# Using Automatic Question Generation to Evaluate Questions Generated by Children

**Wei Chen[1], Jack Mostow[1], Gregory Aist[2]**

[1]Project LISTEN, School of Computer Science, Carnegie Mellon University, USA

[2]Communication Studies and Applied Linguistics, Department of English, Iowa State University, USA

{weichen, mostow}@cs.cmu.edu, gregory.aist@alumni.cs.cmu.edu

**Abstract**

This paper shows that automatically generated questions can help classify children's spoken responses to a reading tutor teaching them to generate their own questions. We use automatic question generation to model and classify children's prompted spoken questions about stories. On distinguishing complete and incomplete questions from irrelevant speech and silence, a language model built from automatically generated questions out-performs a trigram language model that does not exploit the structure of questions.

## Introduction

Why generate questions automatically? There have been rich discussions on how to generate and evaluate questions, but only a few papers have demonstrated its practical use, such as in dialog generation (Piwek and Stoyanchev, 2010) and writing support (Liu et al., 2010). This paper introduces a new use: classifying children's responses to self-questioning prompts. The National Reading Panel (NRP, 2000) identified self-questioning as the single most effective reading comprehension strategy to teach – that is, teaching children to ask themselves about text as they read it, as opposed to teachers asking questions, except as examples to demonstrate the self-questioning strategy. In this paper, we generate questions to model and score children's responses to self-questioning prompts.

Many researchers in the question generation community have identified taxonomies of questions (Boyer et al., 2009; Forăscu and Drăghici, 2009; Graesser et al., 2008; Kalady et al., 2010; Nielsen et al., 2008). These taxonomies suggest guidelines both for generating questions and for evaluating questions asked by humans. Ultimately we hope to classify children's spoken questions according to these taxonomies. However, before classifying questions, we need to know whether a response contains speech; if so, whether the child uttered a question; furthermore, whether the question is complete. Therefore, we classify children's responses into four categories: a complete question (e.g., *I wonder how the cool cat survives in the cold and snowy place*), a partial question (e.g., *I wonder how Tony will*), off-task speech (e.g., *I can't wait till this book is over please tell me this book is over now*), and no response.



### Research Platform

The data for this paper come from a self-questioning activity in Project LISTEN's Reading Tutor (Mostow and Beck, 2007). The activity proceeds as follows. As a child reads a story, our Reading Tutor occasionally intervenes before displaying the next sentence and prompts her to ask questions about what she has read. This spoken prompt, prerecorded by an adult (like all Reading Tutor prompts), is *What are you wondering about now?* The child then responds by speaking into a close-talking, noise-cancelling headset microphone. The Reading Tutor records the child's speech and provides rudimentary feedback based on the amount of speech, storing the child's utterances in a database for future analysis. Our ultimate goal is to provide tutorially valuable feedback. As a step towards this goal, we classify children's responses based on the linguistic content and acoustic features of the recorded utterance.

The rest of the paper is organized as follows. We first describe our approach to question generation and how we use the generated questions to model children's questions. We next describe how we use this language model (LM) to classify children's self-questioning responses. We then present evaluation methods and results. Finally we conclude and point out future work.

## Generating Questions to Model Children's Self-Questioning Responses

We generate questions to predict children's responses to the self-questioning prompt. Like Gates (2008), we use off-the-shelf natural language processing tools to annotate text and generate questions from the annotations. In

particular, we generate questions by filling in question templates. Since the prompt includes the word *wonder*, we expect children to follow similar phrasing too. So all the question templates begin with the phrase *I wonder* or *I'm wondering*. The remainder of the question templates depend on the information requested.

Template 1:
`I wonder | I'm wondering`
`how|why|if|when <THING> <VERB-PHRASE>.`
Example:
*I wonder how wind makes electricity.*

Template 2:
`I wonder | I'm wondering`
`who|what <VERB-PHRASE>.`
Example:
*I wonder what lives on Mars.*

We generate items to fill in `<THING>` and `<VERB-PHRASE>` by running the ASSERT semantic role labeler (Pradhan et al., 2008) on the story text. We extract text marked by the tag [ARG0] (verb argument in front of the verb) to fill in `<THING>`. We combine text marked by [TARGET] (verb) and [ARG1] (verb argument after the verb) to fill in `<VERB-PHRASE>`.

To predict children's speech, we need a LM to set constraints on vocabulary and word order. We do not have sufficient training data to train the LM on children's spoken questions. Therefore, we use the automatically generated questions as a synthetic corpus to build the LM. In particular, we construct a probabilistic finite state grammar (PFSG) that incorporates the generated questions, with equal probabilities for the transitions from each state.

The coverage of the PFSG is limited. We deal with this problem along three dimensions. First, to improve coverage of the LM, we added the Dolch list (Dolch, 1936) of 220 words common in children's books. We expected children's questions to be about story text, so we added all the story words. We used a morphology generator to add all inflections of each verb. We use the resulting vocabulary for the interpolated LM which we describe now. Second, to make the LM for children's questions more robust, we interpolate the PFSG with part of speech (POS) bigrams. We train the bigrams from a POS corpus generated from 673 children's stories from Project LISTEN's Reading Tutor. The stories contain 158,079 words. We use the Stanford POS tagger (Toutanova et al., 2003) to find the POS of the words in the stories. We first train a bigram model using the SRILM toolkit (Stolcke, 2002). To incorporate this model in the PFSG, we add a state for each POS tag. We add a transition from states immediately preceding `<VERB-PHRASE>` to the VB (verb) state and assign it a heuristic probability .0001, and transitions between POS states their POS-bigram probabilities. We tag each word with its most frequent POS. Thus this model approximates Pr(drink the milk) as .0001 * Pr(DT | VB) * Pr(NN | DT). Third, to cover responses that are not questions (i.e., off-task speech), we interpolate the LM with trigrams trained from a corpus of off-task speech with back-off. We assign the interpolation weight a heuristic probability 0.001. Therefore the model approximates Pr(go to the bathroom) as .001 * Pr(go) * Pr(to | go) * Pr(the | go to) * Pr(bathroom | to the). The off-task speech corpus consists of transcriptions of children's off-task speech during oral reading. To avoid overfitting, we restrict the vocabulary for the trigrams to the 200 most frequent words (of which the top 10 are *I*, *you*, *it*, *to*, *the*, *what*, *on*, *go*, *this*, and *that*) in the off-task corpus. We refer to them as "off-task words" and to the rest of the vocabulary as "on-task words."

## Classifying Children's Self-Questioning Responses

We use a hierarchical approach to classify children's responses to *What are you wondering about now?*: First decide whether an utterance involves self-questioning. If yes, then decide whether the utterance contains a complete question. Otherwise decide whether there is any speech response at all. Figure 1 shows this classification hierarchy.

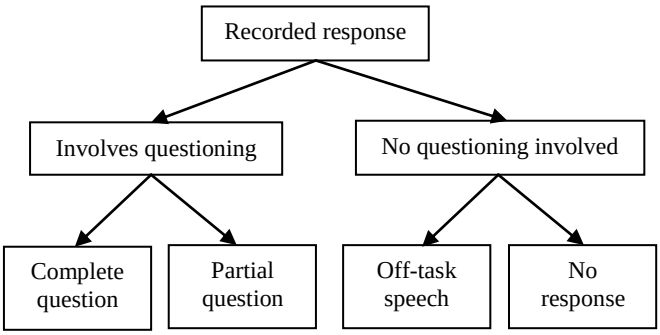


**Figure 1**. Hierarchical classification of children's responses to self-questioning prompts.

To determine whether a child's response involves questioning, we describe each response with a feature vector, and use a support vector machine (SVM) to classify the response. The feature vector has two types of features: lexical features characterize the linguistic content of the response; acoustic features characterize its speaking style.

We extract lexical features from the output of automatic speech recognition (ASR) of a child's response using the Sphinx3 speech recognition system (CMU, 2010), which outputs a time-aligned sequence of hypothesized words, each with a acoustic confidence score whose value ranges from a large negative integer to a large positive integer. Two important components of the speech recognizer are

the LM and the acoustic model. To predict the content and word order of children's questions, we use the LM described in the previous section. To map sound to phonemes, we use an acoustic model trained on 43 hours of 495 children's oral reading to the Reading Tutor.

To capture difference of word distribution in questioning and off-task speech, we extract three features from the ASR output: (1) the percentage of words on our list of off-task words, (2) the percentage of off-task words whose ASR confidence scores fall above a threshold, and (3) the percentage of *on*-task words whose confidence scores fall *below* the threshold. We set the threshold to zero because on the development set of 200 held-out oral reading utterances by 8 children we used to tune it, a value of zero achieved equal error rates for rejecting correctly recognized words and accepting misrecognized words.

We select 50 acoustic features described elsewhere (Chen and Mostow, 2011). We concatenate the lexical features and the acoustic features to form the feature vector describing a response. Because we do not have enough children's self-questioning responses to train the SVM, we train it on 36,492 manually transcribed oral reading utterances. The 2-level classifier operates as follows.

If the first level classifies a response as self-questioning, classify it as a complete question if either of the following two conditions holds: (1) the ASR output for the response contains a question word (*who*, *what*, *when*, *why*, *how*, or *if*) followed by a word sequence marked as a verb phrase by running ASSERT on the ASR output; or (2) it contains a question word followed by a variant of the word *be* (*is, was, are, were, be, being, been*). If ASSERT fails to parse the ASR output, classify the utterance as a question fragment.

If the first level does not classify the response as self-questioning, and the ASR output contains at least one word whose confidence score exceeds the threshold, then classify the response as off-task. Otherwise, classify it as silence (i.e., no response), based on the assumption that it consists solely of background speech or noise.

## Evaluation

Our test data consists of 250 responses to the self-questioning prompt by 34 children ages 7-10 while reading 10 stories. The median number of responses by a child is 5.

The coverage of our LM directly affects ASR accuracy. We measure the coverage using the out-of-vocabulary (OOV) rate computed as the percentage of transcribed word tokens not included in the LM. In general, an OOV word causes at least one ASR error, typically two. The OOV rate for our LM is 19.2%. In comparison, fully 32.7% of the word tokens in the spoken responses do not occur in any of the generated questions.

To analyze accuracy of our response classifier, we classify the transcribed responses by hand into the four categories defined earlier. Of the 250 responses, 146 contain complete questions, 11 contain question fragments, 69 are off-task speech, and 24 contain no speech response at all. As a comparison, we train a trigram LM from 673 stories and the off-task corpus described above. The vocabulary consists of story words and 200 off-task words. Table 1 shows the accuracy of our response classifier according to the human classification. Recall of a category is defined as the percentage of responses in the category that are classified correctly. Precision is defined as the percentage of responses classified as belonging to a category that actually do. On the first level of classification, the LM built from automatically generated questions out-performs the baseline trigram LM in both recall and precision. The errors made by the classifier on the first level propagate to the second level classification. Therefore we present only the accuracy of the four-way classification using the LM built from generated questions.

**Table 1**. Classification accuracy

a. Accuracy in distinguishing questioning vs. non-questioning.

| | LM | Recall | Precision |
|---|---|---|---|
| **Involving questioning** | **QG+interpolation** | 0.59 | 0.85 |
| | ***General trigram*** | 0.55 | 0.80 |
| **Off-task speech and no response** | ***QG+interpolation*** | 0.83 | 0.55 |
| | ***General trigram*** | 0.76 | 0.50 |

b. Accuracy on the four way classification

| Evaluation criterion | Complete question | Partial question | Off-task | No response |
|---|---|---|---|---|
| **Recall** | 0.45 | 0.36 | 0.84 | 0.75 |
| **Precision** | 0.94 | 0.10 | 0.53 | 0.58 |

## Conclusion and Future Work

This paper describes a new application for question generation, namely to serve as an intermediate tool for modeling and categorizing children's spoken responses, at least when their responses are supposed to be self-questions about the text they are reading. We addressed two research challenges: (1) How to automatically assess children's spoken questions? and (2) How to evaluate automatically generated questions?

The answer to the first question is still in its rudimentary stage. For example, we do not attempt to classify which responses are inferential questions, which require reasoning about the text and hence are likely to boost reading comprehension. But we should at least tell which utterances are valid responses to the self-questioning prompt. In particular, our classifiers decide whether an utterance involves questioning at all; if so, whether the question is complete; and if not, whether the response is off-task or absent altogether. These decisions are made

harder by the ungrammaticality of children's speech and by the noise and background speech in school environments.

Because our question generation is tied to a specific application, so is our evaluation of generated questions. In particular, we tested the coverage of the generated questions on children's responses, and we tested the overall classification accuracy based on lexical features of the ASR output and acoustic features on the speech signal.

Future work includes improving classification accuracy on children's responses, especially in distinguishing among finer-grained categories of questions. Our ultimate goal is to use automated classification of children's spoken responses to an intelligent tutor in order to inform its decisions about what feedback to give, so as to teach more effectively such useful skills as a self-questioning strategy to improve reading comprehension.

## Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070458. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute and the U.S. Department of Education. We thank Joseph Valeri and Donna Gates for annotating our data, and the schools and children who provided it by using the Reading Tutor.

## References

Boyer, K. E., Lahti, W., Phillips, R., Wallis, M., Vouk, M., and Lester, J. 2009. An Empirically Derived Question Taxonomy for Task Oriented Tutorial Dialogue. In *Proceedings of the 2nd Workshop on Question Generation*, Brighton, UK.

Chen, W. and Mostow, J. 2011. A Tale of Two Tasks: Detecting Children's Off-Task Speech in a Reading Tutor. In *Proceedings of the Interspeech*, Florence, Italy.

CMU. 2010. CMU Sphinx: Open Source Toolkit For Speech Recognition, from http://cmusphinx.sourceforge.net/

Dolch, E. 1936. A basic sight vocabulary. *Elementary School Journal, 36*, 456-460.

Forăscu, C. and Drăghici, I. 2009. Question Generation: Taxonomies and Data. In *Proceedings of the 2nd Workshop on Question Generation*, Brighton, UK.

Gates, D. 2008. Generating Look-Back Strategy Questions from Expository Texts. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, NSF, Arlington, VA.

Graesser, A., Rus, V., and Cai, Z. 2008. Question Classification Schemes. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, NSF, Arlington, VA.

Kalady, S., Elikkottil, A., and Das, R. 2010. Natural Language Question Generation Using Syntax and Keywords. In *Proceedings of the Third Workshop on Question Generation*, Pittsburgh PA.

Liu, M., Calvo, R. A., and Rus, V. 2010. Automatic Question Generation for Literature Review Writing Support. In *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems*, Pittsburgh, PA.

Mostow, J. and Beck, J. 2007. When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. *Scale-Up in Education, 2*, 183-200.

Nielsen, R., Buckingham, J., Knoll, G., Marsh, B., and Palen, L. 2008. A Taxonomy of Questions for Question Generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, VA.

NRP. 2000. Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: http://www.nichd.nih.gov/publications/nrppubskey.cfm.

Piwek, P. and Stoyanchev, S. 2010. Question Generation in the CODA Project. In *Proceedings of the Third Workshop on Question Generation*, Pittsburgh, PA.

Pradhan, S., Ward, W., and Martin, J. H. 2008. Towards Robust Semantic Role Labeling. *Computational Linguistics Special Issue on Semantic Role Labeling, 34*(2), 289-310.

Stolcke, A. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the Intl. Conf. Spoken Language Processing*, Denver, Colorado.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the HLT-NAACL*, 252-259.