ON THE USE OF AN ON-LINE FREE-TEXT SCORING SYSTEM INDIVIDUALLY OR COLLABORATIVELY

Diana Pérez-Marín, Ismael Pascual-Nieto and Pilar Rodríguez
Computer Science Department, Universidad Autónoma de Madrid
Francisco Tomás y Valiente, 11, 28049, Madrid, Spain

Keywords: Computer Assisted Assessment, Collaborative learning, Natural Language interface, e-Learning.

Abstract: Willow is an adaptive web-based application that allows students to review course material. The system analyzes the students' free-text answers providing immediate feedback to the students. In the past, Willow has been used by individuals working alone. However, the trend of improving learning performance by allowing students to cooperate inspired us to develop a collaborative version of Willow. Our hypothesis was that students working together can reach understanding of ideas better than working individually with Willow. Therefore, in this paper, we explore the collaborative use of the system. We describe from a computer-science perspective, the minimum changes that have to be done to the system in order to permit a collaborative review. Furthermore, we provide the preliminary results of an experiment in which 22 students were given the possibility of using the individual or collaborative version of Willow.

1 INTRODUCTION

Automatic assessment of open-ended questions has been studied since the sixties (Page, 1966). In spite of the critics that the idea of automatically generating students' free-text answers has received (Hearst, 2000), the progress made in the Natural Language Processing (NLP) field, has made possible the uptake of computer-based free-text scoring.

The goal of many of the currently available free-text scoring systems is not to replace the teachers, but to complement them. Automatic free-text scorers can serve as double-checkers of the human scores, or to provide more training to the students before their exams (Valenti et al., 2003).

The core idea is usually to compare the student answer to a set of correct answers provided by the teachers, or other type of reference material (books, Internet, etc.). The more similar they are, the higher the score provided by the system is. Nevertheless, given that the goal is not to replace the teacher score, but to train the students to pass the final exam giving them more possibilities to review, it is usually only an orientative score.

In fact, typical feedback pages of free-text scoring systems include not only the numerical score, but also comments indicating the strong and weak aspects of the answer (according to the comparison performed by the system between the student answer and the teachers' correct answers).

Some free-text scoring systems that include these possibilities are: E-tester (Guetl et al., 2005), the extension of Didalect (Hermet & Szpakowicz, 2006), or SPEBC (Aguilar & Kaijiri, 2007). However, these systems have traditionally been designed to be only used individually.

Given the trend in e-learning systems to foster collaborative work, we think that it could also be interesting to have automatic collaborative free-text scoring systems.

In this paper, we present how the Willow free-text scoring system (Pérez-Marín et al., 2006), that we have developed, can be used not only individually but collaboratively. Our hypothesis is that students working together can reach understanding of ideas better than working individually with Willow.

To test this hypothesis, we performed an experiment with 22 volunteer non-technical students in a lab using the system, 12 working individually, and 10 in small groups.

We then analyzed the logs gathered to find out how the students differed in their use of the system. It has been noted that although in collaborative use the students are able to answer less questions, this is because they take longer to answer each question,

PÃ©rez-MarÃ■n D., Pascual-Nieto I. and RodrÃ■guez P.
ON THE USE OF AN ON-LINE FREE-TEXT SCORING SYSTEM INDIVIDUALLY OR COLLABORATIVELY.
DOI: 10.5220/0001820305010506
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

and we could observe that this extra time was spent discussing the concepts involved in the question, which is beneficial and, thus supports our initial hypothesis.

This paper is organized as follows: Section 2 focuses on the benefits of collaborative work; Section 3 describes the automatic and adaptive individual free-text scoring system; Section 4 details the procedure to transform the individual version of Willow to its collaborative version from a computer-science perspective; Section 5 reports the experiment performed and the results found; and, finally Section 6 provides the main conclusions of the paper together with some lines of future work.

2 BENEFITS OF COLLABORATIVE LEARNING

As defined by Gokhale (1995), collaborative learning refers to an instruction method in which students at various performance levels work together in small groups toward a common goal.

According to Vygotsky (1978), students are capable of performing at higher intellectual levels when they are asked to work collaboratively than when they work individually.

Moreover, students have declared that the computer environment facilitates collaboration, and a research study carried out with 48 technical university students have concluded that collaborative learning fosters the development of critical thinking through discussion, clarification of ideas, and evaluation of others' ideas (Cicognani, 2000).

There is also evidence that cooperative teams retain information longer than students who work individually (Johnson & Johnson, 1986); and, that the shared learning gives students an opportunity to engage in discussion, and take responsibility for the others and their own learning (Smith et al., 2005).

It can also be highlighted the benefits of using distributed student models (Puntambekar et al., 2003; Zapata-Rivera & Greer, 2001; Rueda et al., 2004), which have been extensively studied when the models are represented using concept maps, according to the principles of the Ausubel's Meaningful Learning Theory (Ausubel, 1963).

Some applications that are able to manage group student models represented using concept maps are: COMPASS (Puntambekar et al., 2003), ConceptLab (Zapata-Rivera & Greer, 2001), and DynMap+ (Rueda et al., 2004).

COMPASS (Puntambekar et al., 2003) is an Adaptive Educational Hypermedia System. It supports the assessment as well as the learning process. Each concept is displayed on a separate page, and the user is allowed to navigate between the concepts using textual hyperlinks.

COMPASS student models are very simple only based on their navigation behavior: which concepts they have visited and in what order. These student models can represent individuals or groups.

ConceptLab (Zapata-Rivera and Greer, 2001) is a knowledge construction and navigation system that uses XML-based concept maps to represent the student's view of the domain. It has three main goals: to assess the student's knowledge (it can be done by comparing different maps visually or through queries), to determine problems in the learning process of a student or a group of them and to promote reflection among a group of students in a topic. The student model is based on a bayesian network and a concept map.

According to the authors, the concept map has been included as part of the student model in order to facilitate sharing of knowledge among students and assessment of students' knowledge by teachers.

The concept maps are collaboratively built by the students who can be helped by a guide concept map. By clicking on a particular concept, it is possible to access a variety of links, added by the teacher or classmates, related to the concept of interest.

The knowledge built with ConceptLab can be represented with the VisMod system (Zapata-Rivera, 2004). VisMod allows students and teachers to experiment with the creation of Bayesian what-if scenarios; providing not only a visualization tool, but also an interactive tool for inspection of and reflection on Bayesian student models.

DynMap+ (Rueda et al., 2004) is a graphical tool to display the student model as a concept map. Students introduce the concept map in the computer using the Concept Map Editor provided. DynMap+ can show models not only of individuals but also of groups. Both are overlay models that can be shown to students and instructors.

The general purpose of showing the map to instructors is to provide them with a view of the knowledge and evolution of the students. The general purpose of showing the map to students is to foster reflective thinking about their own learning.

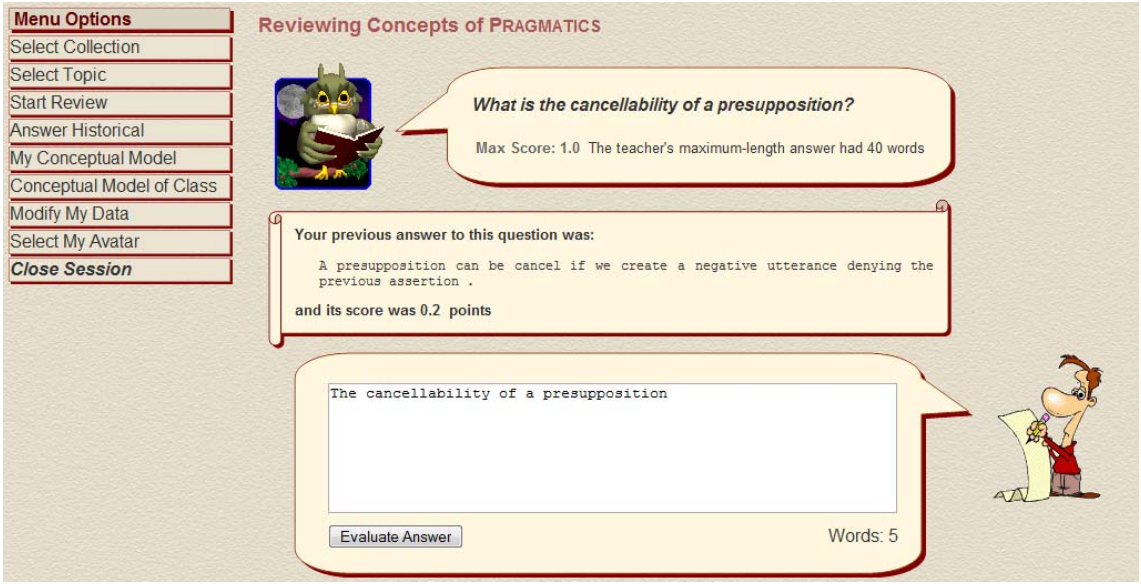

Figure 1: A snapshot of Willow's interface.

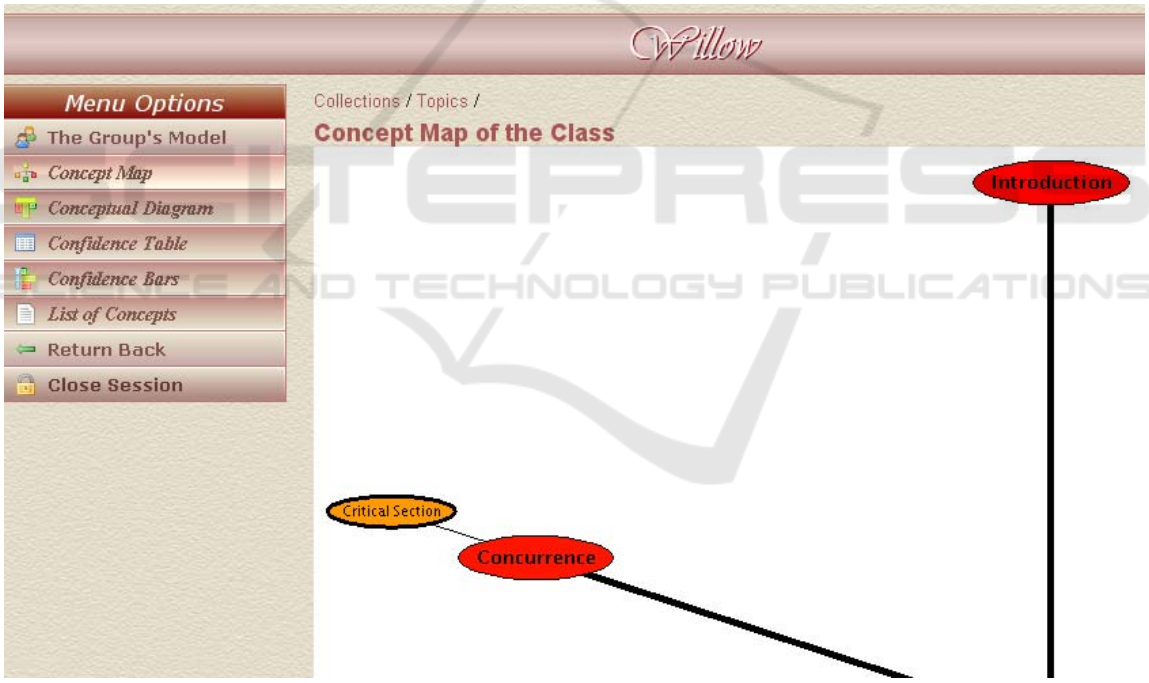

Figure 2: A snapshot of Willow's collaborative interface.

3 WILLOW

Willow is a free-text Adaptive Computer Assisted Assessment (ACAA) system (Pérez-Marín, 2007). That is, it is able to automatically and adaptively assess students' answers written both in Spanish and English languages. See Figure 1 for a snapshot of Willow's interface.

As can be seen, Willow's interface follows a dialogue metaphor in which Willow is represented by an owl, and the student can choose one avatar from several available to represent himself/herself.

The question asked by Willow is one of the set previously introduced by the teachers according to the student's level (as assessed by the system). For this reason, the teachers also need to provide a level of difficulty for each question: easy, medium or difficult.

Additionally, students are promoted or demoted a level based on their answers to a set of questions, and thus teachers need to specify the percentage of questions answered correctly or incorrectly which lead to the student being promoted or demoted a level of difficulty.

The system does not make assumptions as to the student's level of knowledge, and thus the first time a student logs into the system for a particular lesson, s/he is presented with questions of low difficulty.

When the student correctly answers the specified percentage of questions, s/he is promoted to a higher level of difficulty. Similarly, if the student fails a sufficient proportion of the questions, s/he is demoted to a lower level, and will thus receive easier questions.

By keeping questions at a level the student can handle, without being too easy, the level of engagement of the student with the system is maximized.

Whenever a student answers a question in Willow, s/he is presented with immediate feedback. If s/he passes the question, s/he is shown the feedback page, and the question is recorded as correctly answered. Otherwise, the system tries to help the student to pass the question with a set of clarification questions. If, even with this help, the student is not able to pass, the feedback page is displayed, and the question is marked to be asked later.

Furthermore, Willow keeps track in the students' answers of a set of concepts indicated by the teachers to automatically generate each student individual concept map, and the class concept map as shown in Figure 2. The procedure is described in detail in Pérez-Marín (2007).

4 FROM INDIVIDUAL TO COLLABORATIVE

The individual version of Willow asks each student to create an account the first time that s/he logs into the system. That way, Willow is able to keep track of how each student answers the questions, to automatically update each student model and to choose the most suitable question for him or her.

Therefore, the first change to allow the collaborative use of Willow (and, in general of any free-text scoring system) is to allow the creation of a group account. That is, an account that represents not only one student but a group of them. All members of the group can have the same user and password. Groups can be formed using:

- Self-selection: students themselves choose the members of the group. Thus, they have to introduce their names into the computer when registering within the same group account.
- Random assignment: the free-text scoring system randomly chooses the number of students indicated to create the group account. Students can be notified by an automatically generated mail of the names of their group mates.
- Criterion-based selection: a criterion to group the students is decided by the teacher and introduced into the free-text scoring system. That way, the system only chooses the students who meet the criterion. As before, students can be notified by mail of their group mates.

In Willow, we have implemented the self-selection option. It is because we did not want to randomly group the students as we consider too important the choice of group mates to leave it at random decision. On the other hand, we did not want to ask the teachers to think about a criterion to group the students, and they did not provide one voluntarily.

A minimum and maximum numbers of students per group should also be established. We have fixed the minimum as 2 to permit pairs, and the maximum as four to maximize the possibility that all students contribute to the discussion with their own ideas.

Moreover, students can be given the possibility of choosing to work simultaneously, or at different times. In any case, even if they work at different times, given that the system stores the questions already passed, and the previously given answers, students can read the answers of their mates, and look at the feedback generated for them.

When the students work together simultaneously, the interface should be distributed so that it permits all students to read together the question. As can be seen in Figure 1, only one question is presented each time (in the case of Willow it is not a change from the individual use of the system as the interface was already like that to focus all the students on the same question).

It is advisable to record in a log how long the students take to answer the question, until they ask for feedback to find out the time they have devoted

to discuss the most adequate answer to the question (it cannot be guaranteed, when the students are at home, that they have not devoted that time to any other task, but in general it is a good indicator).

It is expected that all of the students contribute with their answers. In Willow, we have not implemented any penalization factor in case that the teacher observes that one of the students has not participated, because it was the first time that the individual version was being transformed to a collaborative version. On the other hand, a role management procedure has been used in the free-text scoring system.

This procedure assigns a role to each member of the group so that all of the students have an specific work to do. For instance, one student can be the reader of the question provided by the system, a second student can be in charge of typing the agreed answer, a third student can read aloud the feedback generated by the system, and a fourth student can act as a moderator and coordinator of the group.

That way, none of the students has the same role than any other student. Provided that there are not enough students to create 4-student groups, then the same student can do more than one of the previously mentioned tasks. The roles should be changed each time that they log into the system. Similarly, if in the previous session, two roles were assigned to the same student, it should be implemented a restriction that prevents the same student to take so many roles in the next session.

5 THE EXPERIMENT

In the 2007-2008 academic year, we asked a group of 45 students of the English Studies degree at our university to voluntarily use Willow to review their Pragmatics course, in their own time.

The teachers of the subject introduced 4 lessons with 49 questions and an average of 3 correct answers per question. They encouraged their students to use the system as a means of reviewing the course. In total, 22 students (49%) agreed to take part in the experiment.

First of all, before they began using the system, the students were asked whether they prefer to review the course alone or in group. Of the 22 students, 12 (55%) opted for individual use while 10 worked in groups (2 groups of 3, and 2 of 2). To avoid influencing the manner of use of the system, only a brief (5 minute) introduction to Willow was given.

Although Willow is an on-line system, students were asked to work during one class (50 minutes) in the lab, because it was the first time that the collaborative possibility was used, and we wanted to directly observe the students.

In fact, observing the students using the system was very useful as we noted that students working collaboratively got more involved than students working individually.

Students who were in a group discussed how to answer each question and, upon receiving feedback, they discussed it among themselves.

Moreover, as Willow has a log system that registers how long each student or group of students have been working on each question, and how many questions have answered, we could analyze the logs gathered to contrast what we have perceived with the data registered by the computer.

Willow's logs confirmed our observations. Students working collaboratively answered fewer questions than students working alone. In fact, the average number of questions answered by individual students was 25, while the average number of questions answered by the group students was 8.

This is because the groups have indeed spent more time on each question. While a student working individually on average has taken around two minutes to answer a question and review the feedback, students working collaboratively have spent on average 6 minutes. In fact, in one case, a group spent 11 minutes just on one question.

6 CONCLUSIONS AND FUTURE WORK

This paper has contrasted collaborative and individual use of an automatic and adaptive free-text scoring system. Our goal was to discover whether students review better working individually or collaboratively with an on-line free-text scoring system.

We devised the list of changes necessary to implement the collaborative version of Willow. All of them were quite simple to apply: to permit the creation of group account, to establish a minimum and maximum number of students per group, and to check that only one question is shown at each time.

Previous experiments with the system have all involved individual use. In the present experiment, with students from a Pragmatics course at our university during the 2007-2008 academic year, both collaborative and individual use was allowed. 22

students participated in a session during one of their lessons in a computer lab. 12 students worked individually and 10 worked in groups of 2 or 3.

We observed that students working alone spent less time on each question than students working in groups and thus, they could complete more questions.

On the other hand, it was also noticed that the collaborative use of the system increased the level of reflection given to the construction of each answer, and also led to discussion of the generated feedback reports. These results support our initial hypothesis that collaborative use of Willow is beneficial.

As future work, we plan to repeat the experiment with more students in order to do a more complete evaluation. We would also like to analyze whether the collaborative versus individual use of Willow has an impact in the final exam scores.

Provided that the results achieved are satisfactory, our plan is to devise a new extension of Willow, which allows new communication possibilities between groups of students. For instance, an on-line chat or forum to talk about how to answer a certain question, or whether they agree with the automatic feedback provided by the system.

ACKNOWLEDGEMENTS

This work has been sponsored by Spanish Ministry of Science and Technology, project TIN2007-64718.

REFERENCES

Aguilar, G., Kaijiri, K., 2007. Design overview of an adaptive computer-based assessment system. *Interactive Educational Multimedia*, 116-130.

Ausubel, D. (1963), *The Psychology of meaningful verbal learning*, New York: Grune and Stratton.

Cicognani, A., 2000. Concept Mapping as a Collaborative Tool for Enhanced Online Learning, *Educational Technology & Society* 3(3), 150-158.

Gokhale, A., 1995. Collaborative learning enhances critical thinking, *Journal of Technology Education* 7(1).

Guetl, C., Dreher, H., Williams, R. 2005. E-tester: A computer-based tool for autogenerated question and answer assessment. *E-Learn*, AACE.

Hearst, M., 2000. The debate on automated essay grading. *IEEE Intelligent Systems* 5(15), 22-37.

Hermet, M., Szpakowicz, S., 2006. Symbolic assessment of free text answers in a second language tutoring system. *Proceedings of the 10th Computer Assisted Assessment conference*, Loughborough University.

Johnson, R., Johnson, D., 1986. Action research: Cooperative learning in the science classroom. *Science and Children* 24, 31-32.

McCafferty, S.; Jacobs, G. & Iddings, A. 2006, *Cooperative Learning and Second Language Teaching*, Cambridge University Press.

Page, E., 1966. The imminence of grading essays by computer. *Phi Delta Kappan* 47(1), 238-243.

Pérez-Marín, D., Alfonseca, E., Rodríguez, P. & Pascual-Nieto, I. (2006). Willow: Automatic and adaptive assessment of students free-text answers, in *Proceedings of the 22nd International Conference of the Spanish Society for the Natural Language Processing (SEPLN)*, 367-368.

Pérez-Marín, D. 2007. *Adaptive Computer Assisted Assessment of free-text students' answers: an approach to automatically generate students' conceptual models*, PhD thesis, Universidad Autónoma de Madrid, Spain.

Puntambekar, S.; Stylianou, A. & Hubscher, R. 2003. Improving Navigation and Learning in Hypertext Environments With Navigable Concept Maps, *Human-Computer Interaction* **18**(4), 395--428.

Smith, K.; Sheppard, S.; Johnson, D. & Johnson, R. (2005), 'Pedagogies of Engagement: Classroom-Based Practices', *Journal of Engineering Education* **94**(1), 87-101.

Rueda, U.; Larrañaga, M.; Elorriaga, J. & Arruarte, A. 2004, Validating DynMap as a mechanism to visualize the student's evolution through the learning process, *Lecture notes in computer science* **3320**, 864-866.

Valenti, S., Neri, F., Cucchiarelli, A., 2003. An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2.

Vygotsky, L., 1978. *Mind in society: The development of higher psychological processes*, Cambridge: Harvard University Press.

Zapata-Rivera, J. & Greer, J. 2001. Externalising Learner Modelling Representations, *Proceedings of Workshop on External Representations of AIED: Multiple Forms and Multiple Roles*, 71-76.

Zapata-Rivera, J. 2004. Interacting with Inspectable Bayesian Student Models, *International Journal of Artificial Intelligence in Education* **14**(2), 127-163.