This is an English version of Leacock, C. (2004). Automatisch beoordelen van antwoorden op open vragen; een taalkundige benadering. *Examens*, 1 (3)

Scoring Free-Responses Automatically:

A Case Study of a Large-Scale Assessment

Claudia Leacock, Educational Testing Service

cleacock@ets.org

Abstract

C-rater is an automated scoring engine that measures a student's understanding of content material through the use of natural language processing techniques. We describe the process used for building c-rater models using *Alchemist,* c-rater's model-building interface. Results are given for a large-scale assessment that used c-rater to score 19 reading comprehension and five algebra questions. In total, about 170,000 short-answer responses were scored with an average of 85% accuracy.

1. Introduction

An automated scoring engine, *C*-rater™, has been developed at the Educational Testing Service (ETS), to measure a student's understanding of specific content material without regard for the student's writing skills. It uses automated natural language processing techniques to determine whether a student response contains specific linguistic information required as evidence that the concept has been learned.

C-rater tries to recognize when a response is equivalent to a correct answer, and so is, in essence, a *paraphrase recognizer*. As such, the scoring engine is designed to recognize a correct response when it exhibits the variations that are ordinarily associated with paraphrases, whether they be syntactic variation, different inflections of a word, substitution of synonyms or similar terms, or the use of pronouns in the place of nouns. In addition to these features, which are ordinarily associated with paraphrasing, c-rater recognizes words that are spelled incorrectly – an essential feature for the K-12 market. Table 1 shows examples of these paraphrase variations as they have appeared in student responses. The recognition of the syntactic structure, inflected words, the referent of a pronoun and spelling correction are all fully automated. In the case of synonyms or similar words, a suggested list generated from a corpus of over 300 million words of current fiction, nonfiction, and textbooks (Lin, 1998) is presented to the model-builder, who can select from among them while building a c-rater model answer. A detailed description of the mechanisms that drive c-rater can be found in Leacock and Chodorow (Forthcoming).

Table 1: Types of Variation

Syntactic Variation	Money worries Walter → Walter is worried about money.
Inflectional Variation	dreams, dreaming → dream
Synonymy or Similarity	dreams → wants expensive → costly
Pronoun Reference	Mama disagrees with Walter. He thinks that money is life.
Spelling	Walter → Wlater, Waalter, Walther

Essentially, a model needs to represent the full range of concepts that a response must contain to receive full or partial credit. C-rater looks at each sentence in a student's response and determines whether it is a paraphrase of a sentence in the model. It is important to note that c-rater is not simply matching words – the paraphrases must obey syntactic constraints. For example, if "Peter ate the apple" is in the model, the sentence "The apple ate Peter" will not be recognized as a valid paraphrase.

A question can be scored by c-rater if there is a finite range of concepts that satisfy it. Thus an open-ended question asking for an opinion or for examples from the student's own experience is not a question for c-rater. A sample of questions that the system has successfully scored is shown in Table 2.

Table 2: Questions that c-rater has successfully scored.

Grade	Subject	Question
8	Science	Explain how you would design an experiment that would investigate the importance of light to plant growth. Include the type of organisms required, the control and variable, and the method of measuring results.
8	Math	A radio station wanted to determine the most popular type of music among those in the listening range of the station. Would sampling opinions at a Country Music Concert held in the listening area of the station be a good way to do this? Explain your answer.
11	Reading Comprehension	Compare and contrast what Mama and Walter in *A Raisin in the Sun* believe to be the most important thing in life or what they "dream" of. Support your choice for each character with dialogue from the excerpt of the play.
College	Database Management	Differentiate between logical and physical models.

Although c-rater scoring is fully automated, the process of creating a c-rater model is manual. For this process, we have developed *Alchemist*, a user-interface designed to guide a content expert through the process of model building. The next section describes an ambitious project where *Alchemist* has been used to generate models for c-rater in a large-scale end-of-year assessment study.

2. A Case Study

In the Spring of 2003, the State of Indiana Commission for Higher Education, the Indiana Department of Education, ETS and a subcontractor, Achievement Data, Inc, collaborated to develop and field test an administration of an online end-of-course test for 11th grade students. The two courses selected for this pilot study were 11th grade English and Algebra 1. A truly innovative aspect of the project was that *all* of the items on the test, including the open-ended short-answer responses, essays, and graphical responses were scored automatically.

The scope of the project was ambitious, beginning with designing and developing the test, administering it online via the Internet, developing and implementing automated scoring procedures for all of the item types, score reporting, data analyses, and finally constructing and delivering operational test forms. Here we describe a single phase of the project – the processes that were put in place to develop the scoring models for c-rater. There was a six-week period from the date testing began in Indiana to the time when all of the scores were reported. During these six weeks, models had to be built for 19 reading comprehension and five algebra questions. In the end, models for four of the algebra and 15 of the reading comprehension questions were successfully deployed. By the middle of June, c-rater had scored about 170,000 11th grade student responses.

On the first day of online testing for the Indiana pilot, we started to collect 100 student responses for each of the questions. These 100 responses were used for range finding. Two human readers and their supervisor scored the range-finding sets together, as a team, making decisions on what should and should not receive credit and modifying the scoring rubric when necessary to make distinctions between score points. Once range finding was completed, the scored responses were handed off to be used as the basis for c-rater model building.

When the questions were very difficult, specifically when the 100 range-finding responses provided fewer than 15 responses that were assigned full credit, it was necessary to collect an additional hundred responses. In two cases, for one reading comprehension question and one algebra question, the sample of 200 responses did not yield 15 that received full credit. For these two questions, model building was not attempted due to the lack of examples.

In order to estimate how accurate the scoring would be on unseen responses, an additional 100 responses for each item were scored independently (ie, without consultation) by the same two readers. These data were used for cross-validation of the model and to calculate inter-reader agreement (reader 1 versus reader 2), as well as c-rater agreement with reader 1 and with reader 2.

3. Model Building with *Alchemist*

Before model building begins, the user inspects the scoring rubric in order to establish what *essential elements* are required for a response to receive credit. For example, the 11th grade reading comprehension item shown in Table 2 requires four *essential elements* if the response is to get full credit: identification of Walter's dream, a supporting quotation, identification of Mama's dream, and a second supporting quotation. A response receives partial credit if it contains at least one, but not four, of these essential elements. This question was used in a previous Indiana pilot test, but it is illustrative of the kind of questions that were scored by c-rater in the case study described here. (We cannot show the questions from the 2003 administration because they are currently being used as test questions.)

Once the essential elements are established, the variety of ways in which students express them was found by inspecting the range-finding data. For example, Table 3 shows a sample of student responses that received credit for the first essential element. All of these sentences are identified by one of three syntactic frames shown below:

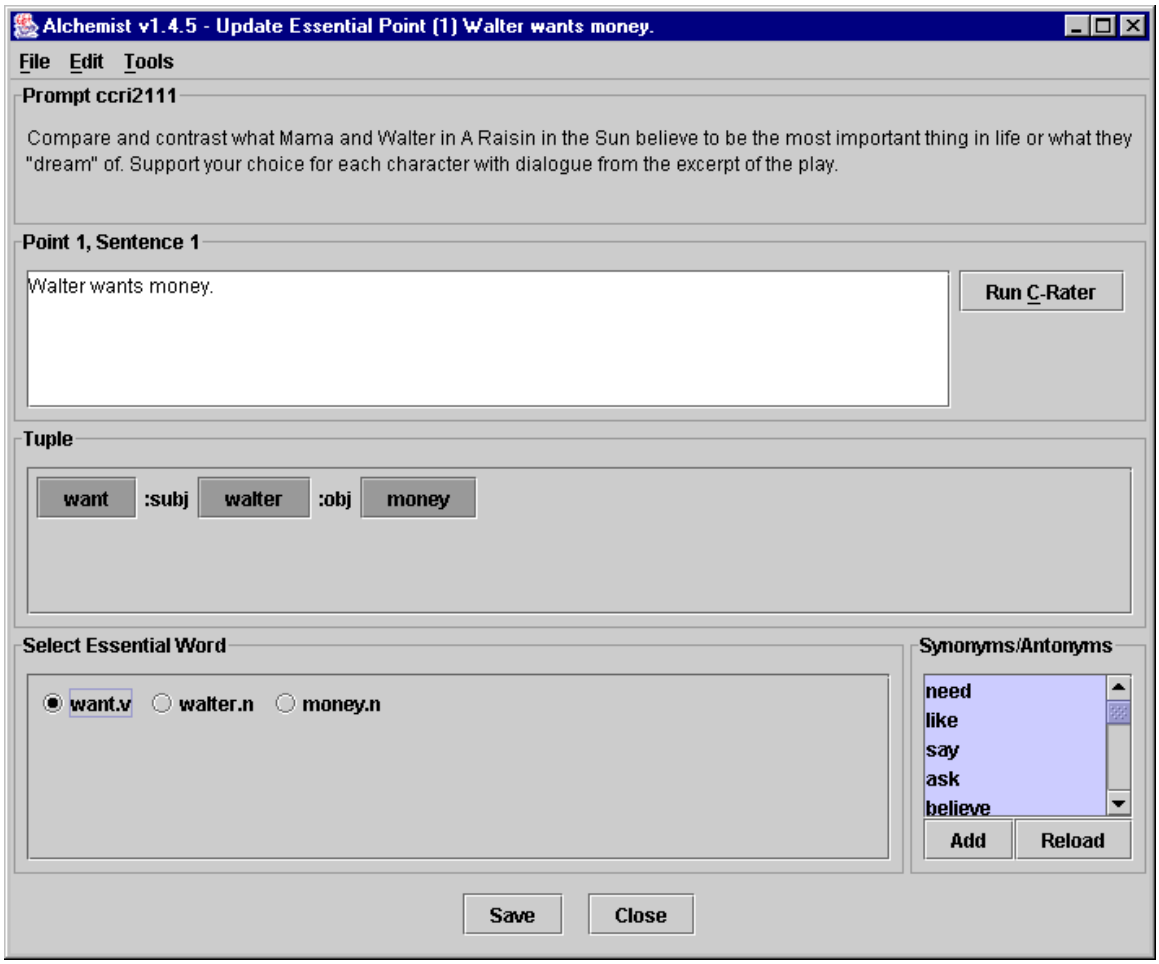
1. Walter wants money.
2. Money concerns Walter.
3. Money is important to Walter.

Table 3: Sentences recognized as being paraphrases

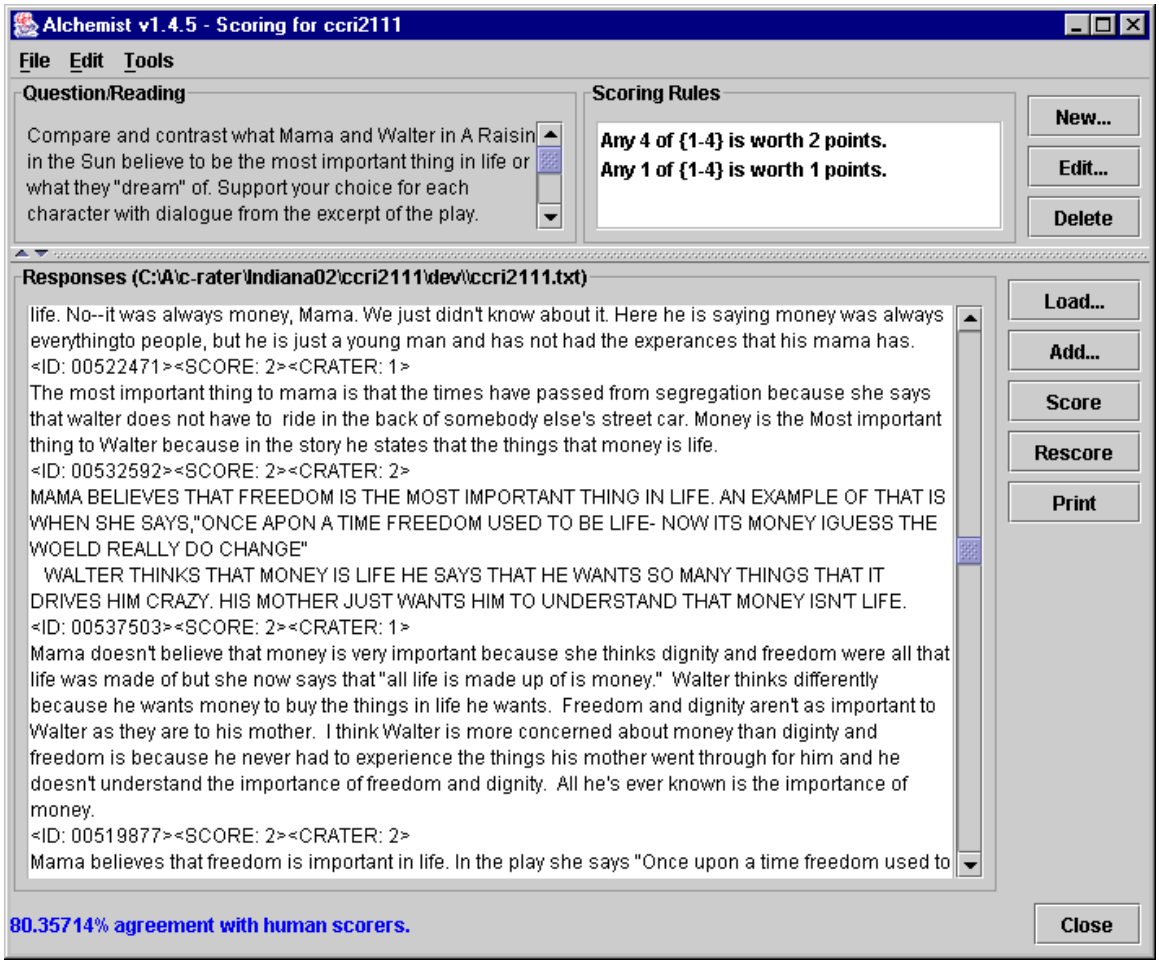
Walter wants money.
Walter thinks that it is money.
... but to Walter money is almost everything.
He most believes that money is important.
Walter is concerned with money.
... he wants material things.
... the son tries to tell his mom that money is the most important thing.
Walter is worried about money.

To enter the first syntactic frame into the model, the user adds a new sentence to essential element 1 by typing it into the window shown in Figure 1. C-rater analyzes the sentence and returns its syntactic and inflectional analyses. In this case, *want* is recognized as the verb, with

Walter as its subject and *money* as its object. The user then selects which words in the sentence are required in order to match the entire concept. In this case, the triple containing subject, verb and object is required, and so the user highlights all three words. For each highlighted word, the user then selects appropriate similar words or synonyms. For example, words similar to *want* include *need, like, say, ask, believe, tell, feel, think, talk, dream, care,* and *explain*. Clearly, these words are not all synonyms for *want*, rather they fit into the syntactic frame where the subject is *Walter* and an object or complement is *money* (or one of its synonyms).



As the user is creating the model, its accuracy can be tested on the scored data. Figure 2 shows *Alchemist*'s scoring window. The window for editing the *scoring rules* is in the upper right corner. The range-finding data appears in the window below, showing the human score and the c-rater score for each response. Overall agreement with the human scores is displayed in the bottom left corner. Thus the *Alchemist* interface allows the user to easily move back and forth, entering and adjusting model sentences and scoring data to check the agreement figures.



When the model is complete, the cross-validation responses are scored to determine whether the model generalizes on unseen data. Model approval is based on these cross-validation results.

4. Results

C-rater scoring models were used for 19 of the 24 open-ended short-answer questions in the pilot test. As shown in Table 4, c-rater agreed with the human scores on cross-validation responses about 85% of the time whereas the humans agreed with one another about 92% of the time. These agreement percentages are accompanied by kappa values, which correct for the level of agreement that is expected by chance. The kappa values for c-rater/human agreement are .77 for both readers. According to Fleiss (1981), "Values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance"

Table 4

	Reading Comprehension percent agreement (kappa)	Algebra percent agreement (kappa)
c-rater & reader 1	84.7 (.777)	85.8 (.738)
c-rater & reader 2	85.1 (.779)	84.9 (.735)
reader 1 & reader 2	93.0 (.902)	91.5 (.862)

As has already been noted, two of the models were not built because there were fewer than 15 full-credit responses in the 200-response samples. Another question was eliminated because agreement between the two human readers was too low – only 75%. In two other reading comprehension questions, the distinctions required for accurate scoring were too subtle for c-rater, even though the human readers could make the distinctions rather well. In one of these, the students were asked to "restate in your own words" the author's point. The question was really asking for a paraphrase – and credit was not assigned when the student entered an exact quotation. Not surprisingly, c-rater is unable to distinguish between the two, recognizing and giving credit to both the quotation and its paraphrase. In addition, this type of question is quite open-ended, allowing for metaphorical interpretations that cannot be fully cataloged. The second question where c-rater could not build a model involved a subtle temporal distinction – where the correct answer often depended on the tense of the verb. As part of c-rater's processing, verbs are replaced with their base form, thereby eliminating tense distinctions and so the information required to score the response had been eliminated.

C-rater/human agreement is consistently lower than human/human agreement. There are several reasons for this. First, there are borderline responses – where on inspection of the c-rater errors, it is not clear whether or why the c-rater score is wrong. Another reason is that humans are better at recognizing misspelled words than c-rater is. For example, one question requires mentioning the concept of *repetition* as a literary device. Both readers accepted *repation* as a variant of *repetition*, whereas the c-rater spelling correction program did not. In addition, some misspellings happen to be perfectly good English words. For example, in looking for the concept of an *odd number* in a response, the readers accepted *add number.* But since *add* is a correctly spelled English word, c-rater did not attempt to "correct" it so that it would match a word in the model.

There are also times when the response is truly original. One accurate response is that a certain kind of window is *too expensive* or *too costly*. One student wrote, idiomatically, that the

windows would "*take a chunk of change*". Of course, the model can be modified to include *chunk of change*, but the chances of that particular idiom ever being encountered again is slight. More importantly, the more open-ended a question is, the more difficult it is to build a model. When the concept being looked for is imprecise, then there are sure to be ways of stating it that are not found in the range-finding sets. At the very least, for these less precisely defined correct answers, the model builder needs access to more that 100 scored responses.

Usually when c-rater errs, it assigns a score that is too high rather than one that is too low, thereby giving more credit than is deserved. This often occurs because a response can contain the appropriate language even though its meaning differs from the concept required by the model. As an example, a concept that c-rater tries to identify is "*it is an old house*". One student wrote that "the author is telling you how *old the house* is", which was not credited by either reader. This becomes more problematic as a model is adjusted to accept sentence fragments as being correct answers. In this adjustment, c-rater imposes fewer requirements in order to allow syntactically incomplete forms that nonetheless embody the elements of the model. The problem seems unavoidable because human readers consistently accept sentence fragments – even very ungrammatical ones.

In general, if a distinction between partial or no credit is difficult for humans to make, as shown by inter-rater agreement, then that distinction is even more difficult for c-rater to make. Similarly for distinctions between partial and full credit.

5. Conclusion

The Indiana pilot has been by far the most ambitious c-rater study to date, and the results have remained consistent with previous studies. (Sandene, et al, forthcoming, Leacock and Chodorow, 2004). The c-rater models that were generated and deployed during this study were built by two people who are familiar with c-rater's mechanisms. It is our goal, during this next year, to continue to develop *Alchemist* so that it can be used by a content expert who is unfamiliar with c-rater's operations. This will involve developing a version of *Alchemist* that imposes constraints on the models and that can query the user about certain structures that are entered.

A drawback for c-rater is the requirement of 100 scored responses that are needed to build the model. As noted, in some cases, we have found that even 200 examples are insufficient information for building a reliable model. We are currently experimenting with an interactive machine learning system that can be used to improve, or extend, the model automatically even when the data sets are small. This will make model building more flexible in terms of the

requirements for preliminary data collection, and this in turn should make it possible to apply c-rater not just to large scale assessments but also to small scale tests of conceptual learning.

Acknowledgements

We am indebted to Eleanor Bolge for her many contributions to c-rater and to John Blackmore for developing the Alchemist interface. I also thank Martin Chodorow, Paul Deane, Ray C. Dougherty and Derrick Higgins for helpful suggestions.

Any opinions expressed here are those of the authors and not necessarily of the Educational Testing Service.

References

Fleiss J. L. (1981). *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons, 212-36.

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

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 898—904, Montreal.

Sandene, B., Bennett, R., Braswell, J., & Oranje, A. (Forthcoming). *Mathematics Online Study: Final Report*. Washington, DC: National Center for Education Statistics.