Enhancing a free-text Adaptive Computer Assisted Assessment system with self-assessment features

Ismael Pascual-Nieto, Diana Pérez-Marín, Mick O'Donnell, Pilar Rodríguez
Computer Science Department, Universidad Autónoma de Madrid, Spain
{ismael.pascual,diana.perez,michael.odonnell,pilar.rodriguez}@uam.es

Abstract

Willow is a web-based application which automatically and adaptively assesses students' free-text answers written in Spanish and English. It is intended to help students review concepts outside of class, and provides an alternative assessment method. However, students tend to be greatly concerned by computer generated scores and, although they have been told that the system is not intended for summative purposes and their teachers will not have access to that scores, they remain worried that the machine could make a mistake and this mistake could have a negative impact on their final course grade. Given that automatic scoring cannot assure 100% approximation to a human teacher's assessment, we have introduced self-assessment features in Willow. Thanks to the combination of automatic free-text scoring and self-assessment, students have more control over their evaluation and, whenever they do not agree with the automatic score assigned by the system, they can change it. This new feature has been tested by 22 students reviewing a course with Willow in the 2007-2008 academic year. We observed that the students did not abuse the self-assessment feature, only modifying the score assigned in 5% of answered questions.

1. Motivation

Computer Assisted Assessment (CAA) is the field that studies how computers can be effectively used to assess student learning. Since its creation, several kinds of assessments have been developed, such as self assessment, in which students score their own exercises (e.g. [1]); and, free-text assessment, which in the opinion of many psychologists and researchers are necessary to assess the highest cognitive skills [2].

In previous work, our focus has been on free-text assessment. We constructed an automatic and adaptive free-text scorer called Willow [3], which is used as a web-based application. The core idea of the system is that the more similar the student's answer is to the answers as provided by the teacher (the 'reference answers'), the higher his or her score should be. The goal is not to replace the teacher or to serve as a summative assessment tool. Rather, the system is meant to serve as an alternative means for the student to assess their progress.

During the 2005-2007 academic years, we asked 56 Engineering degree students to use Willow. Our goal was to find out whether students considered the system useful, and what they liked (or disliked) about it. A questionnaire revealed that in general they welcomed the availability of a system which allowed them to interactively review a course outside of class, and which provided immediate feedback. Observation of their use of the system showed that in terms of feedback, they were interested not only in receiving a numerical score for their answers, but also in viewing more detailed feedback [3].

While the system's assessment correlates reasonably well with the scores assigned by human teachers (54% Pearson correlation), there are cases where the scoring algorithm falls down (e.g., expression of the right ideas but using terms which were not in the teacher's answers).

For this reason, we modified the system to allow students to modify the automatically generated score if they feel it is off the mark.

This paper, will present our work to combine automatic open-ended question assessment with self-assessment techniques, including results of an experiment in which 22 university students have used the system. Section 2 briefly describes the main features of the Willow system; Section 3 introduces the incorporation of self-assessment in Willow, and how it is combined with the automatic free-text scoring; Section 4 focuses on the experiment we performed and the results of that experiment; and, finally, Section 5 provides our main conclusions and lines of future work.

978-0-7695-3167-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICALT.2008.52

2. Willow

Willow is a web-based application which formatively assesses students' free-text answers in an automatic and adaptive way. It is able to process answers written both in Spanish and English.

The system emulates a dialogue between two animated agents: the owl representing the system, and a character representing the student.

Before students use the system, the teachers introduce a set of questions and their correct answers into the system. When a student starts using the system, the system presents a question to the student, and prompts for the student's answer. This free-text answer is compared to the correct answers previously stored, and a score is assigned, based on similarity of concepts covered. The student is presented with this score, and additional feedback, including the teacher's answers, and the student's answer with the concepts which were also in the teacher's answers highlighted in green.

Willow is based on a combination of Natural Language Processing (NLP) [4] and Adaptive Hypermedia (AH) techniques. A student model is maintained, recording, among other factors, how well the student is using a certain set of concepts in their answers and the questions passed.

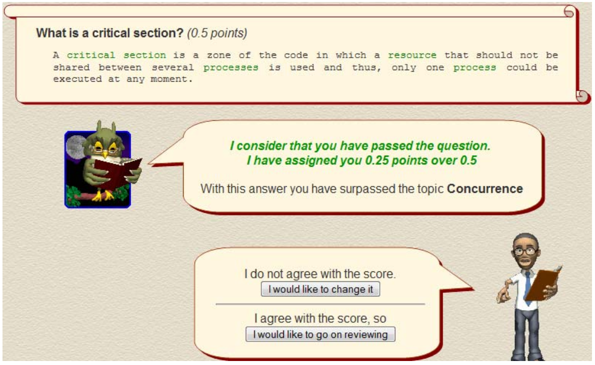
Given that the goal of Willow is not to replace teachers but to support them by providing students with out-of-class training opportunities, teachers are not given access to the scores achieved by the students. However, teachers do have access to various statistics about student use: number of questions answered, how long they have been using Willow each day, etc. [3].

3. Self-assessment in Willow

Students are accustomed to receiving a numerical score for each work they present, or exam they take. Numerical scores are given high importance in traditional summative assessment.

On the other hand, automatic assessment of free-text answers is still not perfect and it will not always be the case that the automatic score corresponds to what a teacher would give. Thus, students faced with an imperfect machine intelligence may stop using the system provided that they feel insecured that the scores are certainly not used with summative goals.

We contemplated the possibility of removing the numerical score provided by Willow, and just providing as feedback the processed student's answer and the teachers' correct answers. Willow would become then a self-assessment system instead of a free-text scoring system.



What is a critical section? *(0.5 points)*

A critical section is a zone of the code in which a resource that should not be shared between several processes is used and thus, only one process could be executed at any moment.

I consider that you have passed the question.
I have assigned you 0.25 points over 0.5

With this answer you have surpassed the topic **Concurrence**

I do not agree with the score.

I agree with the score, so

Figure 1. A snapshot of Willow's self-assessment feature

However, we rejected this idea because even though students are worried about the automatic score given by the system, they also consider this feature one of the most interesting [3]. They just want to have more control over the evaluation, and whenever they feel they have been unfairly marked, they want to have the possibility of changing the assigned score.

Taking this approach, Willow becomes a free-text scoring system combined with self-assessment, as shown in Figure 1. In effect, the interaction between the student and Willow continues as before: a dialogue is emulated between the owl asking a question to the student, the student answering the question, and the owl providing him or her feedback.

In the new system, the dialogue is continued with a negotiation over the score: the owl asks the student whether s/he accepts the automatically generated score. The student can compare his or her answer to the teachers' correct answers and decide whether s/he agrees with that score.

If the student accepts the score, it is stored and Willow asks the student a new question. If, on the other hand, the student rejects the score, the system asks the student to provide a score s/he feels appropriate. This score is then stored.

The student can increase or decrease the automatic score within assigning more than 0 or less than the full score for the question. The student can even assign a passing mark where the computer failed the student, or vice versa.

4. Experiment

After the self-assessment feature was included in the system, we asked 45 English Studies students to use Willow during the 2007-2008 academic year to review their Pragmatics course. Of these, 22 volunteered.

When the students were told about the automatic assessment possibility, they asked us whether this score could have a negative impact on their final score. Thus, when they were told about the self-assessment possibility, they seemed reassured.

Nevertheless, we also warned the students that they should not change the score without having previously compared their answer to the reference answers and truly thought that they deserved a different score than the one provided by the system.

The students claimed that Willow provides a user-friendly way to review their course, and that although the automatic score is not perfect, it is reasonably close, and in most cases they agreed with it.

In fact, from the 215 answers provided by these students and automatically scored by Willow, only in 12 answers (5% of the total) were self-assessed by the student. All 12 of these adjustments were made by just 2 of the 22 students.

Furthermore, in no case did the student lower their assigned score. Usually, the student raised a failing score which was close to passing (e.g. 0.4 increased to 0.6), and in all cases the modification is not very large, usually just two or three decimal points added by the student.

The mean quadratic error between Willow's score and the student's score has been calculated giving a value of 0.06 in a scale 0-1, showing that allowing self-assessment does not greatly change the final mark assigned by the system.

5. Conclusions and future work

The automatic and adaptive free-text scorer, Willow, has been extended to allow self-assessment. The main goal of the system is unchanged: to provide formative assessment to the students, in such a way as to allow students to review the concepts of a course outside of class. Additionally, the system provides immediate feedback.

In the past, the feedback given to the student by Willow consisted of the numerical score, the student's processed answer with the correctly used concepts marked in green and the reference answers. However, from previous experiments with Willow, we observed that students tend to be too concerned with the numerical score given [3]. We believed that students, who are accustomed to being summatively assessed, were worried about the system assigning inappropriate scores and might thus stop using the system.

Additionally, we wanted to give more control to the students over their evaluation, and to help them see that the numerical score given by Willow is just to help guide their reviewing of the course, and is not intended for summative purposes.

Therefore, we implemented a procedure to combine free-text scoring with self-assessment. The new version of Willow asks the same questions and generates the same feedback. The novelty is that now the student is asked whether s/he accepts the score or not.

Where the student accepts the score, it is stored to their student model. If not, the student is asked to provide the score s/he considers more adequate after having compared his or her answer to the teachers' correct answers.

Despite a fear that students would ignore the system generated scores and assign themselves maximum scores, students only modified scores in 12 out of 215 (5%) of the questions, and this was done by 2 out of 22 of the students. These 2 students in most cases only modified their scores by a small margin, generally to push a close fail up to a close pass.

A 0.06 Mean Quadratic Error in the 0-1 scale has been found between the students' and the automatic scores.

In general, the 22 university students who have used Willow claim that they like it and consider it a fun way to review the subject outside of class.

As future work, we would like to ask the students not only whether they agree with the automatic score assigned, but to always provide a degree of confidence in the score given to each question.

Acknowledgment

This work has been sponsored by Spanish Ministry of Science and Technology, project number TIN2007-64718. We would like to thank the teachers and students who have participated in the experiments.

7. References

[1] P. Blayney, and M. Freeman, "Automated marking of individualised spreadsheet assignments: the impact of different formative self-assessment options", in *Proceedings of the 7th Computer Assisted Assessment Conference*, 2003.

[2] M. Birenbaum, K. Tatsuoka, Y. Gutvirtz, "Effects of response format on diagnostic assessment of scholastic achievement", *Applied psychological measurement*, 1992.

[3] D. Pérez-Marín, Adaptive Computer Assisted Assessment of free-text students' answers: an approach to automatically generate students' conceptual models. Ph.D. dissertation, Escuela Politécnica Superior, Universidad Autónoma de Madrid, http://www.eps.uam.es/~dperez, 2007.

[4] E. Alfonseca, A. Moreno-Sandoval, J. Guirao, and M. Ruiz-Casado, "The wraetlic NLP suite", in *Proceedings of the Language and Resources Conference Evaluation*, 2006.