Automatically Generating and Validating Reading-Check Questions

Christine M. Feeney and Michael Heilman

[1] University of Virginia, Charlottesville, Virginia, USA
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Abstract. We describe a method used in the REAP tutor to automatically check whether students are reading practice texts for English vocabulary learning. The number of texts used by the REAP tutor and challenges of automatically generating reading comprehension questions necessitated a simpler type of question that could be computer-generated. We describe an algorithm for generating such reading-check questions for arbitrary texts. Two studies investigating the utility of these questions found reliable, moderate correlations with vocabulary learning and reading comprehension, respectively.

Key words: reading comprehension, question generation, language tutoring

1 Introduction

Intelligent tutoring systems often include a text instruction component. This component may involve a simple introduction or review of basic material prior to the practice tasks that constitute the main instructional program. Texts may also be integrated with practice, especially in language tutoring systems. For example, a tutoring system for the English article system[1] asks students to identify and correct missing or incorrect articles. The REAP tutor for English as a Second Language vocabulary[2][3] also uses texts for instruction by presenting authentic texts containing target words so that students receive implicit information about vocabulary words from the context in which those words appear. The REAP system coordinates this passive implicit information from real texts with explicit information from dictionary definitions as well as interactive input from practice exercises.

It is often desirable for a tutoring system that employs text material to estimate the probability that the student learned from a text he or she has just read. Estimates of comprehension can be useful for many purposes: as a type of experimental data, as input for a learner model, as an indicator that the student should be asked to re-read a text more carefully, or as an indicator that the system should select easier texts. One possible measure is the deep comprehension question. Such questions often require inference and integration of the concepts represented in the text. They often appear on standardized tests

of verbal ability, such as the Scholastic Aptitude Test[3]. Another type of measure is the shallow comprehension question, which requires the recall of facts and ideas from the text, but usually does not require inference. A grade school teacher might use such questions on a reading quiz. A third type of measure, which we call a reading-check question, is simpler than a shallow comprehension question. For this type of question, the student does not have to recall specific facts or events from the text but rather surface features of the text-for example, the specific lexical items that are present. A specific type of reading-check question is described in detail below.

Ideally, a system would employ deep comprehension questions, since they are expected to be the most strongly associated with learning from text. However, generating such questions is extremely challenging, if not impossible, with current human language technologies. Even generating shallow comprehension questions is a challenging task. Systems have been designed for generating comprehension questions (e.g., [4]), but these are not simultaneously error-free, domain-general, and completely automatic. Ideally, systems would have error-free questions, but in many tutoring systems complete automaticity or domain generality can be sacrificed. For instance, if the system only uses a few texts, then a human can easily check for and edit any errors in generated questions.

Some tutoring systems, however, require automaticity because they employ a large number of texts covering a variety of topics. The REAP tutor, for example, employs a corpus of millions of texts across a range of reading difficulty levels in order to provide individualized practice for hundreds of target vocabulary words. It would not be feasible for a human to check and revise reading comprehension questions for all of the texts used by REAP. Therefore, the simpler reading-check questions are employed because they are essentially error-free and thus can be completely automated. This paper outlines a method for generating reading-check questions for texts on various topics. Thereafter, the paper describes an in vivo experiment conducted with the REAP tutor that measured whether student performance on these questions correlates with REAP's primary objective, vocabulary learning. The paper then describes a lab experiment that measured the correlation of performance on reading-check questions to shallow comprehension questions.

1.1 Automatic Generation of Reading-Check Questions

The automatically generated reading-check questions ask the student to choose, from among set of foils, a small set of words that appear in a particular text that he or she has read. Common words such as "the," "this," or "from" are ignored, for two reasons. Such words appear in almost any text and are thus unlikely to have any strong association with the topic being discussed in a particular text. Instead, the method extracts salient content words that are informative about the topic of the text.

[3] http://collegeboard.com/

For a given text, the algorithm first extracts a list of unique words appearing in the text. It then calculates a measure of the salience or information present in each of these words as discussed shortly. The top N words are then chosen for the answer set ($N = 8$ in these experiments). Foils are generated by taking the answer set and randomly replacing half of the words with randomly chosen words that did not appear in the text.

The salience of a word in a given text is measured by the relative frequency of the word in the text minus the relative frequency of the word in general English, divided by the relative frequency of the word in general English. Dividing by the frequency of a word in general ensures that the algorithm avoid very common words such as the. In this work, the relative frequency of a word in general was estimated from a corpus of approximately 50,000 texts. These texts were gathered from the web, cover a variety of topics, and were used by the REAP tutor in previous studies. However, any model of the relative frequency of words in general English would probably suffice. The salience measure is defined in Equation (1), where $S_j(w_i)$ is the salience of the word w_i in the jth text in a collection of texts, V is the size of the set of unique words, D is the number of texts in the collection, and $count_j(w_i)$ is the number of times that w_i appears in the jth text.

$$S_j(w_i) = \frac{\frac{count_j(w_i)}{\sum_{k=1}^{V} count_k(w_i)} - freq(w_i)}{freq(w_i)} \tag{1}$$

where

$$freq(w_i) = \frac{\sum_{m=1}^{D} count_k(w_i)}{\sum_{m=1}^{D}\sum_{k=1}^{V} count_k(w_k)} \tag{2}$$

The REAP tutor applies a few other constraints when choosing the words for answer sets. First, the tutor avoids target vocabulary words that it is explicitly trying to teach. Second, the system only considers the top 5000 most common words in the collection of texts. The tutor thus avoids rarer words including proper names and words that would be unfamiliar to second language learners. Figure 1 shows an example reading-check question from a text about international politics. One thing to note is that these questions are designed for texts approximately 500-2000 words long. The method would probably not apply to very short texts with fewer than 100 words. Also, it would probably have to be extended for longer texts, perhaps by dividing the text into contiguous 1000 word sections.

2 Reading-Check Questions and Vocabulary Learning

2.1 Participants

We conducted a study to measure vocabulary learning with the REAP tutor in the English Language Institute at the University of Pittsburgh through the Pittsburgh Science of Learning Center's English LearnLab. Forty-four students

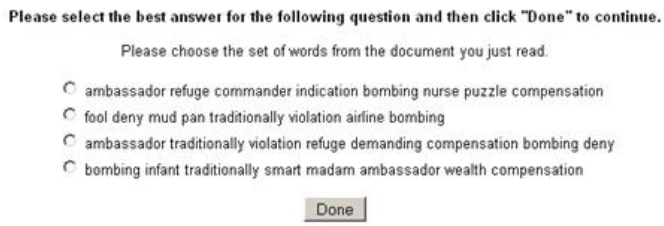

Fig. 1. Example Reading Check Question

at the English Language Institute at the University of Pittsburgh participated in this experiment as part of an intermediate English as a Second Language Reading course in the Fall of 2006. A variety of nationalities were represented among the participants. Eleven students were dropped from the experiment for various reasons, corresponding to an overall attrition rate of 23%. Most students were dropped because they did not attend class on the day of the post-test. A student's work with the REAP tutor did not count for a grade other than a small portion allotted to attendance. Prior to the study, each participant completed the Michigan Test of English Language Proficiency (MTELP)[4], a measure of general English language ability. The test consists of 40 grammar, 40 vocabulary, and 20 reading comprehension questions.

2.2 Instruction with the REAP Tutor

The students attended nine training sessions and working with the REAP tutor for forty minutes in each session. In this study, the REAP tutor provided instruction on a subset of the Academic Word List[5]. In the first session, students took a brief 15-minute self-assessment pre-test to identify which words they already knew in order to provide data for REAP's learner model. During each session, students read texts in which target vocabulary words were highlighted. The texts were between 200 and 2000 words and of an appropriate reading difficulty level for the students, as determined using automatic readability measures[6]. Students spent as much time as they wished on each text.

The REAP tutor facilitated the coordination of implicit information about the target words that is available from context in practice texts[5] with explicit information that is available from dictionary definitions. The tutor allowed students to see definitions by clicking on target words or by typing non-target words into a box at the bottom of the screen. Students completed a series of cloze, or fill-in-the-blank, exercises for target vocabulary words following each text. This interactive practice complemented the passive instruction from the texts and definitions. Brown, Friskoff, and Eskenazi[7] discuss the generation of

[4] http://www.michigan-proficiency-exams.com/mtelp.html

[5] In this paper, the terms "texts" and "practice texts" are used more or less interchangeably.

and rationale for using cloze questions. The sentences used in cloze questions were automatically chosen from texts other than those used by the REAP tutor. ESL teachers from the University of Pittsburgh reviewed and edited, and in some cases rewrote, the questions. The tutor provided immediate feedback on the correctness of student responses. In most cases, between two and five practice questions followed each text.

Following the practice exercises, students completed a multiple-choice reading-check question, generated as described above, for the just-completed practice text. The tutor did not use the results from the reading-check questions to guide instruction. Following this reading check question, students continued with more of these instructional cycles, each consisting of a practice text, vocabulary exercises, and a reading-check question. Heilman, Eskenazi, Collins-Thompson and Callan[3] provide further details about the REAP tutor and its instructional approach.

2.3 Post-test Results and Reading-check Performance

At the end of the semester, students took a post-test consisting of twenty cloze questions and ten sentence production tasks for target words which they had individually identified as unknown through self-assessments at the beginning of the semester. The format of the post-test cloze questions was similar the format of the practice exercises, but the to-be-completed sentences and foils were different. For the sentence production tasks, students wrote a sentence using the target word, demonstrating that they knew the meaning of the word. The produced sentences were graded on a scale from 0-3 by a teacher and a curriculum supervisor. The correlation coefficient between grades assigned by instructors and the curriculum supervisor was 0.68. Disagreement between the teachers and curriculum supervisor was resolved by averaging the scores.

In our analysis, we focused not on the vocabulary learning gains exhibited by students, but rather the association between reading-check question performance and post-test scores. As such, we calculated the Pearson Correlation Coefficients between two pairs of variables: first, proportion of reading-check questions answered correctly by a given student and proportion of post-test cloze exercises correctly answered by that student; second, proportion of reading-check questions answered correctly by a given student and proportion of maximum score for the post-test sentence production tasks. The correlation between reading-check and post-test cloze exercises was $r = .547$, which is statistically reliable (two-tailed test of independent samples, $t(31) = 3.64$, $p = .001$). The correlation between reading-check and sentence production performance was also statistically reliable ($r = .536$, $t(31) = 3.53$, $p = .011$).

We also investigated whether the correlation between reading-check performance and post-test performance was due to the common cause of general English language proficiency-in other words, whether vocabulary post-test scores and reading-check performance were conditionally independent given general English proficiency. The partial correlation between reading-check performance and post-test cloze proportion correct after controlling for MTELP proficiency scores

was statistically reliable ($r = .376$, $p = .034$). The partial correlation between reading-check performance and post-test sentence production scores controlling for MTELP proficiency scores was also statistically reliable ($r = .400$, $p = .012$).

The results indicate reliable associations between reading-check performance and two measures of vocabulary learning, even after controlling for general English proficiency. It therefore seems that students who were able to perform better on reading-check questions following texts were also more likely to successfully coordinate implicit information about target words that was available from context in practice texts. The reading-check questions seem to measure a construct that facilitates vocabulary acquisition while reading practice texts.

3 Correlation of Reading-Check Questions and Comprehension Questions

We conducted a second study to determine whether reading comprehension is the construct measured by the reading-check questions which facilitates vocabulary learning. If so, then better comprehension of the context in practice texts would be the cause of the improved vocabulary acquisition. More specifically, the second study addressed the following question: does performance on automatically-generated reading-check questions correlate with performance on more sophisticated, manually-authored reading comprehension questions?

Reading comprehension questions require readers to recall propositions from a text and possibly make inferences based on those propositions. In contrast, reading-check questions require readers to recall specific lexical items that occurred in a text. Recalling the specific lexical forms in a text does not necessarily lead to recalling propositions, or making inferences. Therefore, tests of shallow reading comprehension and reading-check questions should not correlate strongly with tests of deep reading comprehension. However, in the previous study the reading-check questions were moderately and reliably correlated with vocabulary learning. As such, we expected a similar moderate correlation in this study.

3.1 Experimental Design

The sample consisted of thirty undergraduate students attending summer research programs at Carnegie Mellon University (male = 11). We selected only native speakers of English in order to reduce the number of confounds (e.g., native language) and ensure a high probability of at least partial comprehension of the texts. Participation in the study was voluntary, and participants received ten dollars compensation for their time.

The study employed a between-participants design. The two dependent variables were Percentage of Reading-Check Questions Answered Correctly and Percentage of Reading Comprehension Questions Answered Correctly. The participants read five texts, each of which was followed by four reading-check questions and three to five reading comprehension questions. Participants were asked to read the texts and then answer the associated questions without referring to the

corresponding text. We randomized the order in which both the texts and the questions were presented to each participant.

We originally selected ten texts of between five hundred to one thousand words. This length was comparable to reading comprehension texts that appear on standardized tests, such as the GRE, as well as REAP texts. The set of texts included two biographies, two excerpts from fiction texts, two Scientific America articles, two Wall Street Journal articles, and two excerpts from philosophical essays. This variety of sources ensured varying levels of difficulty. An English as a Second Language teacher who had experience authoring reading comprehension questions wrote three to five reading comprehension questions per text, and the method described above was modified to generate four questions per text. Each in a set of four questions about a text had a unique answer set. The majority of the reading comprehension questions tested recall of facts from the texts but did not require inference. Thus, most of the questions addressed shallow comprehension. A small subset measured deep comprehension.

After pilot testing on five graduate students, we selected one of each type of text. The criteria used to choose these texts included a lack of ceiling effect during pilot testing, avoidance of well-known texts, and the presence of a relatively wide distribution of correct answers-that is, texts which not all participants found to be either very easy or very difficult. The chosen texts were an excerpt from Lois Lowry's *The Giver*, a biography of Frederick Douglass, the Scientific American article "Blowing in the Wind: Arctic Plants Move Fast as Climate Changes," an article entitled "Sony TV Stages a Heavy Online Push" from the Wall Street journal, and an excerpt from Emerson's "Nature." A simple graphical user interface displayed the texts and questions and recorded the participants' responses. Participants were given as much time as they needed to complete the task.

3.2 Results

We calculated correlations for the data among the following conditions: all data, and data for each individual text. For these analyses, we applied the Bonferroni correction for multiple tests. The data were assessed for statistical reliability at $\alpha = .05$, which with the correction became $\alpha = .00625$.

The hypothesized moderate positive correlation between performance on shallow and deep reading comprehension questions was partially supported by the data. When looking at the participants' percentages of correct responses for all the questions, these two variables had a medium positive correlation of .366 ($p < .0005$, one-tailed test). The R^2 value of 0.13 means that overall, thirteen percent of the variance of performance in answering reading comprehension questions can be explained by some mechanism that is at work in answering the reading-check questions.

We then analyzed the relationship between performance on reading-check and reading comprehension questions on each individual text. A statistically reliable correlation was found only for *The Giver* excerpt ($r = .792$, $p < .0005$, one-tailed test), with an R^2 value of 0.63. The other four texts did not have

statistically reliable correlations after the Bonferroni correction, but they are still noteworthy because of their size. Emerson's "Nature" had a correlation of .432 ($p = .009$, one-tailed test), with an R^2 value of 0.19. The Frederick Douglass biography had a correlation of .386 ($p = .018$, one-tailed test), and an R^2 value of 0.15. Performance on the Wall Street Journal article was similar; it had a correlation of 0.359 ($p = .026$, one-tailed test), and an R^2 value of 0.13. Finally, performance on the Scientific American article was slightly degraded, with a correlation of 0.250 ($p = .092$, one-tailed test), and an R^2 value of 0.06.

4 Related Work

Developers of tutoring systems and educational technology have taken a wide variety of approaches to measuring comprehension of texts and conceptual knowledge. Mostow, Beck, Bey, Cuneo, Sison, Tobin, and Valeri[8] have employed automatically generated multiple-choice cloze tasks to measure reading comprehension in Project Listen. They developed a tutoring system for grade school children to practice reading aloud. In that system, a student encounters sentences in the text from which content words have been removed and replaced with blanks. The student selects words that appropriately complete the sentence in a way so that the words fit into the contexts around the blanks. Kunichika and colleagues[4] developed a method for automatically generating reading comprehension questions for computer-assisted language learning. Their approach uses natural language processing methods including syntactic and semantic parsing in conjunction with a thesaurus to generate a variety of question types. Their evaluations found that approximately seven percent of questions were semantically invalid.

Tutorial dialogue systems such as Why2Atlas[9] engage in a conversation with the student in order to assess comprehension of particular concepts, and then to provide instruction on those concepts. While such systems can measure deep comprehension, human authors typically must encode considerable amounts of domain-specific knowledge. Also, the dialogues are not necessarily centered on particular texts. Li and Sambasivam[10] describe a method for generating questions from an ontology, or knowledge base, rather than from a specific text. Such knowledge bases are usually built for a particular domain with considerable human effort. However, such techniques involving knowledge bases might augment methods for generating questions for specific texts.

5 Discussion

Data from the first study showed a significant correlation existed between performance on automatically generated reading-check questions and vocabulary learning in the REAP tutor. As a result, it seemed that the reading-check questions measured a construct associated with vocabulary learning. We conducted a second experiment to investigate the extent to which reading-check questions were correlated with measures of reading comprehension. The second study found a

reliable correlation of reading-check performance and reading comprehension-but it was not an extremely strong correlation. As such, it seems that the reading-check questions measure a construct that is associated with but not equivalent to comprehension.

A possible explanation for the findings of the two studies is that the reading-check questions measure reader attention and engagement. Engagement is necessary but not sufficient for both reading comprehension and vocabulary learning. If students are attending to a text, then they are more likely but not certain to comprehend it. Other factors affect comprehension, including familiarity with the topic, reading ability, and the reading difficulty of the text. Similarly, students attending to a text are more likely but not certain to process implicit information in the text about target vocabulary words and then coordinate this with information from other sources such as definitions. As with comprehension, various other factors affect vocabulary learning, including knowledge of other vocabulary in the context and the reading difficulty level of the text. However, further tests would be needed to verify whether the constructs of attention and engagement are what is measure by the reading-check questions.

Some observations from the second study provide suggestions for improving the reliability and validity of the reading-check questions. A ceiling effect occurred for the reading-check questions in some texts, with most participants achieving a perfect score or only missing one out of the four questions. This finding might be due to the choice of foils: in many cases, the wrong answers contained words that really inappropriate with regards to the text's topic. For example, one set of foil words for Emerson's "Nature" contained the word "ghettoize." The presence of such obviously inaccurate words makes the task somewhat trivial. Automatic techniques for inducing models of semantic spaces (e.g., [11]) might be useful for extracting related words for use as foils.

Further research might examine how learners approach the task of answering reading-check questions. In the second study, one participant reported in a follow-up interview that he or she answered the reading-check question by attempting to comprehend the text and then infer which words would be likely to appear given his or her knowledge of the subject of the text. Future work might investigate the effects on reading behaviors of including reading-check questions after texts. Students might develop strategies that help them answer the reading-check questions correctly but do not promote learning from the text. In the REAP tutor, this does not seem to be the case since, in general, students who correctly answered reading-check questions also learned more vocabulary.

6 Acknowledgements

The researchers would like to thank Gregory J. Mizera and Dr. Maxine Eskenazi. This work was supported by Dept. of Education grants R305G03123 and R305B040063 to Carnegie Mellon University, the National Science Foundation grant 354420 to the Pittsburgh Science of Learning Center, a Siebel Scholarship awarded to the second author, and a National Science Foundation Graduate

Research Fellowship awarded to the second author. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

References

1. Wylie, R., Mitamura, T., Koedinger, K., Rankin, J. Doing more than Teaching Students: Opportunities for CALL in the Learning Sciences. SLaTE Workshop on Speech and Language Technology in Education (2007)
2. Brown, J. and Eskenazi, M. Retrieval of authentic documents for reader-specific lexical practice. Proceedings of InSTIL/ICALL Symposium (2004)
3. Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M. Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. Proceedings of the Ninth International Conference on Spoken Language Processing (2006)
4. Kunichika, H., Katayama, T., Hirashima, T., and Takeuchi, A. Automated question generation method for intelligent English learning systems and its evaluation. Proceedings of ICCE2004 (2003)
5. Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, 34(2): 213-238.
6. Collins-Thompson, K. and Callan, J. Predicting reading difficulty with statistical reading models. Journal of the American Society for Information Science and Technology (2005)
7. J. Brown, G. Frishkoff, and M. Eskenazi. Automatic question generation for vocabulary assessment. Proceedings of HLT/EMNLP 2005. Vancouver, B.C. (2005)
8. Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B., and Valeri, J. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. Technology, Instruction, Cognition and Learning, 2, pp. 97-134 (2004)
9. VanLehn, K., Jordan, P., Rose, C., and NLT Group. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. Proceedings of Intelligent Tutoring Systems Conference, Vol. 2363 of LNCS, pp. 158-167. Springer (2002)
10. Li, T. and Sambasivam, S. Automatically Generating Questions in Multiple Variables for Intelligent Tutoring. The Journal of Issues in Informing Science and Information Technology (2). pp. 471-480 (2005)
11. Curran, J. R. and Moens, M. Improvements in automatic thesaurus extraction. Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), pp. 59-66 (2002)