What a pilot study says about running a question generation challenge

Lee BECKER [a], Rodney D. NIELSEN [a,b] and Wayne H. WARD [a,b]

[a] *The Center for Computational Language and Education Research, University of Colorado at Boulder*
[b] *Boulder Language Technologies*

Abstract.

We present a pilot study, wherein annotators rated the quality of questions produced by our system and human tutors. This study has helped us to evaluate our tutorial dialog system's question generation abilities and has aided us in identifying areas for further refinement of our dialog strategy. Moreover, this experiment has taught us several important lessons and highlighted critical issues related to running a question generation challenge.

Keywords. Question Generation, dialog, intelligent tutoring systems

Introduction

In interactive educational technologies and intelligent tutoring systems the task of question generation (QG) encompasses more than deriving a surface form from a concept; this task includes choosing the appropriate question for the situation. Ideal questions in this domain not only address learning goals and learner knowledge gaps, they exhibit pertinence by grounding themselves in the context of the conversation while simultaneously maximizing learning gains.

Similarly, Vanderwende [7] argues that choosing which question to generate is as important as generating the question itself. She proposes that meaningful evaluation of question quality should focus on judging the *importance* of a question with respect to a larger text. Furthermore, Nielsen [4] states that identifying key concepts is a critical subtask in the QG process.

This problem of choosing an important question bears much similarity to the task of dialog act selection in spoken dialog systems wherein a dialog manager must produce follow-up utterances that are on-topic and aligned with user and system goals. For this paper we approach QG as a dialog management task, and present an experiment in which we assess our tutorial dialog system's question generating capabilities relative to human tutors. In the following sections we describe our motivations, give a brief overview of our dialog-based QG system, detail our question evaluation experiment, present and discuss our results, expound on the difficulties in defining a QG challenge, and close with suggestions for future work.

1. Motivations and Background

Our project's overarching goal is to develop a dialog based tutoring system for elementary school aged students. Our curriculum is based on the Full Option Science System (FOSS) [5] a proven research-based science curriculum system that has been employed in American schools for over a decade. FOSS consists of sixteen diverse science teaching and learning modules covering life science, physical science, earth and space science, scientific reasoning, and technology. For this study, we limit our coverage of FOSS to investigations about magnetism and electricity.

The system's dialog strategies are informed by the Questioning the Author (QtA) [1] teaching method. QtA is an approach to classroom instruction that uses dialog interaction to facilitate development of effective comprehension strategies and deep learning. When applying QtA to FOSS, instructors ask open-ended questions leading to dialogs that encourage the student to make sense of their hands-on experiences. Example questions include "What's going on here?", "What's that all about?", or "How does this connect to. . . ?"

2. System Overview

To generate tutorial dialog moves, we use a modified version of Phoenix [3], a frame and slot-based dialog management system. Target concepts from the lesson are modeled as propositions using the Phoenix frame representation, while sub-concepts are expressed as elements within the frame. During the course of conversation, our system picks manually-authored questions from a pool of questions associated with the frame currently in focus.

Figure 1 shows a simplified frame, its elements, and the pool of questions used to converse about the concept *Electricity flows from negative to positive*. If a concept is only partially addressed by the student, the system will take the next move from the pool of questions corresponding to an empty slot. Additionally, our version of Phoenix has a rule mechanism [2], which allows the system to act based on the content of the elements, and is useful for detecting and addressing misconceptions. The rule shown in Figure 1 triggers when a student is confused about the direction of current in a circuit.

Currently, the general strategy for authoring these frames is to start with an open-ended QtA style question and gradually increase question specificity until the targeted slot is filled. It should be noted that frame actions are multimodal and can include visuals and other multimedia content in addition to spoken dialog.

Though our questions are manually-authored, our system does attempt to choose pertinent or important questions given the conversational context. More importantly the system provides a framework for experimenting with various tutoring styles and questioning strategies.

3. Experiment

Our study was primarily motivated by a desire to see how well questions suggested by our dialog management system compared to questions generated by humans, but it also presented an opportunity to investigate the issues involved in running a QG challenge.

```
Frame: FlowInSeries
[_start]+
    Action: "We've been talking about components in a series circuit
             sharing a single path to a power energy source.
             Tell me about what's going on here"
[Electricity]
[Flow]+
    Action: "What are the blue dots showing?"
    Action: "How can you connect what you notice with the blue dots to
             your thinking about the flow of electricity?"
    Action: "Let's look at this again.  What is going on here?"
[FromTerminal]+
    Action: "You mentioned something about the flow of electricity.
             Tell me more about what is going on with that in this picture"
    Action: "How does the electricity flow in this circuit?"
    Action: "Tell me about how the electricity flows in and out of
             the battery"
[ToTerminal]+
    Action: "Tell me more about where the electricity is flowing to."
    Action: "What side of the battery is the electricity flowing towards?"
[_done]+
    Action: "You've made some good observations.
             Now let's talk more about..."
Rules:
[FromTerminal] ==  "Positive" OR [ToTerminal] == "Negative"
    Action: "Look closely at the flow of electricity.
             Can you tell me again about which way the
             electricity is flowing?"
```

Figure 1. Example Phoenix frame for proposition: *Electricity flows from negative to positive*

Like the Bystander Turing Test conducted by Person and Graesser [6], our study uses snippets of tutorial dialog to provide context for generating questions and then later evaluating them. Unlike the the tests carried out by Person and Graesser, our judges were not asked to ascertain whether the question was generated by a human or computer. Instead our instructions simply asked the evaluator to give a single score to the question taking into account factors like appropriateness to the context and felicity to QtA practices.

3.1. Participants

The six participants in this study are tutors employed by Boulder Language Technologies. Of the six tutors, three are considered experts in Questioning the Author (QtA) style tutoring and in the FOSS curriculum, while the other three have been given training in QtA and FOSS but are not considered to be at an expert level of proficiency. Five of the six participants assisted in producing questions specifically for this experiment, while all six took part in the evaluation of questions. Tutorial sessions were conducted with students in grades 3-6.

3.2. Question Generation / Collection

Transcripts used in this experiment were collected from computer-mediated one-on-one human tutoring sessions. In this environment, the student interacts with a virtual agent

---

- good
- <um> studying magnetism
- <um> there's a scale and some cups and washers p- <um> <side_speech> <uh> magnets and i forget the other thing the yellow things <um> <breath> <um> <fp> you would put the cup in in the scale and then you would put the
- the magnet post <um> on the s- under the cup <side_speech> you put a ma- a magnet in the cup and then you would put the other cup on the left st- side and you'd try and see how many washers you could get in that other cup
- the washers <breath> you when you you aren't gonna try and see how many you can fit in without <um> the force of the magnets breaking <um> since the washers are steel
- <um> yes ((i say)) so go in on the right side of the cup <breath> a- and you put them in softly <breath> then you can fit more

Expert Tutor Question:	what is important about putting the washers in softly?
Non-Expert Tutor Question:	tell me about the spacers
System Question:	what's going on between the two magnets?

Figure 2. Example dialog context (student utterances) and corresponding generated questions

that talks and presents visual materials. Behind the scenes, a human operator controls the agent, deciding which questions to ask and which visuals to display similar to a Wizard-of-Oz experiment. No deception was used; students were told that an experimenter would be available to help the agent. These sessions were conducted by a total of 8 different tutors (2 expert, 6 non-expert).

From these transcripts we randomly sampled 50 expert tutor dialog turns and 50 non-expert turns for evaluation. We then fed the corresponding student dialog, from its beginning to the sampled point, turn-by-turn into our Phoenix system and included the last question generated in our evaluation. Similarly, we generated a third question for evaluation by presenting the same student dialog, approximately one turn every three seconds, to our human tutors, requesting a tutor turn at the sampled point. See section 6 for issues that led to this methodology.

This process yielded 3 questions for each of 100 dialog contexts: an expert tutor question, a non-expert tutor question, and a Phoenix system question – creating a total of 300 questions for evaluation, with half of the expert and non-expert questions being pulled from the actual transcripts and the other half being generated as part of the experiment. An example dialog context and its associated questions are shown in Figure 2.

3.3. Question Evaluation

Each of the 300 questions were evaluated by both an expert tutor and a non-expert tutor. Special care was taken to ensure that evaluators never rated a question for a context where they themselves generated a question. The evaluation environment was similar to the question collection environment, wherein the participant was shown a sequence of student utterances turn-by-turn. After reading the context, a follow-up question was shown, and the participant was asked to give a single score on a scale of 1 (worst) to 5 (best) taking into account factors like how well it followed the context and how well it adhered to QtA principles.

Table 1. Independent Samples t-tests for ratings by evaluator and question generator. Scores were normalized for differences in scoring among individual evaluators, such that scores for a given evaluator had mean = 3.223 (the mean of all evaluations for all questions) and standard deviation = 1.0.

Evaluators	Phoenix v. Expert		Phoenix v. Non-Expert		Expert v. Non-Expert	
All Tutors	t=3.365	p=0.000	t=5.521	p=0.000	t=0.471	p=0.638
Exp. Tutors	t=5.021	p=0.000	t=4.665	p=0.000	t=0.340	p=0.734
Non-Exp. Tutors	t=1.999	p=0.047	t=3.152	p=0.002	t=0.984	p=0.326

Figure 3. Comparisons of normalized mean scores and 95% confidence intervals by evaluator type.

Table 2. Inter-rater Group Correlations. The first column lists the questions' generator, while the other columns shows correlations over questions rated by evaluators from different classes. In columns where the groups are identical (i.e. expert v. expert), the correlations are computed over a subset of questions that were rated in two passes, otherwise the correlations are computed over all questions.

Question Generator	Expert v. Expert	Non-Expert v. Non-Expert	Expert v. Non-Expert
Phoenix	0.490 (p=0.006)	0.311 (p=0.095)	0.341 (p=0.001)
Expert Tutors	0.418 (p=0.021)	0.497 (p=0.007)	0.440 (p=0.000)
Non-Expert Tutors	0.458 (p=0.011)	0.650 (p=0.002)	0.500 (p=0.000)

4. Results

To fairly compare scores between evaluators, question scores were normalized for differences among individual evaluators, so that the mean score for each evaluator was the average score of all evaluations for all questions ($\mu = 3.223$) and the standard deviation was 1.

A series of independent samples t-tests were performed (Table 1 and Figure 4) and found that questions from both classes of tutors significantly outscored the Phoenix system. The difference in ratings of expert-generated questions and non-expert-generated questions was not statistically significant, holding true for both expert evaluators and non-expert evaluators, though one annotator on average gave higher ratings to the Phoenix questions than expert questions.

To get a sense of inter-rater reliability, 90 of the questions were scored in a second pass by another pair of tutors. These were used to compute Spearman rank-order correlation coefficients (Table 2) between tutors of the same group (Expert v. Expert, Non-Expert vs. Non-Expert). The rating from the original 300 evaluations were used to compute correlations across groups (Expert v. Non-Expert). There was positive correlation between all combinations on all groups of questions.

Lastly, Table 3 illustrates the difference in perceived quality between transcript derived human tutor questions, the experimentally generated human tutor questions, and Phoenix generated questions. Independent sample t-tests (Table 3) found significant differences between ratings for experimentally generated questions by human tutors and for questions generated via the other two approaches.

Table 3. Independent Sample t-tests between ratings for questions generated under different conditions (experimental human tutor, transcript derived, and phoenix generated).

Evaluators	Q. Gen. Cond 1	Mean Score 1	Q. Gen. Cond 2	Mean Score 2	t	p-value
All Tut.	Experimental	3.715	Transcript	3.064	6.810	0.000
	Experimental	3.715	Phoenix	2.892	9.062	0.000
	Transcript	3.064	Phoenix	2.892	1.816	0.070

5. Discussion

The most significant finding was the difference in perceived quality between questions extracted from transcripts and questions generated during the experiment. There are at least four factors that could contribute to this difference. First, tutor question asking abilities may have improved since the time the tutorial sessions in the transcripts were conducted. Additionally tutors may not have felt time constrained when writing questions during the experiment like they would during a live tutoring session. Third, observer effects might have played a role, as tutors knew the questions were going to be evaluated. Lastly the evaluation context may be a factor.

In our experiment the evaluation condition perfectly matched the condition for generating the experimental human tutoring question – they each viewed the student dialog turn-by-turn without generating or viewing the tutoring dialog that elicited it, whereas the Phoenix and transcript-derived questions relied on the additional context of their own prior turns. If Phoenix or the transcript tutor already asked a question very similar to the higher-rated experimentally-derived human tutor question, they would likely ask a more specific question, which would be perceived as being of lower quality. As discussed in the next section, this confound could not easily be avoided in the present experiment and has significant implications for a QG challenge task.

Sparse grammar coverage was also a significant factor contributing to the large divergence in ratings between human tutor generated questions and system-generated questions. Without the proper mapping of student utterances to semantic frames, the system is unable to align to the appropriate context, and consequently, is less able to ask a pertinent question.

A confounding factor in this experiment is our inability to synchronize the system state with the focus of the human tutor over the time course of the dialog. Though it is a goal to have the system's state closely match the human tutor's state, there is no way to ensure this. To determine how divergence between system and tutor state could affect question ratings, we analyzed scores by number of student turns taken before arriving at the question of interest; however we found only a minor downward trend in the ratings of system-generated data as the context grew.

6. Issues in Running a Question Generation Challenge

The results above suggest that scores were highest when the evaluation conditions better aligned with the conditions under which the question was generated, indicating that extra effort must be taken to ensure fair evaluation when running a QG challenge. In designing this experiment we debated the merits and drawbacks of several options concerning the dialog context and initialization of our system. The main approaches we considered were as follows:

1. Original transcript dialog context	4. Manually initialize system state
2. Mixed transcript dialog context	5. Combine all student turns
3. System only transcript dialog context	6. Student turns only

In the *original transcript* approach the dialog context presented during QG and evaluation would have both the human tutor and student turns from the original tutoring session. We decided not to use this approach because there would be no way for the system to recognize, model and/or factor tutor turns into the QG process.

The *mixed transcript* approach would present the dialog context as an interleaving of student turns from transcripts with system-generated turns/questions. With this approach student turns would not be in dialog alignment with the system-generated questions and could potentially lead to unnatural and confusing dialogs.

The *system only transcript dialog context* approach follows the approach used in a bystander Turing test, where the transcript utilized is from a pure system tutorial session. This approach would have been the ideal way of evaluating our system, but we could not carry out this experiment because we have yet to collect tutorial session transcripts since incorporating intelligence into our system. This issue presents great difficulty for a QG challenge, where many systems are being evaluated and there is no single system that can provide the dialog context leading back to the problems associated with options one and two above.

With the *manually initialize system state* approach the system's internal state is manually set to reflect gold standard student understanding and dialog context. Evaluation and QG would be carried out using the full dialog context from the original transcript. While this context would have provided a meaningful way to isolate and evaluate the system's question selection capabilities, we did not have the time or infrastructure to carry out an experiment centered around this approach. Requiring knowledge like this in a QG challenge would lead to significant manual intervention and additional system building by participants and would likely invalidate many of the results.

Instead of inputting student utterances turn-by-turn the *combine all student turns* approach would concatenate all the student utterances and feed them into the system as a single turn. The major drawback to this method is the lack of contextual constraint. A combined utterance may cover several concepts meaning there is no single appropriate concept upon which to focus; presumably most of the prior dialog revolved around key concepts.

As stated before, we opted to forego the first five alternatives and use a dialogue context consisting of *student turns only* to avoid many of the issues associated with the other approaches, however in doing so we may have introduced the generating condition biases discussed in section 5.

These confounding factors and our results demonstrate the difficulty in defining a dialog-based based QG task. Systems that make question generating decisions based on their own previous questions will be at a disadvantage for such an evaluation. We believe that requiring systems to add logic to account for input from an arbitrary external QG source would add a significant barrier to participation in a challenge, requiring system logic that, for most, would never be used in an end-user application. Perhaps the only way to fairly evaluate such systems is to utilize extrinsic, application specific metrics, such as student learning gains in an intelligent tutoring task; though a large sample size is needed to account for any potential variability related to such a measure.

7. Future Work

We are very encouraged that there is no statistical difference in evaluation of our system and the transcript-derived human tutoring turns, which we believe has the most comparable generation condition. In fact, the difference between our system's performance and that of well-trained tutors in these circumstances is less than 1/5th of a standard deviation.

These results indicate that there is still significant opportunity to refine our system's QG capabilities. Improving the system's ability to choose appropriate context should allow future evaluations to use the full tutorial contexts instead of only the student turns. Additionally, having gold standard semantic parses would allow us to better evaluate the quality of our dialog system.

Since the participants in this study are familiar with the questions produced by the system, we would also like to conduct a similar experiment using tutors who are not acquainted with our system to provide more independent judgments. Additionally, it is unclear what role the QtA pedagogy played in this evaluation. Conducting this experiment using a different tutorial style may help to clarify this question.

One of the most significant contributions of this paper is the light it has shed on issues involved in running a QG challenge.

Acknowledgements

References

[1] I. L. Beck, M. G. McKeown, J. Worthy, C. A. Sandora, and L. Kucan. Questioning the author: A year long classroom implementation to engage students with text. *The Elementary School Journal*, 96(4):387–416, 1996.

[2] Lee Becker and Wayne H. Ward. Adapting a frame-based dialogue manager for use in tutorial dialogues. Technical report, University of Colorado Boulder, 2009 (forthcoming).

[3] S. Issar and W. Ward. Cmu's robust spoken language understanding system. In *Eurospeech '93*, pages 2147–2150, 1993.

[4] Rodney D. Nielsen. Question generation: Proposed challenge tasks and their evaluation. In Vasile Rus and Art Graesser, editors, *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, September 25-26 2008.

[5] Lawrence Hall of Science. Full option science system (foss). Nashua, NH, 2005.

[6] Natalie K. Person and Arthur C. Graesser. Human or computer? autotutor in a bystander turing test. In *ITS '02: Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, pages 821–830, London, UK, 2002. Springer-Verlag.

[7] Lucy Vanderwende. The importance of being important. In Vasile Rus and Art Graesser, editors, *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, September 25-26 2008.