Educational Adaptive Hypermedia meets Computer Assisted Assessment*

Enrique Alfonseca, Rosa M. Carro, Manuel Freire, Álvaro Ortigosa, Diana Pérez, and Pilar Rodríguez

Computer Science Department, Universidad Autonoma de Madrid,
Carretera de Colmenar Viejo, km. 14,5,
28043 Madrid, Spain
{Enrique.Alfonseca,Rosa.Carro,Manuel.Freire,Alvaro.Ortigosa,
Diana.Perez,Pilar.Rodriguez}@ii.uam.es

Abstract. In this paper we explore the many possibilities that arise when we combine adaptive web-based courses with computer-assisted assessment. We argue that this integration has several advantages, such as the feasibility of getting a better model of the student's progress, which will be used with adaptation purposes, and the possibility of proposing and evaluating open-ended questions in the way that is judged more suitable for each student.

1 Introduction

Adaptive hypermedia has been widely used for the development of adaptive Web-based courses, in which each student is individually guided during the learning process [1]. Most of these systems obtain feedback from the student from two sources: their behaviour browsing the course (e.g. pages visited, time spent in each or navigational path) and test questions (e.g. true-false, multiple-choice or fill-in-the-blank questions). Some authors have expressed their concern that this limited way of assessment may not be really measuring the depth of the student learning [2]. This fact has been the motivation of the field known as Computer-Assisted Assessment (CAA) of student essays. CAA of student essays is a long-standing problem that has received the attention of the Natural Language Processing research community. There are many possible ways to approach this problem, including a study of the organization, sentence structure and content of the student essay [3, E-rater], pattern-matching techniques [4, IEMS], or Latent Semantic Analysis [5, IEA]. Valenti et al. [6] describe the state-of-art of CAA systems.

In order to support adaptive distance teaching and learning, we have developed the TANGOW system, which supports the description of adaptive web-based courses and their dynamic generation, so that their components are tailored to each student at runtime [7,8]. We have also developed, independently, a CAA

* This work has been sponsored by CICYT, project number TIC2001-0685-C02-01.

system called Atenea [9] which is based on n-gram co-occurrence metrics [10]. In this paper we describe the ongoing integration of TANGOW and Atenea and the benefits of this integration.

The paper is structured as follows: in Section 2 we describe separately Atenea and TANGOW, next in Section 3 we describe their integration and finally Section 4 will highlight the conclusions and future work.

2 Atenea and Tangow

Atenea [9,11] is a Computer-Assisted Assessment system for automatically scoring students' short answers. It relies on the combination of shallow natural language processing modules [12] and statistically-based evaluation procedures. The recall is calculated by studying the percentage of those references that is covered by the student's answer and the precision of the student's answer is obtained by calculating the BLEU score [10].

The BLEU algorithm relies on the idea that a text is better when it is closer to a model text written by an expert. It was originally devised for evaluating Machine Translation systems, with the core idea that a Machine Translation is better when it is closer to the human translation. Hence, it looks for coincident n-grams (usually from unigrams to trigrams) between the human translator and the machine translations. In past work [9] we proved that BLEU can also be applied for evaluating students' answers, since the core idea remains: *the more similar a student's answer is to the teacher's reference answer, the better it is.* BLEU has been implemented, with minor modifications, as a module called ERB (Evaluating Responses with BLEU).

The system has already been tested with ten short questions, which are grouped in three collections: Operating Systems, Advanced Operating Systems and Object-Oriented Programming.

The scores provided by ERB are values between 0 and 1, where the upper and the lower bounds depend on the particular question. Nonetheless, we can perform a linear regression to find a correspondence between the interval of possible Atenea's scores and a fixed rank of scores, for instance, between 0 and 10. As expected, the quality of the assessment is very influenced by the kind of question and the references written. When scoring definitions, the correlation between Atenea's scores and the marks assigned manually can reach 0.80 [9]. On the other hand, questions that ask the student to make a reasoned argument or to compare several topics are more difficult to evaluate since they require a deeper linguistics processing.

The feedback that the students get from the system is a numerical score and a copy of their answer where, with color codes, they can observe which were the coincident n-grams and which words did not appear in any reference. From that output they can easily know which are the portions of their answers that are correct and have contributed in incrementing their score. Figure 1 shows an example answer page. In the user profile, students may also indicate whether they just want the score and are not interested in receiving this feedback.

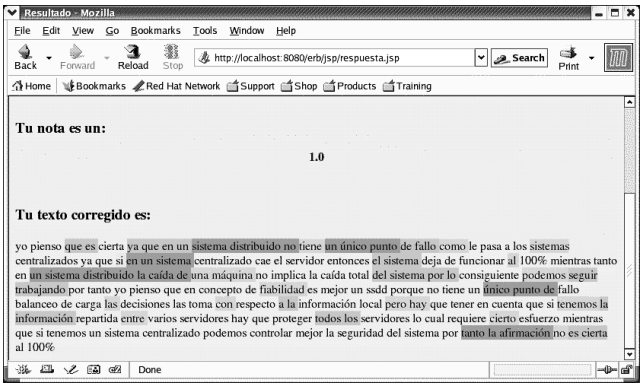

Fig. 1. Feedback that a student gets after answering the question *"Discuss whether distributed systems are more robust than monolithic systems"*, in Spanish. The darker the background, the longer the coincident n-gram.

Although Atenea is currently underpinned by ERB, it is not only limited to it. In fact, more NLP modules are currently being added to the already existing ones, including syntactic analyzers and word-sense disambiguation. A web-based wizard has also been developed to facilitate the task of introducing new data sets of questions and new questions.

The TANGOW system delivers adaptive web-based courses, and has evolved significantly since [7]. Courses delivered by TANGOW are composed of several tasks, that can be accomplished by the students. A task can correspond to either a theoretical explanation, an example, an exercise to be done individually or an activity to be performed collaboratively (problems to be solved, discussions, etc). The set of available tasks is constantly regenerated, tracking changes in the student's profile (static features and dynamic actions). Once a task is chosen, the system generates the corresponding web pages by selecting, among the content fragments and the set of available collaborative tools, those that provide the best possible fit to the current profile.

A rule-based formalism has been developed in order to facilitate the specification of alternative structures for the same course, and to support different teaching strategies, navigational guidance variations and collaboration workspaces for each type of student [8].

3 The integration of Atenea and Tangow

The integration of Atenea and TANGOW will support the inclusion of CAA exercises inside adaptive courses, as a new type of TANGOW task. The process of integrating both systems was expected to be quite easy: Atenea would be launched from TANGOW and, after asking the students and automatically evaluating their answers, it would return the results to TANGOW so that this information could

be used to update the user model and continue with the adaptation process. Atenea is currently configured to show different questions depending on the student's language and experience. The assessment is adapted to the used model, e.g. by showing easier questions to novice students. The information stored in user model also affect the set of reference answers that is chosen. The integration is not complete yet, but our ongoing work on implementation indicates that it will be simple to finish the links between both systems.

An initial step in the integration process was to decide which features from the current TANGOW user model would be used in Atenea in this first experience. We chose to use the *student name* as the login input in order to address the student by his or her name; *age*, because questions should be formulated in a simpler fashion for children than for adults, and different writing styles are expected from them; *language*, because we plan to extend Atenea with multilingual capabilities; *experience*, because the assessing process should be different for advanced students than for novice ones; and *feedback type*, because when formative assessment is used, the feedback should be more detailed than for summative assessment (where the score is the most relevant result).

Concerning the order of the questions, it is possible to take into account the student experience so that advanced students are not asked questions that they have already solved or that are too easy for their level. Moreover, the higher the level of experience, the stricter the system should be when assessing student answers.

The protocol for connecting TANGOW and Atenea is the following:

- TANGOW gathers information about the student's profile and sends it to Atenea, along with the identification of the task the student is going to perform, as well as the type of feedback desired.
- Atenea randomly chooses a question from the dataset corresponding to this task, that has not already been solved by the student (that is, not yet graded or graded less than half of the maximum score). The question is chosen taking into consideration the student profile. The answers submitted by the students are then evaluated by Atenea, and the resulting score and feedback is presented to the student. This process is repeated until the student has answered the required number of exercises. Finally, once the stop condition is satisfied, Atenea returns a holistic student score for the task to TANGOW.

A first consequence of the integration will be a richer set of activities, which can contribute to a more engaging learning process. Secondly, the use of the TANGOW formalism allows course authors to specify different teaching strategies by incorporating CAA activities at different points of the course, depending on each student's evolution. It will be possible for authors to choose the types of users to whom CAA activities will be presented; the places in the course where these exercises will appear; the requirements for a CAA activity to be proposed; and the grading criteria to determine the degree of success of each activity. Each of these adaptations can be made in different ways depending on the user's model. Finally, the formalism also supports the adaptation of CAA activities

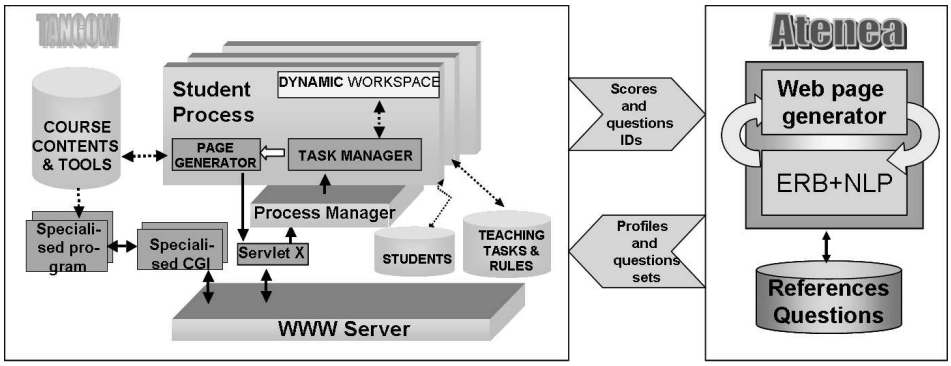

Fig. 2. Architecture of the integration between TANGOW and Atenea

themselves: the questions to be asked and the reference answers can be chosen according to each student's profile.

4 Conclusions and future work

We are implementing the integration of the adaptive hypermedia educational system TANGOW with Atenea, a program for automatic assessment of student answers. Atenea attains a good correlation with respect to teachers' marks, particularly when evaluating definitions and short descriptions [9]. The current implementation allows the students to try out their knowledge, and its complete integration with TANGOW, whose feasibility has been proved, will support the following:

- Atenea will use the description of the user profiles maintained by TANGOW, so it will accept variable profiles.
- The adaptation engine from TANGOW will decide at which time each student should be assessed, depending on his/her profile, knowledge, and actions, and Atenea will choose the most adequate set of questions for this student, resulting in a fairer evaluation.
- TANGOW will benefit not only from the possibility of automatically evaluating free-text answers, but also from the feedback from those questions, which can be used to guide the students during the rest of the course.
- It will be possible to obtain a dataset of student answers related to their profile and performance in the course, which we shall use in further studies to analyze how the adaptation can improve CAA activities.

The interaction protocol between Atenea and TANGOW has already been designed and is being currently implemented. Current work comprises a complete integration of Atenea and TANGOW in the direction described above and the evaluation of the integrated system with real students.

References

[1] Brusilovsky, P.: Adaptive hypermedia. User Modelling and User-Adaptive Interaction, Ten Year Anniversary Issue (Alfred Kobsa, ed.) **11** (2001) 87–110

[2] Whittington, D., Hunt, H.: Approaches to the computerized assessment of free text responses. In: M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough, UK (1999)

[3] Burstein, J., Leacock, C., Swartz, R.: Automated evaluation of essay and short answers. In: M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough, UK (2001)

[4] Ming, Y., Mikhailov, A., Kuan, T.L.: Intelligent essay marking system. In: C. Cheers (ed.), Learners Together, Singapore, NgeeANN Polytechnic (2000)

[5] Laham, D.: Automated content assessment of text using Latent Semantic Analysis to simulate human cognition. Ph.D. Dissertation, University of Colorado, Boulder (2000)

[6] Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. Journal of Information Technology Education **2** (2003) 319–330

[7] Carro, R., Pulido, E., Rodríguez, P.: Dynamic generation of adaptive internet-based courses. Journal of Network and Computer Applications (1999) 249–257

[8] Carro, R., Ortigosa, A., Schlichter, J.: A rule-based formalism for describing collaborative adaptive courses. In: Knowledge-Based Intelligent Information and Engineering Systems. LNAI 2774, Springer-Verlag (2003) 252–259

[9] Pérez, D., Alfonseca, E., Rodríguez, P.: Application of the Bleu method for evaluating free-text answers in an e-learning environment. In: Proceedings of the Language Resources and Evaluation Conference (LREC-2004). (2004)

[10] Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation (2001)

[11] Pérez, D., Alfonseca, E., Rodríguez, P.: Upper bounds of the bleu algorithm applied to assessing student essays. In: Proceedings of the 30th International Association for Educational Assessment (IAEA) Conference. (2004)

[12] Alfonseca, E., Perez, D.: Automatic assessment of short questions with a bleu-inspired algorithm and a shallow semantic representation. Accepted in the 4th International Conference Spain for Natural Language Processing (ESTAL) (2004)