A Taxonomy of Questions for Question Generation

Rodney D. Nielsen[1], Jason Buckingham, Gary Knoll, Ben Marsh and Leysia Palen
[1] Boulder Language Technologies
[1] Center for Computational Language and Education Research, University of Colorado, Boulder
Rodney.Nielsen@Colorado.edu

Abstract

We define a question taxonomy based on prior educational research and the analysis of tutoring transcripts from multiple domains. We discuss how this taxonomy might be useful, with a focus on the application of automatic question generation during tutoring and for the purposes of educational assessment.

1 Introduction

The nature of automatic question generation is different depending on the task within which it is embedded. In summative assessment, questions are intended to evaluate the answerer's knowledge, understanding and skills. In Socratic tutoring, questions should lead students to an "aha" moment, where they understand a concept that they previously did not. Whereas, if the questions are being generated for a help system, say from a technical manual, the intent is for the question asker (the system) to learn what resulted in the request for help. The typical types of questions generated by each of these systems is different and we focus primarily on describing a question taxonomy for tutoring systems, which was validated via human tutoring transcript analyses.

2 Creating the Taxonomy

The goal of the project resulting in our taxonomy was to detail human computer interaction design issues involved in creating an automated Socratic tutor. To achieve this goal, we started with current research on tutoring strategies and then analyzed human tutoring transcripts in six subject areas, elementary school reading comprehension, elementary school binary math, middle school algebra, and college-level research methods, computer programming and conceptual physics. The key design output was a mapping from a classification of learner interactions to an appropriate tutoring response. This included elaborate taxonomies for both the learner and tutor; here we only present the majority of the part of the tutor response taxonomy associated with question generation.

This question branch of our taxonomy started with the list of question types described by Graesser and Person (1994), adapted from Lehnert (1978). We added question types from Collins (1985) and Bloom's Taxonomy of Educational Objectives (1956). Then through an iterative process of analyzing and annotating portions of the transcripts, we revised the taxonomy until we felt each dialog turn was accurately and sufficiently annotated.

The approach involved four annotators analyzing transcripts from the six subject areas and independently defining a Learner Dialogue Act Taxonomy and a Tutor Response Taxonomy. We then met and revised the taxonomies considering redundancy, usefulness, and completeness. Where two taxonomic labels had consistent meanings, we merged them. When categories differed, we re-evaluated their usefulness and sought the most generic categories that would still facilitate effective tutoring. When alternative learner or tutor responses were probable and no label existed, a new category was added. This included adding categories that covered research in pedagogical theory that we deemed valuable for effective tutoring, but that were not necessarily found in the transcripts (e.g., Collins' question types). The key factor involved in deciding to add, change or delete a category was whether the change in information available to the automated tutor would facilitate more effective tutoring and learning. This process was repeated, with the four annotators labeling

separate transcripts according to the taxonomies and revising the taxonomies as necessary to classify the new dialogues, until the taxonomies comprehensively covered all of the tutor and learner responses examined.

Finally, we analyzed the patterns in the labeled transcripts and created a mapping from learner dialog acts to tutor responses based on the taxonomic labels we had assigned to each dialogue turn in the transcripts. The part of the Tutor Response Taxonomy that pertains to questions (Fig. 1) consists of a primary taxonomy of question classes supplemented by several secondary orthogonal taxonomic dimensions (e.g., Collins' Question Type, Bloom's Taxonomy, Content Level, etc.)

Primary Taxonomy

1. Description Questions
1.1. Concept Completion: Who, what, when, where?
1.2. Definition: What does X mean?
1.3. Feature Specification: What features does X have?
1.4. Composition: What is the composition of X?
1.5. Example: What is an example of X?
2. Method Questions
2.1. Calculation: Compute or calculate X.
2.2. Procedural: How do you perform X?
3. Explanation Questions
3.1. Causal Antecedent: What caused X?
3.2. Causal Consequence: What will X cause?
3.3. Enablement: What enables the achievement of X?
3.4. Rationale Questions
3.4.1. Goal Orientation: What is the goal of X?
3.4.2. Justification: Why is X the case?
4. Comparison Questions
4.1. Concept Comparison: Compare X to Y?
4.2. Judgment: What do you think of X?
4.3. Improvement: How could you improve upon X?
5. Preference Questions
5.1. Free Creation: requires a subjective creation.
5.2. Free Option: select from a set of valid options.

Collins' Question Type (Collins, 1985): Form hypothesis, Test hypothesis, Make prediction, Trace consequences, Entrapment, or None.

Bloom's Taxonomy of Educational Objectives, top level (Bloom, 1956): Knowledge, Comprehension, Application, Analysis, Synthesis, or Evaluation.

Content Level: Indicates the amount of answer content included in the question and, thus, the level of hint provided to answer a question, on a scale from 0.0 (no answer content) to 1.0 (question provides the answer).

Example Usage (adapted from Collins, 1985): Positive Paradigm Case, Negative Paradigm Case, Negative Exemplar for a Necessary Factor (near miss), Positive Exemplar for an Unnecessary Factor (near hit), Generalization Exemplar for a Factor (maximal pair), Differentiation Exemplar for a Factor (minimal pair), Exemplar to Show the Variability of a Factor, Exemplar to Show the Variability of the Dependent Variable, Counterexample for Insufficient Factors, Counterexample for Unnecessary Factors, Analogy, Continuation of Example, Reuse of Example, None.

Response Form: This indicates the type and length of the expected response: Boolean, Multiple Choice, Word, Phrase, Sentence or Paragraph.

Question Relation: This indicates the relationship of the current question to preceding questions (the question level also affects the relation information; next at level 2 refers to the next logical question at that level versus the next logical question at the current level): Aside, Initiate Example, Initiate Scaffolding, Initiate Multipart Synthesis, Next, Previously Asked, Recast, Subpart, Subpart of Next, Subpart of Previous.

Connection Question: asked without an expectation of response as a precursor to asking the real question.

Figure 1. The Question Taxonomy

3 Discussion

The most significant deviations from (Graesser and Person, 1994) were the addition of secondary dimensions and the hierarchical structure of the primary taxonomy. We also added question classes and moved some classes to secondary dimensions (e.g., Verification and Disjunctive questions were incorporated into Response Form). In prior taxonomies, it is possible to assign a question to two or more categories, but we felt that where the classification was associated with a very different aspect of the question, such as its expected response form for verification and disjunctive questions, it would be easier to classify and to describe a question using secondary dimensions. This is particularly true when the dimension must always be specified, as with the Response Form, and therefore, would always require two classifications in prior taxonomies.

We attempted to balance the taxonomy's simplicity and power. Too few classes could lead to monotonous ineffective tutoring dialogs, while too many would increase system complexity and lead to difficulties in analyzing corpora and assessing the impact of tutoring response types on learning gains. To satisfy these constraints, we chose to create multiple taxonomy dimensions, each with relatively few class labels. A small number of classes

should lead to easier classification within a given dimension, while the usage of multiple dimensions will lead to a more complete specification of tutoring interactions.

Most current question generation systems (e.g., Mitkov et al., 2003; Brown et al., 2005; Rus et al., 2007) have focused primarily on Bloom's Knowledge level, the subclasses of *Description Questions* described in the Primary Taxonomy, and Boolean, multiple choice and word responses. We would expect early question generation challenges to have a similar focus, since these are the most computationally feasible questions to generate and assess. Subsequent challenges could progress across Bloom's taxonomy and include some Method and Comparison questions, with later challenges incorporating Explanation questions, question series, and eventually questions from Collins' categories and questions that incorporate examples.

While we feel this progression is consistent with the technical challenges involved, we believe all of the question types in the primary taxonomy can, under restricted conditions, be generated based on today's technologies. For example, it is reasonable to explicitly ask a system to generate a Causal Consequent question from the text snippet *The bell produced a sound, because it was vibrating*, since the sentence has strong linguistic cues and it is a given that such a question is reasonable. On the other hand, systems would likely perform very poorly if asked to generate several causal consequent questions given a full document, since even strong linguistic cues such as *because* are often misleading. It is also the case that successfully generating a meaningful question is no guarantee of system effectiveness, as is the case where there is no way to assess the quality of an answer to a judgment question.

This taxonomy could be useful in a number of ways in the question generation challenge. First, it could be used in the main task to specify the type of question that systems should generate from a text snippet (Nielsen, 2008). Second, if the overall question generation task is conceived of as consisting of Concept Selection, Question Type Determination, and Question Construction (Nielsen, 2008), then the output of the Type Determination task could be one or more of the labels from the primary taxonomy. Specifically, given a key concept, an application track, and ultimately a context, the task would be to list the most appropriate types of questions to generate. Such a taxonomy could also be used to classify questions, for example, as part of an automatic question evaluation process.

This taxonomy was generated from analyses and classification of human tutoring dialogues with the intent to utilize it in an Intelligent Tutoring system as a means of defining the most appropriate type of response to a learner turn. The goal is to construct this response from a text on the subject being tutored, but it could similarly be generated from other sources such as a knowledge base describing the key concepts of the subject matter.

Finally, while this taxonomy is geared primarily toward tutoring, it would be almost equally applicable in educational assessment. We also believe most of the *primary* taxonomy is relevant in virtually all areas requiring question generation, though this would require further analysis.

Acknowledgments

We thank Steve Bethard and the anonymous reviewers for helpful comments on this paper. This work was partially funded by IES Award R305B070434.

References

Bloom, B. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain.* New York, NY, Longmans.

Brown, J, Frishkoff, GA, and Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. In *Proc. HLT/EMNLP*.

Collins, A. (1985). Teaching reasoning skills. In S. Chipman, J. Segal, & R. Glaser (eds.) *Thinking and Learning Skills, Vol. 2*. Hillsdale, NJ: Erlbaum.

Graesser, AC. and Person, NK. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104-137.

Lehnert, WG. (1978). *The Process of Question-Answering*. Hillsdale, NJ: Erlbaum.

Mitkov, R. and Ha, LA. (2003). Computer-aided generation of multiple-choice tests. In *Proc. HLT/NAACL, Workshop on Building Educational Applications Using NLP*.

Nielsen, R. (2008). Question generation: Proposed challenge tasks and their evaluation. In *Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, VA.

Rus, V, Cai, Z, and Graesser, AC. (2007). Experiments on Generating Questions About Facts. In *Proc. CICLing*.