Trialog in ARIES: User Input Assessment in an Intelligent Tutoring System

Zhiqiang Cai, Arthur Graesser, Carol Forsyth, Candice Burkett
Institute for Intelligent Systems
The University of Memphis
Memphis, USA
zcai, agraesser, cmfrsyth, cburkett@memphis.edu

Keith Millis, Patricia Wallace
Department of Psychology
Northern Illinois University
Dekalb, USA
kmillis, pwallace@niu.edu

Diane Halpern and Heather Butler
Department of Psychology
Claremont McKenna College
Claremont, USA
diane.halpern@cmc.edu, heather.butler@cgu.edu

Abstract—OperationARIES! **(or ARIES for short) is an intelligent tutoring system that teaches critical thinking and helps learners acquire scientific inquiry skills. One of the core components of ARIES is "trialogs" which are three-party conversations in natural language among a human student and two artificial pedagogical agents (tutor and fellow student). These tutorial interactions were designed to enhance the students' learning experience. Assessing student input is essential to the performance of ARIES. Regular expressions and Latent Semantic Analysis (LSA) were used in models that evaluate students' answers. The resulting computational models were found to be as reliable as human raters.**

Keywords-Operation ARIES!; AutoTutor; intelligent tutoring systems; LSA; natural language processing

I. INTRODUCTION

ARIES is an acronym for *Acquiring Research Investigative and Evaluative Skills*. It is an intelligent tutoring system that teaches critical thinking and scientific reasoning. The system has three modules: interactive text (called *training* in the present study), case study and interrogation. The present study focuses on students' verbal contributions in the training module, but we expect the results to generalize to the case study and interrogation modules.

The ARIES training module utilizes an eBook, multiple-choice questions, and tutorial *trialogs* in its teaching of scientific inquiry skills. The eBook, *The Big Book of Science*, covers 21 topics of research methodology, such as Theory, Hypothesis, Science and Pseudoscience, Operational Definitions, Independent and Participant Variables, and so on. Two multiple choice tests are included in each chapter. Each test contains 3 questions that target the definition, the importance and the application of each topic. The learner can choose to take the first test of the chapter before or after reading the chapter in the eBook. After finishing reading the chapter and the first test, the second test is presented to the reader. Each multiple choice question in the second test is followed by a tutorial trialog. A trialog consists of a 3-party conversation that includes the human student and two conversational agents. Depending on the scores the learner earned in the two tests, the system chooses to launch one of three types of trialogs: vicarious learning (human observes two agents interact), tutoring (a teacher agent teaches the human and a student agent), and teachable agent (the human teaches a student agent, assisted by the teacher agent). The details about the three types of trialogs can be found in Millis et al. [9].

Performance of ARIES relies on the semantic assessment component that analyzes students' verbal contributions in the trialogs. In this paper, we analyze approximately 1000 responses collected from the students' interaction with ARIES and discuss algorithms that were used in the semantic assessments. ARIES cannot respond in an intelligent fashion if there is not an adequate evaluation of what the student expresses.

Section II of this paper provides a brief historical description of conversational systems. Section III elaborates on the script elements in an ARIES trialog. Section IV describes an experiment we conducted using ARIES. Section V presents linear regression models that evaluate student input in ARIES trialogs.

II. FROM ELIZA TO ARIES

The development of conversational systems can be traced back to Weizenbaum's (1966) ELIZA, a conversation system that works like a Rogerian therapist [10]. The following conversation illustrates how ELISA works:

David: *Hi there!*
ELIZA: *How are you today ?*
ELIZA: *What would you like to discuss?*
David: *Let us discuss critical thinking.*
ELIZA: *What does that suggest to you?*
David: *It is knowledge about how to think correctly.*
ELIZA: *Have you asked such questions before?*
David: *No.*
ELIZA: *You are being a bit negative.*
David: *Why do you think so?*
ELIZA: *You're not really talking about me, are you?*

ELIZA uses simple keyword detection and syntactic transformations to create the impression that the system is understanding the student and responding intelligently. In the above conversation, when the user, David, proposed to discuss "critical thinking", ELIZA simply took it as a statement and asked a very general question in response. When David said "It is knowledge about how to think correctly", ELIZA detected the keyword "how" and treated David's input as a question and asked another general question: "Have you asked such questions before?" Since David's input was not a question, ELIZA's response is not adequate. In the rest of the conversation, ELIZA matched the keywords "no" and "you" and asked questions that did not highly match the context. The online ELIZA (www.manifestation.com/neurotoys/eliza.php3) is quite entertaining, but it does not understand the student's contextual meaning nor does it provide any knowledge to the user.

After decades of fast development in computer technologie, artificial intelligence, and computational linguistics, it became possible to integrate large amounts of knowledge into conversational systems. Wallace (1995) and colleagues created ALICE (http://www.alicebot.org), in which knowledge units are represented in AIML (Artificial Intelligence Markup Language). The conversation below shows how conversation with ALICE might go:

David: *Hi there!*
ALICE: *Hello DAVID.*
David: *Let us discuss critical thinking.*
ALICE: *That sounds good to me.*
David: *What do you know about critical thinking, then?*
ALICE: *I have to process that one for a while.*
David: *Should I wait?*
ALICE: *You should do what is best for you.*
David: *Do you know what a theory is?*
ALICE: *It is something that has few practical applications.*
David: *Do you know what a hypothesis is?*
ALICE: *Does it have anything to do with critical thinking?*
David: *Yes. It does.*
ALICE: *Ahuh. How often?*

What is noticeable in this conversation is ALICE's response to "theory", which shows that ALICE has some knowledge about "theory". Another noticeable response is about "hypothesis." ALICE's response linked back to what David mentioned earlier: "critical thinking". That indicates that ALICE has some memory to trace conversation history.

ALICE's performance is limited in three major ways. First, the user input assessment is based on string patterns. The number of patterns can reach extremely large numbers while attempting to match all possible users' input. Second, the knowledge base is mixed with both general and various forms of domain knowledge, making the building process expensive. More importantly, this technique creates difficulty when mapping a user's input into the correct domain. For example, in the above conversation, ALICE responded to "theory" using general mundane world knowledge, instead of the scientific term in critical thinking. Third, the discourse is based on local stimulus-response patterns, which does not scale up to conversing with the user with sophisticated discourse.

Graesser and his colleagues made an important breakthrough in the design of conversational tutoring systems when they created AutoTutor [4,5,6]. AutoTutor teaches conceptual physics, computer literacy and critical thinking skills by holding a conversation in natural language. The following conversation shows a conversation with AutoTutor:

Tutor: *David, can you tell me what theory is?*
David: *Theory is something not of much practical use.*
Tutor: *Well, not really.*
Tutor: *As you try to define theories, think about how theories and reasons are related.*
David: *I think theories give a lot of reasons.*
Tutor: *OK. Let's try this.*
Tutor: *Theories provide explanations for what?*
David: *For why something happens.*
Tutor: *Great job!*
Tutor: *For why things occur.*
Tutor: *A theory is an organized set of principles that explain why something occurs and allow for the creation of new predictions. Scientific theories are tested using strict scientific techniques.*

AutoTutor presents a question to a learner and then directs the learner to actively construct an answer to the question. The underlying *expectation and misconception-tailored* (EMT) mechanism focuses the dialog on learning expectations associated with a specific topic and on correcting students' misconceptions. The conversation is sophisticated enough for the student to acquire the material at deeper levels of comprehension.

AutoTutor uses *Latent Semantic Analysis* (LSA) as an assessment component to compare students' input with stored expectations and misconceptions [2,8]. Using singular value decomposition technique, LSA extracts vector representations for words so that semantically similar words have similar representations. LSA makes it possible to compare the meaning of two text chunks, even if they use very different words.

AutoTutor's EMT mechanism is used in ARIES. However, instead of a two- party (AutoTutor-Learner) dialog, ARIES conversation involves three parties: Quinn, a tutor agent, Glass, a fellow student agent, and the human learner. We refer to this kind of three-party conversation as a *trialog*[1,9]. The following is a typical trialog from ARIES.

Quinn: *David, can you tell me what theory is?*
David: *Theory is something not of much practical use.*
Quinn: *Well, not really.*
Quinn: *As you try to define theories, think about how theories and reasons are related.*
David: *I think theories give a lot of hypothetical reasons.*
Glass: *I think theories and hypotheses are the same.*
Quinn: *No.*
Quinn: *Hypotheses are simply statements about relationship between two variables. Theories are different from hypotheses.*
Quinn: *OK. Let's try this.*
Quinn: *Theories provide explanations for what?*
David: *For why something happens.*
Quinn: *Great job!*
Quinn: *For why things occur.*
Quinn: *A theory is an organized set of principles that explain why something occurs and allow for the creation of new predictions. Scientific theories are tested using strict scientific techniques.*

ARIES uses both regular expressions and LSA to assess student input. One difficulty in this system is to distinguish expectations from misconceptions, because the two are highly related and contain similar words. Consequently, this increases the possibility of misclassifying a student's input as a misconception when it is actually a correct expectation (or vice versa). In the above trialog, the assessment component figured that David had a misconception about theories. However, given that the assessment is not perfect, ARIES uses Glass as a sidekick to articulate a stored misconception that is semantically close to what the user said. This makes it safe for Quinn to give negative feedback and correct a misconception without breaking the flow of the conversation and assigning blame to the human student.

III. TRIALOG SCRIPT ELEMENTS IN ARIES

ARIES trialogs are supported by curriculum scripts prepared by experts. Each script contains a main question, summary, answers of varying quality (ideal, good and partial answers), hints, prompts and misconceptions. Table 1 shows an example script containing these elements.

The main question in the script is about the definition, importance or application of a key conception in the eBook. The main question is asked by one of the two artificial pedagogical agents, Quinn or by Glass, while maintaining the character persona of either teacher or student agent.

The summary about the answer to a question is prepared for Quinn to articulate at the end of each trialog. It includes the ideal answer used to match student's input.

Table 1 Script Elements

Main question	*Could you define what a theory is?*
Summary	*A theory is an organized set of principles that explain why something occurs and allow for the creation of new predictions. Scientific theories are tested using strict scientific techniques.*
Answer	*A theory is an organized set of principles that explain why something occurs.*
Answer key	*organiz|systematic|structure|coherent| methodical|order, expla|clarif, why|reason, occur|happen|event*
Hint	*As you try to define theories, think about how theories and reasons are related.*
Prompt	*Regarding phenomena and events in the world that are not understood, theories provide what?*
Prompt key	*expla|reason|why*
Prompt completion	*they provide explanations*
Misconception	*It's like a person's belief or opinion.*
Misconception key	*belie|opinion|perception*
Misconception correction	*A theory is based on evidence rather than just beliefs or opinions.*

Answers are prepared to assess matches to the student's verbal input. For each answer, a set of key regular expressions is implemented in the curriculum scripts. We use regular expression to handle word variations and alternatives (http://www.regular-expressions.info). For example, "organiz" can match "organize", "organizes", "organizing", "organization", etc. The script writers are faced with the challenge of initially creating multiple good and partial answers to each topic. Such embedded answers are refined and augmented after considering empirical data. However, it is impossible for experts to cover all possible correct answers that may be articulated by potential users. Therefore, regular expressions and LSA help make more flexible matches between the student's verbal input and the expected information.

The human student is taught the information through hints, prompts and misconception correction presented by both artificial agents. A hint provides a clue to the ideal answer on the given topic. If the student is still unable to articulate the correct answer, he or she is given a prompt. A prompt is a focused question which requires only a single word or phrase as an answer. Each prompt has a set of unique regular expressions to compute match scores based on the student's input. In addition, a correct answer is prepared for Quinn or Glass to articulate after the user gives an answer to the prompt.

Possible misconceptions are prepared in the scripts in order to match common misconceptions articulated by the human student. Once a student's input is matched, the prepared misconception will be given by Glass, followed by a correction given by Quinn.

IV. Experiment

We collected data from 21 college students who interacted with ARIES on 11 chapters, each of which contained 3 trialogs. From the interactions, 1296 student answers were collected, including 155 prompt answers and 1141 answers to main questions. Two graduate students independently rated the human input answers on a continuous scale ranging from 1-6, with 1 indicating an incorrect answer to 6 signifying a perfect answer. The correlations of the ratings between the two raters were 0.686 (N=1141) for main question answers and 0.771 (N=155) for prompt answers. The average of the ratings from the two experts was scored and used to train computational models to automatically evaluate the quality of student contributions.

V. Assessment Models

To evaluate student answers, we used three types of measures: (1) number of words, (2) proportion of matched regular expressions and (3) LSA scores. The values of (2) and (3) varied from 0 to 1. Number of words was a simple word count, denoted by "NW".

For main question answers, a student answer was matched with all answer keys (i.e., sentence expectations). The proportion of matched regular expressions was computed and the maximum of all answer keys was taken as the regular expression score for each student answer. For prompt answers, the student answer was matched with the prompt key and the proportion of matched regular expressions was the regular expression score, varying from 0 to 1. The regular expression score is denoted as "RegEx".

We considered two different LSA scores. The first LSA score was the maximum LSA cosine between the student answer and the answers in the script, denoted by "LSAA". The second LSA score was the LSA cosine between a student answer and all answers from other students, denoted by "LSAOUA". This score requires the existence of multiple answers from different students to the same questions.

A. Algorithms for assessing main question answers

For the answers to the main questions, we selected those that had at least 5 answers to the same question. There were 892 such answers in our data set. We split 66% of these for training and 34% for test.

Table 2 Correlation Table

Measure	Correlation
NW	0.337**
RegEx	0.619**
LSAA	0.278**
LSAOUA	0.512**

**p<0.01, N= 892

Table 2 shows the correlation between the measures and the average human rating. The high correlation with RegEx indicates that the regular expressions in the scripts were very well prepared. The regular expressions were prepared by two graduate students who went through several rounds of testing before the experiment was conducted. Nevertheless, the prepared regular expressions cannot perfectly assess all possible student expression.

The performance of the LSA scores was quite interesting. The LSA scores that use the other student answers (to a particular question) had a much higher correlation to the human coders than the LSA that relied on a single ideal answer. This indicates that a mixture of student answers can make up an answer model that is better than well prepared ideal answers by experts. Therefore, when expert-prepared ideal answers are not rich enough to capture all possible student inputs, adding collected student answers can help form a better reference norm for assessment.

Table 3 shows the standardized coefficients and R values for different linear regression models. Models 1, 2 and 3 show performance when using NW and the one of the three match scores (LSAA, LSAOUA, RegEx) whereas models 4 and 5 include 2 of the match scores. Model 5 gives a result R=0.667, which is statistically indistinguishable from two human raters' agreement (R=0.686).

Table 3 Linear Regression Models (Main Question Answers)

Model	NW	LSAA	LSAOUA	RegEx	R
1	0.171	0.270			0.417
2	0.169		0.423		0.580
3	0.153			0.597	0.614
4	0.073	0.133		0.585	0.625
5	0.095		0.222	0.515	0.667

B. Algorithm for assessing prompt answers

The answers to prompts are usually one word or a short phrase. To assess prompt answers, we used 2 measures, *RegEx* and *LSAA*. LSAA is the cosine between the student's answer and the expert prepared prompt answer. Table 4 shows the correlations of these 2 measures to human ratings.

Table 4 Correlation Table (Prompt Answers)

Measure	Correlation
RegEx	0.741**
LSAA	0.536**

**p<0.01, N=155

We did not use number of words as a predictor because the prompt answers are usually one or two words. The LSA cosine to other students' answers was not included because it relies on multiple students' input to each question and our data did not yield enough different prompt answers.

The 155 prompt answers were split to a 66% training set and a 34% test set. The linear regression result is shown in Table 5.

Table 5 Linear Regression Model (Prompt Answers)

Model	LSAA	RegEx	R
1	0.172	0.608	0.770

This model had an R that is virtually equivalent to two human rater' agreement (R=0.771).

These findings can both help us assess and refine the language processing in expectation &misconception tailored natural language conversations such as that found in ARIES. Due to the differences in the length of answer completion, different models were used to assess the prompt completion vs. the answer to the main question.

Assessing prompt completion answers is limited due to the number of answers in the data set. The model that best fits the data includes both the regular expressions and the LSAA; both of these use the pre-prepared scripts rather than other student answers. Perhaps with a larger database, the results to the prompt completion would be similar to those found in the analysis of the answers to the main questions.

In assessing the answers to the overall main question, regular expressions and LSA can increase the effectiveness of the language processing. Specifically, using LSA to match previous students' answers to the current human input (LSAOUA) adds a definitive boost to the matching versus pre-prepared regular expressions alone. The two together create the model most comparable to human ratings. However, there is a disadvantage of LSAQUA compared with LSAA. It is ad hoc in the sense that data need to be collected from a sample of students rather than directly from an ideal curriculum script.

Nevertheless, the ideal system would need some testing and data collection from students, so our recommendation is to have a combination of ideal answers, a family of student answers, and regular expressions. When a new system is built, expert prepared ideal answers and regular expressions should be used to collect student answers. Once enough student answers are collected, there could be rounds of refinement of the assessment model by adding student answers to the model. That will involve human ratings on student answers. The cost of getting the expert ratings is not as high as fine tuning the regular expressions and ideal answers. This final model handles both hints and prompts with an accuracy of semantic matching equivalent to human ratings.

ACKNOWLEDGMENT

This research was support by the National Science Foundation (BCS0904909, DRK-12-0918409) and the Institute of Education Sciences (R305B070349). The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

REFERENCES

[1] Z. Cai, A. C. Graesser, K. K. Millis, D. Halpern, P. Wallace, and C. Moldovan, "ARIES: An intelligent tutoring system assisted by conversational agents," Proc. 14th International Conference on Artificial Intelligence in Education, V. Dimitrova, R. Mizoguchi, B. DuBoulay and A. C. Graesser, Eds., Amsterdam: IOS Pres, 2009, pp.796.

[2] Z. Cai, D. S. McNamara, M. Louwerse, X. Hu, M. Rowe, and A. C. Graesser, "NLS: A Non-Latent Similarity Algorithm," Proc. 26th Annual Meeting of the Cognitive Science Society, K. D. Forbus, Dl Gentner, and T. Gegier, Eds., Mahwah, NJ: Erlbaum, 2004, pp.180-185.

[3] S. D. Craig, A. C. Graesser, J. Brittingham, J. Williams, T. Martindale, and G. Williams, "An implementation of vicarious learning environments in middle school classrooms," Proc. International Conference for the Society for Information Technology and Teacher Education, Chesapeake, VA: AACE, 2008, pp.1060-1064.

[4] A. C. Graesser, X. Hu, and D. S. McNamara, "Computerized learning environments that incorporate research in discourse psychology, cognitive sicence and computational linguistics." In Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer, A. F. Healy (Ed.), Washington D.C.: American Psychological Association, 2005, pp.183-194.

[5] A.C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. Ml Louwerse, "AutoTutor: A tutor with dialogue in antural language," Behavior Research Methods, Instruments, & Computers, vol. 36, 2004, pp.180-193.

[6] A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz, and The Tutoring Research Group, "AutoTutor: A simulation of a human tutor," Journal of Cognitive Systems Research, Vol. 1, 1999, pp.35-51.

[7] G. T. Jachson, E. C. Mathews, D. Lin, A. M. Olney, and A. C. Graesser, "Modeling student performance to enhance the pedagogy of Autotutor," Proc. Conference on User Modeling, Johnstown, PA: Springer, 2003, pp.368-374.

[8] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Eds., "Handbook of Latent Semantic Analysis," Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2007.

[9] K. Millis, C. Forsyth, H. Butler, P. Wallace, A. C. Graesser and D. Halpern, "Operation ARIES!: A serious game for teaching scientific inquiry," in Serous Games and Edutainment Applications, M. Ma, A. Oikonomou, and L. Jain, Eds., Springer-Verlag, UK, in press.

[10] J. Weizenbaum, "ELIZA-A computer program for the study of natural language communication between man and machine," Communications of the ACM, Vol. 9, 1966, pp.36-45.