An Empirically-Derived Question Taxonomy for Task-Oriented Tutorial Dialogue

Kristy Elizabeth BOYER[a], William J. LAHTI[a], Robert PHILLIPS[ab], Michael D. WALLIS[ab], Mladen A. VOUK[a], and James C. LESTER[a]
[a] *Department of Computer Science, North Carolina State University*
[b] *Applied Research Associates, Inc.*
Raleigh, North Carolina, USA
{keboyer, wjlahti, rphilli, mdwallis, vouk, lester}@ncsu.edu

Abstract. Devising an expressive question taxonomy is a central problem in question generation. Through examination of a corpus of human-human task-oriented tutoring, we have found that existing question taxonomies do not capture all of the tutorial questions present in this form of tutoring. We propose a hierarchical question classification scheme for tutorial questions in which the top level corresponds to the tutor's goal and the second level corresponds to the question type. The application of this hierarchical classification scheme to a corpus of keyboard-to-keyboard tutoring of introductory computer science yielded high inter-rater reliability, suggesting that such a scheme is appropriate for classifying tutor questions in design-oriented tutoring. We discuss numerous open issues that are highlighted by the current analysis.

Keywords. Question classification, Question taxonomies, Task-oriented tutoring, Tutorial dialogue

1. Introduction

The automatic generation of questions is an important emerging research area with potential utility for widespread applications [1]. One such application is natural language tutoring, in which questions are generated by an intelligent agent whose primary goal is to facilitate the learner's acquisition and construction of knowledge (*e.g.*, [2-9]). A tutor's pedagogical objectives may be accomplished with dialogue policies designed to enhance the learner's motivation, maintain an emotional state conducive to learning, or help the learner complete specific tasks relevant to the targeted knowledge or skill set.

A dialogue policy for question generation informs decisions about multiple features of an intelligent agent's conversational interactions. It governs decisions about the conditions under which a question should be posed, calibration of the level of content to be included, and choice of the tone of the realized question. Because a central feature involves determining the question type to be selected, devising an expressive question taxonomy is an important step toward high quality, robust automatic question generation [10, 11].

In *Proceedings of the 2nd Workshop on Question Generation*, held in conjunction with the 14th International Conference on Artifical Intelligence in Education, Brighton, UK, 2009.

It is unlikely that a single question taxonomy can meet the needs of question generation for all application areas. In fact, even if we restrict our discussion to question generation for natural language tutorial dialogue, a single taxonomy is unlikely to suffice because, aside from the differences encountered across domains (*e.g.*, qualitative physics, English, mathematics), the format in which tutoring is conducted is likely to result in the need for different types of questions. For example, the tutoring sessions analyzed in this work demonstrate *task-oriented* tutoring, where the primary activity in which the learner engages is problem solving. In task-oriented tutoring, the tutor must be concerned with the quality of knowledge the student attains as expressed through the task at hand, and if a learning artifact is being designed, the tutor may also engage in question-asking specifically to address the quality of the artifact itself.

Question classification research has benefited several other fields of study, including computational modeling of question answering as a cognitive process [12] and answering students' questions with an intelligent tutoring system (*e.g.*, [13]). Recently, question taxonomies have been proposed that begin to address the needs of the question generation community [10, 11]. In this paper, we examine a corpus of human-human task-oriented tutoring and find that existing question taxonomies do not capture all the types of questions posed by the tutors. We propose an empirically-derived hierarchical question classification scheme in which the top level identifies the tutorial goal (*e.g.,* establish a problem-solving plan, scaffold the problem-solving effort through hinting). The second level of the hierarchy consists of annotation for question type; this level shares many categories with classification schemes proposed by Graesser *et al.* [10] and Nielsen *et al*. [11].

2. JavaTutor-Q Corpus

The JavaTutor-Q corpus of questions was collected across two semesters during tutoring studies. Participants were enrolled in a university introductory computer science class titled "Introduction to Computing – Java." The tutors and students interacted via remote keyboard-to-keyboard dialogue, with tutors viewing a real-time display of student problem-solving actions. Seventeen tutors were involved across the two studies; their experience level varied from one year of peer tutoring to several years of full classroom instruction. The majority of tutors, however, did not have any formal training or professional experience in tutoring or teaching; therefore, compared to studies of expert tutors (*e.g.*, [4, 14]), the tutors under consideration here are unskilled. Eighty-two participants interacted for one session each, each session lasting approximately one hour. The complete corpus contains a total of 10,179 utterances. Tutors contributed 6,558 of these utterances, of which 714 were identified as questions during previous dialogue act tagging efforts [15, 16].[1] This corpus of questions serves as the basis for the question classification scheme presented here.

The JavaTutor-Q corpus arose from naturalistic keyboard-to-keyboard human tutoring of introductory computer science: that is, tutors were given no specific

[1] Initially there were 721 questions; however, during the tagging process reported here, 7 of these were identified as non-questions whose original dialogue act tag was erroneous.

instructions regarding tutoring strategies.[2] Qualitative exploration of the corpus revealed an important phenomenon that has shaped the question taxonomy presented here. Table 1 illustrates that tutors in this study often provide hints in the form of questions. This behavior is likely an example of a polite strategy that allows the student to "save face" in the presence of a mistake [17, 18]. Although an indirect approach may not always be ideal for student learning [19], a taxonomy of tutorial questions should capture indirect approaches. The subsequent choice of whether to implement these tactics can then be treated as a higher-level design decision.

Table 1. Excerpts from the JavaTutor-Q Corpus

Tutor 1:	So... parseInt takes a String and makes it into an int... but we only want one digit, so how are we going to get just one digit as a string? *[Proc]*
Student 1:	charAt?
Tutor 1:	Well that would give us a char.
Tutor 1:	There's another String operation that can give us part of a string as a string... I think it's subString? *[Hint]*
Student 1:	Right.
Student 2:	[Declares five distinct variables in the problem-solving window]
Tutor 2:	We can approach this using five distinct variables, but when we work with them in our loops to draw the bar codes, I'm wondering whether making an array will be a better alternative? *[Hint]*
Student 2:	Yeah, we could. Would make looping easier too.

3. Hierarchical Question Annotation

At its top level, the proposed question taxonomy intends to capture the tutorial goal that motivated each question. At its second level, this taxonomy captures the question type, a distinction that is more closely related to the surface form of the question.

3.1. Level 1: Tutorial Goal

It has been recognized that tagging a human-human corpus with tutorial goals can inform the design of the natural language generation component of tutoring systems. For example, the NLG component of CIRCSIM-Tutor was based partly on the

[2] Tutors were provided with a suggested dialogue protocol for maintaining anonymity in the event that the student directly inquired about the tutor's identity.

annotation of tutorial goals [20]. The detailed hierarchical annotations used for CIRCSIM-Tutor were not directly reusable for our purposes because many of the tutoring techniques present in their corpora of expert tutoring were not present in the JavaTutor-Q corpus. In addition, our current goal is to focus specifically on tutorial questions. To that end, we propose a new set of tutorial goals that is intended to capture what goal motivated the tutor to ask each question.

Corbett & Mostow [21] suggest that ideal questions in a reading comprehension tutor should address tutorial goals such as 1) assessing comprehension of text, 2) assessing student engagement, 3) evaluating tutor interventions, 4) providing immediate feedback, 5) scaffolding comprehension of text, 6) improving student engagement, and 7) scaffolding student learning. Some of these goals have direct analogy for task-oriented tutoring; for example, "scaffolding comprehension of text" becomes "scaffolding the student's problem-solving effort." Table 2 presents our set of tutorial goals, which began with analogues to the above goals and then evolved iteratively through collaborative tagging by two annotators. Specifically, an initial set of goals was informed by qualitative inspection of the corpus, and then goals were added or merged until both annotators felt that all tutorial questions in the "training" sample of approximately 400 questions were well-represented. After finalizing the tutorial goal tags, the first annotator tagged all of the remaining questions, for a total of 714 utterances. A second annotator tagged a subset of tutoring sessions, totaling 118 questions, that were not part of the training set. The resulting unweighted Kappa agreement statistic was 0.85, indicating high reliability of the tutor goal annotation scheme [22].

3.2. Level 2: Question Type

The second level of annotation was performed after tutor goal tagging had been completed and disagreements between annotators had been resolved collaboratively. In the second phase, each question was classified according to its type. The question type classification scheme relies heavily on the question taxonomy proposed by Nielsen *et al.* [11], which itself was informed by Graesser *et al*. [10]. The process of formulating the current set of question types was analogous to the formulation of the tutorial goal set. We began with the union of question types from [11] and [10], and through collaborative tagging of a training set of approximately 450 questions, this set was refined until both annotators felt all training questions were adequately classified. Table 2 illustrates the resulting question classification scheme. The first annotator tagged the entire question corpus, while a second annotator applied the question classification scheme to 117 questions that were not part of training. The resulting unweighted Kappa statistic of 0.84 indicates high reliability of the classification scheme.

4. Discussion and Open Issues

Understanding question types is an important step toward the robust automatic generation of high-quality questions. This paper has presented a two-level classification scheme for tutorial questions that occurred as unskilled human tutors worked with novice computer science students who were designing and implementing the solution to an introductory computer programming problem. In the study presented

Table 2. Tutorial Goals (Level 1) and Co-occurring Question Sub-Types (Level 2)

Tutorial Goal	Freq (n_{total} = 714)	Details	Question Sub-Types[3]
Plan	164	Establish a problem-solving plan. Ascertain what the student wants, prefers, or intends to do.	Definition, Free Creation, Feature or Concept Completion, Free Option, Goal, Judgment, Justification, Planning, Procedural, Status
Ascertain Student's Knowledge	282	Find out whether the student knows a specific factual or procedural concept.	Causal Antecedent, Calculate, Causal Consequence, Definition, Enablement, Feature/Concept Completion, Free Option, Improvement, Justification, Knowledge, Procedural, Quantification, Status
Hint	127	Scaffold the student's problem-solving effort.	Hint, Causal Consequence
Repair Communication	34	Disambiguate or correct a previous utterance in the dialogue.	Clarification, Feature/Concept Completion
Confirm Understanding	73	Confirm the student's understanding of a previous utterance in the dialogue or of a previously-scaffolded problem-solving step.	Assess, Backchannel, Causal Antecedent, Confirmation, Status
Engage Student	14	Elicit an utterance from the student, either at the beginning of the tutoring session or after a prolonged period of student silence.	Feature/Concept Completion, Goal, Status
Remind/Focus	20	Focus the student's attention on a previous utterance or problem-solving step for instructional purposes.	Assess, Feature/Concept Completion, Focus, Hint, Procedural

here, the tutors' goals were annotated by researchers in a *post hoc* manner. The informativeness of this tagging might be enhanced by future work in which the tutors themselves indicate the goal of each question either *post hoc* or, perhaps preferably, in real time. Ascertaining the tutors' local goal for each question, along with the state information that motivated that goal, would provide valuable insight for future automatic tutorial question generation systems.

As illustrated in Table 3, several question types from existing taxonomies did not occur in the current corpus. This phenomenon is likely due to the skill level of the tutors; they often utilized very broad question types, such as *Confirmation*, which rely heavily on the student's ability to self-assess [19]. The difference in types of questions asked by experts and novices is an important distinction (*e.g.*, [14, 23]), but because no conclusive differences in effectiveness have been established among question types, it

[3] These sets of question sub-types were not formulated *a priori*; rather, this column displays all question types that occurred in combination with each tutor goal after the question annotation was complete.

Table 3. Question Types (Level 2)

Question Type	Examples	Freq. (n_{total} = 714)	Source		
			Graesser *et al.*	Nielsen *et al.*	New
Assessment	Do you think we're done?	6			●
Backchannel	Right?	6			●
Calculation	What is 13 % 10?	11		●	
Causal Anteced.	Why are we getting that error?	2	●	●	
Causal Conseq.	What if the digit is 10?	8	●	●	
Clarification	What do you mean?	31			●
Composition	*Not present in the current corpus.*	0		●	
Comparison	*Not present in the current corpus.*	0	●	●	
Confirmation	Does that make sense?	60			●
Feature/Concept Completion	What do we want to put in digits[0]?	109	●	●	
Definition	What does that mean?	2	●	●	
Disjunctive	*Subsumed by other tags in current corpus.*	0	●		
Enablement	How are the digits represented as bar codes?	2	●	●	
Example	*Not present in the current corpus.*	0	●	●	
Expectation	*Not present in the current corpus.*	0	●		
Focus	See where the array is declared?	11			●
Free Creation	What shall we call it?	1		●	
Free Option	Should the array be in this method or should it be declared up with the other private variables?	6		●	
Goal Orientation	Did you intend to declare a variable there?	20	●	●	
Hint	We didn't declare it; should we do it now?	128			●
Improvement	Can you see what we could do to fix that?	9		●	
Interpretation	*Not present in the current corpus.*	0	●	●	
Judgment	Would you prefer to use math or strings?	17	●	●	
Justification	Why are we getting that error?	3		●	
Knowledge	Have you ever learned about arrays?	93			●
Procedural	How do we get the i^{th} element?	127	●	●	
Quantification	How many times will this loop repeat?	3	●	●	
Status	Do you have any questions?	17			●
Verification	*Subsumed by other tags in current corpus.*	0	●		

is important for question taxonomies, especially at the formative stages of research in automatic question generation, to be comprehensive. Proceeding from that foundation, investigating the effectiveness of question types given such features as the tutor's immediate goal and knowledge of the student and the problem-solving state will be an important direction for future work. Finally, it is important to note that in the question classification project presented here, questions were tagged in their original dialogue

context. The annotators felt it was often important to consider the surrounding context (usually the previous two or three utterances) for both tutorial goal annotation and question type tagging. A rigorous study of the importance of context for question classification could shed light on how much context is necessary for a question generation system to make sound decisions.

5. Conclusion

Question classification schemes have been the topic of research for several decades, and increased interest in the task of automated question generation raises new issues that highlight the importance of empirically-grounded question taxonomies. We have proposed a hierarchical question classification scheme designed to capture the phenomena that occur during human-human task-oriented tutoring. In the first level of the proposed taxonomy, questions are classified according to the tutor's goal, an approach inspired by previous work using tutorial goal annotation to inform the natural language generation of tutoring systems. The second level of the scheme captures the realized question type using an augmented version of existing question taxonomies. Both levels of question classification were applied with very high inter-rater reliability. This classification scheme represents a first step toward a comprehensive question taxonomy for task-oriented tutoring.

Acknowledgments

This research was supported by the National Science Foundation under Grants REC-0632450, IIS-0812291, CNS-0540523, and GRFP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] V. Rus and A. Graesser. *The Question Generation Shared Task and Evaluation Challenge.* In press.

[2] A. Graesser, G. Jackson, E. Mathews, et al. Why/AutoTutor: A Test of Learning Gains from a Physics Tutor with Natural Language Dialog, *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society,* pp. 1-6, 2003.

[3] C. Zinn, J. D. Moore and M. G. Core. A 3-tier Planning Architecture for Managing Tutorial Dialogue, *Proceedings of the 6th International Conference on Intelligent Tutoring Systems,* pp. 574-584, 2002.

[4] M. Evens and J. Michael. *One-on-One Tutoring by Humans and Computers.* Mahwah, New Jersey: Lawrence Erlbaum Associates, 2006.

[5] V. Aleven, K. Koedinger and O. Popescu. A Tutorial Dialog System to Support Self-explanation: Evaluation and Open Questions, *Proceedings of the 11th International Conference on Artificial Intelligence in Education,* pp. 39-46, 2003.

[6] D.J. Litman, C.P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe and S. Silliman. Spoken Versus Typed Human and Computer Dialogue Tutoring, *International Journal of Artificial Intelligence in Education,* vol. 16, iss. 2, pp. 145-170, 2006.

[7] H.C. Lane and K. VanLehn. Teaching the Tacit Knowledge of Programming to Novices with Natural Language Tutoring, *Computer Science Education,* vol. 15, iss. 3, pp. 183-201, 2005.

[8] E. Arnott, P. Hastings and D. Allbritton. Research Methods Tutor: Evaluation of a Dialogue-Based Tutoring System in the Classroom, *Behavior Research Methods,* vol. 40, iss. 3, pp. 694-698, 2008.

[9] K. VanLehn, P.W. Jordan, C.P. Rose, et al. The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing, *Proceedings of the 6th International Conference on Intelligent Tutoring Systems,* vol. 2363, pp. 158–167, 2002.

[10] A. Graesser, V. Rus and Z. Cai. Question Classification Schemes, *Proceedings of the 1st Workshop on Question Generation,* 2008.

[11] R. Nielsen, J. Buckingham, G. Knoll, B. Marsh and L. Palen. A Taxonomy of Questions for Question Generation, *Proceedings of the 1st Workshop on Question Generation,* 2008.

[12] W. Lehnert. *The Process of Question Answering - A Computer Simulation of Cognition.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1978.

[13] L. Acker, J. Lester, A. Souther and B. Porter. Generating Coherent Explanations to Answer Students' Questions. In H. Burns, J.W. Parlett, et al. (Eds.), *Intelligent Tutoring Systems: Evolutions in Design,* 151-176. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1991.

[14] W. Cade, J. Copeland, N. Person and S. D'Mello. Dialog Modes in Expert Tutoring, *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* pp. 470-479, 2008.

[15] K.E. Boyer, M. A. Vouk and J. C. Lester. The Influence of Learner Characteristics on Task-Oriented Tutorial Dialogue, *Proceedings of the 13th International Conference on Artificial Intelligence in Education,* pp. 365-372, 2007.

[16] K.E. Boyer, R. Phillips, M. Wallis, M. Vouk and J. Lester. Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue, *Proceedings of the Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* pp. 239-249, 2008.

[17] P. Brown and S. Levinson. *Politeness: Some Universals in Language Usage.* Cambridge University Press, 1987.

[18] N. Wang, W. L. Johnson, P. Rizzo, E. Shaw and R. E. Mayer. Experimental Evaluation of Polite Interaction Tactics for Pedagogical Agents, *Proceedings of the 10th International Conference on Intelligent User Interfaces,* pp. 12-19, 2005.

[19] N.K. Person, R.J. Kreuz, R.A. Zwaan and A.C. Graesser. Pragmatics and Pedagogy: Conversational Rules and Politeness Strategies may Inhibit Effective Tutoring, *Cognition and Instruction,* vol. 13, iss. 2, pp. 161-188, 1995.

[20] J.H. Kim, R. Freedman, M. Glass and M.W. Evens. Annotation of Tutorial Dialogue Goals for Natural Language Generation, *Discourse Processes,* vol. 42, iss. 1, pp. 37-74, 2006.

[21] A. Corbett and J. Mostow. Automating Comprehension Questions: Lessons from a Reading Tutor, *Proceedings of the 1st Workshop on Question Generation,* 2008.

[22] J.R. Landis and G. Koch. The Measurement of Observer Agreement for Categorical Data, *Biometrics,* vol. 33, iss. 1, pp. 159-174, 1977.

[23] M. Glass, J. H. Kim, M. W. Evens, J. A. Michael and A. A. Rovick. Novice vs. Expert Tutors: A Comparison of Style, *Proceedings of the 10th Midwest Artificial Intelligence and Cognitive Science Conference,* pp. 43-49, 1999.