Towards an automatic classification system for supporting the development of critical reflective skills in L2 learning

Gary Cheng
Department of Mathematics and Information Technology, The Education University of Hong Kong

This study aimed to develop an automatic classification system, namely ACTIVE, for generating immediate and individualised feedback on students' reflective entries about their second language (L2) learning experiences. It also aimed to explore students' attitudes towards using the system to support the development of their reflective skills in L2 learning. A total of 466 undergraduate students took part in the study. One hundred and twenty-seven participants were involved in the development phase, where their reflective entries were manually annotated according to a classification framework for critical reflection on L2 learning, and the annotated entries were then used to develop the ACTIVE system. The remaining participants were asked to generate automated feedback reports on their reflective entries for improvement by using the system. To solicit their views towards the system, the participants were administered an online questionnaire and some of them were also invited to attend a semi-structured interview. The overall results indicate that the classification accuracy of the system is comparable to that of human annotators. They also suggest that both teacher and machine feedback types have strengths and limitations, highlighting the need to further explore the use of multi-channel, multi-layer feedback in improving students' reflective skills in L2 learning.

Introduction

Reflective writing has been widely used as a pedagogical strategy to help English as a second language (ESL) students develop their critical thinking skills in second language (L2) learning (Hyland, 2003; Scott, 2005). In the process of reflective writing, students are expected to think about and make sense of the connections between what they have learnt and what they are learning, between what they are doing to learn and why they choose to do it and between what they would like to achieve and what they actually get. This practice provides an opportunity for students to critically reflect on their learning experiences and identify their existing strengths and areas for future development. Research suggests that reflective writing can potentially help ESL students to reinforce their understanding of L2 acquisition, to raise their self-awareness with respect to their L2 learning progress and to apply strategies to improve their L2 proficiency (Baturay & Daloğlu, 2010; Chau & Cheng, 2010).

Despite the potential of reflective writing for student learning, students' reflective entries are commonly written at a descriptive rather than an analytical level. For example, Lai and Calandra (2007) studied the problems of preservice teachers in their reflection. The results of their study showed that most reflective entries were merely descriptive of an event or experience, suggesting that the preservice teachers had a very limited understanding and practice of critical reflection. Power (2012) found a similar issue with undergraduate students from a university language program. He suggested that a clear framework for critical reflection should be developed to promote students' understanding of reflection and encourage their participation in reflection through scaffolding. In this light, previous studies proposed a number of multi-level framework to evaluate students' reflective writing and then classify their reflective ability in various professions, such as accounting (Bisman, 2011), healthcare (Smith, 2011), teaching (Granberg, 2010) and L2 learning (Chau & Cheng, 2012).

There are two main challenges facing the use of a classification framework to evaluate students' reflection: a great deal of time and effort put into understanding and applying the assessment criteria by teachers and a need for consistency in assessing a large number of reflective entries across multiple teachers (Wade, Abrami, & Sclater, 2005; Wetzel & Strudler, 2006). Given these challenges, it is worth considering whether existing technology can be adopted to offer immediate and adaptive feedback on students' reflection in a helpful and consistent way. Therefore, the current study was undertaken to meet two objectives: (1) to develop an

automatic classification system for generating immediate and individualised feedback on students' reflective entries about their L2 learning experiences; and (2) to explore students' attitudes towards using the system to support the development of their reflective skills in L2 learning. The findings from this study can provide insights into the design of technology-assisted reflective practice and inform the direction of future research in this area.

Related research

Reflective learning

The definition of reflective learning varies considerably across the literature. Dewey (1933) defined reflection as "active, persistent and careful consideration of any belief or supposed form of knowledge in light of the grounds that support it and the further consequences to which it leads" (p. 9). He argued that people learn more from the process of reflecting upon their former experiences and adapting to the environment than from the experience itself. This concept provides a cognitive basis for reflective learning, but with little consideration of learners' emotional aspects. From the perspective of learners, Boud, Keogh, and Walker (1985) considered reflective learning as "those intellectual and affective activities in which individuals engage to explore their experiences in order to lead to a new understanding and appreciation" (p. 19). In this sense, reflective learning is not just an approach for critically reviewing prior experiences to guide future actions or responses, but also a process of re-examining the experience in order to gain new insights of self and practice (Mezirow, 1998).

The benefits of reflective learning are widely recognised in the field of education and training, especially in L2 education. To promote reflective practice in L2 learning, reflective writing is one of the most widely used pedagogical strategies (Abednia, Hovassapian, Teimournezhad, & Ghanbari, 2013). Reflective writing can be described as a piece of students' written work in which they document their thoughts and feelings in response to their personal learning experience in a specific domain or profession. It can also be viewed as a critical analysis of the learning experience and a discussion of its implications for future applications (Chau & Cheng, 2012). Research indicates that reflective writing offers the potential for students to establish links between new knowledge and existing knowledge (Cochran-Smith & Lytle, 2001), to reinforce the application of critical thinking skills (Kuiper & Pesut, 2004), to promote self-evaluation and professional growth (Lee, 2008) and to develop confidence and competence in their organisational and writing skills (Chang & Lin, 2014).

However, research has also highlighted challenges associated with the implementation of reflective writing in higher education. Chief among them is that students often encounter difficulties in choosing what to write in their reflection. For example, in Lai and Calandra's (2007) study where preservice teachers' difficulties in reflective writing were explored, two themes emerged from the interview data: "struggle in understanding reflection" and "technical and repetitive reflection writing assignments; in most cases, not reflection writing at all" (p. 73). Their study showed that the participants had very little knowledge of reflection and there was a lack of specific requirements and guidance on how to critically examine one's own experiences. The findings are consistent with those of other studies (Greiman & Covington, 2007; Martin, 2005), underlining the need to help students develop a better understanding of reflection. An issue arising from this is to consider whether introducing a classification framework for reflective skills would be beneficial to students.

Classification frameworks for reflective skills

One research area in reflective writing is to design a classification framework for evaluation of students' reflective skills. Some classification frameworks have been developed for specific domains to distinguish lower-level reflective skills from more-advanced reflective skills. Despite differences in their terminology and presentation, the classification frameworks share a common understanding of what constitutes a low level of reflection (e.g., a mere description of learning experiences) and a high level of reflection (e.g., interpretation and transformation of learning experiences into practice). The frameworks are described below.

McNeill, Brown, and Shaw (2010) proposed a framework to evaluate the reflective entries of specialist trainees with varying reflective skills in medical training. The framework classifies a reflective entry into one of the three different levels: the most reflective level (i.e., level 1), the intermediate reflective level (i.e., level 2) and the least reflective level (i.e., level 3). Each level has its own set of descriptors designed by a group of experts in the fields of reflective learning and medical education. Critical thinking, context description, feelings and emotion, literature connection, evidence of learning and action planning are identified as key aspects for assessing the level of a reflective entry. According to the framework, an entry is considered most reflective if it demonstrates a high level of critical analysis with strong evidence of prior learning and with an action plan for future learning. In contrast, an entry is considered least reflective if it is just a mere description of an event with no evidence of learning from the past and no planning for the future. The results of the study indicated that 10% of the trainees' reflective entries were identified as level 1 (most reflective) and nearly one-third were classified into level 3 (least reflective), suggesting that more guidance and support on reflective writing should be available to students.

Hegarty (2011) designed a framework to analyse reflective writing of students undertaking a master's program in teacher education. The framework is represented in a five-level hierarchical structure, where each level corresponds to one of the five categories: descriptive, explanatory, supported, contextual and critical reflection. Descriptive reflection represents the lowest level of the framework. At this level, students simply describe and narrate events in their lives without providing justification for their actions. Explanatory reflection is at the second level, where students analyse their experiences from a personal or professional perspective. The third level, namely supported reflection, refers to a reflective entry demonstrating the connection between an event (or action) and its supporting evidence from the literature. The fourth level, namely contextual reflection, suggests that an entry should include a consideration of different views and a comparison of those views with one's own point of view. Critical reflection is known as the highest level of reflection, which contains in-depth analysis of one's own experiences from multiple perspectives and relates the experiences to future learning. In Hegarty's (2011) study, descriptive reflection and explanatory reflection were found most frequently but the rest were rarely found. This finding is consistent with that of McNeill et al. (2010), pointing to the need for promoting students' understanding of critical reflection.

Ryan and Ryan (2013) proposed a classification framework for reflection, namely 4Rs, intended to enhance students' lifelong learning skills and professional practice in higher education. The framework distinguishes four levels of reflection: reporting and responding, relating, reasoning and reconstructing. The levels vary in complexity of thinking, ranging from lower-order thinking skills (e.g., reporting an incident and responding to it by making observations, expressing opinions or asking questions) to higher-order thinking skills (e.g., reconstructing for reframing future practice or professional understanding). Ryan and Ryan (2013) emphasised that reflective writing is never straightforward, and its success requires pedagogical support with reference to an effective evaluation framework. Ryan (2011) suggested an approach of teaching students how to compare and contrast the features of critical reflection with those of descriptive reflection. Power (2012) adopted a similar approach, where students were asked to individually identify and colour-code critical reflection in their reflective entries, followed by a face-to-face discussion with the class teacher on the color-coded sections. He found that this approach could encourage students to develop self-directed learning skills, and also provide the teacher with an opportunity to give timely, personalised feedback on students' reflection. He also noted that, however, this approach may have a significant impact on teacher workload, especially in a large class. A possible way to address this problem would be to capitalise on technology advances to assist the ongoing process of evaluation and feedback.

Automatic classification of student texts

Latent semantic analysis (LSA) is a computational approach used to determine document similarity in the field of information retrieval. It is known as a method to extract and represent the contextual-usage meaning of words by applying statistical computations to a large corpus of texts (Landauer, Foltz, & Laham, 1998). The underlying concept of LSA is to first aggregate all word contexts where a word does and does not appear, and then to generate a set of constraints that largely determine the similarity in meaning between two words

or groups of words. The constraints will be solved by linear algebra methods like singular value decomposition (SVD), and the solution will be used to guide the judgement of meaning similarity between words or documents.

LSA starts by counting the frequency of each term (word) in each document of a text corpus, and subsequently constructs a term-document co-occurrence matrix where cell(x,y) contains the term frequency-inverse document frequency (tf-idf) weight of the term x in the document y. The next step is to divide the term-document matrix by SVD into three matrices, of which one is a diagonal matrix. Small singular values in the diagonal matrix are eliminated, and the corresponding rows and columns in the other two matrices are ignored. As a result, the original term-document matrix can be transformed into a lower-dimension matrix. This truncated term-document matrix represents the correlational structure between terms and documents in a lower-dimensional semantic space. For automatic classification, a new document will firstly be transformed into a vector of features in the semantic space. The cosine similarity between the vector for the document and that for each known document in the corpus will then be calculated. In extreme cases, the similarity value between two documents is 1 if they are identical and 0 if they are completely different. A new document will be classified into the category of a known document if their similarity is highest among others.

LSA has successfully been applied to text classification for educational purposes such as grading student essays based on their content (Jorge-Botana, León, Olmos, & Escudero, 2010; Lemaire & Dessus, 2001), evaluating brief summaries of narrative and expository texts (Olmos, León, Escudero, & Jorge-Botana, 2011), assessing free-text answers with reference to standard solutions (Pérez-Martin, Pascual-Nieto, & Rodriguez, 2009) and providing adaptive feedback on student summaries in intelligent tutoring systems (He, Hui, & Quan, 2009; Kintsch et al., 2000). The studies reported that there was a high level of agreement between LSA and human judegments (with a coefficient of correlation close to or greater than 0.70), indicating that LSA could provide a good simulation of human performance in classifying student texts.

Proposed approach for classifying reflective skills in L2 learning

Classification framework

Chau and Cheng (2012) introduced a framework called A-S-E-R to classify students' reflective skills in L2 learning by analysing their reflective entries. Four key elements of reflective skills are considered in the framework: analysis, reformulation and future application; strategy application; external influences; and report of events or experiences. Each element is further divided into four hierarchical levels, progressing from level 1 (i.e., developing) to level 4 (i.e., competent). An identified category is denoted by an element's symbol and a level. For example, S4 represents level 4 on the scale assessing strategy application (S), indicating a critical analysis on the effectiveness of applied or alternative strategies for language learning. This study adopted the A-S-E-R framework to discriminate between levels of reflective skills with respect to L2 learning. Details of the classification framework are illustrated in Table 1.

Table 1
The A-S-E-R classification framework (adapted from Chau & Cheng, 2012)

Elements of reflective L2 learning skills	Category	Descriptor
Analysis, reformulation and future application (A): To analyse, reformulate, and refocus the experience; comprehensive discussion of implications of the experience in the context of future applications	A4: Clear ability	Discuss implications of the experience in the context of future applications and explain how the experience would benefit future language learning
	A3: Some ability	Evaluate positive and/or negative effects of the experience on current practices of language learning
	A2: Limited ability	Explain how the experience is connected to the objectives or challenges of language learning
	A1: Very limited ability	Describe and comment on the experience with little or no justification
Strategy application (S): To analyse effectiveness of applied or alternative strategies for language learning	S4: Critical analysis	Analyse and evaluate effectiveness of applied or alternative strategies on improving language learning
	S3: Logical explanation	Identify language problems and explain how applied or alternative strategies can address the problems
	S2: Relevant discussion	Describe choice and application of strategies for language learning
	S1 Superficial description	Briefly describe choice of strategies for language learning
External influences (E): To make comments about external influences (e.g., circumstances, others' perspectives) on the experience	E4: Insightful and constructive comments	Comment on how external influences can help change attitudes and behaviours towards language learning
	E3: Constructive comments	Comment on how external influences can help identify language learning needs and ways to make progress
	E2: Some comments	Comment on how external influences can help identify language learning needs
	E1: Very few comments	Briefly comment on external influences with little focus on language learning
Report of events or experiences (R): To report significant aspects of events or experiences	R4: Detailed and analytical report	Report significant aspects of the experience and make connections between language knowledge and skills gained from the experience
	R3: Detailed report	Report significant aspects of the experience and identify language knowledge and skills gained from the experience
	R2: Coherent report	Report some aspects of the experience and highlight language learning issues arising from the experience
	R1: Disjointed report	Briefly report the experience with little relevance to language learning

Automatic classification system

Figure 1 illustrates the design of the automatic system for classifying students' reflective L2 learning skills in this study. The system was implemented using Natural Language Toolkit (NLTK) in Python (Bird, Klein, & Loper, 2009). It comprises the following four processing stages:

(1) *Compiling a corpus of annotated texts*: The final corpus used in this study was compiled from 748 reflective entries created by 398 students during the academic years 2013–14 and 2014–15. The total length of the reflective entries is 178,772 words (or 10,002 sentences), with an average length of 239 words (or 13 sentences) per entry. Based on the A-S-E-R framework, each entry was manually and independently annotated by two researchers. Differences in annotation results between the researchers were identified and discussed to reach a consensus on the annotation standards. The basic unit of annotation was either a single sentence or an aggregate of consecutive sentences signifying the emergence of a category. If a unit has the properties of more than one category, it will be labelled with multiple categories. Figure 2 shows a sample annotated text where each unit is marked up with more than one category.

(2) *Pre-processing texts*: Data pre-processing is performed to first remove words with little or no semantic value, and then to transform meaningful words into a standardised form for further processing. The tasks include spelling correction (i.e., to correct misspelled words), tokenisation (i.e., to tokenise annotated sentences into bags of words, where each bag represents one category), stop word removal (i.e., to filter out non-essential terms like punctuation marks, numbers, non-letter characters, and meaningless words such as "the", "is", "in" and "on"), and stemming (i.e., to convert terms into their root forms, like converting "connection", "connective" or "connecting" into "connect").

(3) *Selecting features*: The process of feature selection is carried out to extract significant features from each bag of words. It starts by counting the frequency of each word (term) in a bag of words that corresponds to a certain category. LSA is subsequently applied to build a vector of term weights for each category, to generate a term-by-category matrix for representing a N-dimensional feature space where N is the total number of distinct terms, and to truncate the matrix for capturing the latent semantic structure of different categories in a lower-dimensional semantic space.

(4) *Classifying a new text*: Every sentence of a new reflective entry is represented by a vector of term weights in the feature space. At the sentence level, a sentence s is classified into a category c if the cosine similarity between the vectors for c and s is greater than the average similarity between the vectors for c and the training sentences labelled with c. As such, it is possible for a sentence to be labelled with more than one element (e.g., A and S). For each element, however, a sentence will only be assigned a level (e.g., A1 or A2) that yields the highest similarity value. At the entry level, a piece of reflective writing will be assigned a weighted average level for each element based on the classification results of its constituent sentences. For example, Figure 2 shows a sample text containing one occurrence of R2 and one occurrence of R3. The weighted average level of R is computed by adding up all annotated levels of R (i.e., 2 + 3 = 5) and then dividing the sum by the total number of occurrences of R (i.e., 1 + 1 = 2). As a result, the rounded answer is 3 and the sample text will be categorised into R3.

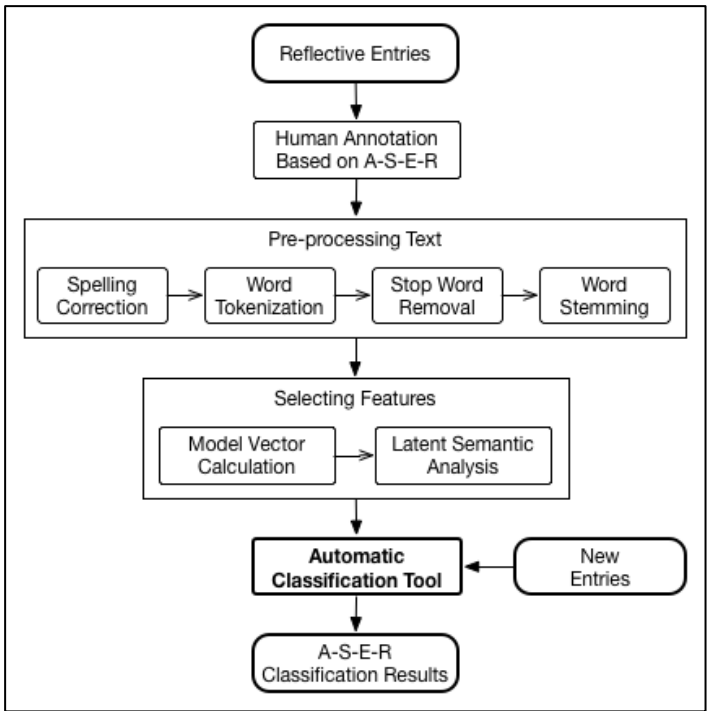

Figure 1. Design of the automatic classification system

{Three international students looked puzzled and said they did not understand what I was saying, it dawned on me something was wrong about my English.} **E2, R2.**

{I thought that the main problem was my English speaking skills, as I didn't know anything about the International Phonetic Alphabet (IPA). I told myself to start from the basic pronunciation. Therefore, I enrolled for a pronunciation course offered by the English Language Centre last year. I believed that the course would help improve my speaking skills.} **S3, R3.**

Figure 2. A sample annotated text

Research method

Context of study

This study was carried out as part of a government-funded project, namely Automatic Classification Techniques in Virtual Environments (ACTIVE). The project was proposed in response to the aforementioned concerns about students' understanding of critical reflective skills as well as teachers' workload in assessing and giving feedback on reflective writing. The main purpose of the project was to explore the effectiveness of using technological methods to address the concerns in the context of L2 learning. This study was specifically designed to meet this purpose with aims of developing a web-based automatic classification system to generate automated feedback on student reflections and examining students' attitudes about using the system to facilitate the development of critical reflective skills. Prior to the study, an ethical review application was approved by the university. An informed consent form was designed with a clear description of the purpose of the study and what each participant was expected to do in the study.

The study took place at the English Language Centre (ELC) of the Hong Kong Polytechnic University during the academic years 2013–14 and 2014–15. Since 2007, the ELC has developed a web-based e-portfolio

platform (http://eportfolio.elc.polyu.edu.hk/) to promote independent L2 learning in its English language enhancement courses, where students have the opportunity to develop, self-evaluate and reflect on their L2 learning experiences. Participants of this study were students enrolled on a 13-week, credit-bearing language enhancement course entitled *Advanced English for University Studies* (AEUS). The AEUS course aims to help learners study more effectively in the university's English medium learning environment and to improve and develop their English language proficiency. In the course, students are required to (1) plan, research for, write and revise a position argument essay; (2) present and justify views effectively in a mini oral defence; and (3) reflect on their English learning experiences and achievement and document the reflection in written form. A written reflective entry should be a minimum of 170 words and be submitted to the e-portfolio platform every 3 or 4 weeks. Failure to meet this requirement results in a grade deduction.

Participants

During the 2013–14 and 2014–15 academic years, 466 students (223 males and 243 females) taking the AEUS course voluntarily accepted an invitation to participate in the experimental study. At the beginning of the study, all participants were asked to sign an informed consent form and fill in a demographic data collection form. Table 2 provides an overview of the demographic information of participants. Of the participants, 82% were freshmen and 18% were sophomores. Their ages ranged from 17 to 26 years (M = 18.6 and SD = 1.1). They came from eight academic disciplines: Applied Sciences, Business, Construction and Environment, Engineering, Fashion and Textiles, Health Sciences, Hotel and Tourism Management, and Social Sciences. The majority attained level 4 or above in the Hong Kong Diploma of Secondary Education (HKDSE) English Language Examination. A benchmarking study (Hong Kong Examinations and Assessment Authority, 2013) showed that level 4 in the HKDSE English Language Examination is equivalent to the band score range between 6.31 and 6.51 in the International English Language Testing System (IELTS).

Table 2
Demographic information of participants

Item	Category	Count (Percentage)
1. Number of participating students	-	466 (100%)
2. Gender	Male	223 (47.9%)
	Female	243 (52.1%)
3. Age	17	35 (7.5%)
	18	232 (49.8%)
	19	132 (28.3%)
	20	47 (10.1%)
	21 or above	20 (4.3%)
4. Year of study	Year 1	380 (81.5%)
	Year 2	86 (18.5%)
5. Academic discipline	Applied Sciences	58 (12.4%)
	Business	72 (15.5%)
	Construction and Environment	47 (10.1%)
	Engineering	84 (18.0%)
	Fashion and Textiles	22 (4.7%)
	Health Sciences	123 (26.4%)
	Hotel and Tourism Management	35 (7.5%)
	Social Sciences	25 (5.4%)
6. HKDSE English language exam result	5**	3 (0.6%)
	5*	29 (6.2%)
	5	65 (13.9%)
	4	252 (54.1%)
	3	1 (0.2%)
	No answer	116 (24.9%)

Measures

Performance evaluation of the automatic classification system
By definition, accuracy measures the proportion of classifications that agree with the manually annotated results in the testing data set (Witten, Frank, & Hall, 2011). It was used to evaluate the effectiveness of the automatic classification system at the entry level. Beside accuracy, Cohen's kappa (*K*) was also used to measure the agreement between human and machine annotation, while taking the possibility of chance agreement into account (Landis & Koch, 1977).

Additionally, five-fold cross-validation was applied to evaluate the overall performance of the automatic classification system. During the cross-validation, all annotated entries in the corpus were randomly split into five equal-sized groups. Four groups were used as training data while the remaining group was used as testing data. This process was repeated five times until all groups were used for both training and testing purposes. The overall result was calculated by averaging the individual results derived from the five iterations.

Students' attitudes towards using the automatic classification system to support reflection
An online questionnaire was designed to solicit students' attitudes towards using the automatic classification system to support reflection. The questionnaire consists of seven self-report items. The first five items are fixed-response questions, of which the first two concern the number of entries submitted and feedback received. The next three measure levels of agreement on the effectiveness of the feedback and willingness to receive the feedback again using a 5-point Likert scale. The last two items are open-ended questions which allow students to provide their written comments. Details of the questionnaire items will be discussed in the ensuing section.

To cross-validate the questionnaire results and to elicit further views on the automatic classification system, participants were invited to attend a focus group interview. The interview protocol comprises four guiding questions:

(1) Do you think that the automatic classification system can help you improve your reflective writing?
(2) What are the main differences between the feedback from the automatic classification system and the feedback from the teacher?
(3) What do you like most and least about the automatic classification system?
(4) Do you have any suggestions to improve the automatic classification system?

Procedure

In essence, this study can be divided into three different phases: the development phase, the implementation phase and the evaluation phase. All participants involved in the study were first introduced to the A-S-E-R framework. A total of 127 students participated in the first phase, namely the development phase. This phase was initiated in the academic year 2013–14 to compile a corpus of manually annotated reflective entries and to build up an online system, namely ACTIVE (http://gp.ied.edu.hk/active), for the provision of automatic classification and feedback. The second phase, namely the implementation phase, involved a total of 339 participants studying the AEUS course during the academic year 2014–15. Participants in this phase were granted access to the ACTIVE system and they were also encouraged to use the system to generate feedback reports on drafts of their reflective entries. A feedback report contains four sections, including annotated entry, scoring and descriptors, remarks on total score and suggestions for improvement (see Figure 3). The final phase, namely the evaluation phase, started at the end of the AEUS course. In this phase, an online questionnaire was administered to the participants involved in the implementation phase and six interview sessions were arranged. Twenty-seven participants were randomly selected to attend the interview in order to share their views and experiences towards the ACTIVE system.

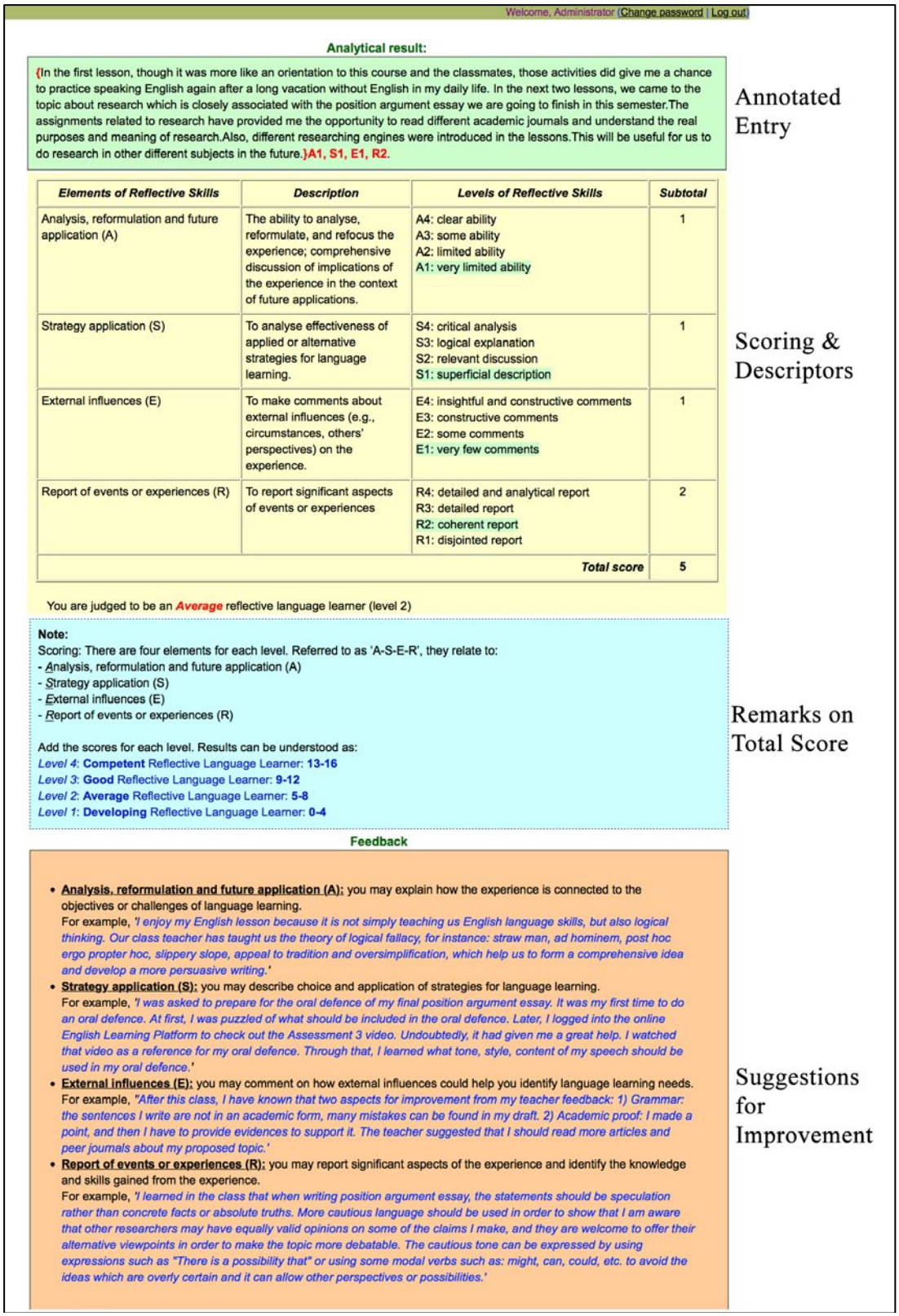

Figure 3. Layout of a feedback report generated by the ACTIVE system

Results and discussion

Performance evaluation of the automatic classification system

As stated earlier, a corpus of reflective entries collected from participants were manually annotated. The quality of the reflective entries varies among participants. The following is a sample reflective entry written by a student with good reflective skills, especially in the aspect of strategy application (S):

> {Most of us may agree that learning a foreign language is a challenging job, especially learning English which is a totally different language system when compared with Chinese. I had a hard time in learning English and disliked English. However, when you find the correct method, you can learn English in an effective way.} *A2*.
>
> {My strategy in learning English mainly follows the rule of *practice makes perfect*. I have read a sentence on a book: *to learn a language, the best way is to soak yourself in it*. This sentence inspired me a lot, as I agree that if you put yourself into the environment with an unknown language, you are forced to learn and speak that language in order to survive in that place. The more you practise that language, the more familiar you will be. The effect must be better than just learning from language books or sitting in classroom.} *S4*.
>
> {Therefore, I listen to the English radio channel every night and watch English movies with subtitles so as to get used to an English environment and practise listening. Moreover, I read English newspaper instead of Chinese newspaper to practise my reading. I have tried various methods to put myself in an English environment.} *S3*, *R3*.

In contrast, a sample of a student's entry shown below does not demonstrate consideration of strategy application (S):

> {As I caught an infectious disease before week 1, I was not supposed to go anywhere and so I did not attend the first English lesson. I was rather nervous before attending the lesson in week 2 since I did not know any classmates. But instead of feeling awkward, I found myself enjoying the lesson. {Although the topic 'research' was not really my favourite type of study, the way that all of us interacted in class was relaxing and enjoyable.} *A1*. }*R1*.
>
> {I had a great time in class but somehow I feel pressure when thinking about the course. The content and assignments of the course are quite challenging. The essay we need to work on is an obstacle to me because every single word I use should be selected very carefully.}*A2*. {But I know that I am not alone. My classmates are doing exactly the same thing as I do. I am going to overcome all the obstacles with unflagging determination.} *E1*. Hope that everyone in the class will work hard and play hard together this year.

As also expected, some students might reflect on their experiences beyond the context of L2 learning. Their entries would thus be unclassified according to the A-S-E-R framework. A representative student sample is given as follows:

> It was finally the last week of my first semester in university. During these 14 weeks, I have been worried, anticipated, moody and even anxious about the quizzes, assignments, presentations and tests of all the subjects. But at the end of this week, I realised that all of these components enriched my university life. I would also like to point out that there are actually quite a lot of differences between my expectation and the reality.
>
> People always claim that university is a place with great freedom and what we have to do is to enjoy our youth there. However, they never tell that when we are enjoying more freedom, we

> ought to take more responsibilities. For instance, while I tried to skip class, I had to spend more time on self-study consequently. While I did not hand in assignment on time, I had to bare the consequence of losing some proportion of marks. There are always consequences to take and be responsible of.
>
> In this semester, I started to get used to the life and systems of university. I might not do all my best in all courses, but the semester was more likely a great experience for me to have better management skill for next semester.

The corpus of reflective entries were automatically annotated by the ACTIVE system. Table 3 shows the confusion matrix and Cohen's kappa (*K*) obtained from the five-fold cross-validation on the corpus to assess the performance of the ACTIVE system. For each element of reflective skills (i.e., A, S, E and R), four hierarchical levels (i.e., 1, 2, 3 and 4) are included in the column and row headings together with an unclassified level (i.e., 0). The value at row *A* and column *B* of the confusion matrix represents the number of reflective entries classified into category *A* by human and category *B* by machine. Hence, the diagonal values of the matrix represent the numbers of matched results between human and machine annotation. As can be seen from Table 3, the results of human annotation are mostly consistent with those of machine annotation. This finding indicates that the performance of machine annotation is comparable to that of human annotation. Furthermore, the agreement between human and machine annotation is found to be substantial because all *K* values in the table are within or close to the range of substantial agreement (0.61–0.80) (Landis & Koch, 1977).

Table 3
Confusion matrix and Cohen's kappa (K) of the ACTIVE system

		Machine annotation					***K***
		A0	A1	A2	A3	A4	
Human annotation	A0	40	4	1	0	0	0.60
	A1	18	115	8	1	0	
	A2	25	23	260	44	4	
	A3	7	2	48	118	6	
	A4	0	0	4	11	9	
		S0	S1	S2	S3	S4	
	S0	102	5	3	0	0	0.73
	S1	40	298	43	0	0	
	S2	6	14	120	10	0	
	S3	1	0	12	91	0	
	S4	1	0	2	0	0	
		E0	E1	E2	E3	E4	
	E0	99	2	1	1	0	0.67
	E1	37	254	29	15	0	
	E2	9	23	167	20	0	
	E3	12	9	15	53	0	
	E4	2	0	0	0	0	
		R0	R1	R2	R3	R4	
	R0	90	0	3	1	0	0.70
	R1	18	142	11	1	0	
	R2	26	11	144	33	0	
	R3	26	0	30	199	3	
	R4	2	0	0	4	4	

Table 4 compares the classification accuracy of the ACTIVE system with two commonly used baseline methods, namely naïve Bayes (NB) and binary relevance (BR). NB is a simple probabilistic classifier based on Bayes' theorem, which assumes that the features in a class are independent of each other. This method is particularly efficient and useful for classifying texts from very large datasets (Sebastiani, 2002). BR is a fundamental approach to address the multi-label classification problem, which constructs a set of binary classifiers to distinguish one single class from all other classes (Tsoumakas, Katakis, & Vlahavas, 2010). From Table 4, it can be seen that the classification accuracy of the ACTIVE system ranges from 72% to 82% over the elements of reflective skills and it outperforms that of the baseline methods. The empirical results demonstrate the feasibility of applying existing technology to the automatic classification of reflective skills in L2 learning with satisfactory performance.

Table 4
Classification accuracy of different methods

Reflective elements	NB	BR	ACTIVE
A	60%	56%	72%
S	52%	66%	82%
E	54%	63%	77%
R	61%	64%	77%

Students' attitudes towards using the automatic classification system to support reflection

Results of the online questionnaire
The online questionnaire was emailed to all participants involved in the implementation phase. A total of 203 completed questionnaires were returned, yielding a response rate of 60%. The results of the fixed-response questions and open-ended questions are summarised in Table 5 and Table 6 respectively.

Table 5 shows that most respondents (93%) submitted one to three reflective entries in the AEUS course, and a similar percentage (89%) used the ACTIVE system one to three times to generate feedback on their reflective entries. Only a minority (less than 8%) did not submit their reflective entries or use the ACTIVE system. When asked about the quality of the feedback offered by the system, about 70% of respondents agreed or strongly agreed that the system's feedback could help them identify the strengths and weaknesses of their reflective skills in L2 learning. Nearly the same number of respondents (67%) agreed or strongly agreed that the system's feedback could help improve their reflective skills in L2 learning. To a large or full extent, a substantial percentage (65%) were willing to continue to receive the system's feedback on their reflective entries in the near future.

Table 5
Results of the fixed-response questions in the online questionnaire

Question	Responses	Number of respondents	Percentage of respondents
(1) How many reflective entries did you submit in the AEUS course?	More than 3	7	3.4%
	3	81	39.9%
	2	62	30.5%
	1	45	22.2%
	0	8	3.9%
(2) How often did you use the automatic classification system to generate feedback on your reflective entries?	More than 3	6	3.0%
	3	54	26.6%
	2	68	33.5%
	1	59	29.1%
	0	16	7.9%
(3) How much do you agree that the system's feedback could help you identify the strengths and weaknesses of your reflective skills in L2 learning?	Strongly agree	10	4.9%
	Agree	133	65.5%
	Neutral	46	22.7%
	Disagree	10	4.9%
	Strongly disagree	4	2.0%
(4) How much do you agree that the system's feedback could help you improve your reflective skills in L2 learning?	Strongly agree	8	3.9%
	Agree	128	63.1%
	Neutral	49	24.1%
	Disagree	15	7.4%
	Strongly disagree	3	1.5%
(5) To what extent are you willing to continue receiving the system's feedback on your reflective entries in the near future?	To a full extent	26	12.8%
	To a large extent	106	52.2%
	To some extent	46	22.7%
	To a small extent	18	8.9%
	Not at all	7	3.4%

In addition to the fixed-response questions, two open-ended questions were asked in the questionnaire to elicit respondents' views on what they like and dislike most about the ACTIVE system. Table 6 shows the questions, students' response categories, descriptive statistics and examples. A total of 58 respondents gave 74 written responses, of whom 57 and 17 were the answers to the first and second questions respectively. As shown in Table 6, the answers to the first question identify students' most favourite features of the system: (1) comprehensive and detailed comments; (2) convenience and ease of use; and (3) facilitation of reflection. On the other hand, the answers to the second question point to students' least favourite features of the system: (1) brief comments and suggestions; (2) incapability to fully understand human language; and (3) unclear scheme for scoring reflection.

Table 6
Results of the open-ended questions in the online questionnaire

Question	Response category	Frequency count	Valid percentage	Example
(1) What do you like most about the system?	M1: Comprehensive and detailed comments	32	56.1	• The generated comments are very detailed. • Some good examples of reflection are given.
	M2: Convenience and ease of use	13	22.8	• It is really convenient to use because feedback could be generated with just a click. • It is easy to generate the feedback. The process is fast and simple.
	M3: Facilitation of reflection	12	21.1	• It enables me to reflect on my language ability. • It can tell me the weaknesses of my reflective journal so I can see where to improve.
(2) What do you like least about the system?	L1: Incapability to fully understand human language	6	35.3	• Some content is misunderstood by the system. • I doubt if the system can really analyse our feelings and then give reasonable marks.
	L2: Brief comments and suggestions	6	35.3	• Suggestions are not concrete enough. • The comments given are a bit too general.
	L3: Unclear scheme for scoring reflection	5	29.4	• I do not know how to mark my reflection. • Sometimes I don't understand how it marks my reflective journals.

Table 7 shows a 3 x 3 contingency table of frequency counts according to the response categories identified in the two open-ended questions. Row percentages of the counts are given in brackets. As illustrated in Table 7, there were 16 respondents answering both questions. Due to this small sample size, the Fisher's exact test was employed to analyse the cross-tabulation. The test result suggests that there is no evidence indicating an association between the response categories for question 1 and those for question 2 ($p > 0.05$). However, students who were in favour of a specific feature tended not to report a particular problem. This finding can be attributed to the contrast between two opposing categories. For example, it can be seen from Table 7 that students who liked M1 (comprehensive and detailed comments) did not raise concerns about L2 (brief comments and suggestions) and L3 (unclear scheme for scoring reflection). This result is reasonable because L2 is exactly the opposite to M1 while L3 is quite inconsistent with M1. A similar explanation can be used to interpret the result that students who liked M3 (facilitation of reflection) did not raise concerns about L3. If students really found that the ACTIVE system could facilitate reflection, it was likely that they could easily understand the scoring scheme and identify the aspects with low scores for improvement in their next reflection.

Table 7
A contingency table of frequency counts according to the response categories identified in open-ended questions

Response categories for Q1	Response category for Q2			Total
	L1	L2	L3	
M1	3 (100%)	0	0	3 (100%)
M2	2 (18.2%)	4 (36.4%)	5 (45.5%)	11 (100%)
M3	1 (50%)	1 (50%)	0	2 (100%)
Total	6 (37.5%)	5 (31.3%)	5 (31.3%)	16 (100%)

As seen in Table 7, students who reported different favourite features (M1 to M3) of the ACTIVE system shared a common problem of L1. This is understandable because using computer technology to interpret human language is never error free. Given that the design of the ACTIVE system was based on the extraction of common features shared by a limited set of reflective entries, the occurrence of uncommon features that cannot be captured by the system may cause misinterpretation. The second and third problems (L2 and L3) identified by some students, however, do not coincide with the favourite feature (M1) identified by some other students. Such a discrepancy warrants further investigation, as discussed in the next section.

Results of the focus group interview

Qualitative data was mainly obtained from six post-event interview sessions with a total of 27 participants (S1 to S27, 17 females and 10 males) chosen randomly. Each session lasted approximately 30 to 45 minutes. All interview sessions were audio-recorded, transcribed verbatim, summarised and thematically categorised using thematic analysis (Braun & Clarke, 2006). After analysing the interview transcripts, four major themes emerged as significant for understanding students' views on the use of the automatic classification system to support their reflection: elements of reflective skills in L2 learning; scoring for reflection; content of the feedback report; and arguments for and against automatic classification of reflective skills. The themes are discussed below along with student quotes from the interview transcripts.

The first recurring theme expressed by the interviewees is about the benefits of using key elements of reflective skills in L2 learning to monitor and evaluate their reflective proficiency. Although the feedback report on each reflective entry was standardised in its format, it was not in content. Different reflective entries prepared by the same student might be classified into different categories of reflective skills. In this regard, most interviewees (S1–S10, S13–S18, S20–S22, S25–S26) considered the feedback report useful in supporting the improvement of their reflection. Such usefulness can be ascribed to the effectiveness of the A-S-E-R framework in identifying the strengths and weaknesses of students' reflective skills. As an interviewee (S1) observed:

> The feedback is effective and helpful because the feedback highlights the strengths and weaknesses in different aspects of my reflective skills in L2 learning. It helps me to see which areas I need to improve on to get into a deeper reflection.

Another interviewee (S2) supported the same view:

> The four aspects [A-S-E-R] and their scores in my feedback provide good directions for me to improve the quality of my next reflection. … It is good to see that the sentences in my reflection are labelled with elements and also graded for their levels. The feedback shows my proficiency in each aspect and which type of reflective student I am.

The second theme expressed by some interviewees is associated with scoring for reflection. While most interviewees recognised the usefulness of the system's feedback in improving their reflection, some (S11, S12, S19, S23, S24) were not satisfied with the scores they attained in the assessment of their reflection. They particularly raised concerns that a low reflection score would undermine their motivation and confidence to

enhance their reflective entries, so they preferred to receive qualitative feedback only. An interviewee (S11) commented on this:

> Actually, I was a little bit disappointed that I got a low reflection score. I did put much effort in drafting my reflection ... I agree that the feedback could point out my weaknesses, but the score could not encourage me to work harder to improve my reflective skills. Given a low score, I just feel that I would not be able to achieve a great improvement in my reflection even though I try to do more.

The third them is about the content of the feedback report. The majority of interviewees (S1–S10, S13–S18, S20–S22, S25, S26) agreed that the feedback report could help identify their strengths and weaknesses in relation to reflective skills in L2 learning. Some interviewees (S2–S5, S13–S16, S25, S26) also appreciated that good examples of reflection were given in their feedback reports. They said that they could learn good reflective practices from those examples and the examples could help them consider how to improve their own reflection. An interviewee (S3) reported:

> I like reading the examples given in my feedback report. From the examples, it was easier for me to understand what constitutes a good reflection and how to effectively reflect on my English learning experiences. I believe that the examples would also help other students improve their reflection.

However, a few interviewees (S11, S12, S23) pointed out that the feedback report should provide more specific suggestions on how to further enhance their reflective skills at sentence level. For example, the following interviewee (S23) noted that strategy application (S) is a key element of reflective skills in L2 learning (see Table 1) and he asked for suggestions about strategies for L2 learning without a specific context (level 1):

> I want to have more feedback about what strategies should be adopted to enhance my English learning ability. The feedback report could suggest something that we can try and do to make progress in developing our reflective skills. For example, talking more with international students may be an effective strategy to improve my English speaking ... I would like to see this type of assistance [effective strategy] offered in the feedback and step-by-step instructions on how to express the suggested ideas in sentences of my reflection.

The following interviewee (S12) echoed the same concern but focused on achieving a higher level of reflective skills:

> It was clear from the system's feedback that some sentences in my reflection were at a low level [of reflective skills]. However, I had no ideas of how those sentences could be revised to attain at a higher level. It would be perfect if the system could give me specific suggestions on how to modify the sentences, just like the instructions given by the teacher during consultation hours.

Both interviewees hoped that the system could provide detailed instructions on how to modify their reflection at sentence level. Their request, however, is beyond the present capabilities of the system and remains a huge challenge for current technology and research areas like automated essay evaluation (Shemis & Burstein, 2013).

The last theme is about the arguments for and against automatic classification of reflective skills. Two-thirds of the interviewees (S1–S6, S8–S10, S13–S17, S21, S22, S25, 26) rendered support to the development of the automatic system for classifying their reflective skills. They noted that their teachers were busy and could not afford to provide timely feedback on their reflective entries. To address this challenge, the interviewees believed that applying automatic classification technology into reflective writing can be a feasible alternative. This view was corroborated by the following interviewee's comment (S3):

> In the past I had to wait for a long time before receiving teacher comments on my reflective entry. This would stall my progress in reviewing and improving my reflective skills. I found that it was a good experience to use the automatic classification system for preparing my reflective entry. Getting immediate feedback on my reflection was just a click away.

Despite the potential for the automatic classification system to improve students' reflective skills, there was a perception among the interviewees (S7, S11, S12, S18–S20) that machine feedback can only supplement but not replace teacher feedback. The following quotation (S7) illustrates this point:

> Although the automatic classification system could provide a timely and systematic evaluation of my reflective skills, I would also prefer to receive teacher feedback because it shows more concerns about my personal feelings and needs. Teacher feedback is also more positive and encouraging, which makes me feel more motivated to keep writing my reflective entries.

Taken together, the analysis of student responses collected by the online questionnaire and focus group interview confirms the benefits of the automatic classification system in supporting the development of students' reflective skills. In particular, the system offers three advantageous features: it is convenient to generate immediate and systematic feedback on student reflection; the feedback highlights one's strengths and weaknesses of reflective skills in L2 learning; and students are provided with good examples of reflection to help them learn reflective skills from others.

On the other hand, the analysis also reveals that there are concerns about the use of the automatic classification systems for self-reflection. They include the demotivating effect of giving a reflective entry a low score; the need for providing more specific suggestions on how to improve reflection; and the indispensable role of teacher feedback in caring about the feelings of students and encouraging them to continue with the practice of reflective writing. The qualitative results clearly show that both teacher and machine feedback have their own strengths and limitations. Teacher feedback could be specific and encouraging, but could also be slow, inconsistent among teachers and occasionally unavailable for many reasons. Machine feedback could be quick, consistent and always available on request. However, it could also be general and rigid. In this connection, there is no single type of feedback (teacher or machine) that can best suit the needs of all students. This issue underscores the need to explore the use of multi-channel, multi-layer feedback in improving students' reflective skills.

Implications and future work

The results of this study have two implications for researchers and practitioners of technology-assisted reflective learning. First, this study contributes to the development of an automatic classification system for generating immediate and individualised feedback on students' reflective skills in L2 learning. It lends further support to the view that technology could play a beneficial role in the process of evaluating the quality of student submissions and then providing automated feedback on them. The findings of this study indicate that the classification accuracy of the ACTIVE system is comparable to that of human annotators. Additionally, they show that the ACTIVE system can provide students with immediate feedback on their reflective entries and with opportunities to make revisions to the entries at an early stage before receiving teacher feedback at a later stage. Researchers can further explore this research direction to gain insights into how the automatic classification technology can support and enhance the feedback process within different learning contexts.

Second, this study also adds evidence to confirm that most students see the potential and benefits of automated feedback on their reflective entries. Specifically, they find the feedback helpful in systematically identifying their strengths, weaknesses and areas for improvement according to the A-S-E-R classification framework. With a high level of classification accuracy, this kind of feedback would furnish students with evaluation results to reflect upon their own learning. Given that the automated feedback can be generated at different points of time, students can monitor and review their progress towards some particular goals. They can also adopt repair strategies to adapt to the attainment of the goals. In the long run, the formative

development of reflective skills would be facilitated through a self-directed, technology-assisted and performance-based learning process. To validate the positive role of the automated feedback, its impact on the quality of students' reflective entries and on their academic achievement will be investigated in future research.

Conclusion

The results of this study indicate that the ACTIVE system could achieve a satisfactory level of classification accuracy over the key elements of reflective skills. This finding not only suggests that the results of machine classification could be substantially consistent with those of human classification, but also demonstrates the feasibility of using existing technology to automatically evaluate and classify the reflective skills of ESL students. In addition, the results show that students were generally satisfied with using the ACTIVE system to receive immediate and systematic feedback on the strengths and weaknesses of their reflective skills. However, the results also indicate that the automated feedback cannot replace teacher feedback. Some students reported that teacher feedback was very important for them to continue with the reflective practice because it could provide them with tailor-made advice, care and encouragement. The overall results of this study validate that both machine feedback and teacher feedback play their unique roles in meeting various needs of students. For this reason, researchers and practitioners should consider how to integrate both feedback types into the process of reflection in order to maximise the improvement of students' reflective skills in L2 learning. In a broader sense, the findings would contribute to stimulating dialogue and advancing further work to explore the use of multi-channel, multi-layer feedback in improving reflective learning.

As with all research, there are limitations in this study and some are noted here. First, the sample was taken from a single university in Hong Kong. In order to generalise the findings, more studies should be carried out across different universities in both English-speaking and non-English-speaking countries. Second, the reflective entries used in this study were collected from an English language enhancement course only. Research can also be undertaken across various professional development programmes (e.g., teacher education and nursing) in which critical reflection is considered a highly valued ability for practitioners. Third, this study focuses on exploring the accuracy of the automatic classification system as well as the benefits and limitations of the automatic classification system perceived by students. Research can be conducted to further investigate the impact of the system on the quantity and the quality of students' reflective entries. The impact of the system on students' persistence in their reflective practice and on their academic achievement can also be explored in future studies.

Acknowledgments

This research is financially supported by General Research Fund of Hong Kong (No. 840913).

References

Abednia, A., Hovassapian, A., Teimournezhad, S., & Ghanbari, N. (2013). Reflective journal writing: Exploring in-service EFL teachers' perceptions. *System, 41*(3), 503–514. doi:10.1016/j.system.2013.05.003

Baturay, M. H., & Daloğlu, A. (2010). E-portfolio assessment in an online English language course. *Computer Assisted Language Learning, 23*(5), 413–428. doi:10.1080/09588221.2010.520671

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastapol, CA: O'Reilly Media.

Bisman, J. (2011). Engaged pedagogy: A study of the use of reflective journals in accounting education. *Assessment & Evaluation in Higher Education, 36*(3), 315–330. doi:10.1080/02602930903428676

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. London: Kogan Page.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. doi:10.1191/1478088706qp063oa

Chang, M. M., & Lin, M. C. (2014). The effect of reflective learning e-journals on reading comprehension and communication in language learning. *Computers & Education, 71*, 124–132. doi:10.1016/j.compedu.2013.09.023

Chau, J., & Cheng, G. (2010). Towards understanding the potential of e-portfolios for independent learning: A qualitative study. *Australasian Journal of Educational Technology, 26*(7), 932–950. doi:10.14742/ajet.1026

Chau, J., & Cheng, G. (2012). Developing Chinese students' reflective second language learning skills in higher education. *The Journal of Language Teaching and Learning, 2*(1), 15–32. Retrieved from http://www.jltl.org/index.php/jltl/article/view/54/21

Cochran-Smith, M., & Lytle, S. L. (2001). Beyond certainty: Taking an inquiry stance on practice. In A. Lieberman & L. Miller (Eds.), *Teachers caught in the action: Professional development that matters* (pp. 45–58). New York, NY: Teachers College Press.

Dewey, J. (1933). *How we think*. Lexington, MA: Heath. doi:10.1037/10903-000

Granberg, C. (2010). Social software for reflective dialogue: Questions about reflection and dialogue in student teachers' blogs. *Technology, Pedagogy and Education, 19*(3), 345–360. doi:10.1080/1475939X.2010.513766

Greiman, B. C., & Covington, H. K. (2007). Reflective thinking and journal writing: Examining student teachers' perceptions of preferred reflective modality, journal writing outcomes, and journal structure. *Career and Technical Education Research, 32*(2), 115–139. doi:10.5328/CTER32.2.115

He, Y., Hui, S. C., & Quan, T. T. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers & Education, 53*(3), 890–899. doi:10.1016/j.compedu.2009.05.008

Hegarty, B. (2011). Is reflective writing an enigma? Can preparing evidence for an electronic portfolio develop skills for reflective practice? In G. Williams, P. Statham, N. Brown, & B. Cleland (Eds.), *Proceedings ascilite Hobart 2011*. *Changing Demands, Changing Directions.* (pp. 580–593). Retrieved from http://www.ascilite.org/conferences/hobart11/downloads/papers/Hegarty-full.pdf

Hong Kong Examinations and Assessment Authority. (2013). *Results of the benchmarking study between IELTS and HKDSE English Language Examination* [Press release]. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/MainNews/press_20130430_eng.pdf

Hyland, K. (2003). *Second language writing*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511667251

Jorge-Botana, G., León, J. A., Olmos, R., & Escudero, I. (2010). Latent semantic analysis parameters for essay evaluation using small-scale corpora. *Journal of Quantitative Linguistics, 17*(1), 1–29. doi:10.1080/09296170903395890

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8*(2), 87–109. doi:10.1076/1049-4820(200008)8:2;1-B;FT087

Kuiper, R. A., & Pesut, D. J. (2004). Promoting cognitive and metacognitive reflective reasoning skills in nursing practice: Self-regulated learning theory. *Journal of Advanced Nursing, 45*(4), 381–391. doi:10.1046/j.1365-2648.2003.02921.x

Lai, G., & Calandra, B. (2007). Using online scaffolds to enhance preservice teachers' reflective journal writing: A qualitative analysis. *International Journal of Technology in Teaching and Learning, 3*(3), 66–81. Retrieved from http://sicet.org/web/journals/ijttl/issue0703/5_Lai_Calandra.pdf

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284. doi:10.1080/01638539809545028

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. doi:10.2307/2529310

Lee, I. (2008). Fostering preservice reflection through response journals. *Teacher Education Quarterly, 35*(1), 117–139. Retrieved from http://www.jstor.org/stable/23479034

Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research, 24*(3), 305–320. doi:10.2190/G649-0R9C-C021-P6X3

Martin, M. (2005). Reflection in teacher education: How can it be supported? *Educational Action Research, 13*(4), 525–542. doi:10.1080/09650790500200343

McNeill, H., Brown J. M., & Shaw, N. J. (2010). First year specialist trainees' engagement with reflective practice in the e-portfolio. *Advances in Health Science Education, 15*(4), 547–558. doi:10.1007/s10459-009-9217-8

Mezirow, J. (1998). On critical reflection. *Adult Education Quarterly*, *48*(3), 185–198. doi:10.1177/074171369804800305

Olmos, R., León, J. A., Escudero, I., & Jorge-Botana, G. (2011). Using latent semantic analysis to grade brief summaries: some proposals. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21*(2/3), 192–209. doi:10.1504/IJCEELL.2011.040198

Pérez-Martin, D., Pascual-Nieto, I., & Rodriguez, P. (2009). Computer-assisted assessment of free-text answers. *The Knowledge Engineering Review*, *24*(4), 353–374. doi:10.1017/S026988890999018X

Power, J. B. (2012). Towards a greater understanding of the effectiveness of reflective journals in a university language program. *Reflective Practice: International and Multidisciplinary Perspectives, 13*(5), 637–649. doi:10.1080/14623943.2012.697889

Ryan, M. (2011). Improving reflective writing in higher education: A social semiotic perspective. *Teaching in Higher Education, 16*(1), 99–111. doi:10.1080/13562517.2010.507311

Ryan, M., & Ryan, M. (2013). Theorising a model for teaching and assessing reflective learning in higher education. *Higher Education Research and Development*, *32*(2), 244–257. doi:10.1080/07294360.2012.661704

Scott, T. (2005). Creating the subject of portfolios: Reflective writing and the conveyance of institutional prerogatives. *Written Communication, 22*(1), 3–35. doi:10.1177/0741088304271831

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47. doi:10.1145/505282.505283

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and future directions*. New York, NY: Routledge. doi:10.4324/9780203122761

Smith, E. (2011). Teaching critical reflection. *Teaching in Higher Education, 16*(2), 211–223. doi:10.1080/13562517.2010.515022

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook*. (pp. 667–685). Berlin: Springer.

Wade, A., Abrami, P. C., & Sclater, J. (2005). An electronic portfolio to support learning. *Canadian Journal of Learning and Technology*, *31*(3), 33–50. Retrieved from http://www.cjlt.ca/index.php/cjlt/article/view/26489/19671

Wetzel, K., & Strudler, N. (2006). Costs and benefits of electronic portfolios in teacher education: Student voices. *Journal of Computing in Teacher Education*, *22*(3), 99–108. doi:10.1080/10402454.2006.10784544

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Corresponding author: Gary Cheng, chengks@eduhk.hk

Please cite as: Cheng, G. (2017). Towards an automatic classification system for supporting the development of critical reflective skills in L2 learning. *Australasian Journal of Educational Technology, 33*(4), 1-21. https://doi.org/10.14742/ajet.3029