Can a Computer Listen for Fluctuations in Reading Comprehension?

Xiaonan ZHANG[1], Jack MOSTOW, and Joseph E. BECK
Project LISTEN, School of Computer Science, Carnegie Mellon University

Abstract. The ability to detect fluctuation in students' comprehension of text would be very useful for many intelligent tutoring systems. The obvious solution of inserting comprehension questions is limited in its application because it interrupts the flow of reading. To investigate whether we can detect comprehension fluctuations simply by observing the reading process itself, we developed a statistical model of 7805 responses by 289 children in grades 1-4 to multiple-choice comprehension questions in Project LISTEN's Reading Tutor, which listens to children read aloud and helps them learn to read. Machine-observable features of students' reading behavior turned out to be statistically significant predictors of their performance on individual questions.

Keywords. Reading comprehension, automated assessment, children's oral reading, cloze questions, speech recognition, Reading Tutor

1. Introduction

Reading has widespread importance in intelligent tutoring systems, both as a means of instruction, and as an important skill to learn in its own right. Thus the ability to automatically detect moment-to-moment fluctuations in a student's reading comprehension would be of immense value in guiding an intelligent tutoring systems to make appropriate instructional decisions, such as estimating student knowledge or determining what points to explain further. This paper explores detection of comprehension fluctuations in Project LISTEN's Reading Tutor [1], which uses speech recognition to listen to children read aloud, and responds with various feedback.

Human tutors track students' comprehension by asking comprehension questions. Similarly, to test students' comprehension, the Reading Tutor occasionally inserts a multiple-choice cloze question for the child to answer before reading a sentence [2]. It generates this question by replacing some word in the sentence with a blank to fill in. Here is an example. The student picks the story *Princess Bertha and the Lead Shoe*. Some time later the Reading Tutor displays the following sentence for the child to read aloud: *"I must go get Herb the Horse before he runs off again," Princess Bertha exclaimed.* The Reading Tutor turns the next sentence into a cloze question, which it displays and reads aloud: *With that, Princess Bertha took off for her* ____. The Reading Tutor then reads the four choices: *gift; friend; lesson; horse*, and waits for the child to click on one of them. It says whether the student chose the right answer,

[1] Corresponding Author: CMU-LTI, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA; E-mail: xiaonanz@cs.cmu.edu.

defined as the original text word, namely *horse*. Then it displays the correctly completed sentence for the student to read aloud. The three distractors are words of approximately the same difficulty as the right answer, selected randomly from the same story. Thus such questions are entirely automatic to generate, administer, and score. Cloze questions test the ability to tell which word fits in a given context. Performance on these automatically generated cloze questions correlates well with scores on a standard test of reading comprehension ability [3].

However, inserting too many comprehension questions wastes time, disrupts the flow of reading, and annoys students. To address this problem, this paper investigates the following question: can a computer listen for fluctuations in comprehending a text? Specifically, can it detect comprehension fluctuations by listening to the student read the text aloud, and tracking help requests in computer-assisted reading?

We answer this question in the context of the Reading Tutor, by using students' observable reading behavior to help predict their performance on individual cloze questions that reflect their fluctuating comprehension. We assume that comprehension, reading behavior, and cloze performance are related as shown in Figure 1.

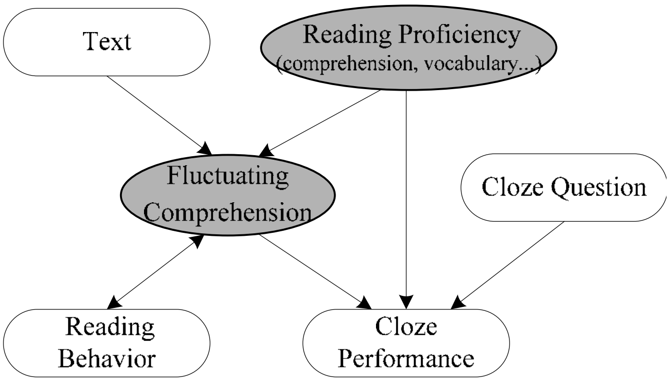

Figure 1: Conceptual model. An arrow from X to Y indicates that X influences Y. Shaded nodes represent hidden entities, while others are observable.

According to this model, comprehension fluctuates as a student with a given level of reading proficiency reads some text whose difficulty for that particular student varies over the course of the text. That is, we model reading proficiency as a relatively stable *trait* that changes only slowly over time, and comprehension as a fluctuating *state*.

The student's fluctuating comprehension is not directly observable, but has observable effects. Comprehension affects (and is affected by) the student's reading behavior, such as speed, prosody, and requests for tutorial assistance. Comprehension also affects the student's performance on cloze questions inserted in the text – but it is not the only influence. Other influences include the student's proficiency as well as the difficulty of the cloze question.

If observable reading behavior really reflects comprehension fluctuations, then it should enable us to predict cloze performance more accurately than if we use only the proficiency of the student and the difficulty of the cloze question. The central task of this paper is to test this hypothesis.

This work is novel because it focuses on detecting fluctuations in text comprehension, in contrast to previous work on assessing students' general reading comprehension skill [2, 4]. Some work [5] has tracked changes in students' reading skills, but focused on decoding rather than on comprehension, and on growth over time rather than on momentary fluctuations. Other work [6, 7] has used speech-based features to detect fluctuations in cognitive load, but not in reading comprehension. Finally, a substantial body of work [8] has studied transitory comprehension processes, but using eye tracking rather than speech input.

The rest of the paper is organized as follows. Section 2 describes a statistical model of student performance in terms of the conceptual model in Figure 1. Section 3 tests this model and explains the results. Section 4 discusses limitations and potential future directions. Section 5 concludes by summarizing contributions.

2. A Statistical Model of Reading and Cloze Performance

We build on a previous statistical model [9] that focused only on the rightmost three nodes in Figure 1. By accounting for differences in question difficulty, this model was able to infer students' proficiency from their cloze performance more accurately than their raw percentage correct on cloze questions could predict. We extend this model by adding information about students' reading behavior.

The form of our model is Multinomial Logistic Regression, as implemented in SPSS 11.0. Its binary output variable is whether the student answered the cloze question right. We now describe its predictor variables.

Reading proficiency: To capture between-student differences in reading proficiency, we include student identity as a factor in the model, as in [9]. The student identity variable is a unique code assigned by the Reading Tutor when it enrolls a student. This variable also subsumes any other student traits that might affect cloze performance, such as motivation. Including student identity as a factor controls for individual differences between students, and accounts for statistical dependence among responses by the same student instead of treating them as independent [10].

Cloze question difficulty: We adopt the same predictor variables used in [9] to represent features that affect the difficulty of cloze questions. One such feature is the length of the question in words. Another feature is the number of choices with the same part of speech as the right answer. For instance, our example cloze question (*With that, Princess Bertha took off for her__.* [*horse*]) would be easier if all distractors had the wrong part of speech, *e.g.*, *finish*, *come*, and *ordinary*. As a student-specific indicator of difficulty, the model also includes the student's response time to answer the cloze question. Response time reflects student engagement [11] and confidence; hasty response times indicate guesses [2], while long response times reflect uncertainty.

Reading behavior: The new variables in our model characterize students' reading behavior, both their oral reading and the tutorial assistance they receive, when they read the cloze sentence with the right missing word filled in. To derive these variables, we first defined some 30 features that we thought might reflect comprehension fluctuations. These features describe or relate to the words being read, the prosody of the oral reading, the time-aligned output of the speech recognizer, and the assistance given by the Reading Tutor.

The individual features are defined in terms specific to the Reading Tutor and unwieldy to explain. None of them by itself is a dramatically effective indicator of comprehension. They are not independent, and in fact some of them correlate strongly with each other. Moreover, 30 features is enough to pose a risk of overfitting.

To solve this problem, we ran Principal Components Analysis on the 30 features, using the SPSS 11.0 implementation including rotation of components to increase interpretability. We selected the top five components (which together explain about 65% of the variance in the data) because the sixth and subsequent factors account for much less variance. We include these five components (more precisely, standardized component scores of each case on each of these five components) as covariates in our model. As is typical, the principal components do not translate to simple constructs. However, the general nature of each component is easy to describe, and may have analogues in other intelligent tutoring systems. We therefore describe each component in general terms. Space and clarity preclude a comprehensive (and incomprehensible!) list of raw features. Instead, we illustrate each component with one or two underlying features that have loading greater than 0.4 (the component loadings are the correlation coefficients between the features and components). The top five components are:

1. **Sentence coverage:** *e.g.*, the percentage of a text sentence read by the student
2. **Fluency**: *e.g.*, the number of words read per second, and the number of letters in a sequence of words read without pausing longer than 0.5 seconds
3. **Words per utterance**: *e.g.*, the number of text words per utterance accepted by the Reading Tutor as read correctly, and the number of utterances per sentence
4. **Rereading**: *e.g.*, the number of times a sentence is read
5. **Truncation**: *e.g.*, the number of incomplete words output by the recognizer

We also experimented with features related to fundamental frequency (F0), including average, maximum and minimum F0, the slope of the F0 curve, and the pitch range, computed as maximum F0 minus minimum F0. However, they turned out to be insignificant predictors of cloze performance, so we removed them from the feature set.

3. Evaluation

We now evaluate how well the model and various ablated versions of it fit empirical data logged by the Reading Tutor during the 2002-2003 school year. This data includes 7805 cloze responses (72.1% of them correct) by 289 children, ranging from grade 1 to grade 4, at seven public schools in the Pittsburgh area. The logged reading behavior of these students includes the audio files of each student's utterances, the output of the speech recognizer, and the actions of the Reading Tutor, such as help on hard words.

Table 1 summarizes the results of fitting the model to this data set. The table lists the predictor variables in the model, one per row, grouped into the categories discussed in Section 2 above: reading proficiency, cloze question difficulty, and reading behavior. Successive columns of the table show the 8-character variable name used in SPSS; what it represents; a chi-square statistic for the variable (the difference in -2 log-likelihoods between models with and without the variable); the degrees of freedom for the variable; and its statistical significance in the model.

As Table 1 shows, all three groups of variables contain significant predictors, including two of the five composite variables that describe reading behavior. The respective β coefficients of these five variables are .254, .240, .092, -.044, and .008, reflecting their relative strength. The sign represents their qualitative effect on the chances of answering the cloze question right. Only FAC1 and FAC2 are statistically significant. Their β values reflect positive effects for reading more of the target sentence and reading more fluently.

Table 1. Significance of predictor variables (from SPSS Likelihood Ratio Tests)

Category	Predictor variable	Description of predictor variable	Chi-square	df	Sig.
Reading proficiency	USER_ID	Student identity	655.198	288	.000
Cloze question difficulty	Q_RC_LEN	Length of cloze question in characters	3.650	1	.056
	PERC_Q_P	% of sentence preceding deleted word	8.630	1	.003
	DIFFICUL	Difficulty of target word (4 categories)	64.241	3	.000
	TAG_POS	Target word part of speech	17.268	4	.002
	TPOS_INT	# legal parts of speech of target word	0.009	1	.926
	TAG_PR_M	Target has its usual part of speech	4.718	1	.030
	CONF_POS	# distractors with target part of speech	.140	1	.708
	RTBIN9	Response time (one of 10 bins)	89.571	9	.000
Reading behavior	FAC1	Sentence coverage	43.921	1	.000
	FAC2	Fluency	42.570	1	.000
	FAC3	Words per utterance	2.359	1	.125
	FAC4	Rereading	1.842	1	.175
	FAC5	Truncation	0.073	1	.788

Table 2. Comparison of full and ablated models

Model name	Reading proficiency (student identity)	Cloze question variables	Reading behavior variables	Nagelkerke's R^2	Adjusted R^2	Classification accuracy (%)		
						all	right	wrong
Full model	√	√	√	0.247	0.209	75.6	92.6	31.3
ID+cloze [9]	√	√		0.229	0.190	75.5	93.7	28.2
ID+reading	√		√	0.212	0.175	74.9	93.7	26.3
ID-only	√			0.172	0.134	74.2	95.4	19.3
Cloze-only		√		0.073	0.070	72.7	98.2	6.7
Reading-only			√	0.069	0.068	72.1	97.4	6.3

Table 1 shows the significance of the individual variables. But to understand how the groups of variables contribute to the model, we compare the full model shown in Table 1 to various ablated models that omit one or more of these groups. Table 2 summarizes these models, one model per line, starting with the full model and listed in order of decreasing fit. The first four columns of the table describe which groups of variables the model includes. Nagelkerke's R^2 in logistic regression quantifies the explanatory power of the model, much as R^2 does in linear regression, but with lower typical values. Cross-validation to test generality is problematic due to the student identity variable, so to estimate model fit on unseen data, we modify Nagelkerke's R^2 to take into account the sample size and the number of features in the model: we compute adjusted R^2 as *1 - (1 - Nagelkerke's* R^2*) × (N-1) / (N-K-1)*, where N is the number of observations and K is the number of features. Categorical variables with M values count as $M - 1$ features, *e.g.,* student identity counts as 288 features. The low R^2 values here reflect the difficulty of predicting individual responses, compared to predicting aggregate performance averaged over many test items. Finally, classification accuracy is simply the percentage of cases for which the model correctly predicts whether the student got the cloze question right, *i.e.*, number of correct predictions / number of cloze items. We disaggregate by whether the student was right.

3.1. Is reading behavior informative?

Comparing the models in Table 2 shows how much reading behavior helps in predicting cloze outcomes. To penalize overfitting, we use adjusted R^2 to measure model fit. Comparing the full model to the ID+cloze model used in [9], we see that reading features uniquely explain 0.019 of the variance, increasing model fit by 10% relative (from 0.190 to 0.209). Comparing Reading-only and Cloze-only shows that reading features have almost as much explanatory power on their own as cloze features do (0.068 versus 0.070). In other words, listening to the student read (and get help) is roughly as informative as looking at the cloze question to predict cloze performance.

Comparing the fit of the ID+reading and ID-only models shows how much variance is uniquely explained by reading behavior, *i.e.*, not by student identity. To compute this non-overlapping portion of explained variance, we subtract the 0.134 adjusted R^2 for the ID-only model from the 0.175 adjusted R^2 for the ID+reading model. The result, 0.041, constitutes the bulk of the 0.068 adjusted R^2 explained by the reading behavior variables alone. Thus they capture something that student identity does not.

3.2. What construct do reading features measure, and how does it vary?

The analysis in Section 3.1 above shows that reading features pick up something. But what is this "something?" We hope that it's local fluctuation of comprehension, yet it might be some other construct that we do not want to measure. To characterize the construct captured by our reading behavior variables, we analyze how it varies.

First, does this construct really vary over time, or not? That is, is it a state or a trait? There are at least two reasons to believe it's a state. One reason is that the logistic regression model already includes student identity, which ought to capture any between-student differences that affect cloze performance – that is, student traits. Another reason is the relatively small overlap (0.027) in the variance explained by ID-only (0.134) and by Reading-only (0.068). We would expect almost complete overlap if the reading behavior variables just measured a student trait.

Second, does the construct fluctuate from one sentence to the next? Yes. We compared the "cloze sentence" Reading-only model to a variant that used reading features from the sentence just before the cloze question. We tested both models on a subset of 6901 cloze questions preceded by least three sentences in the same story. Adjusted R^2 was only 0.026 for the "previous sentence" model, versus 0.069 for the cloze sentence model. Reading factors uniquely explained only 0.018 of the variance, versus 0.042 in the cloze sentence model. The second and third sentences before the cloze question were even weaker predictors than the sentence just before the cloze question. Thus reading behavior on the cloze sentence itself predicts performance on the cloze question better than behavior on some other sentence.

Third, does the construct measure comprehension – or merely some artifact of answering the cloze question? For example, a student who gives a wrong cloze answer expects a different sentence completion than the right one, and might read it differently (*e.g.*, less fluently) for that reason. We cannot entirely rule out this possibility. However, the fact that reading behavior on sentences preceding the cloze question is a significant (albeit weak) predictor of cloze performance provides hope that reading behavior on the completed cloze sentence also reflects comprehension.

4. Limitations and Future Work

We have used the rather loaded word "comprehension" to refer to the hidden state reflected by our reading behavior variables. The justification for using this word is that the state fluctuates from one sentence to the next, and explains variance in cloze performance not explained by the identity of the student or the difficulty of the cloze question. An alternative possibility is that the variables reflect some other state that affects cloze performance, such as student engagement in reading the story. However, engagement seems intuitively less likely than comprehension to fluctuate markedly from one sentence to the next, especially since a prior analysis of cloze behavior found that student engagement appeared relatively stable across a five-minute time scale [11].

Although we have shown that students' reading behavior variables explain variance in their cloze performance, explaining 7% of the variance is not enough for us to rely solely upon these reading features to detect comprehension fluctuations. Thus the work presented here is only a proof of concept, not a demonstration of feasibility. To make the method practical, it would be necessary to increase the model accuracy. The low R^2 may be attributed to errors in the speech recognition from which we derive some of our features, or to the inherently obscure relationship between reading and comprehension, which may be especially tenuous for more skilled readers. Possible solutions include exploiting additional prosodic clues such as accentuation and intonation, or combining reading features with other observations about the student.

However, perhaps the most glaring limitation of the current approach is the nature of the "training labels" that relate reading behavior to comprehension. A cloze question tests comprehension of a specific sentence – but it is a destructive test. Turning a sentence into a cloze question eliminates the opportunity to observe how the student would have read the sentence otherwise. The student's reading behavior might be scaffolded by seeing and hearing the cloze question first – or worse, affected differentially depending on whether the cloze answer was right or wrong.

Eliminating this limitation will require some other way to test comprehension of individual sentences. An obvious solution is to write comprehension questions by hand, and determine their difficulty empirically by administering them to a norming sample of students. However, this approach is labor-intensive. Project LISTEN uses multiple choice cloze questions because we can generate and score them automatically [2], and predict their difficulty [9]. The work reported here used the resulting data because it was available, not because it was ideal.

5. Conclusion

This paper is about analyzing students' reading behavior to detect fluctuations in their text comprehension automatically. We showed that machine-observable features of reading behavior improved a statistical model of students' performance on cloze questions inserted in the text. Behavior on the completed cloze sentence was a stronger predictor than behavior on the sentences preceding the cloze question, suggesting that the reading behavior features are sensitive to fluctuations in comprehension.

This paper makes several contributions. First, we define the problem of detecting local fluctuations in text comprehension, which differs from the previously studied problem of assessing students' overall reading comprehension ability. Second, we propose a simple but useful conceptual model of the relationships among students' overall proficiency, reading behavior, fluctuations in text comprehension, and performance on cloze questions. Third, we translate this conceptual model into a statistical model by using Principal Components Analysis to derive useful predictors from system-specific features of reading behavior. Both the model and the methodology are potentially applicable to speech-enabled intelligent tutoring systems in other domains. Fourth, we evaluate the model on real data from Project LISTEN's Reading Tutor. Specifically, for this dataset, we show that behavioral indicators of comprehension predict student performance almost as well as question difficulty does.

This paper is a step toward future intelligent tutoring systems that detect comprehension fluctuations by listening to their students read aloud, thereby improving their estimates of what their students know. Such a capability could enhance the quality of tutorial decisions about what and how to teach.

Acknowledgements

This work was supported in part by the National Science Foundation under ITR/IERI Grant No. REC-0326153. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government. We thank the educators, students, and LISTENers who generated and processed our data

References

1. Mostow, J. and G. Aist. Giving help and praise in a reading tutor with imperfect listening -- because automated speech recognition means never being able to say you're certain. *CALICO Journal*, 1999. *16*(3): p. 407-424.
2. Mostow, J., J. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning*, 97-134. 2004.
3. Woodcock, R.W. *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. 1998, Circle Pines, Minnesota: American Guidance Service.
4. Beck, J.E., P. Jia, and J. Mostow. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, 2004. *2*: p. 61-81.
5. Beck, J.E. and J. Sison. Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *International Journal of Artificial Intelligence in Education*, 2006. *16*(2): p. 129-143.
6. Berthold, A. and A. Jameson. Interpreting symptoms of cognitive load in speech input. *UM99, 7th International Conference on User Modelling*, 235-244. 1999. Wien, New York: Springer.
7. Müller, C., B. Groβmann-Hutter, A. Jameson, R. Rummer, and F. Wittig. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. *UM 2001, 8th International Conference on User Modeling*, 24-33. 2001. Sonthofen, Germany.
8. Rayner, K., K.H. Chace, T.J. Slattery, and J. Ashby. Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading*, 2006. *10*(3): p. 241-256.
9. Hensler, B.S. and J. Beck. Better student assessing by finding difficulty factors in a fully automated comprehension measure. *8th International Conference on Intelligent Tutoring Systems*, 21-30. 2006. Jhongli, Taiwan.
10. Menard, S. Applied Logistic Regression Analysis. *Quantitative Applications in the Social Sciences*, 1995. *106*.
11. Beck, J. Engagement tracing: using response times to model student disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005)*, 2005: p. 88-95.