Automating Comprehension Questions: Lessons from a Reading Tutor

Albert Corbett and Jack Mostow
Project LISTEN (www.cs.cmu.edu/~listen)
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{corbett, mostow}@cmu.edu

Abstract

How can intelligent tutors generate, answer, and score text comprehension questions? This paper proposes desiderata for such questions, illustrates what is already possible, discusses challenges for automated questions in Project LISTEN's Reading Tutor, and proposes a framework for evaluating generated questions.

1 Desiderata for Automated Questions

Experience with Project LISTEN's Reading Tutor (Mostow & Aist, 2001) suggests ideal properties for comprehension questions in intelligent tutors:

D1. Serve tutorial functions such as:
  a. Assess comprehension of text.
  b. Assess student engagement.
  c. Evaluate tutor interventions.
  d. Provide immediate feedback.
  e. Scaffold comprehension of text.
  f. Improve student engagement.
  g. Scaffold student learning.

D2. Generate questions automatically.
D3. Find correct answers automatically.
D4. Generate incorrect answers automatically.
D5. Score students' answers automatically.
D6. Predict question difficulty automatically.
D7. Target specific skills or knowledge.
D8. Be psychometrically valid and reliable.
D9. Maximize informativeness.
D10. Minimize time to ask, answer, and score.

Project LISTEN has automated questions with some of these properties to assess comprehension, vocabulary, and interventions (D1.a, D1.c) (Aist, 2001; Hensler & Beck, 2006; Mostow, Beck, Bey et al., 2004; Zhang, Mostow, & Beck, 2007), detect disengagement (D1.b) (Beck, 2005), and assist comprehension (D1.e) (Beck, Mostow, & Bey, 2004). This work explored various types of multiple-choice questions; we now discuss two.

2 Automatic Cloze Question Generation

To insert *cloze* (fill-in-the-blank) questions (D2), the Reading Tutor deleted a random word (D3) in the next sentence and chose three random, similarly difficult distracters from the text (D4): "*And the very next day, the ____ had turned into a lovely flower. – grain; lily; walnut; prepare."* The student had to identify the original word (D5); then the tutor showed the original sentence (D1.d).

Mostow, Beck, Bey, *et al.* (2004) established that such automatically generated cloze questions were valid measures of comprehension (D1.a): students' performance on the cloze questions, weighted by word and text difficulty, correlated significantly with a standard measure of reading comprehension, r = 0.85 (D8) – even though many of the randomly selected distracters violated syntactic and semantic constraints on the blank to fill in. Beck (2005) also found that student response times on cloze questions were a reliable indicator of student task engagement (D1.e).

3 Generic *wh*- Questions

Although cloze questions *measured* comprehension skill, Beck, Mostow & Bey (2004) found that they did not *scaffold* students' comprehension of text. However, another type did scaffold comprehension (D1.e) – text-independent multiple-choice *wh*- questions that can be introduced at any point in any text. For example, "*When does this take*

place? – in the present; in the future; in the past; it could happen in the past; I can't tell."

We did not use *wh-* items to test comprehension (D1.a) because we lacked a mechanism to compute the correct answers automatically (D3). Such a mechanism would suffice to score students' multiple-choice answers automatically (D5).

4 Question Generation Challenges

Our automatically generated cloze questions are ill-suited for deeper, fine-grained analysis of comprehension (D7). For this purpose we are designing questions by hand – multiple choice questions, so as to allow automatic scoring (D5) – using criteria that may inform automatic question generation. These design criteria (D1.a), shown in bold below, reflect models of comprehension processes and surface properties of the text.

Answers should depend on students' comprehension of the text during reading (D7). Students should not be able to eliminate distracters based only on syntactic knowledge or real world knowledge at test time (Keenan & Betjemann, 2006). For instance, the context "*the ____ had turned into a lovely flower"* enables ruling out *prepare* based on syntax, and *walnut* by knowing which plants flower.

A related principle is to **identify key information in the text** so that posttest questions hinge on comprehension rather than memory ability. **Questions should measure comprehension, not just the comprehender.** Testing what a reader gleaned from the text differs from testing reader attributes such as working memory (Duke, 2005).

We need literal questions that tap comprehension of explicit propositions in the text, and we need inferential questions that tap various processes (D7) that are part of comprehension, at both the lexical and clausal levels (Duke, 2005). **These text comprehension questions should tap reading time processes, not inferences at test time.** Suppose the student reads the sentence "*The cup tipped over and the ants sipped the bubbly sweetness."* To test whether the student inferred that the *bubbly sweetness* is *pop*, a later question asks "*What did the ants sip? – nectar; honey; juice; pop."* But if the question first repeats the sentence to refresh the student's memory, it might reflect inferences made at test time rather than while reading the original text.

Multiple choice distracters should be designed to **provide additional information** (D9), such as the extent of student miscomprehension. Having students rate all multiple choice alternatives for plausibility may measure deeper comprehension than the conventional procedure of picking a single correct answer (Pearson & Hamm, 2005). But even in the conventional procedure, distractors should be constructed so that errors provide information on the degree of misunderstanding. For example, consider a multiple choice question for a story about a moose: "*His antlers got in the way when he ___. -- slept; swam; ate; pulled things."* The correct answer is *slept*. The distracter *swam* was selected to be plausible, since the text mentions swimming, but not in conjunction with antlers. The distracter *ate* is less plausible, since it does not even occur in the text. The distracter *pulled things* reflects the worst comprehension failure, because the text explicitly says that antlers are useful for pulling things. Generating such questions requires the ability to **determine if text mentions, implies, omits, or contradicts an answer or distracter**.

Finally, to prevent frustration (D1.f), a tutor should **avoid asking questions that are too hard** for the given student.

Thus work on question generation can benefit by considering the processes whereby humans answer existing and prospective automated questions.

5 A Framework for Question Evaluation

Appropriate evaluation criteria for a question (whether human or automatic) depend on its purpose. In learning environments, the purpose of a question may be to test comprehension, to assist comprehension, to encourage reflection, to provide entertainment, to provoke discussion, or to improve learning. In other systems, questions may serve to guide diagnosis, to elicit user preferences, or to obtain information needed to perform a task. These are just a few examples; there are doubtless others. So on its face the enterprise of articulating common criteria to evaluate any question seems doomed, since criteria appropriate for one purpose may be altogether inappropriate for another.

Nonetheless, disparate evaluation criteria may share some underlying commonality. We propose the following framework as one way to think about how to evaluate questions. We assume that the question occurs in the context of some activity

with one or more goals. We can then evaluate the question by the extent to which it is expected to help the activity achieve a given goal. How much better (or likelier, or faster, or ...) should the goal be achieved than if the question had not been asked? To clarify, here are some examples.

In a reading comprehension test, the activity consists of reading some text, and the goal is to assess the reader's comprehension. Thus a question can be evaluated by its informativeness for that assessment. How much does the question increase the psychometric validity and reliability of the assessment? Of course this contribution depends on what other questions have been asked. For example, a question is unlikely to supply information about the reader's comprehension if the same question was just asked a moment ago.

If the purpose of a question is to assist reading comprehension, the activity still consists of reading some text, but the goal is to increase comprehension, not just assess it. So the question should be evaluated by how much better the reader understands the text if asked the question than if not.

If the purpose of a question is to improve learning, a question should be evaluated based on how much more learning occurs with it than without.

To take an example outside the realm of education, consider a spoken dialogue system intended to efficiently perform some task, such as planning a trip. A question can be evaluated based on the expected change to the overall duration of the dialogue, and to the quality of the resulting plan.

In sum, this framework is based on these ideas:

1. Questions occur in an activity with goals.
2. The value of a question relative to a goal is the expected difference in how likely, fast, or well (etc.) the goal is achieved if the activity includes the question than if it does not.
3. The value of a question may depend on what other questions are asked, and therefore impossible to evaluate in isolation.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070458 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We thank Nell Duke, Donna Gates, and Mike Heilman for comments and questions about the framework, and the LISTENers, educators, and children who contributed over the years to the work cited here.

References (many at www.cs.cmu.edu/~listen)

Aist, G. (2001). Towards automatic glossarization: Automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education, 12*, 212-231.

Beck, J. (2005, July 18-22). *Engagement tracing: using response times to model student disengagement.* Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005), Amsterdam, 88-95.

Beck, J. E., Mostow, J., & Bey, J. (2004, September 1-3). *Can automated questions scaffold children's reading comprehension?* Proceedings of the 7th International Conference on Intelligent Tutoring Systems, Maceio, Brazil, 478-490.

Duke, N. K. (2005). Comprehension of what for what: Comprehension as a non-unitary construct. In S. Paris & S. Stahl (Eds.), *Current issues in reading comprehension and assessment*. Mahwah, NJ: Erlbaum.

Hensler, B. S., & Beck, J. (2006, June 26-30). *Better student assessing by finding difficulty factors in a fully automated comprehension measure.* Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 21-30.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without Reading It: Why Comprehension Tests Should Not Include Passage-Independent Items. *Scientific Studies of Reading, 10*(4), 363-380.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B., & Valeri, J. (2004). Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning, 2*, 97-134.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices -- past, present and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's Reading Comprehension and Assessment* (pp. 13-69). Mahway, NJ: Erlbaum.

Zhang, X., Mostow, J., & Beck, J. E. (2007, July 9-13). *Can a Computer Listen for Fluctuations in Reading Comprehension?* Proceedings of the 13th International Conference on Artificial Intelligence in Education, Marina del Rey, CA, 495-502.