revising only to remove errors, writers try to reconsider and refine the whole text. The findings are consistent with those in Beuningen and Kuiken's [7] study, which showed that both DCF and ICF were an effective means to improve the overall quality of the student writing from an initial writing task to its revision. This result also highlighted the importance of giving feedback on both content (structure, organization, supporting ideas, conclusion) and form (grammar and spelling), which resulted in a better performance [4], [8].

Regarding to the sentence diversity, no significant improvement were found in the both class, $t(55)=0.814$, $p>0.05$ in ICF class, and $t(53)=0.724$, $p>0.05$ in DCF class. Most of the students acknowledged that it was difficult to use the feedback for constructing complex sentences due to their linguistic limitation, such as a lack of knowledge about vocabulary and complex grammar, even with the teacher feedback shown as follows.

Don't use too many short sentences. Try to use more complex sentences like compound sentences and subordinate clauses.

Although this feedback is related to a specific content, the student still was not sure how to revise it due to her linguistic limitation.

In sum, these results indicated those two types of feedback were useful to improve the quality of the writing in terms of the essay organization, coherence, supporting ideas and conclusion. Students who received DCF had significantly better improvement than those who re-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2016.2612659, IEEE Transactions on Learning Technologies

****

ceived ICF in relation to the Grammar and Spelling features.

5.2.3 Student Perception of the Feedbacks

After the study, students in both classes were asked to separately identify their level of agreement on that the feedback is helpful on each of seven essay features. The five-point likert scale rate was used to rate the level of agreement. In addition to the questionnaire, informal user feedback was collected from students in both classes.

In the DCF class, more than 73% participants agreed or strongly agreed that the feedback was helpful on *Grammar, Spelling, Conclusion, Supporting Ideas*, *Coherence* and *organization; 62% participants* agreed or strongly agreed that the feedbacks were helpful on *sentence diversity*.

Most of the students interviewed in the DCF class said that the DCF was helpful in the sense that it pointed out their errors in writing, so that student knew what errors they had made. However, when asked if the feedback was useful to improve their writing skills, doubts were expressed: *Sometimes, I don't understand teacher comments, so just quickly accept the changes suggested by the teacher. I might make the same mistakes in the next composition.*

In the ICF class, more than 65% participants agreed or strongly agreed that the feedback was helpful on *Conclusion, Supporting Ideas*, *Coherence* and *Organization.* Forty-eight percent of participants agreed or strongly agreed that the feedback was helpful on *grammar, spelling and sentence diversity.* Some students interviewed said that they liked the ways to explore solutions by themself. However, other students hoped that the feedback could be more specific. Moreover, students complained that the feedbacks were too numerous if four to seven aspects of their essays were required to double check.

Overall, most of students in the DCF class agreed with the usefulness of the feedback on seven aspects of the essay, while those students in the ICF class found that the feedback was particularly useful in supporting ideas, organization, coherence and conclusion. Less positive feedback on *Grammar, Spelling and Sentence Diversity* in the ICF class were received since they were too general. This result is in line with Miceli's study [15], where students felt that ICF was useful in encouraging them to reflect on aspects of their writing and to develop improvements in the content, while DCF to be more helpful when revising syntax and vocabulary.

5.2.4 Observations of The Computer Generated Revision History

The system keeps a revision history of each document. Based on this information, we can analyze how much text changes were made in terms of number of revisions and text edit, when student received the feedback. Text edit refers to create, modified or remove text action. Table 6 shows that students receiving ICF made more text changes than those students getting DCF regarding to the number of edited words (52 in ICF class, 36 in DCF class) and the number of revisions (108 in ICF class, 79 in DCF class). Independent sample t-test results revealed that students in ICF class made significantly more changes than those students in DCF class, $t\ (108)=5.798$, $p<0.001$ in the number of edited words; $t\ (108)=10.600$, $p.<0.001$ in the number of revisions. These results indicated that students in the ICF class spent more time on revising the essay than those students in the DCF class did. The main reason for this is that the ICF encourages students to be more active in their use of feedback [12].

6 Conclusion and Future Work

The most significant findings of this study indicated that the system-generated ICF was useful to improve the quality of writing, particularly in *the structure, organization, supporting ideas, coherence, and conclusion* aspects of the essay in the short term. In addition, the ICF encouraged students to spend more time on performing self-correction than the DCF provided by teachers. These finds were consisted with previous study results, which reported positive impacts of ICF on the ability of students to edit their own composition and to improve levels of accuracy in writing [10], and the effectiveness of the combined both form and content focused feedback in improving the writing development [4], [8]. Although findings on the effectiveness of different feedback types have been conflicting, largely due to the widely varying student populations, types of writing and feedback practices examined [9], this study result implied that the system could be useful for Chinese English-major students with advanced metalinguistic knowledge, particularly in content development. We also observed that some incorrect ICF were generated. But, the incorrect feedback might be particularly beneficial if they promote noticing. The system required the students to consider the weakness of the essay and evaluate them in context to determine whether they are correct.

This study has some limitations. The system-generated feedback could be too general for correcting errors in some aspects of essay, such as Grammar and Spelling. Moreover, some essays triggered feedback on many features, which generated a daunting number of suggestions. This quantity of feedback seemed to undermine our scaffolding goal by targeting too many essay elements at once. Finally, this study did not examine the impact of the tool in non-English major students' writing. Non-English major students do not have good metalinguistic knowledge, so the impact of the feedback might be different.

In the future work, we will focus on how to combine both DCF and ICF together to support different aspects of

TABLE 6: TEXT CHANGES MADE IN THE REVISION

Class	N	Average Number of Edited Words	Std. Dev. Of Edited Words	Average Number Of Revisions	Std. Dev. Of Revisions
ICF	56	52	20.19	108	21.69
DCF	54	36	12.56	79	14.54

****

students' writing. Particularly, we will examine different English parsers, such as Stanford Parser, which it might give more specific feedback on spelling and grammar. We also will study how much amount of feedback is sufficient. Zacharias [57] found that if students had too much feedback they would feel discouraged and were less likely to be motivated to use it for revision. This could be limited based on the severity and/or number of negative features detected in the text. Lastly, the future work will investigate how to effectively use this tool for teaching and encouraging independent editing skills. Ferris [58] and Mu [59] have provided descriptions of procedures for helping students learn to self-edit, and computer-assisted feedback could be included in such procedures.

ACKNOWLEDGEMENT

This work is 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, CQU903005203326, and National Natural Science Foundation of China (61502397) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES:

[1] N. Bureau of Statistis of China, *China Statistical YearBook*. China Statistics Press, 2013.

[2] L. Brannon and C. H. Knoblauch, "On students' rights to their own texts: A model of teacher response," *Coll. Compos. Commun.*, vol. 33, no. 2, pp. 157–166, 1982.

[3] I. Leki, "The preferences of ESL students for error correction in college-level writing classes," *Foreign Lang. Ann.*, vol. 24, pp. 203–218, 1991.

[4] A. K. Fathman and E. Whalley, "Teacher Response to Student Writing- Focus on Form versus Content," in *Second Language Writing: Research Insights for the Classroom*, 1990, pp. 178–190.

[5] D. Ferris, *Response to student writing: Implications for second language students*. Mahwah, NJ : Lawrence Erlbaum Associates, 2003.

[6] I. Lee, "Error correction in L2 secondary writing classrooms: The case of Hong Kong," *J. Second Lang. Writ.*, vol. 13, pp. 285–312, 2004.

[7] C. G. van Beuningen, N. H. de Jong, and F. Kuiken, "The Effect of Direct and Indirect Corrective Feedback on L2 Learners' Written Accuracy," *ITL Int. J. Appl. Linguist.*, vol. 156, pp. 279–296, 2008.

[8] T. Ashwell, "Patterns of Teacher Response to Student Writing in a Multiple-Draft Composition Classroom: Is Content Feedback Followed by Form Feedback the Best Method?," *J. Second Lang. Writ.*, vol. 9, no. 3, pp. 227–257, 2000.

[9] D. R. Ferris, "Does error feedback help student writers? New evidence on the short- and long-term effects of written correction," in *Feedback on Second Language Writing: Contexts and Issues* , K. Hyland and F. Hyland, Eds. New York, 2006, pp. 1–104.

[10] D. Ferris and B. Roberts, "Error feedback in L2 writing classes How explicit does it need to be?," *J. Second Lang. Writ.*, vol. 10, no. 3, pp. 161–184, 2001.

[11] J. Chandler, "The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing," *J. Second Lang. Writ.*, vol. 12, no. 3, pp. 267–296, 2003.

[12] F. Hyland, "Providing effective support : investigating feedback to distance language learners," *Openg Learn.*, vol. 16, no. 3, pp. 233–237, 2001.

[13] D. Little, "Learner autonomy and second/foreign language learning," *Subject Centre for Languages, Linguistics and Area Studies*, 2003. [Online]. Available: http//www.llas.ac.uk/resources/gpg/1409.

[14] D. Ferris, *Treatment of error in second language student writing*. Ann Arbor: The University of Michigan Press, 2002.

[15] T. Miceli, "Foreign Language Students' Perceptions of Reflective Approach to Text Correction," *Finders Univ. Lang. Gr. Online Rev.*, vol. 3, no. 1, pp. 25–36, 2006.

[16] Y. Attali and J. Burstein, "Automated Essay Scoring With e-rater V.2.," *J. Technol. Learn. Assess.*, vol. 4, no. 3, 2006.

[17] S. Elliot, "Intellimetric: From here to validity," in *Automated essay scoring: A cross-disciplinary perspective*, Mahwah, NJ: Lawrence Erlbaum Associates, 2003, pp. 71–86.

[18] J. Burstein, M. Chodorow, and C. Leacock, "Automated essay evaluation: The Criterion online writing service," *AI Mag.*, vol. 25, p. 27, 2004.

[19] Richard Haswell, "The complexities of responding to student writing; or, looking for shortcuts via the road of excess," *Across Discip.*, vol. 3, 2006.

[20] S. C. Weigle, "English as a second language writing," in *Handbook of automated essay evaluation*, 2013, pp. 36–53.

[21] K. Yancey, A. Lunsford, J. McDonald, C. Moran, M. Neal, C. Pryor, D. Roen, and C. Selfe, "CCCC position statement on teaching, learning, and assessing writing in digital environments," *Coll. Compos. Commun.*, vol. 55, no. 4, pp. 785–790, 2004.

[22] J. Villalon, P. Kearney, R. A. Calvo, and P. Reimann, "Glosser: Enhanced Feedback for Student Writing Tasks," 2008, pp. 454–458.

[23] M. Liu, R. A. Calvo, and V. Rus, "Automatic Question Generation for Literature Review Writing Support," 2010.

[24] D. S. McNamara, S. a Crossley, and R. Roscoe, "Natural language processing in an intelligent writing strategy tutoring system.," *Behav. Res. Methods*, vol. 45, no. 2, pp. 499–515, Jun. 2013.

[25] M. Liu, R. Calvo, and V. Rus, "Automatic

****

Generation and Ranking of Questions for Critical Review," *Educ. Technol. Soc.*, vol. 17, no. 2, pp. 333–346, 2014.

[26] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, 2007.

[27] L. M. Rudner, "Reducing Errors Due to the Use of Judges.," *Pract. Assessment, Res. Eval.*, vol. 3, no. 3, 1992.

[28] S. Dikli, "Automated Essay Scoring," *Turkish Online J. Distance Educ.*, vol. 7, no. 1, pp. 49–62, 2006.

[29] M. A. Hearst, "The debate on automated essay grading," *Intell. Syst. Their Appl.*, vol. 15, no. 5, pp. 22–37, 2000.

[30] D. Wade-Stein and E. Kintsch, "Summary Street: Interactive computer support for writing," *Cogn. Instr.*, vol. 22, no. 3, pp. 333–362, 2004.

[31] M. Shermis and J. Burstein, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah NJ: Lawrence Erlbaum Associates, 2003.

[32] P. F. Ericsson and R. Haswell, "Machine Scoring of Human Essays: Truth and Consequences." Utah State University Press, 2006.

[33] G. Gibbs and C. Simpson, " Conditions under which assessment supports students' learning Learning and Teaching in Higher Education," *Learn. Teach. High. Educ.*, vol. 1, no. 1, pp. 3–31, 2004.

[34] R. A. CALVO and R. A. ELLIS, "Students' Conceptions of Tutor and Automated Feedback in Professional Writing.," *J. Eng. Educ.*, vol. 99, pp. 427–438, 2010.

[35] E. C. Thiesmeyer and J. E. Theismeyer, *Editor:A System for Checking Usage, Mechanics, Vocabulary, and Structure*. New York, New York, USA: Modern Language Association, 1990.

[36] J Anderson, *Mechanically Inclined:Building Grammar, Usage, and Style into Writer's Workshop*. Stenhouse Publishers, 2005.

[37] T. J. Beals, "Between Teachers and Computers: Does Text-Checking Software Rea lly Improve Student Writing?," *English J.*, pp. 67–72, 1998.

[38] M. A. Britt, P. Wiemer-Hastings, A. A. Larson, and C. A. Perfetti, "Using Intelligent Feedback to Improve Sourcing and Integration in Students' Essays," *Int. J. Artif. Intell. Ed.*, vol. 14, no. 3,4, pp. 359–374, 2004.

[39] T. Kakkonen and E. Sutinen, "EssayAid: towards a semi-automatic system for assessing student texts," *Int. J. Contin. Eng. Educ. Life-Long Learn.*, vol. 21, no. 2/3, pp. 119–139, 2011.

[40] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: analysis of text on cohesion and language.," *Behav. Res. methods, instruments, Comput.*, vol. 36, no. 2, pp. 193–202, May 2004.

[41] J. W. Pennebaker and M. E. Francis, *Linguistic inquiry and word count (LIWC)*. Mahwah, NJ: Erlbaum, 1999.

[42] R. M. Rufenacht, P. M. McCarthy, and T. A. Lamkin, "Fairy Tales and ESL Texts: An Analysis of Linguistic Features Using the Gramulator," in *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 2011, pp. 287–292.

[43] S. Crossley and D. McNamara, "Predicting second language writing proficiency : The role of cohesion , readability , and lexical difficulty," *J. Res. Read.*, vol. 35, pp. 115–135, 2012.

[44] S. a. Crossley and D. S. McNamara, "Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence," in *The 33rd Annual Conference of the Cognitive Science Society*, 2011.

[45] C. Alderson, *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, UK: Continuum, 2005.

[46] B. North, "Scales for rating language performance: Descriptive models, formulation styles, and presentation formats," *TOEFL Monogr.*, vol. 24, 2003.

[47] U. Knoch, "Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?," *Assess. Writ.*, vol. 16, no. 2, pp. 81–96, 2011.

[48] I. Lee, "Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning?," *Assess. Writ.*, vol. 12, no. 3, pp. 180–198, 2007.

[49] S. Crossley and D. McNamara, "Cohesion, coherence, and expert evaluations of writing proficiency," in *The 32nd Annual Conference of the Cognitive Science Society*, 2010, pp. 984–989.

[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explor.*, vol. 11, pp. 10–18, 2009.

[51] D. S. . McNamara, A. C. . c Graesser, P. M. . McCarthy, and Z. . Cai, *Automated evaluation of text and discourse with Coh-Metrix*. 2012.

[52] W. Kintsch and T. van Dijk, "Towards a model of text comprehension and production," *Psychol. Rev.*, vol. 85, pp. 363–394, 1978.

[53] J. Lafferty, D. Sleator, and D. Temperley, "Grammatical Trigrams: A Probabilistic Model of Link Grammar," in *Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language*, 1992.

[54] S. a. Crossley and D. S. McNamara, "Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing," *Int. J. Contin. Eng. Educ. Life-Long Learn.*, vol. 21, no. 2/3, p. 170, 2011.

[55] L. Jinghong, "The impact of the English Writing Theories on Chinese English Second Language Writing," *Foreign Lang. Teach.*, no. 2, pp. 41–47, 2006.

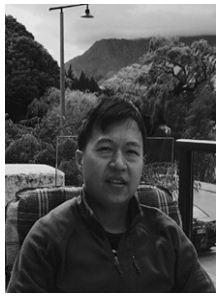
****

[56] J. Hinds, "Quasi-Inductive:Expository Writing in Japanese, Korean, Chinese and Thai," in *Coherence in Writing Research and Pedagogical Perspectives*, 1990.

[57] N. T. Zacharias, "Teacher and Student Attitudes toward Teacher Feedback," *RELC J.*, vol. 38, no. 1, pp. 38–52, 2007.

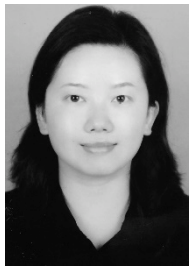
[58] D. R. Ferris, H. Liu, A. Sinha, and M. Senna, "Written corrective feedback for individual L2 writers," *J. Second Lang. Writ.*, vol. 22, no. 3, pp. 307–329, 2013.

[59] C. Mu, "A Taxonomy of ESL Writing Strategies," *Proc. Redesigning Pedagog. Res. Policy, Pract.*, pp. 1–10, 2005.



Dr. Ming Liu is Associate Professor at School of Computer and Information Science, Southwest University, China. He received the PhD in Artificial Intelligence in Education at the School of Electrical and Information Engineering, The University of Sydney, Australia in 2012. His main research interests include question generation, learning analytics and intelligent tutoring system. He participated in national and international projects funded by ARC Linkage (Australia), Young and Well CRC, Office of Teaching and Learning, Google and Chinese National Fund. He is an author of over 20 publications papers in prestigious conferences and journals, such as Intelligent Tutoring Systems, IEEE transactions on Learning Technologies, Journal of Educational Technology and Society.

Dr. Yi Li is an associate professor at Faculty of Education in the Southwest University in China. She has a master in applied statistics and a PhD in educational measurement by the Purdue University in the USA. Her research interest is application of quantitative methodology to explore, understand and influence the technology innovation in the field of education. She has participated in the national and international projects funded by National Science Foundation and National Social Science Foundation of China. Yi is author of many research publications in prestigious international conferences and journals.



Dr. Weiwei Xu is associate professor at College of International Studies, Southwest University in China. She holds a doctorate degree in English from Macquarie University, Australia. She acquired her MA degree in English from Newcastle University, UK. She has been teaching English academic writing for 8 years.

Dr. Li Liu is associate professor at Chongqing University. He is also serving as a Senior Research Fellow of School of Computing at the National University of Singapore. Li received his Ph.D. in Computer Science from the Université Paris-sud XI in 2008. He had served as an associate professor at Lanzhou University in China. His research interests are in pattern recognition, data analysis, and their applications on human behaviors. He aims to contribute in interdisciplinary research of computer science and human related disciplines. Li has published widely in conferences and journals with more than 50 peer-reviewed publications. Li has been the Principal Investigator of several funded projects from government and industry.