Vol.2018-CE-144 No.22
2018/3/18

A design and prototype of a semi-automatic scoring system for English exams

THANH XUAN THI LAM[†1] CUONG TUAN NGUYEN[†1]
MASAKI NAKAGAWA[†1]

Abstract: This paper presents a design and prototype of a semi-automatic marking system for English examinations that integrates an English recognition engine into the system. Following the trend of touch-based and pen-based devices, people use those devices not only for working but also for learning. Language learning is one of the promising applications because writing by hand helps them learn more naturally and easily. However, manual marking by teachers takes a lot of time and energy. It is also error-prone. Therefore, automatic scoring systems are used and being studied to solve these problems and help teachers allocate more time for teaching. There are several kinds of automatic scoring systems, but they mark selections or text typed by keyboard. This paper focuses on marking short-response answers written by students. We present why we choose semi-automatic marking instead of fully automatic marking. Students will write their answers on tablets or digital papers. An English handwriting recognition engine recognizes the answers. If the recognition results are considered reliable, the system will transfer them to the automatic scoring process. If they are uncertain, the system will transfer the answers to teachers for the manual scoring process. For the method of scoring, we propose three methods based on lexicon driven, lexicon free, and their combination. In order to evaluate these methods, we collected data from 15 people with 25 questions. By evaluation experiments, we have confirmed that this system can help teachers save time for scoring. Moreover, its reliability is promising, which is shown by the false positive errors being 0.

Keywords: semi, scoring, English, exam, prototype, marking

1. Introduction

In recent years, devices such as tablets, PDAs and electronic white boards are becoming more and more popular. They allow users to annotate on documents, draw figures, and write notes, and so on more naturally and easily than traditional PCs with keyboards and mousses. A new type of electronic pen and paper devices such as Anoto pen is also available to input handwritings.

Language learning is one of the promising applications because writing by hand helps them learn more naturally and easily. The more we write to practice, the more surely we remember. Moreover, in almost all exams in schools, students usually use their handwritings to answer. This way will help teachers check the knowledge of students better than selection types. However, after the time of learning, we need to take exams to check how many percent we can remember or how many percent of correct knowledge we have.

Exams are necessary to judge or measure the understanding and the ability of students. Traditionally, teachers will score them by themselves. It is called manual scoring. Teachers must spend large time and effort to mark them. This work often makes teachers feel so tired and boring that sometimes the manual scoring cannot be correct. In addition, students need to wait for a while to receive their results. The length of the waiting time will depend on the amount of exams the teachers have to mark. The more exams, the longer students have to wait. Nowadays, we have mark sheets or computer/web based testing immediately returning scores but the questions are limited to selections or text typed in by keyboard.

To resolve this problem, in the Natural Language Processing (NLP) field, many researchers have been focused on researching about automated scoring systems [1]. The limitation

[†1] Graduate School of Engineering, Department of Computer and Information Science, Tokyo University of Agriculture and Technology

of this kind of automated scoring systems is that they score only text tests. When students do tests at schools, they often use their handwritings. Therefore, the current automatic scoring systems cannot be used. The purpose of our research is developing a scoring system that can be used in the reality. It can score handwritten tests. Our proposed system has two parts: handwritten English recognition and semi-automatic scoring.

The proposed semi-automatic scoring system will automatically score students' answers whose recognition results should be confident. Teachers mark manually the remaining answers whose recognition results are not confident. The intension of this task not only reduces the labor of the teachers but also ensure the fair score when scoring. The teachers can save time of evaluation and devote that time for teaching.

However, there are research challenges. Although a good handwritten English recognition engine is available, unfortunately, the recognition rate is not 100% (>90%). If we use directly recognition for the scoring system, it will affect so much on the results of automatic scoring. The scoring system will make mistakes if handwritings are misrecognized. To resolve this problem, we propose a semi-automatic scoring method. After students` handwritings are processed by an English handwriting recognition engine, we use confident scores to evaluate the recognition results. If the recognition results are considered reliable, the system will transfer them to the automatic scoring process. If they are uncertain, the system will transfer the answers to teachers for the manual scoring process. The semi-automatic scoring system includes automatic scoring by machines and manual scoring by teachers.

One more challenge is to increase the automatic scoring completion rate and correct completion rate of the system. The less manual scoring by teachers, the higher efficiency is realized. Therefore, we use lexicon free and lexicon driven [2] based English recognition engines for scoring. For the "given words" type questions, we use lexicon driven to recognize handwritings, and use lexicon free for "translation" type questions. The usage of the two engines together will help improve the scoring rate and performance of the system. We also run lexicon driven and lexicon free alone respectively to compare the performance.

2. Related works

Automating the task of scoring handwritten student essays is a challenging problem of Pattern Recognition (PR) and Artificial Intelligence (AI). The goal is to assign scores which are comparable to those of human scorers even though both human and machine handwriting recognition do not achieve perfect transcription. Srihari et al. proposed the first attempt based on coupling two technologies: optical handwriting recognition and automated essay scoring [3]. The test-bed is that of essays written by children in statewide reading comprehension tests in schools. The process involves several image-level operations: removal of pre-printed matter, segmentation of handwritten text lines and extraction of words. Constraints provided by the reading passage, the question and the answer rubric help recognize handwritten words. The method of essay scoring is based on using a vector space model and a machine learning approach to learn scoring parameters from a set of human-scored samples. System performance is compared with scoring done by human raters.

In the Natural Language Processing (NLP), there are a lot of researches that work in the field of Automated Scoring for English. The Automated Scoring is defined as the computer technology that evaluates and scores the tests. Automated Scoring Systems are mainly used to overcome time, cost, reliability, and generalize issues in writing assessment. The way that automated scores are produced is understandable and substantively meaningful. These researches can be separated mainly into two types. There are the automated scoring systems for short-content responses and the automated essay scoring systems.

Automated Short-Content Response Scoring Systems: Short-text content response (CR) tasks are designed primarily to measure student content knowledge and skills, rather than writing ability. Short CR tasks require a student to respond with short text demonstrating his or her understanding of key concepts [4]. One challenge associated with such systems is to prepare questions with definitive correct answers that the automated scoring system can verify. If the questions call for opinions or other unverifiable discussion, the expected response set becomes less certain and more difficult for the automated scoring system to handle. Thus, a variety of factors influence the success of these systems for scoring, including the number of potential concepts that could be generated in a response, the variety of ways in which these concepts might be expressed, and/or the degree to which there is a clear distinction between correct and incorrect representations of the concept, among others.

Automated Essay Scoring: After the short-content response automatic scoring, many researchers also focus on the automatic essay scoring. This is a trend following the demand of study English for globalization. Many students in Asia countries put their efforts on TOEIC/TOEFL/IELTS tests for studying aboard or finding a good job. A number of studies have been conducted to assess the accuracy and reliability of the Automated Essay Scoring (AES) systems with respect to writing assessment [5]. The results of several AES studies reported high agreement rates between AES systems and human raters. AES systems have been criticized for lacking human interaction, vulnerability to cheating, and their need for a large corpus of sample text to train the system. Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators. One of the best engines is the e-rater of ETS [6].

3. Dataset

To make a practical system, we choose questions from a textbox of Japanese junior high school students [7]. The dataset has totally 25 questions. Because this system is a prototype of automatic scoring English exams, we just choose the immediate level questions. The dataset has three types of questions: fill a sentence with one or two missing words, re-arrange the given words to make a correct sentence, and translation from Japanese to English. Figure 2 shows an example of three types of questions.

Type A: Fill in the blanks with correct words

あなたは何が好きですか。

(_)(_) you like?

Type B: Rearrange given words to make a true sentence

You will (be/it/to/difficult/find) a good leader

Type C: Translate from Japanese to English

この本はとても面白かったです。

Figure 1. An example of three types of questions.

For the type of filling missing words called type A, students have to write the answers with one or two words. Type A is very suitable for students who are at the beginner levels. It helps them take acquaintance with not only the vocabulary but also the grammar of English. In fact, when they are learning a new language, it is important to keep learning new words to improve their vocabularies. However, it is not useful to learn a list of new words. Instead, when they find a new word, they should learn the whole sentence where the new word appears. Don't learn the word in isolation, learn the word in context. Moreover, if they write the phrase down in a notebook, they will find it easier to remember words and how to use them by memorizing an example sentence.

The second is the type of re-arranging the given words to make a correct sentence, which we call type B. This type is a little harder than type A because it inquires students to have enough knowledge about vocabulary and grammar. Students need not only understand the grammar in English but also understand all the meanings of vocabularies in given sentences.

The final is the type of translation from Japanese to English called type C. It requires the highest level. Students have to write a full sentence for answering. This way of study is conducted to examine the effect of using translation from Language 1 (L1) to Language 2 (L2) as a teaching technique on the improvement of EFL learners' linguistic accuracy. Another benefit of using L1 in L2 teaching is psychological value. Contrary to reasons put forth as to why students should be encouraged to use only the target language in class, informal translation in the class can become a form of peer support for the learners. However, there are not only one way to translate from Japanese to English. Japanese is different from English not only grammar but also the structure of sentences. English is S-V-O structure, but Japanese is S-O-V structure [8]. Especially, there are many tenses in English, but there only two tenses in English. They are present and past tense. If an English sentence in present perfect and past perfect tense, it will be really hard to transform 100% meaning from English to Japanese. To refrain this understanding, this system will give a hint about the grammar, and student will follow it.

This dataset targets on collecting 13 people, who are students in our departments. They are Vietnamese, Chinese, and Japanese students. For the structure of the dataset, we use XML format to arrange the information easily

4. Semi-automatic scoring system for handwritten English answers

4.1 Overview of the semi-automatic English scoring system

For the learning process of students, they firstly have to register their personal information. The system should design many levels for the effective and comfortable studying, because different students will have different levels. Each student will have exercises that are suitable for him/her to practice. Finally, they take the test to evaluate their understanding and knowledge. The flow of learning process is described basically in Figure 3. In this paper, we focus on the evaluation step by proposing a semi-automatic scoring system. There are two key features of the semi-automatic scoring system which makes it better than the others. The first thing is that students can answer questions naturally with their handwritings. The second thing is the semi-automatic scoring which saves time and effort of teachers.

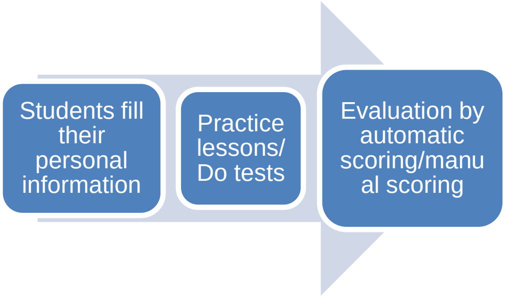
Students fill their personal information
Practice lessons/ Do tests
Evaluation by automatic scoring/manual scoring

Figure 2. Flow of the semi-automatic English scoring system

For the evaluation step, the system includes two sides: the student side and the teacher side. On the student side, students have to fill in their personal data into the system. It will help the system track students` data. This information just shows in a final report of scoring. Next, students will practice writing in some trial tests. They can switch to do tests when they are familiar with handwriting. They have to answer questions in the test and submit them to the system. The system saves the personal information and answers of all students to the particular folder. On the teacher side, the system recognizes handwritten answers and automatically scores them. Tests, which cannot be automatically scored by the system, will be scored manually by teachers. Teachers also add the new answers into the system in the case that the database system does not cover all correct answers. Figure 4 describes the user interface of Semi-Automatic English Scoring System.

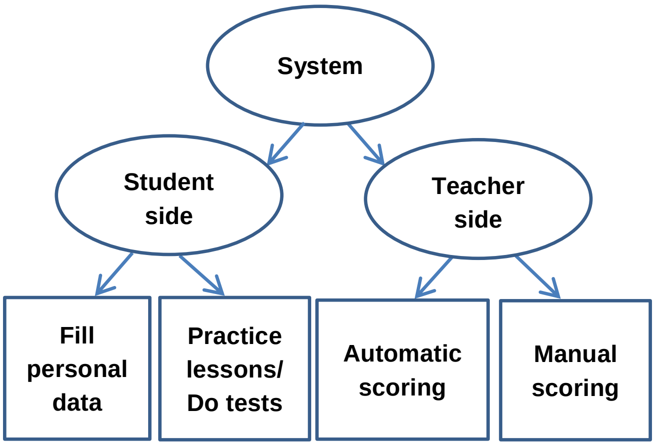

Figure 3. The user interface of Semi-Automatic English Scoring System

4.2 Semi-Automatic Scoring Process

This part describes more details about the semi-automatic scoring method. After students write their answers, the handwritten English recognition engine will recognize and give the recognition result. The recognition result has a score. The higher score is the higher reliability of the recognition result is. Therefore, we set a confident score as a threshold to determine whether the recognition result can be trusted or not. The trusted recognition result is marked automatically by the system while teachers mark the untrusted recognition result manually. For the marking process, we evaluate whether answers are correct or incorrect. In natural language process, they analyze the student’s answers and database answer and consider the grammar and semantics of them to make decision. In this paper, we just consider answers as short sentences or phrases. Therefore, we simplify the marking process by comparing word by word. Figure 5 shows the process of marking in the system. The process of determining confident score and marking process are described in more detail in Section 4.4.

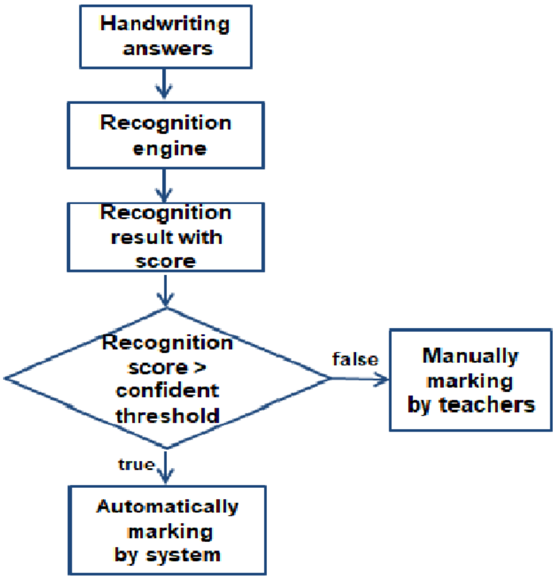
Handwriting answers
Recognition engine
Recognition result with score
Recognition score > confident threshold
false
Manually marking by teachers
true
Automatically marking by system

Figure 4. The Semi-Automatic Scoring Process

4.3 Online handwritten English recognition

We apply the Finite State Machine (FSM) based decoding method to recognize online handwritten text using recurrent neural networks (RNN). RNN receives a sequence of input features and outputs a sequence of class probabilities, and then FSM receives this probability sequence and produces the recognition result. Figure 6 shows the flow of the English recognition engine used in this system.

For feature extraction, we extract a set of point-based features including normalized distance, sine and cosine of the angle between the current line segment and the horizontal line and pen-up/pen-down features.

We use bidirectional Long Short Term Memory (BLSTM) [9], which consists of two LSTM layers for modeling the context from both forward and backward directions. Long Short Term Memory (LSTM) [10], is an advanced architecture of RNN designed to overcome the problem of vanishing or exploding gradients. LSTM could bridge long time delays between relevant input and target events, thus, it could incorporate long-range context for improving handwriting recognition.

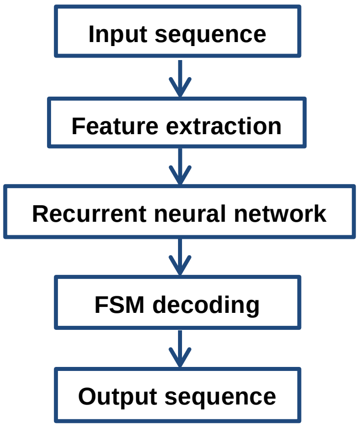

Figure 5. The flow of English recognition engine

For the decoding part, two recognizers can be used. They are lexicon-driven and lexicon free recognizers. With the lexicon-driven, output words are constrained by a predefined vocabulary. It is necessary when we need to score by matching words by words. With the lexicon-free, there will have no any constrains for output words [11]. We will show more details about the lexicon free when it is applied into the automatic scoring using handwriting recognition. In the automatic scoring, we have to compare the pros and cons of lexicon driven and lexicon free to decide which recognizer we should use in the system.

Firstly, we will discuss about the lexicon driven. As for the pros, the recognition rate will be high when it has the language model, but it may cause the false positive due to the constraint decoding. We have two solutions to resolve this case. One is using a threshold to reject low confident recognized words. It can be called the semi-automatic scoring [12]. Another is that we will validate recognized words by using the lexicon free recognition method. In this system, we combine the two methods for high scoring performance.

4.4 Method of scoring

The confident score is an important parameter in our semiautomatic scoring system. It decides how many percentages of answers are scored automatically. We make an analysis as follows. First, we run the recognition engine to recognize all collected answers. Then, we get all recognition scores. We calculate accumulative curve for the recognition score as in Figure 7, 8. Figure 7 shows the accumulative curve for lexicon free recognition engine whereas Figure 8 shows the accumulative curve for lexicon driven recognition engine. The horizontal axis is the confident score while the vertical axis is the completion rate. By using the graphs, we can select the confident score that satisfies the expected completion rate. To keep high accuracy of the scoring method, we select the confident score that the system can score automatically around 50%. Therefore, we select the confident score as -0.275 for lexicon free method and -0.5 for lexicon driven method.

In the dataset, teachers prepare totally correct and partially correct answers with their scores. For recognition results which have scores larger than the confident score, the system scores them automatically. The system calculates the edit distance between the recognition result and the answers provided by teachers. The system gives scores for the prepared answers having 0 edit distance.

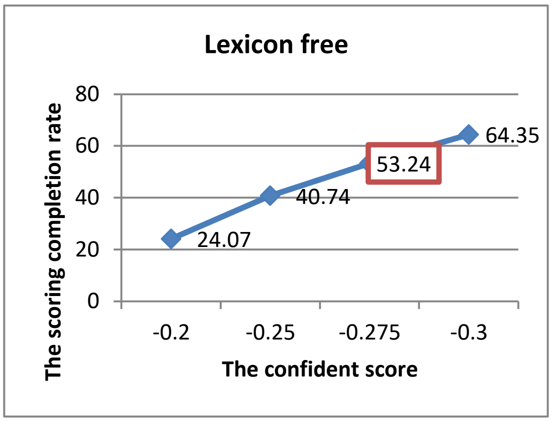

Figure 6. Accumulative curve of confident score for the lexicon free method

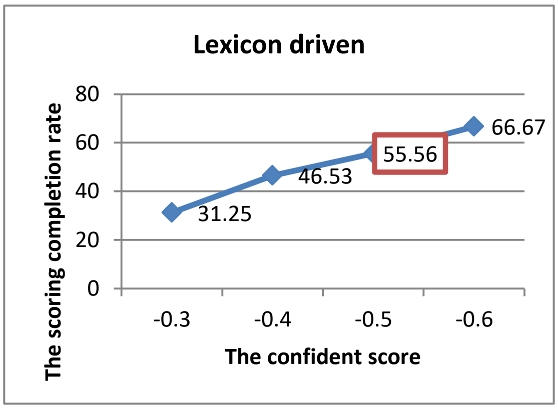

Figure 7. Accumulative curve of confident score for the lexicon driven method

5. Evaluation

We evaluate the system by using three English recognition engines: the lexicon free, lexicon driven, and their combination (lexicon free for recognizing a word/phrase or a sentence from provided words whereas lexicon driven for recognizing translated sentence). We compare the results of the three methods on the following measurements: the automatic scoring completion rate, the automatic correct scoring rate, false negative errors, the automatic scoring time, the manual scoring time, and the total time of manual scoring and checking. They are defined as follows:

$$\textbf{The automatic scoring completion rate} = \frac{\textbf{the number of answers scored by machine}}{\textbf{the number of total answers}} \quad (1)$$

The automatic correct scoring rate =

$$\frac{\text{the number of answers scored correctly by machine}}{\text{the number of answers scored by machine}} \quad (2)$$

When a machine makes a mistake of scoring, there are two kinds of errors: false negative error and false positive error. False negative error is an error that the machine scores an incorrect handwritten answer as a correct answer, while false positive error is an error that the machine scores a correct handwritten answer as an incorrect answer. Students will ask teachers to re-score only the false positive errors. Therefore, the false negative errors are very important measurement. We expect the false negative errors low as much as possible.

Table 1 and table 2 show the accuracy and processing time of the three scoring methods. For the automatic scoring completion rate and the automatic correct scoring rate, the combined method is the best. For the false negative errors, the three methods are 0. They are reliable and can be used in practical applications.

The automatic scoring time is almost similar around 2 hours among the three methods. The total time of manual scoring and checking is less than 30 minutes among them. The combined method incurrs the lowest time because this method scores 61.53% automatically. As the result, teachers will save a lot of time for scoring by using our system.

Table 1. The accuracy of the three scoring methods

Scoring method	Automatic scoring completion rate	Automatic correct scoring rate	False positive errors
Lexicon free based method	56.33%	71.43%	0
Lexicon driven based method	57.33%	78.07%	0
Combined method	61.53%	78.50%	0

Table 2. The processing time of the three scoring methods

Scoring method	Auto scoring time	Manual scoring time	Total time of manual scoring and checking
Lexicon free based method	1 h 55m 47s	13m 12s	28m 18s
Lexicon driven based method	2h 5m	12m	22m 33s
Combined method	1h 55m	10m 29 s	19m 44s

CONCLUSION

In this paper, we presented a prototype of the semi-automatic English exam scoring system integrating a handwritten English recognition engine. For the method of scoring, we proposed three methods based on lexicon driven, lexicon free, and their combination. In order to evaluate these methods, we collected handwritten answers from 13 people with 25 questions. Based on the results of the experiments, we see this system can help teachers save time for scoring. Moreover, its reliability is good, which is shown by the false positive errors of three methods are 0.

However, this system still has two weaknesses. The first limitation is that the speed of the English recognition engine is still slow, so that it effects on the system’s performance. The second weak-point is that the recognition rate of a long sentence is not good, so the scoring performance of long sentences is not high now. Moreover, the method of scoring still has some restrictions for the translation question type. This type usually has multiple answer keys. When the semi-scoring system runs on the first time, it may has some answers that teachers have not covered yet.

In the future work, we would like improve not only the English recognition engine in both recognition rate and speed but also the method of scoring. Now, this system score very well with missing words and short-responses types, so we want to improve the method of scoring for long sentences in the future work by using Computational Vector Grammars [13].

REFERENCES

[1] M. D. Shermis, J. C. Burstein, “Automated Essay Scoring: A Cross. Disciplinary Perspective”, Journal of Educational Measurement Vol. 42, No. 2, pp. 215-218, summer 2005.

[2] Bharath and Sriganesh Madhvanath,“HMM-based lexicon driven and lexicon free word recognition for online handwritten Indic scripts”, Proceedings of the Fourth International Conference on Document Analysis and Recognition Conference, 18-20 Aug. 1997.

[3] Sargur Srihari, Rohini Srihari, Pavithra Babu, Harish Srinivasan ,”On the Automatic Scoring of Handwritten Essays”, 2000.

[4] David M. Williamson,Randy E. Bennett, Stephen Lazer, “Automated Scoring for the Assessment of Common Core Standards”, Sep 2001.

[5] John K. Lewis,“Automated Essay Scoring: Writing Assessment and Instruction”, Sep 2001.

[6] Chodorow, M., & Burstein, J, “Beyond essay length: Evaluating e-rater's performance on TOEFL essays “, TOEFL Research Report No. RR-73,ETS RR-04-04. Princeton, NJ: ETS, 2004.

[7] Evine, Mr. Evine の中学英文法を修了するドリル，アルク

[8] Burstein, J., & Marcu, D.,“Benefits,“Modularity in an automated scoring system ”,Workshop on Using Toolsets and Architectures to Build NLP Systems, 18th International Conference on Computational Linguistics, Luxembourg, July 2000.

[9] A.Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", Neural Networks, 2005.

[10] S. Hochreiter, J. Schmidhuber,”Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification”, vol. 9, no. 8, Nov. 1997, pp. 1735-1780,

[11] Cuong Tuan Nguyen, Masaki Nakagawa,“Finite State Machine Based Decoding of Handwritten Text Using Recurrent Neural Networks”,2016 15th International Conference on Frontiers in Handwriting Recognition.

[12] Sargur Srihari, Rohini Srihari, Pavithra Babu, Harish Srinivasan, “On the Automatic Scoring of Handwritten Essays”.

[13] Richard Socher and John Bauer and Christopher D. Manning and Andrew Y. Ng,“Parsing With Compositional Vector Grammars”, ACL 2013.