An empirically-based, tutorial dialogue system: design, implementation and evaluation in a first year health sciences course.

Jenny McDonald
Higher Education Development Centre
University of Otago
Alistair Knott
Department of Computer Science
University of Otago
Sarah Stein
Higher Education Development Centre
University of Otago
Richard Zeng
Higher Education Development Centre
University of Otago

This paper presents one possible approach to providing individualised and immediate feedback to students' written responses to short-answer questions. The classroom context for this study is a large first-year undergraduate health sciences course. The motivation for our approach is explained through a brief history of intelligent tutoring systems, the philosophical and educational positions which inspired their development and the practical and epistemological issues which have largely prevented their uptake in a higher education context . The design and implementation of a new empirically-based tutorial dialogue system is described along with the results of in-class evaluation of the new system with 578 student volunteers.

Keywords: Tutorial Dialogue Systems, Natural Language Processing, Formative Feedback

Introduction

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. The coordinator of a large first-year health sciences class (1500-1800 students) approached the researchers for suggestions about the ways in which technologies might assist students to practice writing answers to short-answer, or constructed response, questions. Anecdotally, students typically performed poorly on these types of questions in the final exam relative to the multiple-choice questions. The course coordinator hypothesised that this was due to the lack of opportunity during the course for students to practice answering such questions: there were simply not enough teaching staff available to provide formative feedback on all the student responses. Intelligent tutoring systems (ITS) which employ natural language as their interface (tutorial dialogue systems) seemed to offer some promise for supporting and enhancing student understanding of key concepts in the current classroom context. The appeal of tutorial dialogue to both teachers and researchers was that formative questions are embedded in a tutorial plan: the questions arise in a meaningful context and concepts and ideas are linked together in a coherent form.

This paper begins with a brief history of intelligent tutoring systems, the philosophical and educational positions

which inspired their development and the practical and epistemological issues which have largely prevented their uptake in a higher education context . This leads to the motivation for the current research and a description of the design and implementation of a new empirically-based tutorial dialogue system. The results of in-class evaluation of the new system with 578 first year health sciences students are presented and the paper concludes with a discussion of these results and opportunities for ongoing research and development.

Intelligent tutoring systems: past and present

Jaime Carbonell's *Scholar* (Carbonell, 1970) is frequently cited as the earliest intelligent tutor (see for example, Woolf, 2008; Evens & Michael, 2006; Pea 2004; Shute & Psotka, 1994). *Scholar* produced individualised responses to typed student statements in a specific domain (for example, South American geography) using a semantic network. Carbonell's system parsed natural language input (that is, it could break down a sentence into its component parts and analyse the syntactic role of each part) using a system based on case grammar (Fillmore, 1968). *Scholar* then translated the parsed input into a logical form for processing by a semantic network and generated textual output from pre-written templates which matched specific logical outputs from the semantic network.

From a philosophical standpoint, *Scholar* can be thought of as the prototypical, rationalist inspired ITS. The prevailing approach adopted in the development of ITS up until quite recently has been without a doubt, the rationalist one. By the 1990s and through to the present day, the general architecture of ITS has been resolved to include at least some kind of knowledge base, which might include simulations or an expert system, an expert problem solver (these two together constituting a domain model), a student model, some kind of teaching model (Evens & Michael, 2006), authoring tools to allow teachers, and in particular those without specialist programming knowledge, to create the ITS in context (see Murray (1999)) and increasingly, dialogue modules to facilitate two-way natural language communication (for example, *Why2-Atlas* (VanLehn, et al., 2002), *Circsim-Tutor* (Evens & Michael, 2006))

Several issues arise from the rationalist approach to creating an ITS. First, ITS which utilise deep natural language processing techniques (NLP), involving the creation of domain-specific grammars, are typically very limited in their ability to handle language outside their domain of 'understanding' (Jurafsky & Martin, 2009). Second, the domain of 'knowledge' represented by the system needs to be mapped out and represented in some way. Third, some way of modeling student and tutor actions is usually required (Woolf, 2008). Finally, significant analytical effort is required in order to build an ITS even in a very restricted domain and even if authoring tools are available (Murray, 1999).

By contrast, the empiricist approach to building an ITS involves taking linguistic input from the tutee and looking up the most appropriate linguistic response based on empirical evidence about how to respond. Manning & Schutze (1999) characterise the empiricist 'camp' as privileging sensory input over mental organisation and contrast this to the rationalist position which emphasises innate mental structure over sensory input. Statistical or surface-based, NLP techniques are used to "understand" student input. Statistical NLP techniques are increasingly finding utility in practical applications where traditional NLP methods fail (Manning & Schütze, 1999) and are relatively straightforward to incorporate into new applications using standard NLP libraries. In the last 10-15 years surface-based dialogue systems or conversational agents have begun to appear and a few ITS do use surface-based techniques (for example, *Auto-Tutor* (Graesser et al., 2001) or a combination of surface and deep NLP (for example, *CarmelTC* (Rosé et al., 2003)) for natural language understanding. However, even these ITS still retain domain and student modeling. In a "pure" empiricist design, no calculations or assumptions should be made about either mental or machine state; there are no rules. If a particular linguistic pattern or feature-set has been seen before then the machine should respond on the basis of a known response to that pattern; if not, it makes no assumptions and simply says (or types), 'I don't know'.

The motivation for ITS

Even if rarely found in educational practice, ITS have persisted in the research domains of cognitive science and educational psychology. In looking for a reason why this might be the case, no reviewer exploring the ITS literature could fail to notice the impact of Bloom's 1984 study and what has become known as the 2 sigma problem. The Bloom (1984) study (1854 citations according to Google Scholar at June 24, 2013) claimed an effect size (ES) of 2.0 for human 'expert' tutoring and is regularly cited in the ITS literature and textbooks not only as the benchmark against which machine tutors should be compared but also as the reason why the provision of individualised support is a worthy goal. (For example, Woolf, 2008; Evens & Michael, 2006). However, similar studies to Bloom's demonstrate less impressive and highly variable effect sizes. Cohen, Kulik

& Kulik (1982) reviewed 52 studies, found an average effect size of 0.40 and noted that the size of the effect varied widely, the largest being 2.3. More recently, VanLehn (2011) found an average effect size of 0.79 when he reviewed 10 studies comparing human tutoring to no tutoring (ES ranged from -0.24 to 1.95) and an average effect size of 0.76 for step-based (ITS) tutoring compared to no tutoring (29 studies where ES ranged from -0.32 to 1.35). It is not that human tutors cannot be as effective as Bloom claimed; in a few documented instances they have been. Similarly, in some instances, ITS have demonstrated large effect sizes and a few are used, or have been used, in real class settings. However, on the basis of the evidence above it is clear that both human tutors and ITS vary widely in their effects and they do not consistently produce strong positive effects.

Given the wide variability in reported effect size, perhaps it is time to resist the rationalist urge to benchmark ITS against human tutors. The focus could usefully shift to delineating which tutoring or teaching practices or conditions produce the greatest learning effects. Indeed, there are already a number of researchers who are doing just this (see for example, Chi, 2009; VanLehn, Jordan, & Litman, 2007; Chi, Roy, & Hausmann, 2008). In a similar vein, Tamim, Bernard, Borokhovski, et al. (2011) have argued that a more nuanced approach would also be helpful in studies which look at the effect of computer aided instruction (CAI) rather than continuing comparisons between human and automated efforts.

The response of educators to ITS

A specific and important objection from educators, as well as from some educational psychologists and scholars working in the ITS domain, relates to the use of student models in ITS, where the steps taken by the student to solve a problem are compared to those used by an expert and the departure from expert steps or rules is modeled as errors. Laurillard (1988) presents a compelling case for abandoning models of student errors and argues that teaching should move beyond treating problem solving procedures 'as a set of uninterpreted rules'. Scardamalia et al. (1989) suggest that it is "not the computer that should be doing the diagnosing, the goal-setting and the planning, it is the student" (p.53). An entire volume contrasting the "modelers" and the "non-modelers" was published in 1993 (Lajoie & Derry, 1993).

Perhaps because of these doubts which were raised during the 1980s and early 1990s and perhaps because as previously noted ITS seldom find utility in educational practice, it is hard to find much reference to ITS in mainstream educational technology literature, including in the *ascilite* and *AJET* archives. As Reeves & Hedberg (2003) point out, "even the staunchest proponents … of ITS must acknowledge the lack of impact these computer-as-tutor applications have had on mainstream education and training"(p.6). However, if the focus is shifted to educational feedback, of the kind that human teachers and tutors provide, then searching the educational literature provides a good deal of information which is relevant to the design of ITS.

The positive benefits on student performance of formative assessment have been demonstrated in classroom studies since the 1920s (Frederiksen, 1984) and similar positive effects have been demonstrated in psychology laboratory studies since the 1970s (McDaniel et al., 2007). The large scale meta-analysis of studies which investigate the impact of practice tests on student outcomes indicate that on average, practice assessments during a course of study do confer an advantage (Bangert-Drowns, Kulik, Kulik & Morgan, 1991). More recently, a meta-analytic educational study to identify the key mediators of learning outcomes, found that feedback from student to teacher and from teacher to student, are among the top-ranked mediators (Hattie, 2009). In general, non-graded individualised feedback which avoids personal comment (including praise) and which highlights strategies for improvement, results in the largest gains (for example, Hattie & Timperley, 2007; Lipnevich & Smith, 2009; Shute, 2008).

The motivation and teaching context for this research

The first year health sciences course at the University of Otago is a prerequisite for entry into all the professional health science programmes, such as Medicine, Dentistry, Pharmacy, and Physiotherapy. Entry into these professional programmes is highly competitive and is dependent, amongst other things, on students achieving excellent grades in their 1st year courses.

The problem of providing individualised feedback to large numbers of students on free-text answers to formative questions was a key motivating factor for this research. Teaching staff involved in the course were keen to support the research and a bonus was that there was a very large cohort of highly motivated students potentially available each year to work with the system during design, implementation and evaluation. The specific domain selected for researching automated tutorial dialogue was the first year undergraduate study of the human cardiovascular system, in particular, cardiovascular homeostasis. There were two reasons for this

choice. Firstly the domain was the same as at least one other natural language tutor, *Circsim Tutor* (Evens & Michael, 2006), although pitched at a more introductory level. This was helpful in that it provided some confidence that the domain was suitable for automated tutorial dialogue. Secondly, it was a domain familiar to the lead researcher and thus obviated the need to find additional staff for authoring of the tutorial questions and script.

Given the large body of evidence for the beneficial effects of formative feedback, given the practical problems associated with current rationalist inspired ITS, the issues associated with student modeling, and finally, given the desire by educators for individualised, intensive and relevant learning environments, we felt it was worth adopting an empiricist approach to the design and implementation of a new system. The system emphasises utility in practice, no student model and categorisation of actual student responses and is described in the next section.

A new surface-based tutorial dialogue system

Overall design goals

The new tutor had to be responsive and practical in a real class setting. With this in mind, the broad design specification for the new surface-based tutorial dialogue system was as follows:

- The natural language understanding (NLU) component of the new system relies on empirical or statistical NLP techniques rather than deep semantic NLP techniques. This choice, in addition to sitting well with the empiricist philosophical position is also a pragmatic one; statistical NLP techniques which utilise machine learning are increasingly finding utility in practical applications where traditional NLP methods fail.
- The new tutor abandons the idea of explicit student models, pre-ordained teaching models and any formal or logical representation of the knowledge domain. But, it does retain the idea of unrestricted free-text input from the student. The family of dialogue systems or conversational agents closest to it, are those inspired by Weizenbaum's 'psychotherapist', ELIZA (Weizenbaum, 1966). These dialogue systems or conversational agents, which are not necessarily designed for tutoring, take typed natural language input, attempt to classify the input based on either regular expression matching or surface-based NLP techniques and generate typed output from a pre-defined script.
- Ideally, given the difficulty and expense of authoring, or customising ITS for specific contexts, a generic tutorial dialogue structure which is based on existing models of human dialogue should be designed into the new system in order that it can be readily extended or customised in the future.

Prototyping and data collection

The first stage of the project involved producing a detailed set of questions to probe student understanding of key elements of the tutorial domain and evaluating these questions, in the form of a scripted dialogue, with students. The TuTalk dialogue engine from the Learning Research and Development Centre at the University of Pittsburgh (Jordan, 2007) was chosen to pilot the initial script primarily because it was, at the time, one of the few readily available domain-independent tutorial dialogue systems and provided a relatively easy way to author dialogues using only a text editor.

Questions for the initial cardiovascular homeostasis tutorial script were developed in close consultation with course teaching staff and were written using lecture notes, laboratory manuals and self-directed learning material from the course itself. A prototype tutorial system based on the script, and which included 'guessed' student answers to match student responses against, was released to students for use on a voluntary basis at the beginning of their module on the human cardiovascular system. 437 students accessed the system during the course and produced a total of 532 dialogues; several students accessed the dialogue more than once. However from the total number of dialogues, only 242 dialogues were completed through to the half-way point and only 127 dialogues were completed to the end. A handful of dialogues were interrupted because of system-related problems but the majority that terminated before completion did so because because the students simply ended their session. Feedback from course tutors and comments from the students themselves supported researcher inutition that poor system understanding of student dialogue contributions was probably a key reason for the fall-off in use. This was confirmed when accuracy, precision and recall measures for individual questions were calculated: apart from a handful of essentially yes/no questions the majority of these metrics were zero. Nevertheless, the exercise served its purpose in capturing a large quantity of student responses to tutorial questions. These were to serve as training data for the next stage of development.

Creating the dialogue and building the new system

The next stage of the project involved creating categories of student responses from the responses collected during stage one in order to train statistical machine learning classifiers to recognise new student responses. In addition, the script developed in the first stage was refined in order to deal appropriately with the newly created categories of response. This process is described in detail elsewhere (McDonald, Knott & Zeng, 2012) but broadly parallels methods which are very familiar to educational researchers using qualitative research methods, (for example, phenomenography (Marton & Saljo, 1976) for identifying student conceptions or the methods of content analysis (Stemler, 2001)). This approach to creating categories or themes from student responses, is far less common in the realms of ITS development.

The overall architecture of the new system revolves around a dialogue manager which consults a hand-crafted script in order to direct the dialogue. The dialogue manager implements a simple finite-state architecture with a minimalist representation of information state (Traum & Larson, 2003). The script structure is loosely based on the Core & Allen (1997) dialogue coding scheme where each dialogue contribution node is divided into forward and backward functional layers. Each contribution node in the script contains forward and backward elements and each of these contain relevant dialogue acts or directions for action (for example, a request for information or a directive to go to a specified contribution node). The script is an XML file which is defined in the XML schema for the dialogue system and which comprises a series of dialogue contribution nodes. This design is based on a combination of practical and theoretical concerns. First, the finite-state approach is one of the simplest dialogue management models to implement and this was important in terms of developing the system in a timely manner and in such a way that it was robust enough to use with a large class of students. Second, the finite state approach is consistent with the directed nature of a tutorial dialogue where the dialogue purpose is to revise a specific and well-defined area of the curriculum. Third, the single-initiative question-asking approach provides an opportunity for students to practice tasks they are expected to complete as part of their formal summative assessment. (Note that limited mixed-initiative ability – i.e. the ability for students to ask questions – is built into the system but this was not evaluated as part of this research.)

The dialogue manager passes preprocessed text input from the student to the classifier module and receives back the result of classification which is either the name of the class chosen for the current dialogue context or a flag that classification failed. (Input text is preprocessed to isolate meaningful words or tokens and correct for spelling errors, identify abbreviations and so on). The dialogue manager then consults the script to determine what response to send back to the student. Both the preprocessor module and the classfier module utilise NLP libraries which are available in the Natural Language Toolkit (Bird et al. 2006). Specific implementation details are described elsewhere (McDonald, Knott & Zeng, 2012). The next section describes how the system works in practice, describes the types of questions included in the present dialogue script and briefly summarises the results of bench-testing the system for understanding of previously unseen input.

Dialogue walk-through, question-types and classifier performance

Students connect to the system via a simple chat-style web interface. The system types out statements and questions and the student responds by typing into a dialogue box. The system always waits for a student response before continuing. The following is an excerpt from a real dialogue session between the system and a student. The excerpt begins with an assertion from the tutor:

> TUTOR: Stroke Volume (SV) is the volume of blood ejected from the heart with each beat. The average adult heart pumps a total of 5 litres of blood each minute, and the average adult heart rate is 72 beats/minute.

The tutor then asks the student to perform an action which is directly related to this information:

> TUTOR: Please calculate the stroke volume for an average adult.

The student's initial response to this is:

> STUDENT: 5000/72

The student has not included the units for stroke volume which even though not explicitly requested, have been modelled by the tutor. The tutor recognises the student's response is numerically correct and responds with an acknowledgement of the student input followed by some scaffolding:

> TUTOR: OK. Looks like you have the right idea but you have either forgotten to give the units or used the incorrect units. Remember you can work out the correct units to use like this: You are given HR (beats/minute) and CO (l/minute). SV = CO/HR which is l/min divided by beats/min = l/beat. Finally don't forget to multiply the result by 1000 in order to convert l/beat to mls/beat.

Next, the tutor asks the student to try again and repeats the question:

TUTOR: Try again. Please calculate the stroke volume for an average adult.

This time the student responds with the same answer they produced the first time but this time in decimal form:

STUDENT: 69.4

Once more, the tutor recognises that the units are missing but this time responds slightly differently. The problem with the student's answer is again stated explicitly but detailed scaffolding is not included since the tutor also indicates that the problem is the same as the student's previous attempt:

TUTOR: Again, looks like you have the right idea but you have either forgotten to give the units or used the incorrect units.

And, once more the tutor repeats the question:

TUTOR: Try again. Please calculate the stroke volume for an average adult.

This time the student provides the correct answer with appropriate units:

STUDENT: 69 ml

The tutor provides positive feedback before moving on to the next dialogue context:

TUTOR: Well done.

This excerpt from a dialogue session illustrates three important features of tutorial dialogue. First, the student in this situation is not initially responding directly to the feedback given. It seems as though some other thinking is going on which results in the same answer being repeated, but just expressed differently. Feedback has to be noticed in order to be responded to. Second, if the dialogue is to be believable, the tutor has to respond appropriately where the same question is repeated. This finessing is one of the features of human dialogue which sets it apart from machine-speak. Humans typically will not just repeat the same utterance over and over, nor should the tutor. This feature is achieved through keeping track of the current dialogue state. Third, there are many ways that students might express the answer to this particular question – the tutor needs to be able to recognise most, if not all of them, in order to provide appropriate feedback. Another common error in this dialogue context for example, was problems with algebraic manipulation. The feedback if this were the error is different but the tutor action is the same: the student is asked to try again, and the question is repeated.

There are three broad categories of question-type in the tutorial dialogue script. These are binary, multi-part and open. In brief, a binary question requires exactly one response, usually just a word or two, and the response is either there or it is not. Yes/No questions are good examples of a binary type question. By contrast, an open question is one which requires some kind of development of ideas; for example making an inference, justifying a choice, applying a principle, or as in the example above, performing a calculation. It requires much more than a simple yes or no response or restatement of facts. Multi-part questions are those where several specific components or features are required in the response. For example, a question beginning with ‘List 3 variables . . . ’ is likely to be a multi-part question.

In all there are 29, what might be termed, top-level questions in the dialogue script. Requests to repeat questions and fall-back yes/no questions which are used where classification fails, are not included in this number. In laboratory tests which were conducted before the system was released to students, the accuracy on held-out unseen data for 26 top-level question classifiers ranged from 0.75 to 1.00 with the median value at 0.95. Of these classifiers, 13 were for binary-type questions and as might be predicted, the accuracy for recognising previously unseen responses to these tended to be towards the higher end of the range. Conversely, the accuracy of classifiers for open questions, such as, “What is the pulse?” or “Can you explain why you cannot feel a pulse in someone's vein?” tended to be at the lower end of the range (in this case, 0.75 and 0.85 respectively). The three remaining top-level questions were multi-part and the performance of classifiers for these is measured using a different metric (Measuring Agreement on Set-valued Items or MASI distance, where 0 indicates complete agreement between the class labels assigned to the reference and held-out data sets and 1 indicates no overlap. (Passoneau, 2006)). MASI distance for these classifiers ranged from 0.04 to 0.23.

While these “bench-test” results are promising and consistent with results using surface-based NLP techniques to recognise responses to short-answer questions (see for example, Butcher & Jordan, 2010) there is room for improvement. Nevertheless, the results were convincing enough to proceed with an in-class evaluation of the new tutorial dialogue system. Furthermore the design of the dialogue manager was conservative in that it always preferred an “I don't understand your answer” response and a fall-back yes/no question, where the confidence level of the classifier result was doubtful. The results of the in-class evaluation are described in the next section.

In-class evaluation

The goals of evaluating the tutorial dialogue system were twofold: first, to evaluate the system performance in terms of a) its ability to recognise and respond appropriately to student input, and b) the student experience of

using the tutor; second, to formally test a set of hypotheses involving student use of free-text and menu-based versions of the tutor. In order to test the hypotheses a menu-based version of the tutor was created. This was identical in every respect to the free-text version described in this paper except that instead of typing their responses to questions, students chose their preferred option from a menu. The menu reflected exactly the same classes of response that were available to the question classifiers which were used in the free-text version of the system. The hypotheses we tested were:

1. Either tutorial intervention, free-text or menu-based, results in better performance on a post-test than no intervention.
2. Free-text input results in better post-test performance overall than MCQ, because construction of a textual response from scratch requires first, recall of the relevant material and second, active processing of this material. Construction of responses should therefore promote retention and/or understanding better than simply selecting from pre-constructed options.
3. Free-text tutorials lead to increased performance particularly on short-answer questions because of a practise or testing effect.
4. MCQ tutorials lead to increased performance particularly on MCQ questions, also because of a practise or testing effect.

Background and experimental method

One of the researchers and the lecturer for the cardiovascular physiology section of the course introduced students to the experiment and the dialogue system during the last lecture on the cardiovascular system. A recording of the lecture was also available for students to access online from the following day. A web-page link for the tutorial dialogue system was also provided to all students via the course LMS. Prior to logging in to the tutorial dialogue system students could read the background to the research and experiment. Access to the tutorial dialogue system was taken as consent to participate in the study. Students could login any time during a three-week period that began immediately following the lecture in which the system and its evaluation was introduced. The three-week period coincided with the laboratory and self-study periods assigned to the cardiovascular system and ended on the day of a summative multiple-choice terms test designed to examine student understanding of the cardiovascular section of the course.

Evaluation criteria

Appropriate recognition and response to free-text student input was evaluated at the conclusion of the study. Preliminary results are reported in the next section for a small sample of four classifiers. Two human markers independently classified a sample of 100 student responses for each of these four classifiers. Marker classification was checked for inter-rater reliability and compared with system classification.

Student experience of the system was evaluated through a combination of student uptake and completion of the tutorial and administration of a student experience questionnaire which consisted of six 5-point Likert-scale questions, a section for free-text comments and one yes/no question. In addition, any unsolicited e-mail feedback from students was recorded.

In order to test the hypotheses an experimental study design was used. Student volunteers were randomly assigned to one of three conditions:

- A free-text condition where students complete a pre-test, then the free-text version of the tutorial dialogue, and conclude with an immediate post-test;
- a menu-based condition where students complete a pre-test, then the menu-based version of the tutorial dialogue, followed by an immediate post-test, or
- a control condition where they simply complete pre- and post-tests.

Performance in each condition was evaluated by:

1. Normalised score on an immediate post-test (conducted straight after the intervention or the pre-test for the control group) minus normalised score on pre-test. The immediate post-test comprised 7 MCQs and 7 short answer questions.
2. Normalised score on a delayed post-test comprising 3 MCQs, short-answer questions and a mini-essay question from the cardiovascular section of the final examination for the course.

Evaluation results

Overall, during the three week period in which it was available 720 students from a class of 1500 logged into the experimental system. Of these, 578 students completed the session through to the end of the immediate post-test. However, at completion of the study all student data recorded by the system as complete was checked. Following this process, 47 student sessions were removed from the analysis because of web browser or connection timeout issues. The final number of completions included in the analysis therefore was 531. The highest number of completions was 205 in the control condition, followed by 177 in the menu condition and 149 in the free-text condition.

While bench-testing of classifiers yielded promising results across all three types of classifier question (open, binary and multi-part), preliminary evaluation of four representative classifiers used in-class suggests a dramatic performance reduction for all but the binary question classifiers. Accuracy for 2 open question classifiers dropped from 0.8 and 0.9 on bench-testing to 0.61 and 0.65 in-class respectively. By contrast accuracy on a binary question classifier dropped from 1.0 on bench testing to 0.97 in-class. The multi-part question classifier also fared rather poorly increasing from a MASI distance of 0.23 on bench-testing to 0.68 in-class. There are several possible reasons for the drop in performance. An important one is that while the curriculum remained essentially the same between the collection of training data and its application in the tutorial dialogue system, any subtle change in emphasis from teaching staff could result in a drop in classifier performance. In fact, there was one change of lecturer during this time and this would have the potential to introduce new language, new expressions and new emphasis for the students. The reduced performance of the multi-part question classifier was likely due to limited training data (the more parts to a question, the more it is helpful to have larger data sets) and a potential class-imbalance problem (see for example, Japkowicz (2000)).

Nevertheless, in spite of the drop in performance of the free-text classifers, there seemed to be little subjective difference as far as the students were concerned between the free-text and menu-based conditions. A total of 105 responses to a student experience questionnaire were received (23% of the total number of students who logged in to either tutorial condition). Of these responses, 47 were from students who had been assigned the free-text tutorial and 58 were from students assigned to the menu-based condition. This response rate is consistent with large class evaluation response rates processed by the Evaluations Unit at the University of Otago but at the lower end of the range (typically 20-30%). The most striking feature of responses to the questionnaire was that 94% of all those who responded indicated that they would recommend the tutorial to other students. This feed-back is consistent with the 80% completion rate of those who participated in the evaluation, the 78% positive rating of the tutorial as an aid to learning and the 73% positive rating of the tutorial as a revision tool. There were a total of 38 free-text comments provided on the questionnaire. Eight responses related to reasons for non-completion. Three of these cited technical issues and two suggested either the tutorial was too long or that the student had insufficient time to devote to it. One student noted that they did not find the tutorial helpful and one felt that the tutor did not properly understand their answers. There were 30 general comments. These were predominantly complimentary and/or positive about the tutorial (19). Five found the tutor frustrating or felt their responses were poorly understood by the tutor. Other key themes from student suggestions and comments included: supporting media (e.g. video) would be helpful (2) Technical issues (2) More questions and/or more depth to questions (2) Tutor questions hard to understand (2) Tutorial patronising (1) Abbreviations not explained (1) Tutorial too long/lack of time (1). Unsolicited feedback was received from 6 students. Their comments were largely positive and provided useful validation of the feedback solicited via the student evaluation questionnaire.

The dependent variable to test the first hypothesis, which was that either tutorial condition should result in a performance gain over the control condition, was taken as the difference between pre- and post-test performance for each student with the pre-test result serving as a common baseline in each case. The differences between pre- and post-test scores were normally distributed. A between-subjects ANOVA gave an F value of 3.73 and a post-hoc Tukey multiple comparison of means at 95% confidence level showed a significant difference when compared with the control for the menu-based tutorial condition ($p=0.039$) but just outside significance for the free-text condition ($p=0.076$). On this basis, the first hypothesis is only supported for the menu-based condition. However, further investigation revealed that significant differences at the 5% level between both conditions and the control did exist up until 2 days before the end of the experimental period. At this point large numbers of students opted to take part in the evaluation and for this group there was no significant difference between tutorial and control conditions. In other words, scores in the control group increased on average as students studied towards the terms test and the effect of completing either tutorial when combined with intensive study confers no additional advantage. Linear regression analysis of scores in each condition confirmed this ($p < 0.05$ for both slope and intercept).

None of the remaining hypotheses was supported. There was no support for the second hypothesis that free-text input results in better post-test performance overall than menu-based input; comparison between the mean scores for free-text condition and menu-based condition was not significant ($p=0.987$). There was also no demonstrated benefit for free-text tutorials improving scores on free-text questions in the immediate post-test nor multiple-choice questions improving immediate post-test performance on the MCQs. Finally, when looking at the results of the delayed post-test, in this case, MCQ, short-answer and mini-essay questions from the final examination, approximately 3 months later, a between-subjects ANOVA gave an F value of 0.41. A post-hoc Tukey multiple comparison of means at 95% confidence level showed no significant difference when compared with the control for either the menu-based tutorial condition ($p=0.99$) or for the free-text condition ($p=0.66$).

Overall, the most striking result from the experiment was the lack of difference in student performance between the free-text and menu-based groups. This finding is consistent with studies where differences in performance between free-text and menu selection has been specifically examined (Corbett et al., 2006; Aleven et al., 2004). Either of these tutorials has a clear positive benefit on immediate post-test scores but this effect is, perhaps unsurprisingly, diluted by additional study as students work towards a summative terms test. There was no discernible effect on delayed post-test but given the relatively brief nature of the intervention, it would have been remarkable indeed to see an effect as highly motivated students prepared themselves for a critical examination!

A future for tutorial dialogue systems in contemporary educational settings?

The results reported here are the first from what we hope will be many in-class studies conducted using this tutorial dialogue system and others like it. Certainly a request has come this year from teaching staff for the cardiovascular homeostasis tutorial to be made available again. There are many issues to address and some of them have been touched on in this brief paper. Some of these are technical but the purpose of this paper is not to highlight these. There is room to improve the language recognition or text classification part of the system and perhaps this alone may result in greater learning gains from a free-text input system. Another key issue is setting the system up so that teaching staff themselves can create the dialogues. But perhaps most important is the opportunity to automatically classify student conceptions or understandings of aspects of the curriculum so that teachers can identify what these are and teach directly to them. It is interesting to note that this general idea is also gaining prominence through the emerging field of learning analytics.

In addition to the provision of practice and what is, on the basis of this study, an engaging tool for students, there is also the opportunity to give teachers practice at asking deeper and more difficult questions; writing questions and providing feedback to questions which encourage and support understanding rather than the simple repetition of facts. After the fact examination of our immediate post-test questions, which were prepared in consultation with teaching staff, revealed that they arguably only tested the surface recall of facts, even though some were open-ended questions. This is a well documented problem both globally (Frederiksen, 1984) and locally (Walker et al., 2010) and future work will need to address this.

In many large classes, as the teaching staff who approached us suggested, it can be all but impossible to find opportunities to provide feedback to short-answer questions, except perhaps during the final examination by which time it is too late for many. Through use of systems like ours, not only is there the opportunity to manage and evaluate large quantities of question and response data, it is presented and managed in a coherent form. The very nature of a tutorial dialogue does not lend itself to the mere presentation of isolated facts: ideas are presented in context and there is an internal coherence. We have demonstrated in this paper that there is both a benefit to students and an appreciation from students for the two versions of the tutorial dialogue system, free-text and menu-based. It remains to be seen whether there will be a discernible difference between the two.

References

Aleven, V., Ogan, A., Popescu, O., Torrey, C., & Koedinger, K. (2004). Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In J. C. Lester, R. M. Vicario, & F. Paraguaçu (Eds.), *Proceedings of seventh international conference on intelligent tutoring systems, ITS 2004* (pp. 443-454). Berlin: Springer Verlag.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Bird, S. (2006). NLTK: The natural language toolkit. In N. Calzolari, C. Cardie, & P. Isabelle (Eds.), *Proceedings of the COLING/ACL on interactive presentation sessions, COLING-ACL '06* (pp. 69–72). Stroudsburg, PA: Association for Computational Linguistics.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Butcher, P. G. & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, *55*(2), 489–499.

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. *Man-Machine Systems, IEEE Transactions*, *11*(4), 190–202.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*(2), 301–341.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237–248.

Corbett, A., Wagner, A., Lesgold, S., Ulrich, H., and Stevens, S. (2006). The impact on learning of generating vs. selecting descriptions in analyzing algebra example solutions. In S. Barab, K. Hay, & D. Hickey (Eds.), *Proceedings of the 7th international conference on learning sciences, ICLS '06* (pp. 99–105).

Core, M. G. & Allen, J. F. (1997). Coding Dialogs with the DAMSL Annotation Scheme. In D. Traum (Ed.), *Working notes of the AAAI fall symposium on communicative action in humans and machines* (pp. 28–35). Cambridge, MA: AAAI

Evens, M. & Michael, J. (2006). *One-on-one tutoring by humans and computers*. NJ: Lawrence Erlbaum.

Fillmore, C. J. (1968). The case for case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 1-90). New York: Holt, Rinehart and Winston.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202.

Graesser, A. C., Hu, X., Susarla, S., Harter, D., Person, N., Louwerse, M., Olde, B., et al. (2001). AutoTutor: An intelligent tutor and conversational tutoring scaffold. In *10th ICAI in Education*, 47–49.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 international conference on artificial intelligence (ICAI2000), Volume 1* (pp. 111–117), Las Vegas.

Jordan, P. (2007, July). Tools for authoring a dialogue agent that participates in learning studies. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Proceedings of the 13th international conference on artificial intelligence in education, (AIED 2007)* (pp. 43–50), Los Angeles.

Jurafsky, D. & Martin, J. (2009). *Speech and language processing*. New Jersey: Prentice Hall.

Lajoie, S. P & Derry, S. J. (1993). *Computers as cognitive tools.* New York: Routledge.

Laurillard, D. (1988). The pedagogical limitations of generative student models. *Instructional Science, 17*(3), 235–250.

Lipnevich, A. A. & Smith, J. K. (2009). I really need feedback to learn: Students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability, 21*(4), 347–367.

Manning, C. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marton, F. & Saljo, R. (1976). On qualitative differences in learning 1: Outcome and process. *British Journal of Educational Psychology*, *46*(1), 4–11.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494–513.

McDonald, J., Knott, A., & Zeng, R. (2012). Free-text input vs menu selection: Exploring the difference with a tutorial dialogue system. In P. Cook & S. Nowson (Eds.), *Proceedings of the Australasian language technology association workshop 2012, Volume 149* (pp. 97–105), Dunedin.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of state of the art. *International Journal of Artificial Intelligence in Education*, *10*, 98–129.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In N. Calzolari (Ed.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)* (pp. 831–836). Genoa.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences*, *13*(3), 423–451.

Reeves, T. C. & Hedberg, J. G. (2003). *Interactive learning systems evaluation*. Englewood Cliffs, New Jersey: Educational Technology Publications.
Rosé C. P., Roque, A., Bhembe, D., & Vanlehn, K. (2003). A hybrid text classification approach for analysis of student essays. *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing-Volume 2* (pp. 68–75). Association for Computational Linguistics.
Scardamalia, M., Bereiter, C., McLean, R. S., Swallow, J., & Woodruff, E. (1989). Computer-supported intentional learning environments. *Journal of Educational Computing Research*, *5*(1), 51–68.
Shute, V. & Psotka, J. (1994). *Intelligent tutoring systems: Past, present and future*. Technical report, Human Resources Directorate, Manpower and Personnel Research Division, Brooks Air Force Base.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.
Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation*, *17*(7), 137–146.
Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning. A second-order meta-analysis and validation study. *Review of Educational Research*, *81*(1), 4–28.
Traum, D. R. & Larsson, S. (2003). The information state approach to dialogue management. In J. Kuppevelt and S. R. W (Eds.), *Current and new directions in discourse and dialogue* (pp. 325–353). Dordrecht: Springer.
VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.
VanLehn, K., Jordan, P., & Litman, D. (2007, October). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In M. Eskenazi (Ed.), *Proceedings of SLaTE workshop on speech and language technology in education ISCA tutorial and research workshop* (pp. 17–20). Carnegie Mellon University and ISCA Archive: Farmington, PA.
VanLehn, K., Jordan, P. W., Rosé, C. P., Bhembe, D., Böttner, M., Gaydos, A., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the sixth international conference on intelligent tutoring systems*, *Berlin* (pp. 158–167). Berlin: Springer-Verlag.
Walker, R., Spronken-Smith, R., Bond, C., McDonald, F., Reynolds, J., and McMartin. (2010). The impact of curriculum change on health sciences first year students' approaches to learning. *Instructional Science*, *38*(6), 707–722.
Weizenbaum, J. (1966). ELIZA a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.
Woolf, B. P. (2008). *Building intelligent interactive tutors*. MA: Morgan Kaufman: Burlington.

Author contact details:
Jenny McDonald, jenny.mcdonald@otago.ac.nz

Please cite as: 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 H. Carter, M. Gosper and J. Hedberg (Eds.), *Electric Dreams. Proceedings ascilite 2013 Sydney*. (pp.562-572)