Let's talk! A tutorial dialogue system for FYHS

Jenny McDonald (PI), Rebecca Bird, Richard Zeng, Adon Moskal, Angela McLean, Ruth Napper, Ali Knott.

Project overview

In large undergraduate classes, it is time-consuming, costly and impractical for teachers to provide students with individualised feedback on written responses to questions. This project explored one option to address this through the use of a natural language dialogue system specifically developed for the cardiovascular section of HUBS192. An existing dialogue system which had evaluated well during an in-class experiment in 2012 was further enhanced with script revisions and multimedia capability. Free-text and menu-based versions of the system, without multimedia, were evaluated again through a combination of student feedback and an in-class randomised experiment using a pre-post-test design. This time, there was no significant difference in student score gain between student volunteers who used either version of the system and a control group. Student evaluations of the system provided useful insights for the project team. These results raise a number of issues and form the basis for ongoing work on natural language processing techniques and tutorial dialogue systems in in-class settings. The project has been presented at two seminars: one in HEDC and the other at the Centre for Learning and Research in Higher Education, University of Auckland. A journal article is currently in preparation.

Introduction – the context and rationale for the project as well as the project objectives

In large undergraduate classes, it is time-consuming, costly and seldom practical for the teacher to provide students with individualised feedback on their written responses to questions. Typically, computer-based marking of formative tests is used as an alternative and examples of this include Learning Management System (LMS) based multiple-choice quizzes or similar. However, this type of assessment does not provide an opportunity for students to receive feedback on their recall of key concepts (as opposed to recognition) or on their ability to describe how to apply these concepts in familiar or novel contexts.

Tutorial dialogue systems offer some promise for supporting and enhancing student understanding of key concepts in the large undergraduate classroom context (see for example, Evens & Michael, 2006). The appeal of tutorial dialogue is that formative questions are embedded in a tutorial plan: the questions arise in a meaningful context, concepts and ideas are linked together in a coherent form and immediate feedback is provided to the student on their understanding of taught concepts. An important side-effect of this type of dialogue is that a rich and structured natural language dataset is available to both teachers and researchers from which information about and relationships between student conceptions can be extracted. This can be used to inform ongoing course development and teaching practice.

This project utilises a natural language tutorial dialogue system developed by the Principal Investigator. The system has already been trialled extensively in the cardiovascular module of the FYHS paper, HUBS192 in 2012 and 2013.

Work to date demonstrates that input from students engaging with either a free-text or menu-based version of the tutorial dialogue system produces learning gains, on immediate post-test MCQs and short-answer questions, when compared with a control condition (McDonald, Knott, Stein & Zeng, 2013). However, to date there is no discernible difference between students in either of the tutorial conditions.

Project Objectives:

1. To further develop, and evaluate in-class, the natural language understanding component of the free-text version of the dialogue system.
2. To design, develop and evaluate in-class a new media integration component for either version of the dialogue system which utilises the University of Otago UniTube media sharing platform.
3. To develop and validate a series of short-answer questions and MCQs to assess how well students recall and apply specific concepts in both familiar and novel contexts. These questions and the student responses to them are used to evaluate system performance (refered to as pre and post-tests in the in-class described experiment below).

Approach and methods

The existing dialogue script was reviewed by teaching staff and project team members. Several changes were made to encourage deeper responses from students. For example, instead of simply seeking recall of facts, selected questions were designed to provide information and seek an explanation from the student related to the information given:

> *Three main factors which affect contractility of the heart are: ventricular filling, sympathetic stimulation and cardiac muscle mass. Please explain what is meant by contractility of the heart?*

A marking schedule for the questions contained in the dialogue script was created. The marking schedule comprised a set of categories or classes into which student responses could be divided. We already had a corpus of student responses for many questions from previous experiments and trials (2012 and 2013) and these responses were marked (classified)

by hand according to the marking schedule. The marked questions were then used to train the free-text classifier used by the natural language understanding component of the tutorial dialogue system. For newly developed questions, where we had no existing student responses, we updated the dialogue script to provide the student with a model answer and ask them to check their own answer against the model.

Dialogue system classifier results were compared with those obtained during classifier training conducted in 2012 using a standard machine learning training-set/test-set approach for training, followed by scoring using a held-out data set (e.g Manning and Schütze, 1999). Incremental improvements in performance for some questions were obtained through careful preprocessing and feature selection but accuracy results (i.e. the ability of the dialogue system to correctly classify a student response that it has not seen before) were not substantially different and in particular little improvement was obtained for more pedagogically interesting questions like the example given above. We have several ideas to improve feature selection and classification but this work is beyond the scope of this project.

Several questions were selected for media enhancement and an extension to the existing web-based interface for the tutorial dialogue system was created in order to incorporate these. Nevertheless it became clear even from bench-testing that to avoid interfering with the flow of the dialogue, significant script revision would be required. We therefore decided not to include media in the version of the dialogue system used in the in-class experiment. Nevertheless, this media capable version of the system works well from a functional perspective, is a useful addition to the system and with new script development is ready to be evaluated with students in future projects.

The cardiovascular system is covered early in the first few weeks of semester 2 in HUBS192. The tutorial was administered after cardiovascular system lectures and during the three week period when students work through their self-directed study material and labs before the terms test.

Student volunteers were randomly assigned to one of three groups:

1. **Control condition group** (no tutorial),
2. **Menu-based condition** where students respond to questions posed in the dialogue by selecting from a menu of options. They receive feedback according to the option selected and this also determines the next question posed.
3. **Free-text condition** where students respond to questions posed in the dialogue by typing their answers. They receive feedback according to the answer they typed and this also determines the next question posed.

All versions of the tutorial were made available to all HUBS students for revision immediately following the in-class experiment and prior to the terms test. Students in each condition sat a pre-test and immediate post-test. In the case of the control group, this was just a pre-test followed immediately by a post-test. Student performance for each condition on the pretest, immediate post-test and delayed post test was analysed to determine the impact of the tutorial and the different versions of the tutorial on student understanding.

A student perception questionnaire was administered to each student within a few days after they completion of the tutorial.

In the end, a full quantitative evaluation of the performance of the natural language understanding component of the free-text version of the dialogue system could not be conducted within the resources available for the project. Nevertheless, given minimal improvements in performance on bench-testing we did not anticipate any improvement in the classroom setting over 2012 or 2013. On the basis of spot checks and student evaluations (see below), this proved to be the case.

Key findings from the in-class experiment

362 student volunteers logged in and started the pre-test for the in-class experiment. 210 students completed the pretest and were randomly assigned to the control group (68), free-text group (71), or menu group (71). Of these students, 127 completed all questions including the immediate post-test. (control=57, dialogue=31, menu=39) . The number of students completing was disappointingly low compared to the 2012 in-class experiment where 531 students completed through to the end of the post-test (< 25%). A one-way ANOVA to test pre and post-test mean score differences between experimental conditions revealed no significant difference:

Free-text - Control	$p = 0.9468038$
Menu - Control	$p=0.5248661$
Menu - Free-text	$p=0.4164697$

This too was a disappointing result compared with the 2012 experiment where mean score gain did improve for both the free-text and menu conditions compared with the control group. The number of students who completed the delayed post-test (1 week later) was too low for inferential tests to be meaningful.

A total of 52 students completed evaluation questionnaires. Again, a low overall response rate (~14% of the total number of students who at least logged into the system). A number of students expressed frustration - their own lack of preparedness to respond to questions, the other pressures they were under, and the timing of the experiment in the course. A few had real frustrations with the tutorial itself including problems with the free-text version of the system failing to correctly classify student responses. Many comments related to not knowing how far through the tutorial they were and not being able to save and return later. Some students found the questions hard and there was one comment about being questioned on material not yet covered. Nevertheless, among those who at least engaged with all or part of the tutorial (free-text or menu-based), overall there was an acknowledgement that it was useful and most would engage with it for revision for the cardiovascular and other parts of HUBS if it were made available and included images, diagrams or animations. The newly developed questions where students were asked to check their answer again a model answer just seemed to confuse, particularly in the menu-based condition.

Discussion and implications

The project team had a number of discussions to try to work out why the response rate on this occasion was so low compared to 2012. HUBS192 teaching staff did note substantial differences in the student cohort for 2014 compared to previous years across a number of areas of the paper including performance in graded work. The fact that this is an in-class experiment is probably also part of the problem: to prove its usefulness it has to be run in a way that is useful for the students, which is to do with the timing in relation to the course and some of the comments they made in their evaluations. However, because it is an experiment with rigid parameters that aren't necessarily 'useful' for students, like timing, the pre and post-tests, inconsistent or confusing responses from the free-text version of the system, it undermines its own chance of success. If students don't engage, give up or put others off using it because of some of the experiments parameters, it is less likely to succeed or get highly positive feedback. The system is not actually being tested in the real way that students would ultimately engage with it. It is possible that simply making the dialogue system available as a resource for students to use as part of the course or for their own study or revision could result in far greater engagement. Certainly we have agreed that there is little point in running further in-class experiments using a pre-post-test design. We feel it will be far more productive to focus on improving the text classification and making a menu-based version available throughout the course which appropriately incorporates multimedia elements to reinforce and practice taught concepts. We hope to continue this work in 2015.

In terms of our original project objectives it is fair to say that this project has been successful only in part. Nonetheless, some valuable lessons have been learned and some interesting opportunities for further research and application of the dialogue system in the classroom have been identified.

Summary of spending

The full amount budgeted for this project of $10,176 was spent in 2014. Refer attached activity centre report. (NB: Some unrelated salary costs came through this account through a payroll error last year. This was corrected and is shown as the $523 credit at the top of the financial statement)

Other outputs

- Multimedia enhancement for dialogue system completed
- HEDC seminar given in September, 2014
- Invited presentation on the project given at the University of Auckland, Centre for Learning and Research in Higher Education, 27th June, 2014.
- An allied project designed to examine the nature of free-text responses captured during this project has been funded as one of a number of case-studies funded through Ako Aotearoa NPF-0008 in 2015/16
- A paper describing this project and results is currently in preparation and planned for submission to the Australasian Journal of Educational Technology.

References

Evens, M & Michael, J. (2006). *One-on-one Tutoring by Humans and Computers*. NJ: Lawrence Erlbaum Associates

McDonald, J., Knott, A., Stein, S. & Zeng, R. (2013). An empirically-based, tutorial dialogue system: design, implementation and evaluation in a first year health sciences course. In M. Gosper, J. Hedberg, H. Carter (Eds.) *Electric Dreams. Proceedings ascilite Sydney 2013.*

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.