A Generative Model for Scoring Children's Reading Comprehension

Joseph Tepperman
Signal Analysis and Interpretation Laboratory
University of Southern California
Los Angeles, USA
tepperma@usc.edu

Matteo Gerosa
FBK
Fondazione Bruno Kessler
Via Sommarive 18
Trento, Italy
matgero@gmail.com

Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory
University of Southern California
Los Angeles, USA
shri@sipi.usc.edu

ABSTRACT

The use of speech technology in children's reading assessment can help teachers to diagnose reading difficulties and plan appropriate interventions for a large number of students. We present a Bayesian Network model of student reading comprehension that can be used to estimate automatic scores for a child's spoken answers to open-ended questions about a text. Through the use of features derived from language models capturing different degrees of comprehension, we found that on the TBALL dataset we could achieve 0.8 correlation with reference comprehension scores derived from teachers, exceeding the teachers' own correlation with this same reference. This student model also proved to perform without bias due to a speaker's native language, which was not the case for a comparable baseline method, nor for the teachers themselves.

1. INTRODUCTION

As children grow into more advanced readers, their teachers place a stronger emphasis not just on the fundamental skills of decoding sounds from text, but on the longer-term goal of all literacy education: comprehension of what a text and its sounds together mean. The latter skill is to some degree a consequence of the former, of course - children who struggle through sounding out individual words are likely to miss the point of a passage as a whole [5]. Some children, however, in spite of normal word-level decoding skills, still have trouble understanding main ideas in a text, and even when listening to spoken material [8]. To diagnose independent problems in these related areas would be the goal of every conscientious reading teacher.

Students can demonstrate reading comprehension independently of word reading in many ways, but perhaps the most telling is for them to respond to open-ended questions about a passage in their own words. Based on the child's pronunciation, choice of words, and manner and rate of speaking, teachers can infer the child's degree of comprehension, and can adjust their judgments according to prior knowledge of the child's background, the difficulty of the questions, notions of expected right and wrong answers, and many other factors. However, the way this inference is synthesized from all the available cues is not only very complex but not entirely transparent to the teacher who does it. The problem remains poorly-defined from a perceptual point of view, and is subjective enough to prohibit complete agreement among teachers [8].

This paper presents an extension of the work in [3] toward automatically assessing a child's answers to these open-ended reading comprehension questions. Due to the complex interaction among the various sources of evidence and prior knowledge, we propose using a generative model for student comprehension based loosely on the Bayesian Network classifier first proposed in [7]. The point of using automatic methods is not to replace teachers, but to provide them with the tools to plan individualized interventions for a large group of students over a number of discrete reading skill sets, with a minimum of teacher time or energy invested in actually conducting these assessments. Our goal, then, is to infer a child's degree of comprehension as we hypothesize a teacher would, so that our automatic scores agree with teachers as well as teachers can agree among themselves.

2. CORPUS

The children's speech data used in this study was composed of native and bilingual English speakers recorded at Los Angeles public schools as part of the TBALL project [1]. Under real classroom conditions, 33 first-graders and 37 second-graders were prompted by an animated user interface to read a short paragraph out loud. The material was the same for all students, though unique to each grade level. To test their comprehension of this paragraph, students were asked to answer 8 yes/no questions and 3 open-ended questions. Each of five elementary school teachers then listened to recordings of these 210 open-ended answers and scored each one on a three-point scale: complete comprehension (1), zero comprehension (0), or partial comprehension (0.5). The reference score for each particular answer was taken as the quantized mean of the scores from all five teachers: mean scores below 0.25 were quantized to 0, those above 0.75 were quantized to 1, and everything in between was set to 0.5. The average inter-teacher correlation in these scores was 0.752, and each teacher's average correlation with the quantized mean score from the other four was 0.794 - this justifies using the quantized mean as a reference, since it correlates better than a sixth teacher would. Each child's utterance was also transcribed on the word level, and we collected statistics about each child's native language (L1) through surveys of their parents - 39 of the children were reported as native Spanish speakers, 21 as native English speakers, and for the remaining 10 this statistic was missing, possibly because the

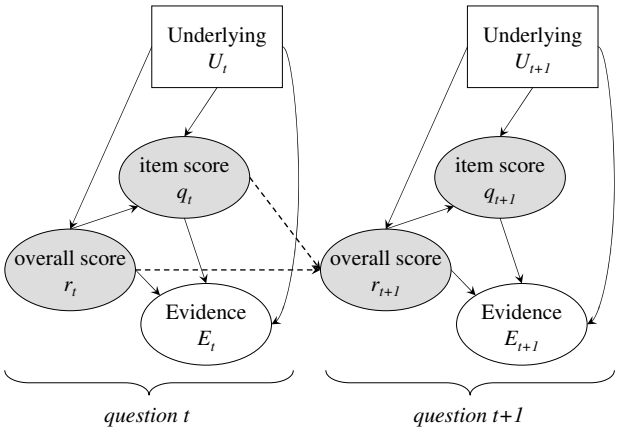

Figure 1: A high-level graphical illustration of the open-ended question student comprehension model. Shaded nodes denote Hidden variables. The dashed lines are not probabilistic relations, but indicate how the overall comprehension score for question $t+1$ is derived from the combined scores of the previous question.

parents chose not to respond.

3. THE STUDENT MODEL

In this section we describe the student model we used to calculate a score representing a child's degree of reading comprehension. It consists of two components: one, the standard automatic speech recognition (ASR) models that estimate the Evidence to be used by the second component, the Bayesian Network that generates an overall score based on that Evidence and other knowledge about the child and the test. The ASR acoustic and language models used to estimate this Evidence are essentially models for teacher perception and prior notions of "correct" answers - the Bayesian Network that combines them is the student model proper.

3.1 Acoustic and Lanuage Models

Our acoustic models were standard three-state left-right Hidden Markov Models for context-dependent phonemes. With 32 Gaussian mixtures per state, these models were estimated using standard 39-dimensional MFCC features including first- and second-order derivatives. These were trained both on the TBALL data (different speakers from those in Section 2's open questions dataset) and about 16 hours of children's speech from a comparable corpus available from OGI [4]. The models were further refined to speaker-adapted ones using a text-independent method based on CMLLR. We trained a baseline bigram language model on text from the two read passages and all transcripts of answers to the questions. This model was then adapted three separate times - with transcripts of the answers demonstrating complete comprehension, those demonstrating zero comprehension, and those for partial comprehension - resulting in three different bigram models, one for each type of answer. Combining these acoustic models with the baseline language model, Viterbi decoding of the open-ended questions' answers resulted in 14.1% WER compared to the manual transcripts.

3.2 Comprehension Model

The available variables that teachers might use in making a judgment of a student's reading comprehension were divided into three categories: Evidence, Underlying variables, and Hidden variables. A *Hidden* variable is one that can only be inferred by the teacher - in this case, the child's degree of comprehension - and is modeled as one causal root of the Evidence seen as clues to that comprehension. We propose two Hidden variables: q_t, the item-level comprehension for question t - a discrete variable of cardinality 3 (complete, partial, and zero comprehension, trained on the answers elicited from teachers as described in Section 2). The other is r_t, a continuous-valued running score that represents overall comprehension and is estimated at each test question as the mean of the scores from all previous questions (and is initialized as $r_1 = 0$). The item-level Hidden state was modeled as conditionally dependent on this overall comprehension variable, based on the assumption that the child's comprehension on, for example, the third question, may be informed by their performance on the previous two.

By *Evidence*, E_t, we mean anything the teacher might observe directly at the time of the child's response, e.g. their rate of speaking, their word choice, etc. Using the acoustic and language models described in Section 3.1, we calculated three Evidence features by first using Viterbi decoding to estimate likelihood scores of the observed MFCC features, O, given each of the three language models: $P(O|M_1)$, $P(O|M_{0.5})$, and $P(O|M_0)$. We then used these likelihoods to estimate three posterior probabilities, one for each language model M_L:

$$P(M_L|O) = \frac{P(O|M_L)P(M_L)}{\sum_n P(O|M_n)P(M_n)} \quad (1)$$

Here the priors for each model, $P(M_L)$, were assumed to be equal. These posteriors represented an individual answer's distance from the three sets of comprehension levels captured in the language models. This use of language model posteriors as Evidence of comprehension was based on the simple idea that different types of answers (complete, zero, or partial comprehension) would exhibit unique distributions of n-gram word strings, and so models trained separately on each could be used to automatically score unknown text. The model with the highest likelihood estimated from Viterbi decoding served as another Evidence variable:

$$\hat{M} = \underset{L}{\operatorname{argmax}}\, P(O|M_L) \quad (2)$$

This was a discrete value representing the best comprehension level if a teacher were asked to pick just one of the three. One last Evidence variable was the child's rate of speaking (ROS), defined as the number of phonemes recognized per second.

Underlying variables, U_t, are the ones that might influence our expectations of the child's Hidden comprehension states and also perhaps of the Evidence, e.g. their grade level, demographic information, or the difficulty of the test question. Included in our Underlying variables were the following discrete statistics: the child's grade (1st or 2nd), gender, L1 (either Spanish or English), and the assessment text (one of two) and question index for that text (out of three), assuming that the questions and texts might vary in difficulty.

Table 1: Item- and student-level overall score correlation between automatic results and reference teacher scores, over various scoring methods

	baseline	*Bayes Net Student Model*					*teacher*
	recognition	*only* $\hat{M}$	*no* E_t	*no* U_t	*no* r_t	*all variables*	*agreement*
Item-level correlation	0.784	0.777	0.401	0.800	0.717	0.693	0.794
Student-level correlation	0.846	0.829	0.301	0.809	0.701	0.684	0.828

Our hypothesized model dependencies, shown graphically in Fig. 1, propose to unite all these variables in a way that reflects how we hypothesize a teacher would conceive of reading comprehension. The child's cognitive degree of comprehension, q_t, is the source of the observed Evidence, E_t, but Underlying variables, U_t, might inform our expectations of the cognitive state and the Evidence as well. Inference on the degree of comprehension for item t was then calculated as follows:

$$\begin{aligned} & P(q_t|E_t, U_t, r_t) \\ = \; & P(q_t, E_t, U_t, r_t)/P(E_t, U_t, r_t) \\ = \; & \frac{P(q_t|U_t, r_t)P(E_t|q_t, U_t, r_t)P(r_t|U_t)P(U_t)}{P(E_t|U_t, r_t)P(r_t|U_t)P(U_t)} \\ = \; & \frac{P(q_t|U_t, r_t)P(E_t|q_t, U_t, r_t)}{P(E_t|U_t, r_t)} \end{aligned} \quad (3)$$

where a final score for answer t was defined as

$$\begin{aligned} Sq_t \; = \; & 1 * P(q_t = 1|E_t, U_t, r_t) \\ & +0.5 * P(q_t = 0.5|E_t, U_t, r_t) \end{aligned} \quad (4)$$

This proposed network structure is similar to (and inspired by) the Knowledge Tracing model of procedural knowledge acquisition [2], in which a student's observed answers to test questions (the Evidence) may be the result of their knowledge state (a Hidden variable) or scaffolding on the part of a tutor (an Underlying variable), but this tutoring can affect future inference on the Hidden knowledge state as well, if the information is actually taught beyond just scaffolding the answer.

The network in Fig. 1 was implemented using BNT [6] with all continuous nodes modeled as Gaussian distributions, all discrete nodes with no continuous parents modeled as probability tables, and all discrete nodes with continuous parents modeled as softmax functions. All Evidence variables and the item-level Hidden comprehension state, q_t, were modeled as conditionally dependent on all Underlying variables, but the overall comprehension variable r_t was modeled as only dependent on the Underlying variables that applied globally: grade, gender, L1, and the text. Table 2 shows these hypothesized dependencies in more detail.

4. EXPERIMENTS AND RESULTS

Our experiments were intended to address the following questions:

- How does the proposed student comprehension model compare to a simpler recognition-based baseline?
- What is the relative importance of the novel features proposed in this work?
- Is it possible to improve results by automatically refining the hypothesized network structure?
- How do these automatic results compare to teacher agreement with the reference scores?

To address these questions we divided the available data into 7 subsets separated by speaker - all results are reported in terms of a 7-fold crossvalidation on the entire dataset. Features and scores were estimated using the methods outlined in Section 3. The baseline recognition-based method was simply to take the Evidence variable $\hat{M}$ representing the Viterbi-decoded maximum-likelihood estimate of comprehension and use that for a score as it was elicited from the teachers in Section 2: 1 for complete comprehension ($\hat{M} = M_1$), 0.5 for partial comprehension ($\hat{M} = M_{0.5}$), and 0 for no comprehension ($\hat{M} = M_0$). For comparison with this baseline, we generated automatic scores from the Bayesian Network student model using different subsets of features: just the baseline $\hat{M}$ as the only feature, and using the whole set but separately leaving out each of the novel feature categories (E_t, U_t, and r_t). Correlation of these automatic scores with the reference scores is reported in Table 1, both on the item and student levels, where student-level scores were defined as the sum of their three item-level scores.

Some of the variables thought to be dependent may not in fact be, and with finite training instances and an overly complex model there is always the possibility that true dependencies might not be estimated properly due to a dearth of training instances representative of all combinations of dependent variables - this proved to be the case for the hypothesized structure outlined in Section 3. For these reasons we propose a forward-selection greedy search algorithm to refine the network structure. The algorithm begins with just one arc in the network representing a baseline dependency: the arc from q_t to the comprehension recognition Evidence variable, $\hat{M}$. Then it proceeds in a random order to add each hypothesized arc individually, keeping an arc if it improves the likelihood of the training variables given the Bayesian Network. This process is looped until it has been shown that adding any remaining hypothesized arc will decrease the model likelihood. This method is also useful in that analysis of the refined network may reveal the true dependencies present in the data. The likelihood of the training set given the model was defined as the log-likelihood after EM convergence, and convergence was defined as either 10 iterations of EM or the number of iterations required to make the following inequality true:

$$\frac{|LL(i) - LL(i-1)|}{\text{mean}\{|LL(i)|, |LL(i-1)|\}} < 0.001 \quad (5)$$

Table 2: Using the forward selection procedure outlined in Section 4, these are the total number of times each of the hypothesized network dependencies was selected for the final refined network, summed over 7 crossvalidation training sets. Cells left empty were not part of the original hypothesized structure - these parent/child combinations were assumed to be independent

		Children						
				E_t				
Parents		r_t	q_t	*ROS*	M_1	$M_{0.5}$	M_0	$\hat{M}$
U_t	*index*		7	7	7	7	7	7
	text	7	7	7	4	7	7	3
	L1	7	7	7	7	7	7	7
	gender	7	7	7	7	7	7	7
	grade	5	4	7	7	5	5	7
	r_t		0	7	7	7	7	0
	q_t			7	7	7	7	7

Here $LL(i)$ is the log-likelihood after iteration i. Ordinarily, EM on our data met this inequality within 7 iterations.

5. DISCUSSION

The first thing to notice about Table 1 is that the baseline method's correlation is already at the levels of teacher agreement - the differences in these correlation coefficients was not statistically significant using the one-tailed test for difference in correlation. This was due partly to the use of speaker adaptation in the acoustic models - without it, this baseline method had 0.713 item-level and 0.741 speaker-level correlation, and these are significantly less than the baseline results reported in Table 1 with $p \leq 0.05$. The Bayesian Network with only $\hat{M}$ as a feature performed almost identically to the baseline, as we might expect.

Including only the rest of the Evidence and Hidden variables in the network (the "no U_t" column in Table 1) did not change the correlation results significantly. We did find that leaving out the Underlying variables, U_t, was beneficial, probably because the relative sparsity of the data would not allow for these nodes' parameters to be trained reliably. Omitting just the overall comprehension variable, r_t, did significantly degrade performance compared to the "no U_t" set ($p \leq 0.02$), but this is possibly also from including the U_t features. However, leaving out the Evidence, E_t, worsened performance most dramatically - clearly the Evidence were the most important of the cues to reading comprehension proposed in this work.

Interestingly, the refined version of the full network structure did not improve correlation at all. This is curious since the forward selection procedure explained in Section 4 only allowed for hypothesized conditional dependencies in the final network structure if having them improved the likelihood of the feature set given the network. This suggests that each of the 7 crossval training sets was too small for an improvement in their likelihoods to translate to improved correlation in their respective test sets. Table 2 gives the counts for how many times each hypothesized dependency was selected for the network, and in that we see a couple of trends. First off, any dependencies that would have required softmax distributions (i.e. discrete children with continuous parents) were avoided entirely - continuous r_t was not allowed as a parent of discrete q_t or $\hat{M}$. However, r_t was not omitted as a parent variable entirely, so this must speak to a general problem with adequately estimating the parameters of the softmax distribution. We also see that the grade and the text variables were sometimes excluded as parent variables. This can be explained in that they are probably redundant when used together, since each text passage was generally given exclusively to only one grade level of student.

The baseline method correlated with the reference scores more for native Spanish-speaking than for native English-speaking students, and for 1st graders more than for 2nd graders (both with $p \leq 0.05$). The latter was true for the "no U_t" results, but the difference in performance based on native language was not significant at this confidence level. The same was true for possible score bias: both the baseline method and the teacher evaluators gave significantly higher scores to native English over native Spanish speakers, and to 1st graders over 2nd graders ($p \leq 0.05$ using a one-tailed test for difference in proportions). The "no U_t" Bayesian Network mirrored this difference in proportions based on grade level, but not on L1. Thus the student comprehension model is, in terms of native language, less biased than both the baseline method and expert evaluators.

6. CONCLUSION

This work presented a student model intended for automatically scoring children's reading comprehension based on their recognized responses to open-ended questions about a reading passage. With three different language models - each adapted to answers demonstrating complete, partial, and zero comprehension, respectively - we found that we could use maximum-likelihood Viterbi decoding to estimate reading comprehension scores that correlate with teacher-derived reference scores as well as teachers themselves do. Using extra features (including pronunciation evidence, child demographics, and prior knowledge of the test) in a Bayesian Network student model framework did not improve upon these results, possibly due to a sparsity of training data. We did find, however, that the best-performing student model's automatic scores were not biased in favor of either native language, whereas this could not be said for the baseline method of comprehension recognition, nor for the teacher evaluations. Future work in this area would see the most benefit from using more features related to pronunciation and language model Evidence - the few proposed here were found to have much more predictive power in automatic scoring than any of the Underlying demographic or test-related variables.

7. REFERENCES

[1] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang. A system for technology-based assessment of language and literacy in young children: the role of multiple information sources. In *Proceedings of International Workshop on Multimedia Signal Processing*, Chania, Crete, October 2007.

[2] K.-M. Chang, J. Beck, J. Mostow, and A. Corbett. A Bayes Net toolkit for student modeling in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, pages 104–113, Jhongli, Taiwan, 2006.

[3] M. Gerosa and S. Narayanan. Investigating automatic assessment of reading comprehension in young children. In *Proceedings of ICASSP*, Las Vegas, Nevada, April 2008.

[4] J. H. K. Shobaki and R. Cole. The OGI KIDS' speech corpus and recognizers. In *Proceedings of ICSLP*, Beijing, China, October 2000.

[5] G. R. Lyon. Towards a definition of dyslexia. *Annals of Dyslexia*, 22:3–30, 1995.

[6] K. Murphy. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, 33, 2001.

[7] J. Tepperman, M. Black, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan. A Bayesian Network classifier for word-level reading assessment. In *Proceedings of InterSpeech ICSLP*, Antwerp, Belgium, August 2007.

[8] B. Wise and L. Snyder. *Identification of Learning Disabilites: Research to Practice*. Lawrence Erlbaum, Mahwah, NJ, 2002.