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 er-

- *M.Liu is with the Education Research Institute, Faculty of Education and the School of Computer and Information Science, Southwest University, Chongqing 400715, China. E-mail: mingliu@swu.edu.cn*
- *Y.Li is with the Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University and 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

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

****

ror has been made but does not provide a correction, thereby leaving the student to diagnose and correct it. 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

****

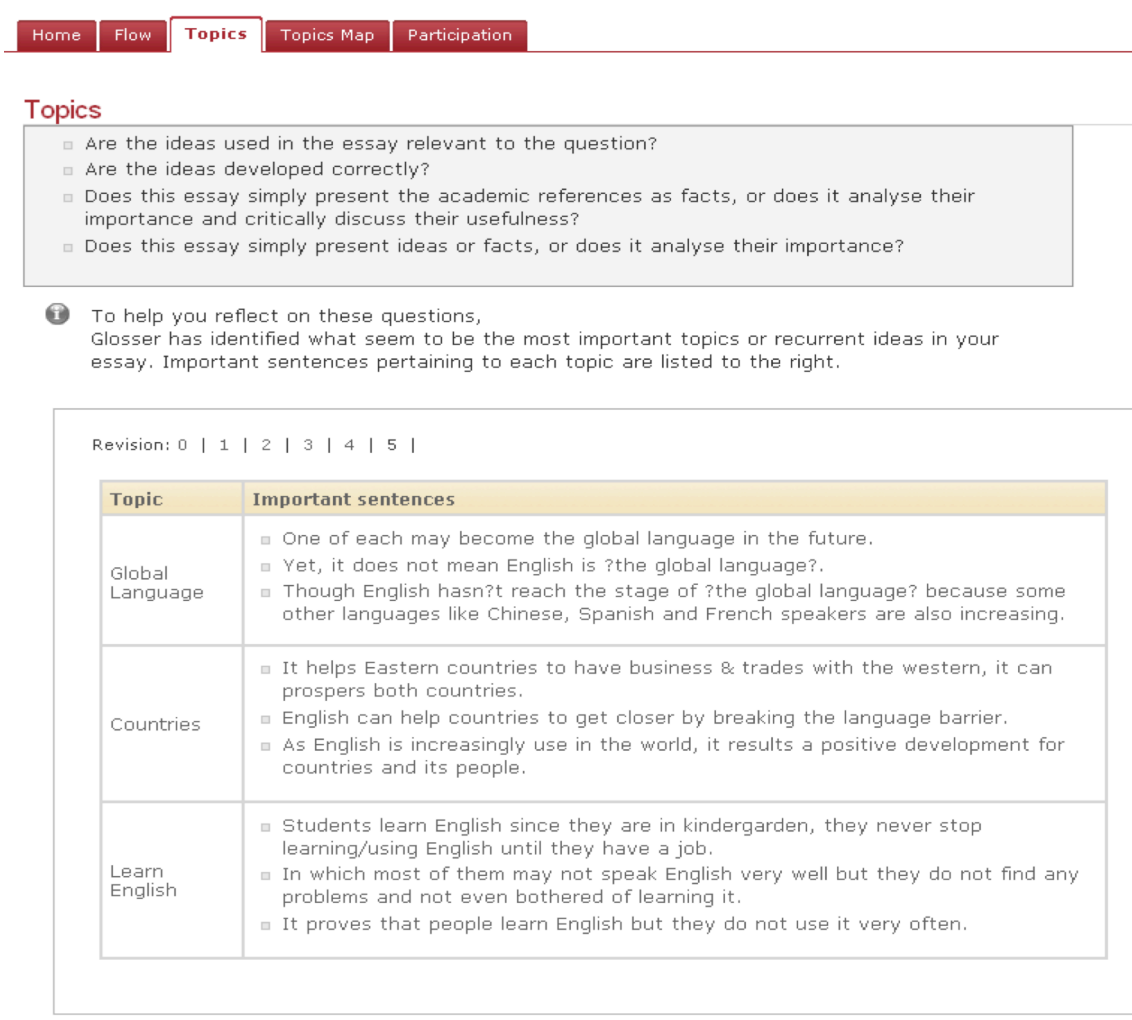

Figure 1: The user interface of the topics tool in Glosser. The figure is reproduced from Calvo et al. [22]

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-