Automated Scoring of a Summary Writing Task Designed to Measure Reading Comprehension

Nitin Madnani, Jill Burstein, John Sabatini and Tenaha O'Reilly

Educational Testing Service
660 Rosedale Road, Princeton, NJ 08541, USA
{nmadnani,jburstein,jsabatini,toreilly}@ets.org

Abstract

We introduce a cognitive framework for measuring reading comprehension that includes the use of novel summary writing tasks. We derive NLP features from the holistic rubric used to score the summaries written by students for such tasks and use them to design a preliminary, automated scoring system. Our results show that the automated approach performs well on summaries written by students for two different passages.

1 Introduction

In this paper, we present our preliminary work on automatic scoring of a summarization task that is designed to measure the reading comprehension skills of students from grades 6 through 9. We first introduce our underlying reading comprehension assessment framework (Sabatini and O'Reilly, In Press; Sabatini et al., In Press) that motivates the task of writing summaries as a key component of such assessments in §2. We then describe the summarization task in more detail in §3. In §4, we describe our approach to automatically scoring summaries written by students for this task and compare the results we obtain using our system to those obtained by human scoring. Finally, we conclude in §6 with a brief discussion and possible future work.

2 Reading for Understanding (RfU) Framework

We claim that to read for understanding, readers should acquire the knowledge, skills, strategies, and dispositions that will enable them to:

- learn and process the visual and typographical elements and conventions of printed texts and print world of literacy;
- learn and process the verbal elements of language including grammatical structures and word meanings;
- form coherent mental representations of texts, consistent with discourse, text structures, and genres of print;
- model and reason about conceptual content;
- model and reason about social content.

We also claim that the ability to form a coherent mental model of the text that is consistent with text discourse is a key element of skilled reading. This mental model should be concise but also reflect the most likely intended meaning of the source. We make this claim since acquiring this ability:

1. requires the reader to have knowledge of rhetorical text structures and genres;
2. requires the reader to model the propositional content of a text within that rhetorical frame, both from an author's or reader's perspective; and
3. is dependent on a skilled reader having acquired mental models for a wide variety of genres, each embodying specific strategies for modeling the meaning of the text sources to achieve reading goals.

In support of the framework, research has shown that the ability to form a coherent mental model

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 163–168,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

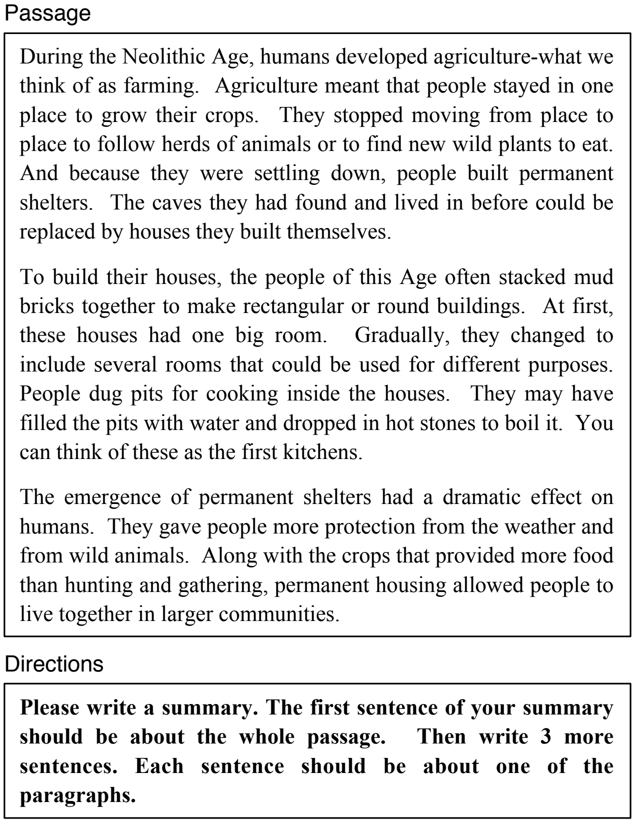
is important for reading comprehension. Kintsch (1998) showed that it is a key aspect in the process of construction integration and essential to understanding the structure and organization of the text. Similarly, Gernsbacher (1997) considers mental models essential to structure mapping and in bridging and making knowledge-based inferences.

2.1 Assessing Mental Models

Given the importance of mental models for reading comprehension, the natural question is how does one assess whether a student has been able to build such models after reading a text. We believe that such an assessment must encompass asking a reader to (a) sample big ideas by asking them to describe the main idea or theme of a text, (b) find specific details in the text using locate/retrieve types of questions, and (c) bridging gaps between different points in the text using inference questions. Although these questions can be multiple-choice, existing research indicates that it is better to ask the reader to write a brief summary of the text instead. Yu (2003) states that a good summary can prove useful for assessment of reading comprehension since it contains the relevant important ideas, distinguishes accurate information from opinions, and reflects the structure of the text itself. More specifically, having readers write summaries is a promising solution since:

- there is considerable empirical support that it both measures and encourages reading comprehension and is an effective instructional strategy to help students improve reading skills (Armbruster et al., 1989; Bean and Steenwyk, 1984; Duke and Pearson, 2002; Friend, 2001; Hill, 1991; Theide and Anderson, 2003);
- it is a promising technique for engaging students in building mental models of text; and
- it aligns with our framework and cognitive theory described earlier in this section.

However, asking students to write summaries instead of answering multiple choice questions entails that the summaries must be scored. Asking human raters to score these summaries, however, can be time consuming as well as costly. A more cost-effective and efficient solution would be to use an automated scoring technique using machine learning and natural language processing. We describe such a technique in the subsequent sections.



Passage

During the Neolithic Age, humans developed agriculture-what we think of as farming. Agriculture meant that people stayed in one place to grow their crops. They stopped moving from place to place to follow herds of animals or to find new wild plants to eat. And because they were settling down, people built permanent shelters. The caves they had found and lived in before could be replaced by houses they built themselves.

To build their houses, the people of this Age often stacked mud bricks together to make rectangular or round buildings. At first, these houses had one big room. Gradually, they changed to include several rooms that could be used for different purposes. People dug pits for cooking inside the houses. They may have filled the pits with water and dropped in hot stones to boil it. You can think of these as the first kitchens.

The emergence of permanent shelters had a dramatic effect on humans. They gave people more protection from the weather and from wild animals. Along with the crops that provided more food than hunting and gathering, permanent housing allowed people to live together in larger communities.

Directions

Please write a summary. The first sentence of your summary should be about the whole passage. Then write 3 more sentences. Each sentence should be about one of the paragraphs.

Figure 1: An example passage for which students are asked to write a summary, and the summary-writing directions shown to the students.

3 Summary Writing Task

Before describing the automated scoring approach, we describe the details of the summary writing task itself. The summarization task is embedded within a larger reading comprehension assessment. As part of the assessment, students read each passage and answer a set of multiple choice questions and, in addition, write a summary for one of the passages. An example passage and the instructions can be seen in Figure 1. Note the structured format of summary that is asked for in the directions: the first sentence of the summary must be about the whole passage and the next three should correspond to each of the paragraphs in the passage. All summary tasks are structured similarly in that the first sentence should identify the "global concept" of the passage and the next three sentences should identify “local concepts” corresponding to main points of each subsequent paragraph.

Each summary written by a student is scored according to a holistic rubric, i.e., based on holistic criteria rather than criteria based on specific dimensions of summary writing. The scores are assigned on a 5-point scale which are defined as:

Grade 4: summary demonstrates excellent global understanding and understanding of all 3 local concepts from the passage; does not include verbatim text (3+ words) copied from the passage; contains no inaccuracies.

Grade 3: summary demonstrates good global understanding and demonstrates understanding of at least 2 local concepts; may or may not include some verbatim text, contains no more than 1 inaccuracy.

Grade 2: summary demonstrates moderate local understanding only (2-3 local concepts but no global); with or without verbatim text, contains no more than 1 inaccuracy; OR good global understanding only with no local concepts

Grade 1: summary demonstrates minimal local understanding (1 local concept only), with or without verbatim text; OR contains only verbatim text

Grade 0: summary is off topic, garbage, or demonstrates no understanding of the text; OR response is “I don’t know” or “IDK”.

Note that students had the passage in front of them when writing the summaries and were not penalized for poor spelling or grammar in their summaries. In the next section, we describe a system to automatically score these summaries.

4 Automated Scoring of Student Summaries

We used a machine learning approach to build an automated system for scoring summaries of the type described in §3. To train and test our system, we used summaries written by more than 2600 students from the 6th, 7th and 9th grades about two different passages. Specifically, there were a total of 2695 summaries – 1016 written about a passage describing the evolution of permanent housing through history (the passage shown in Figure 1) and 1679 written about a passage describing living conditions at the South Pole. The distribution of the grades for the students who wrote the summaries for each passage is shown in Table 1.

Passage	**Grade**	**Count**
South Pole	6	574
	7	521
	9	584
Perm. Housing	6	387
	7	305
	9	324

Table 1: The grade distribution of the students who wrote summaries for each of the two passages.

All summaries were also scored by an experienced human rater in accordance with the 5-point holistic rubric described previously. Figure 2 shows the distribution of the human scores for both sets of summaries.

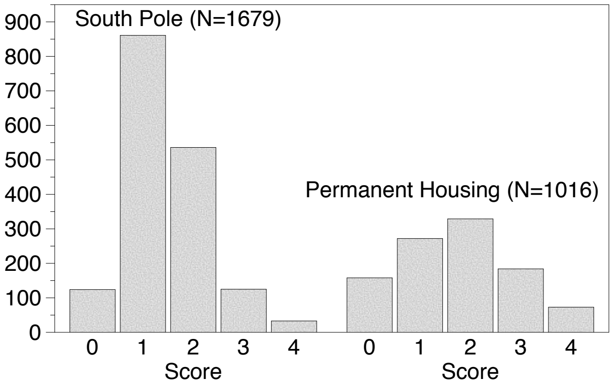


Figure 2: A histogram illustrating the human score distribution of the summaries written for the two passages.

Our approach to automatically scoring these summaries is driven by features based on the rubric. Specifically, we use the following features:

1. **BLEU**: BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) is an automated metric used extensively in automatically scoring the output of machine translation systems.

It is a precision-based metric that computes n-gram overlap (n=1 ... 4) between the summary (treated as a single sentence) against the passage (treated as a single sentence). We chose to use BLEU since it measures how many of the words and phrases are borrowed directly from the passage. Note that some amount of borrowing from the passage is essential for writing a good summary.

2. **ROUGE**: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003) is an automated metric used for scoring summaries produced by automated document summarization systems. It is a recall-based metric that measures the lexical and phrasal overlap between the summary under consideration and a set of "model" (or reference) summaries. We used a single model summary for the two passages by randomly selecting each from the set of student summaries assigned a score of 4 by the human rater.
3. **CopiedSumm**: Ratio of the sum of lengths of all 3-word (or longer) sequences that are copied from the passage to the length of the summary.
4. **CopiedPassage**: Same as CopiedSumm but with the denominator being the length of the passage.
5. **MaxCopy**: Length of the longest word sequence in the summary copied from the passage.
6. **FirstSent**: Number of passage sentences that the first sentence of the summary borrows 2-word (or longer) sequences from.
7. **Length**: Number of sentences in the summary.
8. **Coherence**: Token counts of commonly used discourse connector words in the summary.

ROUGE computes the similarity between the summary S under consideration and a high-scoring summary - a high value of this similarity indicates that S should also receive a high score. **CopiedSumm**, **CopiedPassage**, **BLEU**, and **MaxCopy** capture verbatim copying from the passage. **FirstSent** directly captures the "global understanding" concept for the first sentence, i.e., a large value for this feature means that the first sentence captures more of the passage as expected. **Length** captures the correspondence between the number of paragraphs in the passage and the number of sentences in the summary. Finally, **Coherence** captures how well the student is able to connect the different "local concepts" present in the passage. Note that:

- Although the rubric states that students not be penalized for spelling errors, we did not spell-correct the summaries before scoring them. We plan to do this for future experiments.
- The students were not explicitly told to refrain from verbatim copying since the summary-writing instructions indicated this implicitly ("... *about the whole passage*" and "... *about one of the paragraphs*"). However, for future experiments, we plan to include explicit instructions regarding copying.

All features were combined in a logistic regression classifier that output a prediction on the same 5-point scale as the holistic rubric. We trained a separate classifier for each of the two passage types.[1] The 5-fold cross-validation performance of this classifier on our data is shown in Table 2. We compute exact as well as adjacent agreement of our predictions against the human scores using the confusion matrices from the two classifiers. The exact agreement shows the rate at which the system and the human rater awarded the same score to a summary. Adjacent agreement shows the rate at which scores given by the system and the human rater were no more than one score point apart (e.g., the system assigned a score of 4 and the human rater assigned a score of 5 or 3). For holistic scoring using 5-point rubrics, typical exact agreement rates are in the same range as our scores (Burstein, 2012; Burstein et al., 2013). Therefore, our system performed reasonably well on the summary scoring task. For comparison, we also show the exact and adjacent agreement of the most-frequent-score baseline.

It is important to investigate whether the various features correlated in an expected manner with the score in order to ensure that the summary-writing construct is covered accurately. We examined the weights assigned to the various features in the classifier and found that this was indeed the case. As expected, the **CopiedSumm**, **CopiedPassage**, **BLEU**,

[1] We used the Weka Toolkit (Hall et al., 2009).

Method	Passage	Exact	Adjacent
Baseline	South Pole	.51	.90
	Perm. Housing	.32	.77
Logistic	South Pole	**.65**	**.97**
	Perm. Housing	**.52**	**.93**

Table 2: Exact and adjacent agreements of the most-frequent-score baseline and of the 5-fold cross-validation predictions from the logistic regression classifier, for both passages.

and **MaxCopy** features all correlate negatively with score, and **ROUGE**, **FirstSent** and **Coherence** correlate positively.

In addition to overall performance, we also examined which features were most useful to the classifier in predicting summary scores. Table 3 shows the various features ranked using the information-gain metric for both logistic regression models. These rankings show that the features performed consistently for both models.

South Pole	Perm. Housing
BLEU (.375)	BLEU (.450)
CopiedSumm (.290)	ROUGE (.400)
ROUGE (.264)	CopiedSumm (.347)
Length (.257)	Length (.340)
CopiedPassage (.246)	MaxCopy(.253)
MaxCopy (.231)	CopiedPassage (.206)
FirstSent (.120)	Coherence (.155)
Coherence (.103)	FirstSent (.058)

Table 3: Classifier features for both passages ranked by average merit values obtained using information-gain.

5 Related Work

There has been previous work on scoring summaries as part of the automated document summarization task (Nenkova and McKeown, 2011). In that task, automated systems produce summaries of multiple documents on the same topic and those machine-generated summaries are then scored by either human raters or by using automated metrics such as ROUGE. In our scenario, however, the summaries are produced by students—not automated systems—and the goal is to develop an automated system to assign scores to these human-generated summaries.

Although work on automatically scoring student essays (Burstein, 2012) and short answers (Leacock and Chodorow, 2003; Mohler et al., 2011) is marginally relevant to the work done here, we believe it is different in significant aspects based on the scoring rubric and on the basis of the underlying RfU framework. We believe that the work most directly related to ours is the Summary Street system (Franzke et al., 2005; Kintsch et al., 2007) which attempts to score summaries written for tasks not based on the RfU framework and uses latent semantic analysis (LSA) rather than a feature-based classification approach.

6 Conclusion & Future Work

We briefly introduced the Reading for Understanding cognitive framework and how it motivates the use of a summary writing task in a reading comprehension assessment. Our motivation is that such a task is theoretically suitable for capturing the ability of a reader to form coherent mental representations of the text being read. We then described a preliminary, feature-driven approach to scoring such summaries and showed that it performed quite well for scoring the summaries about two different passages. Obvious directions for future work include: (a) getting summaries double-scored to be able to compare system-human agreement against human-human agreement (b) examining whether a single model trained on all the data can perform as well as passage-specific models, and (c) using more sophisticated features such as TERp (Snover et al., 2010) which can capture and reward paraphrasing in addition to exact matches, and features that can better model the "local concepts" part of the scoring rubric.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100005 to the Educational Testing Service as part of the Reading for Understanding Research Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We would also like to thank Kelly Bruce, Kietha Biggers and the Strategic Educational Research Partnership.

References

B. B. Armbruster, T. H. Anderson, and J. Ostertag. 1989. Teaching Text Structure to Improve Reading and Writing. *Educational Leadership*, 46:26–28.

T. W. Bean and F. L. Steenwyk. 1984. The Effect of Three Forms of Summarization Instruction on Sixth-graders' Summary Writing and Comprehension. *Journal of Reading Behavior*, 16(4):297–306.

J. Burstein, J. Tetreault, and N. Madnani. 2013. The E-rater Automated Essay Scoring System. In M.D. Shermis and J. Burstein, editors, *Handbook for Automated Essay Scoring.* Routledge.

J. Burstein. 2012. Automated Essay Scoring and Evaluation. In Carol Chapelle, editor, *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

N. K. Duke and P. D. Pearson. 2002. Effective Practices for Developing Reading Comprehension. In A. E. Farstrup and S. J. Samuels, editors, *What Research has to Say about Reading Instruction*, pages 205–242. International Reading Association.

M. Franzke, E. Kintsch, D. Caccamise, N. Johnson, and S. Dooley. 2005. Summary Street: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33:53–80.

R. Friend. 2001. Effects of Strategy Instruction on Summary Writing of College Students. *Contemporary Educational Psychology*, 26(1):3–24.

M. A. Gernsbacher. 1997. Two Decades of Structure Building. *Discourse Processes*, 23:265–304.

P. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

M. Hill. 1991. Writing Summaries Promotes Thinking and Learning Across the Curriculum – But Why are They So Difficult to Write? *Journal of Reading*, 34(7):536–639.

E. Kintsch, D. Caccamise, M. Franzke, N. Johnson, and S. Dooley. 2007. Summary Street: Computer-guided summary writing. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of latent semantic analysis*. Lawrence Erlbaum Associates Publishers.

W. Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

C. Leacock and M. Chodorow. 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, 37(4):389–405.

C.-Y. Lin and E. H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of HLT-NAACL*, pages 71–78.

M. Mohler, R. Bunescu, and R. Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of ACL*, pages 752–762.

A. Nenkova and K. McKeown. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 311–318.

J. Sabatini and T. O'Reilly. In Press. Rationale For a New Generation of Reading Comprehension Assessments. In B. Miller, L. Cutting, and P. McCardle, editors, *Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension*. Brookes Publishing, Inc.

J. Sabatini, T. O'Reilly, and P. Deane. In Press. Preliminary Reading Literacy Assessment Framework: Foundation and Rationale for Assessment and System Design.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2010. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23:117–127.

K. W. Theide and M. C. M. Anderson. 2003. Summarizing Can Improve Metacomprehension Accuracy. *Educational Psychology*, 28(2):129–160.

G. Yu. 2003. Reading for Summarization as Reading Comprehension Test Method: Promises and Problems. *Language Testing Update*, 32:44–47.