On the dynamic adaptation of Computer Assisted Assessment of free-text answers*

Diana Pérez-Marín*, Enrique Alfonseca*† and Pilar Rodríguez*

*Computer Science Department, Universidad Autonoma de Madrid
†Precision and Intelligence Laboratory, Tokyo Institute of Technology
{Diana.Perez,Enrique.Alfonseca,Pilar.Rodriguez}@uam.es

Abstract. To our knowledge, every free-text Computer Assisted Assessment (CAA) system automatically scores the students and gives feedback to them according to their responses, but, none of them include yet personalization options. The free-text CAA system Atenea [1] had simple adaptation possibilities by keeping static student profiles [2]. In this paper, we present a new adaptive version called Willow. It is based on Atenea and adds the possibility of dynamically choosing the questions to be asked according to their difficulty level, the students' profile and previous answers. Both Atenea and Willow have been tested with 32 students that manifested their satisfaction after using them. The results stimulate us to continue exploiting the possibilities of incorporating dynamic adaptation to free-text CAA.

1 Introduction

Computer Assisted Assessment (CAA) studies how to use effectively computers to automatically assess students' answers. Traditionally, it has been done just with objective testing questions. However, it is considered a quite limited type of assessment [3]. Hence, several other kinds have been proposed. In particular, in the mid-sixties, the possibility of assessing free-text answers was presented [4]. Since then, advances in Natural Language Processing (NLP) have made possible a favorable progress of this field [5].

The approach described in this paper is based on the free-text scoring system called Atenea [1] and its new version called Willow able to dynamically adapt the assessment process for the first time. Willow considers the students' personal profiles in the evaluation section and adjusts the difficulty level of the questions to the students' knowledge. Two experiments have been done with 32 students of our home university to study how well the adaptation in the assessment is appreciated and which adaptive techniques are more valuable.

The article is organized as follows: Section 2 describes Atenea and Willow. Section 3 details the experiments performed with the students. Finally, the conclusions and the open lines for future work are drawn out in Section 4.

* This work has been sponsored by Spanish Ministry of Science and Technology, project number TIN2004-0314.

2 Atenea and Willow

Atenea [1] is an on-line CAA system for automatically scoring free-text answers [1]. It is underpinned by statistical and NLP modules. Its main aim is to reinforce the concepts seen during the lesson with the teacher. It compares the student's answer to a set of correct answers (the references) by using the wraetlic toolkit[2]. The more similar a student's answer is to the references, the higher the score the student achieves.

Atenea randomly chooses the questions to ask the student until the end-of-session condition is fulfilled as a fixed number of questions has been completed or as a limited amount of time has expired. Recently, simple adaptation capabilities based on stereotypes were added to the system [2]. However, this kind of adaptation was very limited as it does not allow the system to dynamically adapt the assessment. Thus, we have created Willow, a new version of Atenea that, keeping all previous features, modifies dynamically the order in which the questions are presented to the students.

During the assessment session, as the students answer the questions of the different topics, which are chosen according to their difficulty levels, the values are modified to adjust the level of the questions to the level of knowledge that each student has in each topic addressed in the collection. When the students successfully answer a certain (configurable) percentage of the questions in a collection they are promoted to a higher level. On the other hand, a certain percentage of failures will demote them to the lower level. A topic is considered successfully passed when a student is in the highest level and has exceeded the percentage necessary to be promoted even further. In this way, a session may finish as soon as the student is considered apt in all the chosen topics.

3 Experiments with students

Atenea and Willow have been used in two different experiments by the students in the *Operating Systems* course, in the Telecommunications Engineering degree, Universidad Autonoma de Madrid[3]. The teachers of that subject (none of whom was involved in the development neither of Atenea nor of Willow) introduced twenty different questions of different levels of difficulty and topics from real exams of previous years. The use of the system was voluntary, but the teachers motivated the students by telling them that the questions had been taken from previous exams and that the practise would positively help them towards the final score in the subject.

A total of 32 students took part in the first experiment, from which two subgroups were randomly created each one with 16 students: group A that used Atenea, and group B that used Willow. The score to pass a question was set to

[1] Available at http://orestes.ii.uam.es:8080/ateneaAdaptativa/jsp/loginAtenea.jsp
[2] Available at www.ii.uam.es/∼ealfon/eng/download.html
[3] The authors would like to thank to Manuel Cebrián, Almudena Sierra, and Ismael Pascual for their collaboration in the experiments with the students.

Question	group A	group B
Familiarity with on-line applications	4.3	3.8
Difficulty of use	4.1	4.1
Intuitiveness of the interface	4.0	3.5
System's answer time	4.1	3.8
Fitness of students' needs	3.4	3.2
Order of the questions	3.2	3.4
Level of difficulty	2.3	2.9
Number of references	3.0	3.0
Number of questions answered	7.0	8.5
Time to study this course	less than 5 h.	less than 5 h.
Recommendation of using Atenea/Willow	yes	yes

Table 1. Average results for the first experiment.

50% of the maximum score, and the percentage to be promoted or demoted was set to 40% of the total number of questions. At the beginning all the students received a brief talk (5 minutes) about Atenea and Willow, its aim and how to use the system. Next, they were required to take a 5-minute test with five multiple-choice questions corresponding to the five topics under assessment. In a 0–5 scale, the average score was 2.8 for group A, and 3.2 for group B. Once finished the test, the students were allowed to start using the indicated version of the system during 20 minutes. After that, they were asked again to complete the same test to check if they had acquired new knowledge during the assessment session. The average score for the group A did not change at all, whereas the average score for the group B increased slightly up to 3.4. Finally, the students were asked to fill a non-anonymous Likert-type scale items satisfaction questionnaire. The results are summarized in Table 1.

In the second experiment students could use Atenea and/or Willow during a week from anywhere, at anytime, and feel free to choose any option. In particular, they were asked to compare Atenea and Willow and fill a non-anonymous comparison questionnaire at the end of the week. In total, seven students (22%) volunteered to take part in the experiment and six of them filled the questionnaire. The results are as follows: all the students agree that Willow fits better their needs; they think that the promotion-demotion feature is quite good; and, in general, they agree with the schema of starting with easy questions and next having them increasingly harder.

4 Conclusions and future work

The free-text CAA systems Atenea and Willow have been tested in two different experiments. The students were mostly familiarized with on-line applications but none of them had used before a system that automatically scores open-ended questions. The adaptation was focused on the dynamic selection of the questions

according to the procedure of promotions and demotions of difficulty levels as described in Section 2.

According to the comments given by the students, it can be confirmed that they like the idea of having an interactive system with questions from exams of previous years and the teachers' answers. 91% of the students would recommend to use the system to other friends in Operating System and other subjects. 80% of the students with Internet access at home prefer to log into the system from their home because they feel more comfortable. All the students find easy to use the system irrespectively of the version. Besides, they think that it is very useful to review concepts.

The students who used Willow were able to lightly increase their score the second time the test was presented after using the system just 20 minutes, whereas the students who use Atenea kept the same score. As expected, in average, the students of the first experiment who used Atenea answered less questions and they felt that the questions were more difficult than those who used Willow who declared that the order of the questions were more adequate.

When the students are directly asked if they prefer Atenea or Willow, there is not a clear answer. However, when they are asked about the system's features one by one, it can be seen that most prefer Willow, because it fits better their needs, the order of the questions is more adequate, and they feel more satisfied as the system controls their progress. In particular, the students who use Willow find its use more amusing and they feel more engaged to keep answering questions. On the other hand, some students say that they feel that they were learning less because the questions presented were less varied and that they find the interface less intuitive.

Including more dynamic adaptation in the system is a promising line of work that could be further exploited by updating dynamically the level of difficulty of each question according to the answers given by most of the students; giving the option of moving freely between the questions, with a color code to warn the students whether each question belongs to their knowledge level or not; and repeating the experiment with more students, maybe as a compulsory and anonymous experiment, to gather more results.

References

1. Alfonseca, E., Pérez, D.: Automatic assessment of short questions with a BLEU-inspired algorithm and shallow NLP. In: Advances in Natural Language Processing. Volume 3230 of Lecture Notes in Computer Science. Springer Verlag (2004) 25–35
2. Pérez, D., Alfonseca, E., Rodríguez, P.: Adapting the automatic assessment of free-text answers to the students profiles. In: Proceedings of the CAA conference, Loughborough, U.K. (2005)
3. Birenbaum, M., Tatsuoka, K., Gutvirtz, Y.: Effects of response format on diagnostic assessment of scholastic achievement. Applied psychological measurement **16** (1992)
4. Page, E.: The imminence of grading essays by computer. Phi Delta Kappan **47** (1966) 238–243
5. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. Journal of Information Technology Education **2** (2003) 319–330