Available online at www.sciencedirect.com
ScienceDirect
Computers & Education xxx (2008) xxx–xxx
COMPUTERS & EDUCATION
www.elsevier.com/locate/compedu

Assessing creative problem-solving with automated text grading

Hao-Chuan Wang [a,d,*], Chun-Yen Chang [a,b,*], Tsai-Yen Li [c]

[a] Science Education Center, National Taiwan Normal University, Taipei, Taiwan
[b] Department of Earth Sciences, National Taiwan Normal University, Taipei, Taiwan
[c] Department of Computer Science, National Chengchi University, Taipei, Taiwan
[d] School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

Received 1 October 2007; received in revised form 9 January 2008; accepted 29 January 2008

Abstract

The work aims to improve the assessment of creative problem-solving in science education by employing language technologies and computational–statistical machine learning methods to grade students' natural language responses automatically. To evaluate constructs like creative problem-solving with validity, open-ended questions that elicit students' constructed responses are beneficial. But the high cost required in manually grading constructed responses could become an obstacle in applying open-ended questions. In this study, automated grading schemes have been developed and evaluated in the context of secondary Earth science education. Empirical evaluations revealed that the automated grading schemes may reliably identify domain concepts embedded in students' natural language responses with satisfactory inter-coder agreement against human coding in two sub-tasks of the test (Cohen's Kappa = .65–.72). And when a single holistic score was computed for each student, machine-generated scores achieved high inter-rater reliability against human grading (Pearson's $r = .92$). The reliable performance in automatic concept identification and numeric grading demonstrates the potential of using automated grading to support the use of open-ended questions in science assessments and enable new technologies for science learning.

Keywords: Computer-aided assessment; Automated grading; Creative problem-solving; Science learning assessment; Machine learning application

1. Introduction

Contemporary science education envisions that students should learn to solve real-world problems, which can be viewed as classes of problem-solving tasks that are relatively more complex, ill-defined, and qualitative in nature, as well as creatively more demanding in contrast to those utilized in traditional textbooks. Since assessment is an integral component in education for various instructional, institutional, and administrative purposes, the question of how to evaluate students' authentic problem-solving abilities in science classrooms

* Corresponding authors. Address: Science Education Center, National Taiwan Normal University, 88 Sec. 4, Ting-Chou Rd, Taipei 116, Taiwan.
E-mail addresses: haochuan@cs.cmu.edu (H.-C. Wang), changcy@ntnu.edu.tw (C.-Y. Chang).

0360-1315/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2008.01.006

and in large-scale ability tests is thus an important but under-investigated topic. Among various assessment approaches, the use of open-ended questions that seek for students' constructed responses that reveal their ability to integrate, synthesize, design, and communicate their ideas in natural language is considered as a powerful tool (Osterlind & Merz, 1994; Zenisky & Sireci, 2002). But as prior works have pointed out, grading and analyzing students' open-ended answers may be rather difficult and time-consuming (Shermis & Burstein, 2003).

Automated essay grading as an application of computational and statistical text processing has been proposed to address the cost of essay grading in the area of language learning, especially writing (Shermis & Burstein, 2003). Many systems have been successfully developed and deployed in evaluating the holistic quality of students' essays and traits such as creativity, organization, style, etc. in writing (Shermis, Koch, Page, Keith, & Harrington, 2002). Satisfactory inter-rater reliability has been achieved in the grading regular essays automatically (Hearst et al., 2000; Shermis & Burstein, 2003; Shermis et al., 2002), and commercialization has been realized for many years (Hearst et al., 2000). Science education may benefit from automated grading technologies as well. Natural language, in particular, has been observed to be one of the most natural and important media for students to use in communicating/elaborating their ideas and qualitative understandings without the artificial constraints as posed in traditional assessments.

In this article, we present our efforts in developing automated grading technologies and applications for the assessment of a creative problem-solving task called the Debris flow hazard (DFH) task. This task requires students to generate ideas creatively about the factors underlying the occurrence of the hazard and possible solutions to prevent the hazard from occurring. Students are also required to provide explanations to justify their ideas. Essentially, two components related to creative problem-solving are tapped in this task, creative idea generation (Nijstad & Stroebe, 2006; Wang et al., 2007) and self-explanation (Chi, De Leeuw, Chiu, & Lavancher, 1994). By examining the properties of this task, three main possible automated grading methods have been identified and evaluated. These three methods, namely pure heuristics-based grading (PHBG), data-driven classification with minimum heuristics grading (DCMHG), and regression-based grading (RBG), are conceptually aligned with the current general discussions in the different approaches to the building of intelligent agents and applications. These computer systems are designed to behave intelligently relying on either explicit knowledge authored by human experts or implicit knowledge induced from empirical data by statistical machine learning algorithms (Fogel et al., 1993; Mitchell, 2006). The three grading methods appear to have their own advantages and disadvantages related to tradeoffs between building cost and grading performance. From the empirical evaluation, automated grading has been found to be reliable and effective in grading creative problem-solving in the Earth science domain and may be further generalized to other similar scientific topics.

2. Background

This work is closely related to prior research in two research fields. The first is how to design and bring into effect methods of assessment to evaluate students' ability in transferring what they have learned to real-world situations in relation to educational and psychological research concerning the characteristics of human problem-solving in classroom and everyday life situations. And the second is how computational and statistical methods can be employed to develop educational technologies that help to handle the complexity of natural language communication and enable new applications in instruction and assessments. Since this work aims to develop automatic technologies to support the grading of natural language responses from a creative problem-solving assessment, we discuss the limitations of multiple-choice tests in this section, including the characteristics of this class of problem-solving task and the rationale of using open-ended questions, and also existing works related to the automatic grading and analysis of natural language responses.

2.1. General limitations of multiple-choice items

It is generally considered that traditional multiple-choice tests may suffer from the following shortcomings:

(1) Lower reliability due to students' guessing (Singley & Taft, 1995). In multiple-choice tests, students have to make choices among alternative responses. This approach makes the assessment highly scorable and also easy for data analysis. But when a simpler framework of analysis is adopted, such as in classical

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

testing theory, we may not account for the influence of the guessing strategies used by students. Although modern testing theories may better model the guessing behaviors of students, it is not easy to implement them in classrooms for the purpose of everyday diagnosis due to their data-driven and statistical nature.

(2) Lower validity due to inadvertent hints associated with alternative responses (Singley & Taft, 1995). The presence of alternative responses in multiple-choice tests may inadvertently give hints. This may change the nature of the problem-solving and reasoning and thus decrease the validity of the test items in relation to the constructs of interest. Another related phenomenon is the learning effect associated with recognition tests (Roediger & Karpicke, 2006). Taking a multiple-choice test may not only increase the retention of correct knowledge, but may also teach the students false knowledge associated with the alternative responses (Roediger & Marsh, 2005). The learning effect associated with testing may be utilized as a foundation for the design of instruction, but raising students' false knowledge is not favored either in terms of assessment or in instruction.

(3) Lower validity when the construct cannot be measured through the selection of option(s). In many cases, it is inappropriate to measure the constructs of interest in a multiple-choice manner. A trivial example is in the area of language learning, where students' writing proficiency can hardly be assessed by means other than open-ended essays. In science education, higher-level integrated process skills, such as hypothetical reasoning, idea generation, and self-explanation, are considered more important than ever. The interest in and practical need to use open-ended essays as science assessments are increasing.

2.2. Real-world creative problem-solving and assessment design

Cultivating students' abilities in solving real-world problems is considered as one of the most important goals in science education. Recent science education standards in the US proposed that science teaching must involve students in engaging in problem-solving, planning, decision-making and group discussions (National Research Council, 1996). The most recent curriculum guidelines in Taiwan (where the work was conducted) also stated that the science curriculum should develop students' abilities in independent thinking and problem-solving as well as stimulate their creativity and potential (Chang, 2005; Ministry of Education, 2001).

In educational practices, people call for systematic ways to evaluate students' learning and to find out how close students are to educational goals. And in many educational settings, it may be inappropriate or impossible to ask students to work on actual problems as means of evaluation due to limitations of time, space, and context. It is thus useful and crucial to find evidence that students possess the ability in solving particular classes of problems through a suitable assessment format. What has often been ignored in related discussions is discrimination in the classes of the problem-solving in relation to the nature of the required knowledge/expertise, the task characteristics, and their interactions. Such discrimination may lead to very distinct design concerns and results when designing assessments that aim at measuring the problem-solving ability of students in a systematic way.

Traditional problems like balancing a chemical reaction equation or finding out solutions for algebraic systems are characterized by having unambiguous problem-solving goals and formal operational models that can be followed to reach the goals (Chi and Glaser, 1985; Simon, 1973). These characteristics provide rationales about how to instruct students in the required domain knowledge to be able to perform well-defined problem-solving (Anderson, Corbett, Koedinger, & Pelletier, 1995) and also about how objective assessments should be designed to evaluate formal knowledge and cognitive skills in this regard. For example, it is reasonable that multiple-choice items may be devised to assess students' well-defined and quantitative problem-solving due to the existence of deterministic answers. In some experimental contexts, evaluating students' ability in identifying and selecting the correct operations to be applied is essential for the successful execution of procedural problem-solving plans.

However, real-world scientists and problem-solvers solve not only well-defined or constrained problems, though those could be fundamental to higher-order thinking (Chang, Barufaldi, Lin, & Chen, 2007). They very often have also to deal with problems that do not seem to have formal models or even correct answers (i.e. problem-solving goals). Authentic tasks such as making scientific discoveries against observed phenomena

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

and making dynamic decisions according to incomplete or changing information have long been ignored in classrooms. Multiple-choice recognition-oriented tests may not seem to be an appropriate form of evaluation for the real-world problem-solving involved in these tasks. This is due to the lack of absolutely correct answers or that the problem-solving processes require students' active construction, such as creative idea generation (Wang et al., 2007) and self-explanation (Chi et al., 1994). A creative problem-solving task that poses demands for students' divergent and convergent thinking (Basadur, 1995; Chang & Weng, 2002) can be viewed as a combination of idea generation and explanation (or justification) behaviors. And therefore, the design of appropriate assessment for creative problem-solving in a real-world context would need formats other than multiple-choice items.

Open-ended questions that elicit students' constructed responses and give students a higher degree of freedom in reasoning may serve as a better foundation for the design of authentic science assessments (Singley & Taft, 1995; Chang & Chiu, 2005). Nevertheless, the presence of natural language data in assessments may raise the cost in grading and analyzing students' answers. Some highly unconstrained responses like essay writing may even require mechanisms such as blind and peer reviewing in order to maintain reliability and objectivity in grading.

2.3. Automatic text processing in education

Two series of technological developments are closely related. One is in regard to the development of automated essay grading methods originating in the area of language learning and writing (Shermis & Burstein, 2003). Language learning has consistently had an unvarying need to use essays as a means of evaluation or instruction for students to practice their skills in language usage and communication. Models and techniques in grading essays through text processing have been developed in this area. In general, the task of automated grading can be viewed as a regression problem in which the objective is to find a set of features that represent the essays and serve as inputs of the regression methods. Regression algorithms are utilized to estimate the weights of each term (i.e. feature) in the regression equation so that the prediction performance can be optimized with regard to the actual values of the variable to be predicted/explained by the model (Hastie, Tibshirani, & Friedman, 2001, chap. 3). For open-ended science learning assessment that aims to evaluate constructs like creative problem-solving, general ideas are useful but need further adaptation according to task characteristics. This is because the foci of classic automated essay grading are more on students' writing skills but not on their cognitive or knowledge status regarding the particular science topics that underlie the constructed responses.

Another line of research that is informative is work on natural language-based computerized tutoring systems (Grasser, VanLehn, Rosé, Jordan, & Harter, 2001). The objective of this area is to identify students' intended meaning from their natural language discourses, infer their cognitive status in relation to the particular problem-solving tasks, and then offer feedback adaptively according to the inferred state. This line of research has accumulated knowledge and experience in the development and evaluation of natural language understanding systems for computerized tutoring. One of the most challenging issues here is how to enable computer systems to identify students' concepts in texts accurately. Several techniques have been developed in this field, such as sophisticated knowledge-based natural language understanding (Popescu & Koedinger, 2000), statistical latent semantics indexing and analysis (Grasser, Chipman, Haynes, & Olney, 2005), and automatic text categorization (Donmez, Rosé, Stegmann, Weinberger, & Fischer, 2005).

These natural language understanding or concept identification methods are useful in different aspects depending on the characteristics of the target educational domains. For example, in the domain of geometry, it is important for a computerized tutoring system to detect semantic and logical subtleties embedded in students' explanations of the steps which they take in solving the problem (Aleven, Popescu, & Koedinger, 2001). Knowledge-based approach turns out to be useful in that it may transform natural language responses into formal representations and sophisticated inferences may be made accordingly, while it could be more difficult for machine learning algorithms to learn complete inference models of this sort from examples. But in domains consisting of rich and probably unbounded concepts like environmental protection issues in Earth science, in contrast, it could be laborious and ineffective to author hard rules to capture lexical and linguistic variance required in the identification of domain concepts. What is likely to be more practical and has been shown to be effective is the utilization of machine learning algorithms in text categorization (Wang et al., 2007).

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

In the present work, the objective is to develop reliable automated grading systems for creative problem-solving responses. As we will show later, a framework is proposed to integrate ideas and methods developed in the automated essay grading and computerized tutoring communities.

3. Assessment design

In this section, we discuss the general characteristics of the open-ended assessment employed in this study. Then we describe in greater detail the method employed in the manual grading of students' responses in the assessment and the design of the coding schemes.

3.1. Task characteristics

The assessment consists of a creative problem-solving task in the domain of Earth sciences, namely the Debris flow hazard (DFH) task, at its core. A story-telling scenario describing the context of the occurrence of a debris flow hazard in Taiwan was first presented to the students, "*Imagine that you and your friends traveled to a famous scenery site in a mountainous county. When you arrived the hotel and were about to take some rest, you happened to see through the windows that amount of rock and mud mixing with water were moving rapidly toward this region. Many houses in this region were destroyed by this unexpected debris flow hazard. Fortunately, the hotel was not affected much. Since you are a famous young scientist who witnessed the occurrence of this hazard, you are asked to investigate possible factors underlying this hazard.*" There were then two stages for the students to work on. In the first stage, the students were asked to give their responses to the following questions: "*What are the possible factors underlying the occurrence of debris flow hazards? Why do you think so?*" Subsequently, at the second stage, students were asked to answer "*What are possible solutions for preventing debris flow hazards from happening? Why do you think so?*" The first question can be viewed as being at a *problem-finding* stage and the second as at a *problem-solving* stage. Note that the two-stage design is connected to the stage-wise model for creative problem-solving popular in the literature (Basadur, 1995; Osborn, 1963). In this type of problem-solving model, a divide-and-conquer strategy is proposed to divide the whole task into stages focusing on specific sub-goals of the task, such as divergent idea generation or the selection of plausible ideas. In the DFH task, the design of the two stages, problem-finding and problem-solving, are based on the characteristics of the scientific domain and also possess parallels to the discussions in the creativity literature (Basadur, 1995).

In each stage of the DFH task, we first asked students to generate ideas (i.e. factors underlying the hazard or solutions for preventing the hazard) and then to explain them. The combination of an idea-generation activity (Nijstad & Stroebe, 2006; Wang et al., 2007) and a self-explanation activity (Chi et al., 1994) shapes the task more as a science assessment than as a pure creative problem-solving task. Such a design may possess a higher diagnostic power in providing teachers and researchers with richer information regarding students' scientific reasoning. For example, some students may be able to produce ideas but be unable to give reasons for those ideas based on scientific knowledge and valid reasoning. It is expected that the integration of idea-generation and self-explanation may trigger not only creative thinking but also opportunities for deeper reasoning.

Fig. 1 shows examples of one student's responses collected from the study that we will describe later. As these examples show, the task required students to provide explanation for each of their ideas. Students may leave blank if they cannot think of a reason to support their idea.

3.2. Manual coding and grading procedures

In this section, we first formally formulate the system that the human graders followed to code the students' responses and then to give numeric total grades. In the next section, we describe the coding schemes we developed and employed.

The DFH task consists of two stages, problem-finding (PF) and problem-solving (PS). The students produced two different types of responses, ideas, in the first stage, and the reasons for each idea, in the second stage. Human grading was performed based on the following procedures. For the PF stage, we first defined

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

Idea	***Reason***
Problem Finding (PF) "*What are the possible factors underlying the occurrence of debris flow hazards? Why do you think so?*"	
Grew betel palms in mountainous areas	Betel palms are shallow rooted and cannot solidify soil
Over-developed mountainous areas	(Blank)
Intense rainfall lasted for several days	Intense rainfall may loosen soil and rock
Problem Solving (PS) "*What are possible solutions for preventing debris flow hazards from happening? Why do you think so?*"	
Grow deep-rooted plants	To better solidify soil
Constrain urban development in mountainous area. Emphasize on water-soil conservation	This would retain natural vegetation and prevent debris flows less likely from happening after rainfall

Fig. 1. Examples of one student's responses in the PF and PS stages (translated from original Chinese responses).

a set of creditable ideas for that stage, $C^{\mathrm{PF}} = \{c_1^{\mathrm{PF}}, c_2^{\mathrm{PF}}, \ldots, c_i^{\mathrm{PF}}\}$, and a set of conceptual knowledge in the domain that may be used as justification, $K^{\mathrm{PF}} = \{k_1^{\mathrm{PF}}, k_2^{\mathrm{PF}}, \ldots, k_i^{\mathrm{PF}}\}$. *Concept coding* or *concept identification* are the tasks to assign a code $c \in C^{\mathrm{PF}}$ to each student idea and a code $k \in K^{\mathrm{PF}}$ to each reason that the students drew upon to explain their ideas. Note that the human coders only identified codes for ideas and reasons separately at this point and did not consider whether each idea-reason pair proposed by students was reasonable. Thus human coders further perform binary *pair coding* to determine whether each (c, k) pair is reasonable. If a pair is reasonable, a code "true" is given, otherwise, "false" is given. For the PS stage, another set of creditable ideas C^{PS} and conceptual knowledge K^{PS} was defined, and the same coding procedures were performed.

Concept coding and pair coding provided rich and analyzable information regarding the performance of the students in the DFH task. But numeric total grades are often required for ease in performing various quantitative analyses typical in educational research. Therefore, systematic quantification was performed to map the coding results to holistic grades. Functions were defined to map each concept or reason to numeric scores in order to differentiate the value of each idea or reason, $f : C^{\mathrm{PF}} \rightarrow \mathcal{N}$. The ranges of numeric grading for the four sub-tasks, including the PF-idea, PF-reason, PS-idea and PS-reason, were also defined. Then we simply mapped the idea set and reason set of a student to numeric grades within the grading ranges. Numeric grades for all the sub-tasks were combined and aggregated based on given heuristics to derive the total grades for each student. For the grading of the reason set, only reasons that had a "true" value in the pair coding were given credits. The purpose was to give scores only to reasons associated with correct *idea-reason pairings* that provide evidence on the ability of generating reasonable justifications. Duplicate ideas and idea-reason pairings received credits only once. Note that students might use the same reason to justify several ideas. All of the justifications might all receive credits as long as they were valid according to the coding schemes and results of pair coding. Finally, the grades of the four components were summed linearly with equal weights to derive a single holistic grade.

In summary, human grading is an extension of the concept coding and pair coding tasks. Human graders did not directly assign holistic grades to each student. Instead, they were instructed to perform coding at the level of ideas, of reasons, and of the connections between the two. Systematic quantification was then undertaken to compute the total grade. The approach of coding detailed concepts is argued to improve the quality and reliability of numeric grading and also retain valuable information regarding what concepts are known and unknown by students for pedagogical diagnostics and other instructional applications.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

3.3. Coding schemes

The way in which students' responses are coded and graded is a crucial aspect in the design of coding schemes. A coding scheme for a specific sub-task (e.g. PF-idea) consists of a set of creditable ideas/reasons identified by a panel of domain experts including an Earth science professor and graduate students possessing teaching experience of the subject matter in middle/high schools. Possible literal variations for each creditable idea or reason were enumerated whenever possible. The human coders were instructed to categorize each student's idea/reason by matching the answers to the category in the coding scheme with which the idea/reason appeared to have the closest semantics. If nothing seemed to have a creditable match in the coding scheme, the coders were instructed that they might label the response as "others". Fig. 2 shows examples of categories in the coding scheme for the PF-idea. Table 1 provides an overview of the number of categories in the coding scheme and the ranges of numeric grading for each of the sub-tasks.

4. Automated grading schemes

We have identified two general directions in automated grading and three specific designs in implementing automated graders. The first direction is the so-called heuristics-based approach. By heuristics-based approaches, we refer to the concept that operational knowledge regarding how to perform the coding/grading was offered by domain experts or invented by system designers and then represented or programmed in the form that can be executed by computers. The second direction is the data-driven approach, in which we did not elicit operational knowledge about coding/grading from human experts directly and explicitly. Instead, the major human labor here is to prepare a set of coded or graded responses as training data. Computational–statistical machine learning algorithms were then employed to induce regulations and operational knowledge from the training data empirically. Data-driven approaches can be thought of as acting in a way similar to that of human experts if they were to provide coded/graded examples as a means to teach the machines.

Three specific automated grading system designs were proposed according to the two general approaches. As shown in Fig. 3, path 1 represents the pure heuristics-based grading (PHBG) method, which requires

Category ID	Sample Answers
C1-1	Steep mountain slope; the slope is too steep; the degree of the mountain slope; the mountain area is too steep
C2-1	Planting shallow-rooted plants for economic purposes; planting betel palm; planting betel palm on the mountain area; did not plant deep-rooted plants; planting tea trees; growing vegetables on steep slope
C3-0	Typhoon or heavy rainfall; great amount of precipitation; intense and lasting rainfall; heavy rain in the mountainous area

Fig. 2. Example categories in the coding scheme for the idea generation part in the problem-finding (PF) stage. The original coding scheme was authored in Chinese. Here shows the translation.

Table 1
Number of categories in the coding scheme and ranges of numeric grading for each sub-task

	PF-idea	PF-reason	PS-idea	PS-reason
Number of categories[a]	20	17	16	11
Ranges of numeric grading	[0, 28]	[0, 28]	[0, 28]	[0, 30]

[a] Creditable categories plus 'others'.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

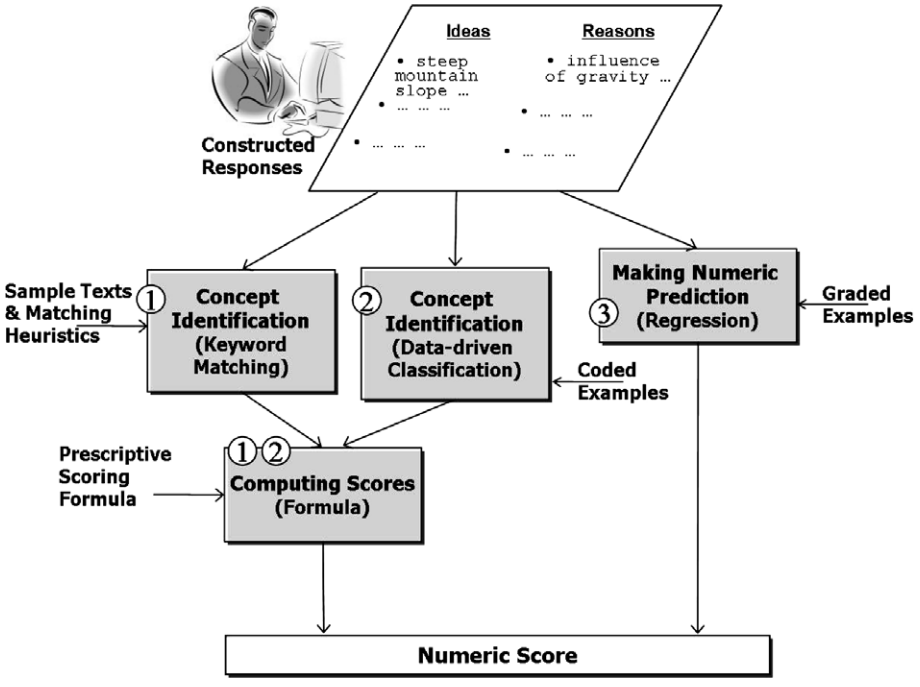

Fig. 3. Three automated grading methods developed and compared in this work.

excessive knowledge engineering but no training data. Path 2 uses data-driven classification algorithms to perform concept identification. But when performing numeric grading, a prescribed mapping function or scoring heuristic is required. Thus the method in path 2 is called the data-driven classification with minimum heuristics grading (DCMHG) method. Coded examples and heuristics for numeric grading were both needed. Path 3, the regression-based grading (RBG) method, utilized regression algorithms to make predictions as to the students' numeric grades holistically. This can be viewed as a pure data-driven approach that needs no heuristics regarding coding or grading. The sole output of path 3 is numeric grades, and only numerically graded examples were required by this method. Table 2 shows a summary of the general techniques adopted in each path. Next we present each path in greater detail.

4.1. Path 1: pure heuristics-based grading (PHBG)

4.1.1. Concept identification

PHBG performed concept identification by first representing text objects, including students' responses and the ideal answers associated with each category in the coding scheme as *word vectors*. That is, each text object D_i was represented by a vector in a t-dimensional vector space: $\vec{D}_i = (w_{i1}, w_{i2}, \ldots, w_{it})$, in which w_{ij} represents the feature value the jth term in D_i, which could be binary (e.g. the value is 1 if the term occur, and 0 if it does not) or weighted using various metrics (e.g. term frequency). Here we adopted the term frequency-inverse document frequency (TF-IDF) weighting scheme as the metric (Salton & McGill, 1983). Qualitatively speaking, the design of the TF-IDF weighting scheme originated from the observation that the semantic salience of a

Table 2
Technical approaches of each automated grading method

	Path 1	Path 2	Path 3
Concept identification	H	D	–
Numeric grading	H	H	D

H – Heuristic-based and D – data-driven.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

term in a document often correlates positively with its local frequency in the document and inversely with its overall popularity in the corpus.

The dimensionality of the vector space, t, is determined by the union of words occurring in all students' responses as well as all of the ideal answers in the coding scheme. This type of representation is known as the vector space model (VSM), and is widely-used in language technologies fields like information retrieval (IR) and text categorization (Raghavan & Wong, 1986; Salton, Wong, & Yang, 1975).

After representing text objects appropriately, the concept of each unknown response was determined by selecting the category in the coding scheme that was most similar to the response in the vector space. The similarity score between two text objects (e.g. a student response and a conceptual category) was derived by computing the cosine metric between two word vectors. That is

$$\text{cosine}(D_1, D_2) = \frac{\vec{D}_1 \cdot \vec{D}_2}{\|\vec{D}_1\|\|\vec{D}_2\|}$$

For each student response P, we computed the cosine similarity scores between P and a list of categories in the coding scheme, C. The concept label assigned to P is determined by

$$\text{label}(P) = \arg\max_{c \in C} \text{cosine}(P, c)$$

That is, the conceptual category with the highest cosine similarity score was selected for response P. In PHBG, we used an open source toolkit, SecondString (Cohen, Ravikumar, & Fienberg, 2003), for feature weighting and computing text similarity.

4.1.2. Numeric grading

After identifying the concept associated with each student response, as shown in Fig. 3, numeric grading was performed by mapping the concepts produced by the students to numeric scores using the prescribed scoring heuristics. The scoring heuristics are similar to the ones employed in manual grading that quantified identified concepts. Note that since PHBG only performed concept identification but did not perform pair coding, an independent human expert was asked to author and prescribe valid connections between ideas and reasons for both the PF and the PS parts. With these heuristics, PHBG was able to compute numeric total grades for each student.

4.2. Path 2: data-driven classification with minimum heuristics grading (DCMHG)

4.2.1. Concept identification

DCMHG performed concept identification by casting the task as a text categorization problem, and used classification algorithms developed in the machine learning community. Similar to PHBG, DCMHG first represented each text object as a word vector. But DCMHG did not then employ the notion and heuristic of document similarity to assign categorical labels to students' responses as in PHBG, where such a categorization model can be viewed as a manually-crafted one. Instead, the categorization model in DCMHG was induced from empirical data by machine learning algorithms. In such type of data-driven paradigm, a formalism formulating the input/output and required parameters of a classification model is typically delineated first. A set of labeled examples (i.e. coded responses) were prepared and divided into two subsets, training data and testing data. A set of features, such as the set of words occurring in the texts, describing the dataset was constructed and crafted. Corresponding training algorithms were then utilized to build or learn a model that may classify input instances into output categories automatically according to the training data. Testing data that had not been used in the training phase were employed to test the trained model in order to perform an unbiased evaluation. See Sebastiani (2002) for a general introduction and comprehensive survey of the use of machine learning methods to perform automatic text categorization.

In our DCMHG, a specific machine learning method, the support vector machine (SVM), was used to perform concept identification. The SVM method was proposed to handle a binary classification task by finding an optimal decision hyper-plane in the t-dimensional space that best discriminates the positive examples from the negative (Bishop, 2006; Burges, 1998; Cristianini & Shawe-Taylor, 2000; Vapnik, 1995). Geometrically, the

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

optimal decision hyper-plane is operationalized as the one that possesses the maximum margin between those closest positive and negative examples. SVM may also be used to handle multi-class classification by applying some techniques, such as training and making an appropriate combination of multiple binary classifiers. Training SVM models involves solving a quadratic programming problem. Efficient optimization algorithms have been developed and made available (Cristianini & Shawe-Taylor, 2000; Platt, 1998).

When the classification problem is not linearly separable, prototypical SVM may not be ideal. One class of increasingly popular and important techniques in machine learning is the so-called kernel method. The core idea is to project the data from the original space to one in a higher dimension in which the problem may appear more linearly separable (Bishop, 2006; Burges, 1998; Cristianini & Shawe-Taylor, 2000). Characteristics of SVM make the kernel method particularly applicable. SVM has been shown to be fairly effective in many areas of application. For example, SVM is known to be one of the best-performing classifiers in text categorization, and has several advantages, such as being robust in the situation, where there may be an overfit to training data and requiring only very little parameter tuning (Joachims, 1998; Sebastiani, 2002).

Note that in the data-driven paradigm, the construction and adjustment of the feature set have be shown to be influential factors affecting the performance of the model (Guyon & Elisseeff, 2003), in addition to the selection of specific machine learning algorithms. In research on automated essay grading, it was also proposed that the validity and reliability of an automated grader were much influenced by the set of features constructed and employed to represent the responses (Shermis & Burstein, 2003). In DCMHG, the feature space (i.e. elements in the word vectors) representing students' responses consisted of a set of unigrams extracted from students' responses, which was simply the set of single token words in the texts (e.g. "science", "education", "is", "important" from the sentence "science education is important"). The unigram model can be viewed as a special case of the *n*-gram feature model popular in the areas of IR and text categorization (D'Amore and Mah, 1985).

4.2.2. Numeric grading

As illustrated in Fig. 3, DCMGH employed the same set of heuristics used by path 1-PHBG. The procedure for computing numeric total grades was identical to that in PHBG.

4.3. Path 3: regression-based grading (RBG)

4.3.1. Concept identification

RBG did not perform concept identification as a mediating stage for holistic grading. By using regression algorithms, RBG is able to make numeric predictions directly based on training data. The advantage of the approach is that it almost requires no heuristics and is thus fairly cost effective, given that compiling complete heuristics manually could be much more laborious and difficult than grading students' responses manually for training purposes. Besides possible tradeoffs in grading reliability, a further limitation is that this method cannot provide a detailed profile about the ideas/reasons that the students have generated in a diagnostic sense.

4.3.2. Numeric grading

Similar to DCMHG in following the data-driven paradigm, RBG trained the regression models to make numeric predictions empirically from the training data. The numeric prediction of grades $\hat{y}$ made by a generic regression model can be expressed in the following linear form:

$$\hat{y}(X, W) = \omega_0 + \omega_1 x_1 + \cdots + \omega_k x_k$$

where $X = (x_1, \ldots, x_k)^{\mathrm{T}}$ are features employed to represent the data, and $W = (w_0, \ldots, w_k)$ are weights to be learned. The standard way to estimate parameters is by using the least squares (LS) or the least median squares (LMS) methods to search for the parameters that minimize the loss function, such as the summation of squared residuals in LS. The LMS method has been argued to be more robust against outliers (Rousseeuw, 1984).

The problem formulation and optimization techniques used in SVM classification can be extended to regression problems. A technique called SVM regression can be used to replace the loss function with Vapnik's ε-insensitive loss function (Bishop, 2006; Vapnik, 1995), which tolerates a certain amount of deviation between

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

the actual data and the predictions. The goal is to minimize accumulated errors that were not tolerated by the loss function. It has been shown that the minimization problem can be formulated as a quadratic optimization problem similar to that in SVM classification. Also, kernel projection methods for SVM classification that we mentioned previously can be applied in SVM regression to achieve non-linear regression.

5. Evaluation

5.1. Dataset

A dataset consisting of 226 Taiwanese high school students' responses (expressed in Chinese) on the DFH task was used in the evaluation. There were a total of 1560 ideas and reasons for the PF part, and 1138 ideas and reasons for the PS part.

Since students' responses were expressed in the Chinese language, specific pre-processing was required before automated grading. Chinese sentences are characterized by an absence of delimiters (e.g. space) to mark word boundaries that is available in English. Chinese word segmentation is thus essential for most language processing tasks that utilizes techniques originally developed for the English language. We employed the automatic Chinese word segmenter developed by the Institute of Information Science, Academia Sinica (Ma & Chen, 2003), which has been shown to be highly accurate in segmenting students' responses.

5.2. Evaluation method

The main objective of the evaluation is to empirically examine the reliability of the three automated grading methods in terms of concept identification and holistic grading with respect to human coding/grading.

A general principle in evaluating data-driven automated grading methods is the separation of the training and the testing data. Training data should be used only to train the model, while testing data are employed to evaluate the performance of the model. This is to ensure that the testing data is new and unseen to the trained model, and thus unbiased evaluative reports can be derived. When the amount of labeled data is limited, it is a common practice to perform n-fold cross-validation (Kohavi, 1995; Witten & Frank, 2005). The core idea of this procedure is to divide labeled examples into k mutually exclusive folds randomly. Each time one untested fold from the k folds is held out as the testing data, and the remaining $(k-1)$ folds are combined as the training data to train a new predictive model. After training, the held-out testing data is used to evaluate the performance of the trained model. Therefore, such paired training and testing steps are undertaken for k-times, and at each round, the held-out testing data is switched. In this way, even with a limited sample size, we may still evaluate a particular machine learning algorithm in an impartial manner.

For the two automated grading methods involving data-driven techniques, DCMHG and RBG, we trained data-driven models by using the implementation of the machine learning algorithms in an open source software package called Weka (Witten & Frank, 2005). We also used an open source toolkit called TagHelper (Donmez et al., 2005), designed for supporting data coding in educational research, to help construct the feature sets.

6. Results

6.1. Concept identification

First, we look at the inter-coder reliability that human coders may achieve for the DFH task. This can be viewed as a baseline for comparison. Two independent human coders were recruited and instructed in how to perform the coding task on the dataset by using the coding scheme as described previously. As shown in the first row of Table 3, inter-coder reliability using Cohen's κ between the two human codes ranged from .59 to .77 in concept identification, and .57 to .58 in pair coding.

Note that we had two sets of labeled results from the two independent human coders. In order to perform the most critical evaluation of data-driven concept identification in DCMHG, we trained the DCMHG model using labeled data from only one of the human coders, and then evaluated PHBG and DCMHG against the

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

Table 3
Performance of concept identification using various methods[c]

	PF-idea ($n = 1560$)	PF-reason ($n = 1560$)	PF-pair coding ($n = 1560$	PS-idea ($n = 1138$)	PS-reason ($n = 1138$)	PS-pair coding ($n = 1138$)
Human	.77	.68	.57	.75	.59	.58
Path 1-PHBG[a]	.65	.33	–	.58	.30	–
Path 2-DCMHG[b]	.72	.47	–	.65	.40	–

Inter-coder reliability measure: Cohen's Kappa.
[a] Text matching measure: Cosine similarity with TF-IDF weighting.
[b] Data-driven classification method: SVM (feature set: unigrams).
[c] Path 3 did not consist of the concept identification phase and thus was not included in the comparison.

labels assigned by the other independent human coder. Table 3 and Fig. 4 show the results. DCMHG was found to achieve near-human performance in PF-idea and PS-idea with Cohen's $\kappa = .72$ and .65, respectively. DCMHG outperformed PHBG in all of the four parts of the assessment. Note that neither DCMHG and nor PHBG performed well on the PF-reason or the PS-reason. But the human coders' coding of the two reason-giving sub-tasks was also not as reliable as the coding on the two ideation sub-tasks ($\kappa = .68$ and .59, respectively). One explanation may be that students did not always seem clear about what might constitute good and reasonable explanations of the ideas generated and this situation might make it more difficult to perform accurate classification. Also, the coders reported that it took them more time and effort to code the reason-giving parts. Improving the clarity of the task instruction and coding scheme for the reason-giving tasks may help enhance the reliability of manual and automatic coding in the future.

6.2. Numeric grading

When determining the choice of regression algorithm and features to be used in RBG, less reference was available as guidance for such a decision in this special context. Therefore, we first experimented with several regression algorithms along with different configurations of the feature space using the method of 10-fold cross-validation. Three regression algorithms were compared: LMS regression, standard SVM regression, and SVM regression with radial basis function (RBF) kernel mapping. The RBF kernel is a well-established mapping technique that deals with non-linearly separable cases in SVM (Burges, 1998). The three feature configurations included in the comparison were: unigrams, unigrams + bigrams, and unigrams + bigrams + part-of-speech

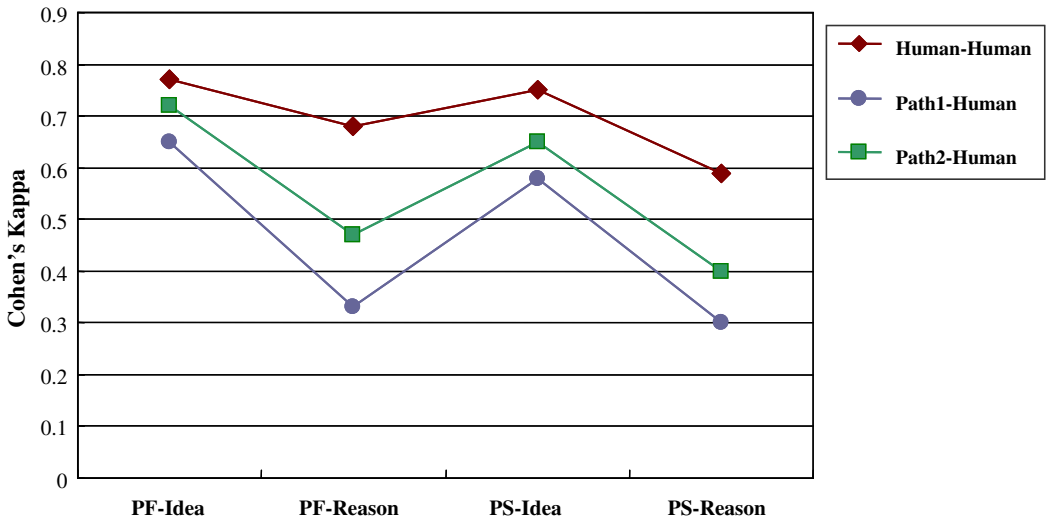

Fig. 4. Comparison of concept identification reliability.

bigrams. Bigrams refer to a special case of the *n*-gram feature model, which combines two consecutive words in a sentence as a feature. For example, we may extract features like "science_education", "education_is", "is_important", etc. from the sentence "science education is important". The type of feature may possess the utility for capturing the semantics realized by word order to an extent. Part-of-speech (POS) bigrams further take syntactic information associated with words into account. Automatic POS tagging was performed by using the Academia Sinica Chinese word segmenter (Ma & Chen, 2003).

As shown in Table 4, the configuration of using the feature set, unigrams + bigrams + POS bigrams, to train an SVM regression model with the RBF kernel appeared to be the most preferred one because of the outperformed reliability (Pearson's $r = .81$ for PF and .86 for PS) and conciseness of the feature set (number of features = 1193 for PF and 989 for PS). Note that although using another feature set consisting of only unigrams to train an SVM regression with RBF kernel model actually achieved a higher reliability for the PF part (Pearson's $r = .83$), this configuration was not preferred because it doubled the size of the feature space (number of features = 2913 for PF and 2883 for PS) but did not enhance the performance in PF to

Table 4
Performance of different regression methods with various feature set design for path 3 (RBG) using 10 cross-fold validation

	Word unigrams		Unigrams + bigrams[a]		Unigrams + bigrams + POS bigrams[a]	
Task (# features)	PF(2913)	PS(2883)	PF(830)	PS(649)	PF(1193)	PS(989)
LMS regression	.71	.76	.72	.77	.74	.81
SVM regression	.79	.82	.74	.78	.77	.83
SVM regression with RBF kernel	.83	.84	.80	.82	.81	.86

Inter-rater reliability measure: Pearson's *r*; $n = 226$.
[a] Rare features (number of occurrence <5) were removed.

Table 5
Performance of numeric grading using various approaches

	PF task	PS task	Total (PF + PS)
Human	.95	.93	.96
Path 1-PHBG[a]	.87	.84	.90
Path 2-DCMHG[b]	.90	.89	.92
Path 3-RBG[c]	.80	.86	.86

Inter-rater reliability measure: Pearson's *r*; $n = 226$.
[a] Text matching measure: Cosine similarity with TF-IDF weighting.
[b] Data-driven classification method: SVM (feature set: unigrams).
[c] Regression method: SVM regression with RBF kernel (feature set: unigrams + bigrams + POS bigrams).

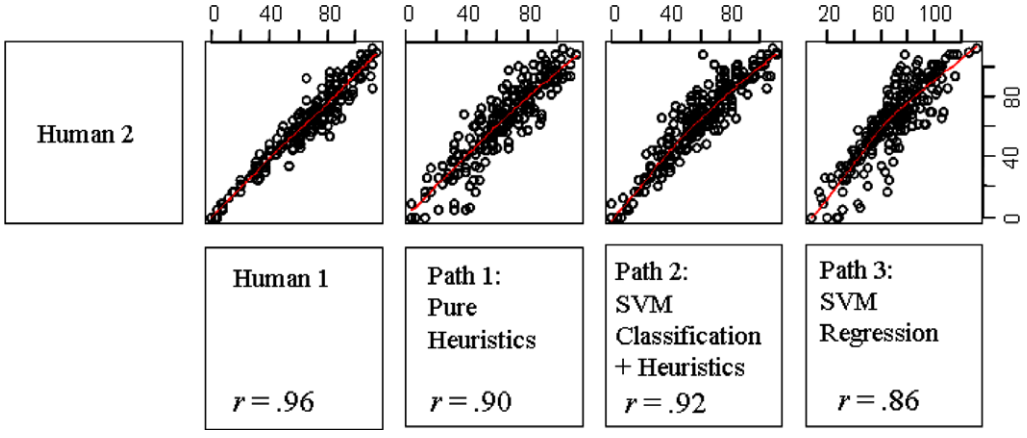

Fig. 5. Scatter plots comparing the reliability of various grading methods.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

any great extent. A concise feature space is more efficient in training and also more elegant from the theoretical point of view that prefers parsimony in model selection (cf. Mitchell, 1997).

After determining the configuration of RBG, a comparison of numeric grading reliability in the three automated grading paths was performed. Similar to concept identification, we employed the grading results of only one human coder to train data-driven models and then used labels assigned by another independent human coder in the evaluation. Table 5 summarizes the results of the evaluation, and Fig. 5 shows the scatter plots visually presenting the relation between human grading and various automated grading approaches. The inter-rater reliability between the two human graders achieved Pearson's $r = .96$. All of the three automated grading methods appeared to be reliable in numeric grading with a varied performance, ranging from $r = .86$ to .92. DCMHG was found to outperform the others with $r = .92$. RBG showed the least good performance, but the inter-rater reliability, $r = .86$, was already satisfactory.

7. Discussion and implications

The article has introduced approaches to automated text grading for the assessment of creative problem-solving in science education. The evaluative results have demonstrated the coding and grading reliability of the proposed automated grading methods. Among the three automatic approaches, DCMHG was found to have the highest reliability in both concept identification and numeric grading with a near-human performance in most of the coding and grading sub-tasks. The PHBG approach, which requires extensive prescribed heuristics but no labeled examples, was ranked in the middle according to its performance in concept identification and holistic grading. Finally, the RBG approach provided no information about what students knew or did not know at the concept level. The performance of RBG in numeric grading was the least good, but the grading reliability, $r = .86$, remained quite satisfactory according to regular data coding and grading practices in behavioral research.

In further considering the time and effort required by each grading approach, recommendations as to the design may be offered based on resources available in the specific context of application. It is not unusual that human experts may code or grade students' responses without any problem, but they suffer from providing detailed operational knowledge or heuristics about how they perform the coding or grading tasks. It is even harder and more tedious to encode or "translate" such knowledge to a format that is computer-reasonable and executable for automated grading purposes. Data-driven approaches thus offer plausible solutions to avoid the bottleneck of knowledge elicitation. DCMHG may be a fruitful path to follow when it is viable to manually code or assign labels on students' responses to an amount enough for the training of a reliable machine learning classification model. The function of detailed concept coding provided by DCMHG may help teachers identify what students have learned and what they should learn next. But if the goal of automated grading is simply to get a "fast-preview" of the results of an assessment, such as when seeking to compare the scores of one class of students to another, a fast-prototyped automated grader can be built using the RBG approach. The RBG approach requires only numerically labeled examples (i.e. there is no need for detailed concept coding for training purposes) to train the model. RBG also appeared to be reliable enough in the context of DFH grading to provide preview information. However, in situations, where abundant manually coded or graded examples are not available, one may still consider implementing a heuristics-based automated grader, like PHBG. As we have demonstrated, it is sometimes possible to enumerate lexically varied expressions of each domain concept and identify concepts embedded in students' responses by computing their semantic similarities to domain concepts using suitable heuristics.

Another criterion for choosing the appropriate grading approach is to consider whether the nature of the assessment is formative or summative. For formative assessments that aim to collect detailed information about students' learning status for planning instructional feedback, profiles containing students' known/unknown concepts and component grades are demanded. PHBG and DCMHG are more appropriate for formative assessments since they may both provide information of this sort. In the opposite, RBG lacks the function for providing detailed information required by formative assessments. RBG may be more suitable for the purpose of fast-prototyping automated graders for summative assessments.

The proposed framework for automated grading methods is intended to be flexible and applicable to another creative problem-solving domain, where the formalism and assumptions of assessment design

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

proposed in this work are followed. However, it is inevitable that a new domain would involve domain- and context-dependent factors. The reliability performance reported in the article should not be taken literally. Educational practitioners interested in developing similar automated graders for creative problem-solving in a new context are therefore advised to *cautiously* and *rigorously* evaluate the inter-rater reliability in the specific context of application before further employing the automatic methods to obtain critical information for research, instructional, screening and other decision-making purposes. Another related issue is the feasibility of transferring and applying the proposed methods in other language contexts. The technical framework is designed at the level independent of specific languages. Techniques like vector space modeling, cosine similarity and machine learning have been shown to be generally applicable to different languages by the language technologies community. It is thus feasible to apply the proposed methodologies to develop automated graders in other languages. There are still some minor language-specific issues to handle, such as locating suitable stemming programs and part-of-speech taggers for preprocessing the data, but these issues are general in the development of language technologies but not specific to automated grading. But once again, we stress the importance of rigorous evaluation when applying the proposed methods in new language contexts.

There has been promising work in using data-driven methods to grade essays for language learning purposes (Shermis & Burstein, 2003) and to support the analysis of educational discourses, such as collaborative learning dialogues (Donmez et al., 2005). Machine learning as a timely approach for the development of these tools virtually roots its power in the analysis of empirical data. One may consider the required efforts for preparing training data make data-driven approach less cost-effective. In many application fields, this characteristic is more of a benefit rather than a detriment. In educational assessments, to empirically evaluate reliability and validity of newly developed test items, it is not unusual in standard practices to construct large datasets of manually graded responses. The development of data-driven automated grading in such contexts that utilizes general machine learning algorithms and existing datasets actually requires little extra work. Once reliable grading models for targeted assessments are trained upon representative samples, the assessments and the models can be reused repeatedly. One possible future direction is to explore how to further reduce costs required for preparing training data in data-driven approaches. Research on semi-supervised learning and model selection may be closely related.

The functionality of automated grading is much desired in distance education and online learning applications. For example, the approach has recently been applied to the development of a web-based online testing system with the functionality of performing rapid data visualization for researchers and teachers to obtain a fast online preview of students' testing results (Huang, Wang, Li, & Chang, 2006). An intelligent tutoring system that traces students' idea generation and provides adaptive feedback was also developed. The mechanism of providing adaptive feedback for ideas generated in the performance of the problem-finding task has shown to be effective (Wang, Li, Rosé, Huang, & Chang, 2006). The framework for automated grading may also be further applied to the investigation of fundamental research questions in creativity research and science education. It would be interesting to examine creative problem-solving theories and models by looking at how an earlier problem-solving stage (e.g. thinking about factors for a natural hazard) may affect a later one (e.g. thinking of solutions for the hazard). We may also study misconceptions and the processes of conceptual change (Tsai & Chang, 2005) by developing coding schemes and automatic grading systems specific to the detection of misconceptions in science domains.

Acknowledgments

We thank Chun-Chieh Huang and Sophia Wang for their assistance on the project. We thank Carolyn P. Rosé, Yuen-Hsien Tseng, and Chao-Lin Liu for their advices on techniques of language processing and machine learning. We thank Keh-Jiann Chen and the Chinese Knowledge and Information Processing Group in Academia Sinica for providing the Chinese word segmentation toolkit and technical support.

References

Aleven, V., Popescu, O., & Koedinger, K. R. (2001). Toward tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of the international conference on artificial intelligence in education*. IOS Press.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of Learning Sciences, 4*(2), 167–207.

Basadur, M. (1995). Optimal ideation-evaluation ratios. *Creativity Research Journal, 8*(1), 63–75.

Bishop, C. M. (2006). *Pattern recognition and machine-learning*. New York: Springer.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

Chang, C.-Y. (2005). Taiwanese science and life technology curriculum standards and earth systems education. *International Journal of Science Education, 27*(5), 625–638.

Chang, C.-Y., Barufaldi, J. P., Lin, M.-C., & Chen, Y.-C. (2007). Assessing 10th-grade students' problem solving ability online in the area of Earth sciences. *Computers in Human Behavior, 23*, 1971–1981.

Chang, S.-N., & Chiu, M.-H. (2005). The development of authentic assessment to investigate ninth graders' scientific literacy: In the case of scientific cognition concerning the concepts of chemistry and physics. *International Journal of Science and Mathematics Education, 3*, 117–140.

Chang, C.-Y., & Weng, Y.-H. (2002). An exploratory study on students' problem-solving ability in Earth sciences. *International Journal of Science Education, 24*(5), 441–451.

Chi, M. H., De Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–488.

Chi, M. T. H., & Glaser, R. (1985). Problem-solving ability. In R. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 227–257). San Francisco: W.H. Freeman & Co.

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 18th international joint conference on artificial intelligence, workshop on information integration on the web* (pp. 73–78).

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

D'Amore, R. J., & Mah, C. P. (1985). One-time complete indexing of text: Theory and practice. In *Proceedings of the 8th annual international ACM SIGIR conference on research and development in information retrieval*. New York: ACM Press.

Donmez, P., Rosé, C. P., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In *Proceedings of computer supported collaborative learning conference*. International Society of the Learning Sciences.

Fogel, D., Hanson, J. C., Kick, R., Malki, H. A., Sigwart, C., Stinson, M., et al. (1993). The impact of machine-learning on expert systems. In *Proceedings of the 1993 ACM conference on computer science* (pp. 522–527).

Grasser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612–618.

Grasser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39–51.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Hearst, H., Kukich, K., Landauer, T. K., Lahan, D., Foltz, P., Hirschman, L., et al. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22–37.

Huang, C.-C., Wang, H.-C., Li, T.-Y., & Chang, C.-Y. (2006). An online testing and analysis system for creative problem-solving ability in sciences. In *Proceedings of the 10th annual global Chinese conference on computers in education*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European conference on machine-learning (ECML)*. Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the international joint conference on artificial intelligence (IJCAI)*.

Ma, W. & Chen, K. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing*.

Ministry of Education (2001). *The 1–9 grades science and life technology curriculum standards*. Taipei: MOE.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Mitchell, T. (2006). *The discipline of machine learning*. Technical Report, Machine-learning Department, Carnegie Mellon University, CMU-ML-06-108.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review, 10*(3), 186–213.

Osborn, A. (1963). *Applied imagination: Principles and procedures of creative problem-solving*. New York: Charles Scribner's Sons.

Osterlind, S. J., & Merz, W. R. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment, 2*(2), 133–148.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In C. B. B. Schoelkopf & A. Smola (Eds.), *Advances in kernel methods – Support vector learning*. MIT Press.

Popescu, O., & Koedinger, K. (2000). Towards understanding geometry explanations. In *Building dialogue systems for tutorial applications. Papers of the 2000 AAAI fall symposium* (pp. 80–86). Menlo Park, CA: AAAI Press.

Raghavan, V. V., & Wong, S. K. M. (1986). A critical analysis of the vector space model for information retrieval. *Journal of the American Society for Information Science, 37*(5), 279–287.

Roediger, H. L., & Karpicke, J. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155–1159.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM, 18*(11), 613–620.

Sebastiani, F. (2002). Machine-learning in automatic text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. NJ: LEA.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(1), 5–18.

Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4*, 181–201.

Singley, M. K., & Taft, H. L. (1995). Open-ended approaches to science assessment using computers. *Journal of Science Education and Technology, 4*(1), 7–20.

Tsai, C.-C., & Chang, C.-Y. (2005). Learning effects of instruction guided by the conflict map: Experimental study of learning about the causes of the seasons. *Journal of Research in Science Teaching, 42*(10), 1089–1111.

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Wang, H.-C., Li, T.-Y., Rosé, C. P., Huang, C.-C., & Chang, C.-Y. (2006). VIBRANT: A brainstorming agent for computer supported creative problem-solving. In *Proceedings of the 8th intelligent tutoring systems conference*.

Wang, H.-C., Rosé, C. P., Cui, Y., Chang, C.-Y., Li, T.-Y., & Huang, C.-C. (2007). Thinking hard together: The long and short of collaborative idea generation in scientific inquiry. In *Proceedings of the computer-supported collaborative learning conference*. International Society of the Learning Sciences.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine-learning tools and techniques*. San Francisco: Elsevier.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362.

Please cite this article in press as: Wang, H.-C.et al., Assessing creative problem-solving with automated text grading, *Computers & Education* (2008), doi:10.1016/j.compedu.2008.01.006