Macro-Adaptation in Conversational Intelligent Tutoring Matters

Vasile Rus, Dan Stefanescu, William Baggett, Nobal Niraula, Don Franceschetti, Arthur C. Graesser

The University of Memphis, Memphis, TN, 38152, USA
vrus@memphis.edu

Abstract. We present in this paper the findings of a study on the role of macro-adaptation in conversational intelligent tutoring. Macro-adaptivity refers to a system's capability to select appropriate instructional tasks for the learner to work on. Micro-adaptivity refers to a system's capability to adapt its scaffolding while the learner is working on a particular task. We compared an intelligent tutoring system that offers both macro- and micro-adaptivity (fully-adaptive) with an intelligent tutoring system that offers only micro-adaptivity. Experimental data analysis revealed that learning gains were significantly higher for students randomly assigned to the fully-adaptive intelligent tutor condition compared to the micro-adaptive-only condition.

Keywords: macro-adaptation, intelligent tutoring systems, assessment

1 Introduction

We address in this paper the role of macro-adaptivity in ITSs. We study the role of macro-adaptivity in the context of conversational or dialogue-based ITSs (Rus et al.; 2013). These ITSs interact with the students primarily through conversation although other elements, such as images associated with instructional tasks, may accompany the dialogue. Our target domain is conceptual Newtonian Physics and our target population is college students taking an introductory course in Physics, (e.g. nursing, engineering students, or even Physics majors).

Currrent state-of-the-art ITSs are quite effective. An extensive review of tutoring research by VanLehn (2011) showed that the effectiveness of computer tutors (d = 0.78) is as high as the effectiveness of human tutors. Furthermore, it was found that the effectiveness of human tutoring is not as high as it was originally believed (effect size d = 2.0) but much lower (d = 0.79). Relevant questions arise from these findings. Where does the effectiveness come from and how can it be further increased? The conventional wisdom of the last decade or so has speculated that as interactivity of tutoring increases, the effectiveness of tutoring should keep increasing. However, VanLehn (2011) reported that as interactivity of tutoring increases, the effectiveness of human and computer tutors plateaus.

There are several aspects of state-of-the-art conversational ITSs that may explain their plateau in effectiveness. First, they do not emphasize macro-adaptation through selection of learner-specific content and tasks, which is needed when students begin a tutoring session with different backgrounds. Second, while tutorial strategies are somehow understood, that is not necessarily the case for tutorial tactics that control tutors' actions at micro-level, e.g. decisions about step in a solution to a problem (VanLehn, Jordan, & Litman, 2007). Third, existing conversational ITSs emphasize mostly cognitive aspects. Other aspects of learning, such as affect and motivation, are less considered. Researchers have started to address at least two of the above three aspects that could lead to further increases in ITSs effectiveness: tutorial tactics (VanLehn, Jordan, & Litman, 2007) and affect (Lehman et al., 2011). We investigate in this paper the role of the less studied aspect, i.e. macro-adaptivity. Therefore, our research complements existing efforts towards better effectiveness of ITSs.

It should be noted that the role of macro-adaptation was noted early on (Brusilovsky, 1992). Attempts to handle macro-adaptivity have been made but their exact impact on learning gains has not been pursued to the best of our knowledge. For instance, while the intelligent tutor ANDES (VanLehn et al., 2005) relies on a student model which could be used for macro-adaptation, it was never used for this purpose (Conati, Gertner, & VanLehn. 2002; VanLehn et al., 2005). In fact, there is one ITS that focuses exclusively on macro-adaptation. Indeed, the mathematics tutor ALEKS offers macro-adaptation only. Once a task has been selected for a learner, the learner sees an identical worked-out solution to the task as any other student that was assigned the same task. That is, within a task all students see same information following a one-size-fits-all approach (no micro-adaptivity). Interestingly, a recent study showed that ALEKS can offer significant learning gains comparable to other ITSs (Sabo, Atkinson, Barrus, Joseph, Perez, 2013). This result emphasizes the importance of macro-adaptation in intelligent tutoring.

Our work here offers further support for the important role of macro-adaptation in tutoring. In particular, we offer a glimpse at the important role of macro-adaptation in conversational ITSs. To achieve our goal, we compared a fully-adaptive conversational ITS that offers both macro- and micro-adaptivity, i.e. a fully-adaptive system, with a micro-adaptive-only ITS. In the fully-adaptive ITS, instructional tasks for a particular student were selected based on the knowledge level of the student. We defined four distinct knowledge levels based on a global analysis of the performance on the pre-test of our subject sample. Each individual student was then placed at a corresponding knowledge level based on his performance on the pre-test. The selection of instructional tasks for each knowledge level was based on the idea that tasks should target concepts that students in a knowledge level are just beginning to understand ("green shoots", i.e. concepts ready to emerge) while students at the immediately higher (and even higher) knowledge levels already show proficiency (to them, these look like full-grown concepts).

2 Data-driven Macro-Adaptation

The basis of our data-driven macro-adaptation is a multiple-choice test that participants were given prior to undergoing training. The pre-test consists of 24 multiple-choice questions from Force Concept Inventory (FCI; Hestenes, Wells, & Swackhamer, 1992), 8 multiple-choice questions from Alonzo and Steedle (2009; (A&S), and 7 multiple-choice questions of our own (total=39 questions). Students took the pre-test 2-3 weeks before the actual training in order to mitigate tiring effects during the actual training session and for logistical reasons. The training session consisted of about 1 hour of training with one of our ITSs, followed by 30 minutes of post-test taking (post-test was identical to the pre-test taken weeks before).

Once the student responses (*n=49*) on the pre-test were available, we selected critical concepts that students were struggling with based on Item Characteristic Functions (Wang and Bao, 2010) and defined knowledge levels based on this analysis. There is an Item Characteristic Function for each pre-test question which indicates the probability of answering the question correctly for various levels of student proficiency. In our case, instead of using directly student proficiency levels as given by, for instance, an Item Response Theory (IRT) analysis, we relied on the overall pre-test score. Due to the small *n*, an IRT analysis would have not been possible in our case. The use of the overall pre-test score as an approximation of proficiency level is reliable as explained next. Wang and Bao (2010) conducted an IRT analysis of FCI and confirmed the correctness of the unidimensional assumption needed for IRT analysis, i.e. a factor analysis revealed that existence of a dominant factor explaining college students' abilities to answer FCI questions. Furthermore, they showed a correlation of 0.994 between the overall FCI score (#correctly-answered/total-questions) and IRT proficiency levels.

In order to facilitate the selection of targeted concepts for training, we divided the space of proficiency levels into four knowledge levels: low knowledge, medium-low knowledge, high-medium knowledge, and high-knowledge. These knowledge levels offer a more fine distinction among students than the typical binary categorization (low vs. high knowledge) but less than the finest-grain categorization based on actual proficiency levels derived based on an IRT analysis (or its approximation through the overall pre-test score). Grouping the 39 proficiency levels into four groups (low, medium-low, medium-high, high) was regarded as a good compromise between cost (authoring effort) and performance (effectiveness). Using this method, the following four proficiency/knowledge levels were obtained based on the average pre-test score (13.95/39) and standard deviation (3.97): low knowledge (score≤10; n=7), medium-low knowledge (11≤score≤14; n=17), medium-high knowledge (15≤score≤18; n=14), and high knowledge (score≥19; n=11). For instance, students in the medium-low knowledge level had scores within one standard deviation below the average. Of the 49 students who were present for pre-test, 30 participated in training.

Once the knowledge levels were assigned, we proceeded with identifying the concepts that should be targeted during training for each level group. The basic idea was to use the pre-test as a source of identifying concepts that are "green-shoots" (ready to

emerge) for students at particular knowledge level. We have two criteria for identifying promising "green shoots" for a particular knowledge-level: students at that level begin to show some understanding (e.g., 10-30% of students at that level answer correctly questions related to a concept) and students at higher levels master it (e.g., >80% of the students show profficiency). Both criteria are important because there may be misleading "green shoots." Misleading "green shoots" are concepts that seem to emerge at one knowledge level (k; i.e., 10-30% of students answer correctly questions related to a concept) and are still in an emerging state (instead of becoming fully-grown concepts) for students at the higher-up level (k+1). We conclude that such green-shoots are not yet ready for "full-growth" for students at level k because students at the immediately higher level (k+1) are still struggling with such concepts.

Once we detected the ready-for-growth "green shoots" for a knowledge-level, appropriate instructional tasks were developed aiming at exposing students to the emerging concepts. There is one exception for the highest knowledge level for which there is no immediately higher level. That is, the second criterion of selecting concepts already mastered by students at the immediately higher knowledge level cannot be applied. In this case, we simply selected concepts with the highest learning potential.

3 Experiment and Results

As already mentioned, students attending a college-level conceptual Physics course were recruited for this experiment. This was an introductory course opened to all college students. The course provided the pre-requisite kind of training that seems to be important for experiments of the type we are describing here. Subjects were randomly assigned to one of the two training conditions: micro-adaptive-only vs. fully-adaptive.

Condition 1 (Micro-adaptive Only). In this condition students interacted with a dialogue-based ITS that used a fixed, predefined set of instructional tasks for all students. That is, there was a one-size fits all approach in terms of adapting instructional tasks to students. The set of predefined tasks included two tasks associated with each of the four knowledge levels defined for the other condition (uniform selection of tasks from all four knowledge levels) plus one additional task selected at random for a total of nine tasks (the number of tasks is the same in both conditions). Once working on a task (problem solving), students were scaffolded as needed through hints in the form of increasingly informative questions. That is, there was micro-adaptation.

Condition 2 (Fully-Adaptive: Macro- and micro-adaptive): In this condition students interacted with the fully adaptive system. The system would categorize students to different levels of understanding based on their pre-test score and then select appropriate tasks that were deemed most conducive of learning at that level of understanding. Tasks were selected for each knowledge level using the data-driven method presented earlier. A total of nine tasks were selected for each knowledge level. Once a task was selected for the students to work on, the micro-adaptation within the task was identical to the micro-adaptation in the micro-adaptive only condition.

The distribution of students into the four knowledge levels was: (Low=2, MediumLow=5, MediumHigh=5, High=2) for the Fully-Adaptive condition and (Low=5, MediumLow=3, MediumHigh=7, High=1) for the Micro-Adaptive condition.

Procedure. After signing a consent form, students took a pre-test under supervision. Students were all present in the same room and were given the pre-test at the same time (on paper). After they took the pre-test (39 multiple choice questions), students were given the opportunity to sign up for free tutoring sessions. Students who chose to participate were given extra credit in the course. Students participated in training sessions in a lab in small groups. Each student individually interacted with the tutoring system over the Internet from a personal computer. Each training session was about 1.5-hour long and consisted of approximately 1-hour of training (9 Physics problems) followed by a 0.5-hour for a post-test. There was a time span of about 3 weeks between the time students took the pre-test and the time they participated in training (and the post-test). Pre-test and post-test were identical.

Results. A number of 30 students participated in the training experiment with 16 of them in the micro-adaptive-only condition and 14 of them in the fully-adaptive condition. There was no significant difference in pre-test scores (percentage correct on the test) between the two conditions (t[28]=-.343, p=.734). A mixed ANOVA analysis was conducted with a pre-post-test within-subjects variable and the condition as a between-subjects variable. The ANOVA revealed a significant test*condition interaction ($F(1,28)=6.793$; $p=0.015$; see Figure 2). Adjusted post-test scores were compared between conditions by running an ANCOVA with the pre-test scores as covariate. A significant difference was found ($F(1,27)=11.974$; $p=.002$). A pre-post test comparison, revealed that the fully-adaptive condition had an effect size of (Cohen's) $d= 0.786$, $r=0.366$ (computed using means and pooled standard deviations). This is as good as human tutors. VanLehn (2011) reported an average human tutor effect of $d= 0.79$ (across many domains).

4 Conclusions

The positive results of our study in favor of macro-adaptivity indicate that improvements in this area hold the promise of increasing the effectiveness of tutoring systems beyond the interaction plateau if coupled with advanced tutorial tactics that boost micro-adaptation.

One weakness of our method stems from the IRT-style analysis based on which we defined our knowledge levels. A standard IRT analysis treats each wrong answer, i.e. distractor in a multiple-choice question, on equal footing. There is plenty of evidence that students of different proficiency levels react differently to different distractors (Dedic, Rosenfield, & Lasry, 2010). We will address this issue in order to further improve the level of macro-adaptivity by exploring recent advances proposed by the science education research community, e.g. learning progressions (Rus et al., 2013), and using polytomous IRT analysis.

Acknowledgments.

This research was supported by the Institute for Education Sciences (IES) under award R305A100875 to Dr. Vasile Rus. All opinions and findings presented here are solely the authors'.

References

1. Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. Science Education, 93, 389-421.
2. Brusilovsky, P. (1992). A Framework for Intelligent Knowledge Sequencing and Task Sequencing. In: Proceedings of Intelligent Tutoring Systems, (1992), pp. 499–506.
3. Dedic, H., Rosenfield, S. Lasry, N. (2010). Are All Wrong FCI Answers Equivalent?, Proceedings of the Physics Education Research Conference, Portland, Oregon, July 21-22, 2010.
4. Hestenes, D., Wells, M., and Swackhamer, G. (1992). "Force concept inventory," Phys. Teach. 30, 141–158 (1992).
5. Evens, M. and Michael, J. (2006). One-on-One Tutoring by Humans and Computers, Lawrence Erlbaum Associates, Inc., 2006.
6. Graesser, A. C.; VanLehn, K.; Rose, C. P.; Jordan, P.; and Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. AI Magazine, 22(4), 39–41.
7. Lehman, B., D'Mello, S., Chauncey, A., Gross, M., Dobbins, A., Wallace, P., Millis, K., & Graesser, A. C. (2011). Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In S. Bull & G. Biswas (Eds.), Proceedings of the 15th International Conference on Artificial Intelligence in Education (pp. 171-178). New York / Heidelberg: Springer.
8. Rus, V., D'Mello, S., Hu, X., & Graesser, A.C. (2013). Recent Advances in Conversational Intelligent Tutoring Systems, *AI Magazine*, 34(3):42-54.
9. Sabo, K.E., Atkinson, R.K., Barrus, A.L., Joseph, S.S., Perez, R.S. (2013). Searching for the two sigma advantage: Evaluating algebra intelligent tutors Computers in Human Behavior. 2013;29(4):1833-1840.
10. VanLehn, K.: The Behavior of Tutoring Systems. International Journal of Artificial Intelligence in Education: 16**,** 227-265. (2006)
11. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons Learned. International Journal of Artificial Intelligence and Education, 15 (3).
12. VanLehn, K., Jordan, P., and Litman, D. (2007). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed, Proceedings of SLaTE Workshop on Speech and Language Technology in Education (ISCA Tutorial and Research Workshop).
13. VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems, Educational Psychologist, 46:4, 197-221.
14. Wang, J. & Bao, L. (2010). "Analyzing Force Concept Inventory with Item Response Theory," Am. J. Phys., 78 (10), 1064-1070 (2010).