Recent Advances in Conversational Intelligent Tutoring Systems

Vasile Rus, Sidney D’Mello, Xiangen Hu, Arthur C. Graesser

Abstract

We report recent advances in Intelligent Tutoring Systems with conversational dialogue. We highlight progress in terms of macro- and micro-adaptivity. 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. The advances in macro- and micro-adaptivity that are presented here were made possible by the use of Learning Progressions, deeper dialogue and natural language processing techniques, and by the use of affect-enabled components. These improvements extend the bandwidth of possibilities for tailoring instruction to each individual student which is needed for maximizing engagement and ultimately learning.

Introduction

The vision of one-on-one human tutoring being the most effective solution to instruction and learning has attracted the attention of many for decades. Encouraged by the effectiveness of one-on-one human tutoring (Bloom, 1984), computer tutors that mimic human tutors have been successfully built with the hope that a computer tutor could be available to every child with access to a computer. An extensive review of tutoring research by VanLehn (2011) showed that computer tutors are as effective as human tutors. VanLehn reviewed studies published between 1975 and 2010 that compared the effectiveness of human tutoring, computer-based tutoring, and no tutoring. The conclusion was that the effectiveness of human tutoring is not as high as it was originally believed (effect size $d = 2.0$) but much lower ($d = 0.79$). The effectiveness of computer tutors ($d = 0.78$) was found to be as high as the effectiveness of human tutors. So, there is something about the one-on-one connection that is critical, whether the student communicates with humans or computers. Graesser et al. (1995) argued that the remedial part of tutorial interaction in which tutor and tutee collaboratively improve an initial answer to a problem is the primary advantage of tutoring over classroom instruction. Chi, Siler, and Jeong (2004) advanced a related hypothesis: tutoring enhances students’ capacity to reflect iteratively and actively on domain knowledge. Furthermore, one-on-one instruction has the advantage of engaging most students’ attention and interest as opposed to other forms of
instruction such as lecturing/monologue in which the student may or may not choose to pay attention (VanLehn et al., 2007).

Intelligent Tutoring Systems (ITSs) with conversational dialogue form a special category of educational technologies. These conversational ITSs are based on explanation-based constructivist theories of learning and the collaborative constructive activities that occur during human tutoring. They have proven to promote student learning gains up to an impressive effect of 1.64 sigma when compared to students learning the same content in a canned text remediation condition that focuses on the desired content (VanLehn, Graesser, et al., 2007). Of course, not all meta-analyses and research reports are so rosy, as indicated in the Dynarsky et al. (2007) report on commercial learning systems in K-12 environments and also some experimental conditions investigated by VanLehn and colleagues (2007). For instance, when intermediate students are given intermediate content and when novice students are given novice content, then tutoring is no better than canned text remediation. The true impact of conversational tutoring on learning is still not settled empirically.

The conventional wisdom of the last decade 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. VanLehn’s finding challenged ITSs developers to find new approaches to further increase computer tutors’ effectiveness. Indeed, novel approaches to ITSs development, such as the ones presented here, are needed to push their effectiveness beyond the interactivity plateau.

There are several aspects of state-of-the-art conversational ITSs that may explain the plateau in their effectiveness. First, they do not emphasize macro-adaptation through selection of learner-tailored content and tasks, which is needed when students begin tutoring sessions with different backgrounds. Second, conversational ITSs rely on dialogue and language processing algorithms to guide the interaction between the tutor and tutee. The quality of these algorithms have a direct impact on core ITSs tasks such as summative and diagnostic assessment, i.e. the detection and tracking of students’ knowledge states, and providing formative feedback. Assessment and feedback are critical components of any tutoring system that is fully adaptive. There is room for improvement when it comes to dialogue and language processing in conversational ITSs. Third, existing conversational ITSs emphasize mostly cognitive aspects when providing micro-adaptation. Other aspects of learning, such as affect and motivation, are typically not considered to guide micro-adaptation.

In this article, we present advances in conversational ITSs that address the above weaknesses and which have been proposed over the last decade, since the 2001 AI Magazine report by Graesser et al. (2001) on the state-of-the-art in conversational ITSs. We highlight (a) the addition of the Learning Progressions framework to increase macro-adaptivity (LPs; Corcoran, Mosher, & Rogat, 2009); (b) deeper dialogue and natural language processing to increase accuracy of student input assessment and quality of tutor feedback (Rus & Graesser, 2006; Rus & Lintean, 2012); this improves macro- and micro-adaptivity of ITSs; and (c) modeling students’ affective states in addition to their cognitive states (D’Mello & Graesser, 2012); affect-enabled components improves the micro-adaptivity capabilities of ITSs. It should be noted that LPs can have a significant impact on micro-adaptivity as well but we do not emphasize these aspects in this article. Furthermore, affect-enabled components may impact macro-adaptivity as well but we do not emphasize these aspects either.

Taken together, these advances presented in this article form one promising approach to increasing the effectiveness of tutoring systems beyond the interaction plateau (VanLehn et al., 2007; VanLehn, 2011).

**Intelligent Tutoring Systems with Animated Conversational Agents**

Conversational ITSs have several advantages over other types of ITSs. They encourage deep learning as students are required to explain their reasoning and reflect on their basic approach to solving a problem. Conceptual reasoning is more challenging and beneficial than mechanical application of mathematical formulas. Furthermore, conversational ITSs have the potential of giving students the opportunity to learn the language of scientists, an important goal in science literacy. A student associated with a more shallow
understanding of a science topic uses more informal language as opposed to more scientific accounts (Mohan, Chen, & Anderson, 2009).

The impact of conversational ITSs allegedly can be augmented by the use of animated conversational agents that have become more popular in contemporary advanced learning environments (Graesser, Rus, Jackson, D’Mello, 2008). The animated agents interact with students and help them learn by either modeling good pedagogy or by holding a conversation with the learners. Both single agents and ensembles of agents can be carefully choreographed to mimic virtually any activity or social situation: curiosity, inquiry learning, negotiation, interrogation, arguments, empathetic support, helping, and so on. Agents not only enact these strategies, individually or in groups, but can also think aloud while they do so.

Examples of successful conversational ITSs are AutoTutor (Graesser, Rus, Jackson, D’Mello, 2008), Why2 (VanLehn et al., 2007), CIRCSIM-Tutor (Evans & Michael, 2005), and GuruTutor (Olney et al., 2012). DeepTutor (www.deeptutor.org), described here, is an emerging conversational ITS.

We will focus next on AutoTutor, a conversational ITS described in the 2001 AI Magazine report by Graesser and colleagues (2001), and on VanLehn’s (2006) two-loop framework for describing ITSs. The advances in conversational ITSs that we emphasize in this article can be best understood with respect to AutoTutor’s basic dialogue and pedagogical framework and VanLehn’s (2006) two-loop framework for describing ITSs.

**AutoTutor Overview**

AutoTutor is a conversational ITS that helps students learn science topics by holding a mixed-initiative dialogue with students in natural language (Graesser, Rus, Jackson, D’Mello, 2008). The structure of the dialogue in both AutoTutor and human tutoring follows an expectation and misconception tailored (EMT) dialogue. EMT dialogue is the primary pedagogical method of scaffolding good student answers. The dialogue moves and the problems AutoTutor can tutor on are stored in a curriculum script. The curriculum script is a knowledge structure employed by novice tutors that largely determines the content and flow of a tutoring session. AutoTutor promotes active construction of knowledge by providing explanations only when the learner is floundering. Moves at the beginning of the tutorial interaction for each new problem (i.e., pumps and hints) provide less information to the student than later moves (i.e., prompts and assertions).

The behavior of AutoTutor and that of any ITS, conversational or not, can be described using VanLehn (2006)’s two-loop framework. According to VanLehn, ITSs can be described in broad terms as running two loops: the outer loop, which selects the next task to work on, and the inner loop, which manages the student-system interaction while the student works on a particular task. AutoTutor’s outer loop is usually just an iteration over the set of existing tasks (i.e., main questions or problems; Graesser et al., 2003). That is, all students go through the same set of tasks in the same order. Some level of adaption does exist as the set of tasks can be chosen based on known student conceptions and misconceptions. However, once the tasks are designed all students see the same tasks regardless of their individual differences. It should be noted that sometimes this one-size-fits-all approach of selecting training tasks is required by experimental design constraints which dictate that all students must be exposed to same content. Ideally, students should work on tasks that best suit their background. The use of LPs in the emerging conversational ITS DeepTutor will enables achieving this goal.

The inner loop of an ITS monitors students’ performance through embedded assessment, updates its model of students’ levels of understanding (i.e., the student model), and uses the updated student model to provide appropriate scaffolding in the form of feedback and other scaffolds. In dialog-based ITSs, embedded assessment relies heavily on language understanding algorithms as students’ responses are natural language utterances. AutoTutor’s language processing algorithms rely on co-occurrence methods such as Latent Semantic Analysis (LSA; Landauer et al., 2007), inverse weighted word frequency overlap, and regular expressions. Another important aspect of the inner loop is to provide appropriate scaffolding while the student is working on a task, e.g. correct a misconception immediately through appropriate feedback. We discuss in this paper advances about how to better assess students’ natural language inputs
and how to better tailor scaffolding using affect-sensitive components while the student is working on a task.

The three major advances in conversational ITSs described in this article will have direct impact on both the outer and inner loops of ITSs and will lead to improvements in core tasks handled by ITSs: modeling the task domain, tracking students’ knowledge states, selecting appropriate learning trajectories, and the feedback mechanisms. Advances in these core tutoring tasks will move state-of-the-art ITSs closer to implementing fully adaptive tutoring which implies tailoring instruction to each individual student at both macro- and micro-level.

**Advanced Macro-Adaptivity in DeepTutor**

DeepTutor is a conversational ITS that is intended to increase the effectiveness of conversational ITSs beyond the interactivity plateau by promoting deep learning of complex science topics through a combination of advanced domain modeling methods, deep language and discourse processing algorithms, and advanced tutorial strategies. DeepTutor has been developed as a web service and a first prototype is fully accessible through a browser from any Internet-connected device, including regular desktop computers and mobile devices such as tablets. A snapshot of the learner view is shown in Figure 1. DeepTutor currently targets the domain of conceptual Newtonian Physics but it is designed with scalability in mind (cross-topic, cross-domain). The spin-off project of AuthorTutor (www.authortutor.org) aims at efficiently porting DeepTutor-like ITSs to new domains by investigating well-defined principles and processes as well as developing software tools that would enable experts to efficiently author conversational computer tutors across STEM disciplines. Another authoring tool, called SEMILAR (derived from SEMantic simILARity toolkit; Rus et al., 2013), is being developed as well to assist with authoring algorithms for deep natural language processing of student input in conversational ITSs. More information about the SEMILAR toolkit is available at www.semanticsimilarity.org.

Figure 1. Snapshot of DeepTutor showing a Physics problem on the top right pane and Dialogue history on the left pane. The Multimedia box synchronizes with the dialogue showing identified information, e.g. velocity and force vectors, visually.
It is beyond the scope of this article to describe all the novel aspects of DeepTutor or related projects. Instead, we focus on two components of DeepTutor: domain modeling based on LPs and deeper dialogue and natural language processing.

Learning Progressions

Learning Progressions (LPs) have been developed by the science education research community as a way forward in science education. The NRC (2001) report called for better descriptions of how students learn based on models of cognition and learning. Based on such descriptions of how students learn, “assessments can be designed to identify current student thinking, likely antecedent understandings, and next steps to move the student toward more sophisticated understandings” (NRC, 2001, p. 182). This was basically a call for developing Learning Progressions (Corcoran, Mosher, Rogat, 2009).

LPs adopt a learner-centric view of a topic by modeling students’ successful paths towards mastery as opposed to paths prescribed by domain experts following a logical decomposition of the big ideas of a domain. The logical decomposition provided by experts could be useful as a starting point that needs be re-organized based on evidence of how students actually develop mastery of the big ideas. These actual, developmentally-proven paths must be documented and guide instruction.

Learning Progressions in DeepTutor. LPs are the central theme in DeepTutor around which everything else (domain modeling, assessment, and instructional tasks) is organized and aligned. This centrality of the LP can be seen in Figure 2. Inside the circle in the middle of the figure, we show a snapshot of the LP. The shown LP is a partial view, for illustration purposes, of our Newtonian Physics LP.

The LP in DeepTutor is organized in a set of strands along the horizontal axis with strands more to the right signifying more sophisticated topics. For instance, the Circular Motion strand (rightmost column in the LP) is more complex compared to the Mass-and-Motion strand (third column). Along the vertical axis, each strand is organized more like a traditional LP (Alonzo & Steedle, 2009), in levels of sophistication. Each level corresponds to a set of coherent ideas or models that students use to reason about the domain. The higher the level in the LP, the stronger the model, i.e. the model explains more phenomena of the domain. We call our LP a two-dimensional or 2-D LP due to its organization in two dimensions. A typical LP, e.g. Alonzo & Steedle’s, is unidimensional (single-strand) showing only levels of sophistication for a major theme. The hierarchal structure of the LP levels within a strand can be accomplished by following several methodologies, such as how close a model is to the best model (the mastery model is the top level in an LP, also called the Upper Anchor), item difficulty, and also based on developmental and cognitive considerations. It should be noted that LP developers acknowledge that there is no unique, perfect progression or hierarchal structure of ideas but rather a family of such progressions. The goal is to document the alternatives and the most frequent progressions or levels of understanding that students experience in their journey towards the Upper Anchor.

Our 2-D LP is the most comprehensive and fine-grained LP compared to any other existing Physics LP. The actual LP has seven strands and up to nine levels for certain strands. It should be noted that a complex and finely-grained LP is needed to drive the operations of an effective educational technology such as DeepTutor.

There is an interesting interplay among assessment, LPs, instructional tasks, and advanced tutoring strategies that is finely orchestrated by DeepTutor. The LPs are aligned with the initial assessment instrument (i.e., pretest - shown on the left side of Figure 2) which students must complete before they interact with the system. Based on this first summative assessment, an initial map of students’ knowledge levels across all strands in the LP is generated. This corresponds to the solid, wavy line (shown in red, if colors available) across the middle layers of the LP in Figure 2. Basically, we get a first impression where students are with respect to the Upper Anchor of each LP strand. Based on recent research by others (Bao & Redish, 2006; Alonzo & Steedle, 2009) and our own experience, this first assessment of a student’s knowledge state is just an approximation. In fact, the two wavy, dotted lines below and above the thick line indicates a range of levels that a student might be at. We are moving towards a probabilistic model (called a cloud model, inspired from the cloud model of the electron in modern physics) of assessing students’ levels of understanding in which we can only assert with a certain probability at which level students are. The cloud model assumes that students can be at multiple levels in the LP, i.e. have multiple models.
simultaneously active in their minds, some stronger than others, which they activate depending on various factors. For instance, Bao and Redish (2006) studied different student models for Newton’s Third Law and identified the probability with which a student activates a particular model based on three features of instructional tasks that target Newton’s Third Law.

Our cloud model for assessing students’ knowledge states is work in progress as of this writing. A glimpse of it can be seen in Figure 3 where we show three strands of our LP and an actual student model indicating his performance with respect to the LP levels. The leftmost column of the indicates the level in the LP. Each of the other columns represents an LP-strand. Each strand is further divided into three inner-columns: left – number of potential answer choices in the pre-test that map to the corresponding LP level, middle – number of correct student-chosen answer choices, right – number of incorrect student-chosen answer choices. Because each multiple-choice question in the pre-test has 5 choices of which only one is correct and because the answer choices of a question can map to different levels and strands in the LP the relationship among the numbers shown in Figure 2 is more complex. A simple pattern does exist: the number of (correct and incorrect) student answer choices should add up to the number of questions in the pre-test. From the figure, we can notice that this particular student has two relatively persistent models for the Force and Motion topic. These two models correspond to the LP levels most chosen by the student: level 1 (4 incorrect answer choices picked by the student) and level 2 (6 student choices of which 5 are incorrect and 1 correct). These are weak models because they are in the mid-lower part of the Force and Motion strand. Based on the placement of the student in the LP, instructional tasks that are appropriate for levels 1 and 2 in the Force and Motion LP strand are selected for this particular student.

Tailored Learning Trajectories. The set of instructional tasks that are selected at the outer loop should, ideally, be tailored to each individual learner because learners start interacting with an ITS system at different levels of understanding and also they learn at different paces. We call the set of instructional tasks a student is assigned a learning trajectory. It should be noted that there is a distinction between the hypothetical learning trajectory, which is first computed based on a student’s pretest performance, and the actual learning trajectory, which is a dynamically updated version of the hypothetical learning trajectory based on students’ actual performance while working on tasks (Battista, 2011).
In DeepTutor, we select the hypothetical learning trajectory based on the alignment between the pre-test and embedded/formative assessment, on one hand, and the LPs, on the other hand, and between the LPs and instructional tasks. We have tasks corresponding to each cell in our 2-D LP. In fact, for each cell we have a set of tasks that enable us to implement advanced tutoring strategies to address some of the illusions of tutoring (Graesser, D’Mello, and Person, 2009). For instance, consider the “illusion of mastery”, the unwarranted assumption that the student has mastered much more than the student has really mastered. Indeed, one provocative finding in the tutoring literature is that there is sometimes a positive correlation between a student’s knowledge of the material (based on pretest or posttest scores) and their likelihood of saying no rather than yes to the tutors’ comprehension-gauging questions (Graesser et al., 1995). Thus, it is the knowledgeable tutees who tend to say “No, I don’t understand.” We detect the illusion of mastery in DeepTutor and activate a tutoring strategy to penetrate the potential illusion. For instance, one way to detect it is to identify students who keep saying ‘I understand’ when asked if they understood a previously discussed idea or concept but actually have a low pre-test score. The tutoring strategy triggered for such students consists of challenging the students to solve a similar problem in order to demonstrate that he or she truly understood the concept. This is why each cell in the 2D LP has a set of corresponding instructional tasks (see the layered tables on the right side of Figure 2). Each table indicates a set of LP-aligned tasks. Tasks in corresponding cells in the different tables are equivalent. Sets of equivalent tasks for same cell in the LP are needed such that DeepTutor can trigger as many tasks at a particular level of understanding as needed.

There are many possible learning trajectories that can be chosen for a particular student. For instance, drilling tasks which offer training on one big idea or theme covered in one LP strand will more likely help students move up the level of understanding within that strand while not making progress on ideas in other strands. For some strands, e.g. Newton’s 3rd law, for which the correct answer to many problems is the same (the tasks are isomorphic to some degree), such a drilling training strategy is less effective. Students may learn the jingle (“the action and reaction forces are equal and opposite”) after seeing the solution to few problems and just recite the jingle when prompted to solve subsequent tasks, without actually developing a deep understanding of Newton’s 3rd law. Smarter sequencing of problems must be adopted as isomorphic problems lead to copying and therefore shallow learning (VanLehn, 2011; Renkl, 2002).

DeepTutor is still under development as of this writing, but components of the system have been validated. For instance, we have validated our Hypothetical Learning Progressions (HLP) designed by our experts based on data collected from students in six different high-school classes that form a developmental progression: Physics, Honors Physics, IB Physics 1, IB Physics 2, AP Physics C 1, and AP Physics C-2. The data was collected at one time across all these classes. That is, different students were accounted for in different classes. Ideally, a longitudinal study is needed to observe how a group of students’ taking these classes progress over time. The adopted procedure is the best approximation within the constraints of collecting data in one semester.

Deeper Dialogue and Natural Language Processing

In this section, we present DeepTutor’s advanced dialogue and natural language processing algorithms. We distinguish between dialogue processing components that handle aspects of managing the dialogue and components that deeply understand students’ contributions as a way to assess their level of understanding.
Dialogue Processing. The dialogue manager in DeepTutor implements conversational goals used in existing dialogue-based tutoring systems, such as coaching students to articulate expectations, correcting students’ misconceptions, and attempting to answer student’s questions when they are sufficiently inquisitive to ask questions. However, DeepTutor has additional conversational goals that attempt to achieve accurate grounding at each turn (i.e., the system and tutor perfectly understand each other at every turn and over many turns), accurate feedback on students’ contributions, error-handling (for cases when the system cannot accurately interpret what the student is saying), naturalistic dialogue, and optimized knowledge transfer. To achieve the new goals, we added components that explicitly handle dialogue moves associated with these goals to the core dialogue management module. For instance, we have a component to detect the need for establishing common ground and another to initiate and handle the dialogue moves necessary to establish common ground. Consider the scenario in which the tutoring system presents a rare or unseen word X to a student and the student replies with “What is X?” That is an indication of a request for grounding. By the same token, we can imagine a student replying “Yes” to a comprehension-gauging question (“Do you understand?”) asked by the tutoring system. The system would then skeptically challenge the student with a verification statement (such as giving the student an isomorphic problem) to double-check the student’s understanding. If the student stumbles, it is an indication of knowledge transfer failure and thus the system must activate a component to optimize the transfer of knowledge.

Speech Act Classification. An important component of the dialogue manager is the identification of students’ intentions based on their utterances, i.e. the task of speech act classification (SAC; Rus et al., 2012b). The SAC uses a multi-leveled taxonomy of speech act categories with 4 major categories at the top (meta-cognitive, meta-communicative, question, and contribution) and 35 categories total at the second and third level. In case a student utterance is labeled as being a contribution, which is a content-rich statement, it is passed on to the semantic processing component that is described in the next section. Contributions from students can be relevant or irrelevant (e.g., a content-rich statement that talks about food or friends when the topic is Physics). Only relevant contributions are passed on for further semantic analysis.

Advanced Algorithms for Understanding Student’ Natural Language Essays and Contributions. Algorithms are needed to interpret the meaning of students’ natural language contributions at each turn in the dialogue as well as assessing the more comprehensive essay-type answers that students are required to provide immediately after being prompted to solve a problem. This section describes advances including the addition of negation handling and syntactic information as well as proposing algorithms that incorporate optimized semantic matching solutions.

Semantic similarity is the underlying principle for understanding student contributions. We assess how similar a student contribution is to an expert answer. The expert answer is deemed correct and therefore the student contribution is deemed correct if it is semantically similar to the expert answer (and incorrect otherwise). More nuanced assessments are made (e.g., partially correct or partially correct and partially incorrect at the same time) but we focus here on simple binary judgments (correct versus incorrect). The alternative to the semantic similarity approach is the true or full understanding approach in which the student response is fully interpreted. The full understanding approach is intractable for real-world, sizeable applications such as ITSs because it requires vast amounts of world and domain knowledge. Domain and world knowledge are captured at some extent by semantic similarity approaches.

It should be noted that because student contributions can vary in length, from very short, e.g. including just one content word such as equal, to a sentence or even a paragraph, methods that apply across different granularities of texts are needed. The methods presented next are generally applicable to texts of various sizes although some are more suited for a certain granularity level. For instance, when syntactic information is used in a particular method then the method cannot be applied at word level directly.

We present next one method addressing the task of semantic similarity in the context of dialogue-based ITSs which we explored and embedded in our SEMILAR tool (Rus et al., 2013) for exploring semantic similarity methods. Several broad categories of semantic similarity methods were investigated and are included in our SEMILAR toolkit: vectorial methods including LSA (Landauer et al. 2007; Lintean, Moldovan, Rus, & McNamara, 2011), probabilistic methods including Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003; Niraula et al., 2013), greedy methods, optimal methods (Rus et al., 2012a; Rus & Lintean, 2012), and some others. Due to space reasons, we only discuss one of the most recent and most advanced methods we developed.

Optimal Word-to-Word and Syntactic Matching via Quadratic Assignment (QAP). This method looks for an optimal global assignment of words in one sentence (e.g., a student response) to words in the other sentence (e.g., the expert answer) based on their word-to-word similarity, while simultaneously maximizing the match between the syntactic dependencies. Accounting for the syntactic dependencies among words is the
primary advantage of the Quadratic Assignment Problem (QAP; Koopmans-Beckmann, 1957) formulation versus the job-assignment formulation of the student response assessment task (Rus & Lintean, 2012). We formulate first the Quadratic Assignment Problem and then show how we model the semantic similarity task based on it.

The goal of the Koopmans-Beckmann (1957) formulation of the QAP problem, first proposed for economic activities, is to minimize the objective function QAP (see below) where matrix F describes the flow between any two facilities, matrix D indicates the distances between locations, and matrix B provides the cost of locating facilities to specific locations. F, D, and B are symmetric, non-negative matrices.

\[
\min QAP(F, D, B) = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} d_{\pi(i)\pi(j)} + \sum_{i=1}^{n} b_{i,\pi(i)}
\]

The \( f_{ij} \) term denotes the flow between facilities \( i \) and \( j \) which are placed at locations \( \pi(i) \) and \( \pi(j) \), respectively. The distance between these locations is \( d_{\pi(i)\pi(j)} \). In our case, \( F \) and \( D \) describe dependencies between words in one sentence (or the other) while \( B \) captures the word-to-word similarity between words in opposite sentences. Also, we have weighted each term in the above formulation and instead of minimizing the sum we are maximizing it, resulting in the formulation below.

\[
\max QAP(F, D, B) = a \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} d_{\pi(i)\pi(j)} + (1 - a) \sum_{i=1}^{n} b_{i,\pi(i)}
\]

The \( f_{ij} \) term quantifies the syntactic relation between words \( i \) and \( j \) in text A which are mapped to words \( \pi(i) \) and \( \pi(j) \) in text B, respectively. The distance \( d_{\pi(i)\pi(j)} \) quantifies the syntactic relation between words \( \pi(i) \) and \( \pi(j) \). For words \( i \) and \( j \) that have a direct dependency relation, i.e. an explicit syntactic relation among two words, such as subject or direct object, the "flow" \( f_{ij} \) is set to 1 and 0 otherwise. Similarly, the distance \( d_{\pi(i)\pi(j)} \) between words \( \pi(i) \) and \( \pi(j) \) is set to 1 in case there is a direct dependency relation among them and 0 otherwise. We also experimented with a variant in which we enforced that the dependency between words \( i \) and \( j \) and the dependency between the corresponding words in the other text, \( \pi(i) \) and \( \pi(j) \), be identical. That is, we prefer matchings between words in texts A and B, respectively, that not only lead to direct dependencies among words in A and the corresponding matched words in B but those dependencies must be of the same type.

A brute force solution to the QAP problem, which would generate all possible mappings from facilities (words in a sentence) to locations (words in the other sentence), is infeasible because the solution space is too large. For example, when considering all possible pairings of words between sentence A, of size \( n \), and sentence B of size \( m \), where \( n < m \), and we pose no limitations on the type of pairings that can be made, there are \( m!/(m-n)! \) possible solutions. For sentences of average size \( n = m = 20 \) words, there are \( 2.4 * 10^{18} \) possible pairings.

An efficient branch-and-bound algorithm has been developed to reduce the explored space in search for the optimal solution. This is possible by defining a bounding function that always overestimates or underestimates solutions, depending on what type of optimal solution is sought, maximum or minimum cost, respectively. In our case, a maximum cost solution is desired. Our solution yielded the best performance (accuracy=77.6%) reported so far on a benchmark text-to-text similarity task at sentence level, i.e. the Microsoft Research Paraphrase corpus (Dolan et al., 2004).

**Designing for Affect**

Another important advancement of the ITSs’ inner loop focuses on techniques that target students’ affective states. Although ITSs and other advanced learning technologies (ALTs) have traditionally paid less attention to affect, the tide is gradually shifting as affect-enabled (affect-aware, affect-sensitive, or affective) learning technologies are coming online. These systems are motivated by a preponderance of research that suggests that cognition and affect are inextricably coupled, and meaningful learning always encompasses some blend of the two. In this section, we briefly discuss some of the research in the growing area of affect-enabled learning technologies and closely examine one system that takes a somewhat unconventional route towards leveraging the affect-cognition relationship to increase learning gains.
Types of Affect-Enabled Learning Technologies

One simple categorization of affect-enabled technologies broadly distinguishes between proactive and reactive systems. Proactive systems aspire to increase the incidence of affective states deemed beneficial (e.g., interest, curiosity, engagement), while simultaneously decreasing the incidence of certain negative states (e.g., boredom). For example, some ALTs aspire to leverage the facilitating effects of games on engagement by implementing some of the competitive and motivational features of games. Others take this a step further by carefully embedding the learning content in games that support narrativity, realism, and immersion.

Reactive systems make no notable a priori attempt to up-regulate or down-regulate positive and negative affect, respectively. Instead, they simply detect and respond to affective states as they arise. This involves fully automated detection of learner affect, which is accomplished with top-down predictive models independently or in conjunction with bottom-up sensor-driven models. Once the learner’s affective state is detected with reasonable accuracy, the system then dynamically alters its pedagogical plan in response to the sensed state.

A number of reactive affect-enabled ALTs have been developed and tested. The Affective AutoTutor is a dialog-based ITS that monitors contextual cues (e.g., learner response accuracy, response time, tutor feedback), body movements, and facial features to detect when a learner is confused, bored, or frustrated, and responds with motivational dialog moves to encourage learners to persist in their learning despite these negative emotions (D’Mello et al., 2010). Forbes-Riley and Litman (2011) have recently endowed a speech-enabled dialog-based physics ITS with the ability to tailor its actions on the basis of the uncertainty and correctness of learner responses. Uncertainty is detected by fusing an acoustic-prosodic analysis of the learners’ spoken responses with features extracted from the ensuing tutorial dialog (e.g., turn number). Yet another example is Caze Tutor, a learning environment for biology that monitors eye movements to infer when learners are disengaged or zoning out and launches gaze-reactive dialogs in an attempt to reengage the learners (D’Mello et al., 2012).

Although the proactive and reactive strategies represent the major research thrusts in the area of affect-enabled ALTs, there are additional possibilities as well. One possibility is to intentionally induce certain affective states that have a somewhat counterintuitive relationship with learning. Confusion is a particularly compelling example of such a state. Though most would consider confusion to be a negative affective state, both in terms of its subjective experience (i.e., most people do not like being confused) and its assumed impact on learning (i.e., intuition suggests that confusion is harmful to learning), there is some correlational evidence that suggests a positive relationship between confusion and learning gains (D’Mello et al., in press). We were intrigued by this finding, so we developed a learning environment to investigate if there are any benefits to intentionally inducing confusion.

Case Study of an Affect-Enabled ALT: Using Confusion to Promote Clarity

The idea that some forms of confusion can be beneficial to learning is grounded in theories that highlight the benefits of impasses, cognitive conflict, and cognitive disequilibrium to learning at deeper levels of comprehension. These theories suggest that confusion signals that something is wrong with the learner’s state of knowledge and can engender deeper modes of processing to the extent that the learner engages in effortful cognitive activities to resolve the confusion. Importantly, it is not the confusion itself, but the cognitive activities that accompany confusion resolution that lead to any improvement in learning gains.

A model of confusion induction and resolution. Figure 4 depicts some of the key processes of our model. ALTs, by and large, strive to promote clarity and understanding by providing hints, explanations, elaborated feedback, and other cognitive scaffolds. This might lead to illusions of clarity and understanding, but not necessarily to deep conceptual understanding. An alternate approach is to temporarily suspend clarity by interleaving contradictions, erroneous opinions, inaccurate feedback, and other confusion-inducing events. The idea is that the induced confusion will inspire learners to actively engage in deliberation, problem solving, and other forms of sense making in order restore clarity by resolving their confusion. The learning environment can also assist by providing hints and targeted explanations when confusion-resolution fails and the learner risks being hopelessly confused.

The success of this strategy will ultimately depend on the extent to which confusion has been induced and effectively resolved. Learning is not expected to be impacted if the induction fails or if the induced confusion is simply ignored. Learning might also be negatively affected if the induced confusion is too severe, the pedagogical scaffolds are insufficient, and learners are hopelessly confused. On the other hand, some form of deep learning is expected if learners effortfully engage in confusion resolution processes, even if they do not fully resolve their confusion. Learners might also experience a form of conceptual change if
problematic misconceptions are identified and corrected during confusion resolution. This is, of course, the sine qua non of learning.

Testing the Model. We tested the model by developing a learning environment that strategically induced confusion at critical points during the learning of scientific research methods, such as learning about the importance of random assignment to substantiate causal claims. The multimedia learning environment, modeled after the educational game Operation ARA (Halpern et al., 2012), taught research methods concepts by presenting example cases of studies (including the research design, participants, methods, results, and conclusions) that were frequently flawed. Learners were instructed to evaluate the merits of the studies and point out problems over multiple trials. The critiques were accomplished by holding multi-turn conversations in natural language with embodied conversational agent(s) and the human learner.

One rendition of the system (contradiction-only version) included a tutor agent who led the tutorial lessons and served as an expert on scientific inquiry and a peer agent who simulated a peer of the human learner. Confusion was induced by having the animated agents contradicting each other by occasionally disagreeing on ideas and voicing inaccurate information, and asking the human learner to intervene and decide which opinion had the most scientific merit. There was also a delayed-contradiction version in which the agents initially agreed on a concept but abruptly contradicted each other as the conversations progressed. A third version of the system (false-feedback version) implemented dialogs between the tutor agent and the human learner (there was no peer agent). This version used a false-feedback manipulation to induce confusion in which the tutor provided positive feedback to incorrect learner responses and negative feedback to correct responses.

The interface for contradiction-only and delayed-contradiction versions shown in Figure 5 consisted of the tutor agent (A), the peer agent (B), a description of the research case study (C), a text-transcript of the dialogue history (D), and a text-box for learners to enter and submit their responses (E). The agents delivered the content of their utterances via synthesized speech while the human learner typed his or her responses. Text-transcripts of the trialogues were stored in log files for offline analysis.
The different versions also varied the amount of help provided to learners to resolve their confusion. The delayed-contradictions and the false-feedback versions provided learners with a short explanatory text to read after the confusion inductions. The contradiction-only version provided no explicit scaffold, with the exception that all misleading information (contradictions and false feedback) was corrected at the end of the discussion for each case study (this was the same in all versions).

We conducted four experiments (total N = 474) in which college students learned between 4 to 8 research methods concepts by discussing the scientific merits of sample research studies with the animated agent(s). Experiments 1 and 2 used the contradiction-only version (D'Mello et al., in press), Experiment 3 used the delayed-contradiction plus explanatory text version (D'Mello et al., in press), and Experiment 4 used the false-feedback plus explanatory text version (Lehman et al., prep.). All four experiments used a within-subjects design in that learners analyzed some case studies with the confusion induction methods enabled and other studies with these methods disabled.

Confusion was tracked via offline cued-recall procedures (Experiments 1 and 2), online self-reports (Experiments 3 and 4), and learner behaviors (e.g., response times after confusion induction, accuracy on probing questions immediately following the manipulations - all experiments). Learning was measured via multiple choice retention tests (all experiments) and a transfer task involving detection of flaws in new case studies (Experiments 3 and 4).

The results of all four experiments generally indicated that the manipulations had the desired effect of inducing confusion. Importantly, learning was not affected by the manipulations alone or when the manipulations failed to induce confusion. However, performance on the learning measures was substantially higher in the experimental conditions compared to the control conditions when the manipulations were successful in inducing confusion.

The results of these four experiments are significant because they constitute some of the first experimental evidence on the moderating effect of confusion on learning. The most obvious implication is that there might be some benefits for ALTs that intentionally perplex learners. Of course, this is only if learners have
the requisite knowledge and skills to resolve the confusion and the learning environment provides appropriate scaffolds to help with the confusion resolution process. But, when done right, confusion is one road to clarity.

Concluding Remarks

Much has changed in the last decade since the Graesser and colleagues article (2001). We now know that conversational ITSs are more effective than non-interactive text controls, their effectiveness depends on the extent of the alignment between the content and students’ prior knowledge, they are as effective as non-conversational ITSs, and they rival or outperform novice human tutors on comparable content (Olney et al., 2012; VanLehn et al., 2007; VanLehn, 2011). We also know that the content of the tutorial dialogs is the primary driver of learning gains, at least when compared to other aspects of the interaction, such as whether the tutors or students communicate via speech or text or if animated pedagogical agents are present or absent.

Though several of the earlier predictions pertaining to the hypothesized effectiveness of conversational ITSs have generally been supported, two counterintuitive findings, one positive and one negative, stand out. On the positive front, a recent meta-analysis by VanLehn (2011) revealed that the so called "expert" human tutors rarely achieve the much touted two sigma effect noted by Bloom (1984) and that conversational ITSs are equally effective as these human tutors. This positive finding is tempered by the fact that the effectiveness of conversational tutors plateaus at about one sigma (approximately a letter grade), a point at which increased interactivity has little to no additional impact in increasing learning.

This article has focused on some of our recent research on next-generation conversational agents that aspire to ascend the interactional plateau into the land of two sigma effects. We have described DeepTutor, a conversational ITS that implements Learning Progressions and deeper natural language understanding. We have also discussed affect-aware ITSs that take the student-tutor interaction to the next level by modeling both the affective and the cognitive aspects of learning.

Our hope for the next decade is a world in which these and other conversational ITSs autonomously help hundreds of thousands of students develop content mastery, learning strategies, critical thinking, writing proficiency and other 21st century skills in a manner that effectively integrates cognition, motivation, and emotion. This is a tall order indeed, but the task of educating the next generation of students is not for the faint of heart.

ACKNOWLEDGMENTS

The authors would like to thank members of DeepTutor project (www.deeptutor.org), the Tutoring Research Group (http://www.autotutor.org/), and the Emotive Computing Group (http://emotion.autotutor.org/) at the University of Memphis. This research was supported by the Institute for Education Sciences (IES) under award R305A100875 to Dr. Vasile Rus and National Science Foundation (NSF) under awards ITR 0325428, HCC 0834847, and DRL 1235958 to Drs. Graesser and D’Mello. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the IES or NSF.

References

Battista, M.T. 2011 Conceptualizations and issues related to learning progressions, learning trajectories, and levels of sophistication. The Mathematics Enthusiast, 8, 507-569.
Bloom, B. 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.


Graesser, A.; Moreno, K.; Marineau, J.; Adcock, A.; Olney, A.; and Person, N. 2003 AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head? In U. Hoppe, F. Verdejo & J. Kay (Eds.), Proceedings of the 11th International Conference on Artificial Intelligence in Education (pp. 47-54). Amsterdam: IOS Press.

Graesser, A. C.; Rus, V.; D’Mello, S.; and Jackson, G. T. 2008 AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), Current perspectives on cognition, learning and instruction: Recent innovations in educational technology that facilitate student learning (pp. 95-125). Information Age Publishing.


Rus, V.; Moldovan, C.; Graesser, A.; & Niraula, N. 2012b. Automatic Discovery of Speech Act Categories in Educational Games, 5th International Conference on Educational Data Min-ing (EDM12), June 19-21, Chania, Greece.


**Vasile Rus** is currently an associate professor of Computer Science in the Institute for Intelligent Systems at the University of Memphis. His current research focuses on building conversational Intelligent Tutoring Systems. He is the Principal Investigator of the DeepTutor project funded by the Institute for Education Sciences. Dr. Rus was involved in various roles in the development of other Intelligent Tutoring Systems including MetaTutor, W-Pal, iSTART, and AutoMentor.

**Sidney D’Mello** is an Assistant Professor holding joint appointments in the departments of Computer Science and Psychology at the University of Notre Dame. His interests include emotional processing, affective computing, artificial intelligence in education, human-computer interaction, speech recognition and natural language understanding, and computational models of human cognition. He is an associate editor for IEEE Transactions on Affective Computing and serves as an advisory editor for the Journal of Educational Psychology.

**Xiangen Hu** is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and visiting professor at Central China Normal University. Dr. Hu’s primary research areas include Mathematical Psychology, Research Design and Statistics, and Cognitive Psychology. More specific research interests include General Processing Tree (GPT) models, categorical data analysis, knowledge representation, computerized tutoring, and advanced distributed learning.

**Art Graesser** is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is an Honorary Research Fellow at the University of Oxford. His primary research interests are in cognitive science, discourse processing, emotion, tutoring, and the learning sciences. He and his colleagues have developed learning technologies with animated conversational agents (such as AutoTutor and Operation ARA) and automated analyses of texts at multiple levels (such as Coh-Metrix, and Question Understanding AID -- QUAID).