Automatic Assessment of Constructed Response Data in a Chemistry Tutor

Scott Crossley
Kristopher Kyle
Georgia State University
Atlanta, GA 30303
scrossley@gsu.edu
kkyle@student.gsu.edu

Jodi Davenport
WestEd
San Francisco, CA 94107
jdavenp@wested.org

Danielle S. McNamara
Arizona State Univ.
Tempe, AZ, 85287
dsmcnama@asu.edu

ABSTRACT

This study introduces the Constructed Response Analysis Tool (CRAT), a freely available tool to automatically assess student responses in online tutoring systems. The study tests CRAT on a dataset of chemistry responses collected in the ChemVLab+. The findings indicate that CRAT can differentiate and classify student responses based on semantic overlap with student input and indices related to word frequency, text content, and lexical sophistication. Overall, the findings suggest that more accurate student responses show greater overlap with the content learned, include more academic function words, contain greater content that is descriptive, and includes more specific and familiar words.

Keywords

Natural language processing, on-line tutors, constructed response scoring

1. INTRODUCTION

For science education to be more effective, students should move beyond memorizing facts and procedures and toward gaining deeper conceptual understanding that allows them to both apply scientific knowledge to explain new phenomena and to design investigations. The Next Generation Science Standards [1], offer a new vision of science instruction that integrates science practices, disciplinary core ideas, and cross cutting concepts, such as scale, energy, and patterns that unify different fields. However, assessing learning of these interconnected strands is challenging using traditional, multiple-choice items. Constructed responses, as well as more novel types of assessments provide students with important opportunities to demonstrate reasoning, explanation, and inquiry skills and are thus an important educational tool [2].

One problem with constructed responses are associated scoring costs [3]. A possible solution to these costs can be found in automated scoring tools that can reduce the need for human scoring and potentially increase scoring consistency [4]. In this study, we introduce a freely available natural language processing (NLP) tool called the Constructed Response Analysis Tool (CRAT) that can automatically score constructed responses in domain specific learning environments. We conduct a pilot study that tests the efficacy of CRAT to score student responses to a domain specific question in an on-line chemistry tutoring system by comparing scoring models developed by CRAT to human ratings of constructed responses.

1.1 Assessing student understanding

Simulations and games provide rich environments for students to learn science and demonstrate their understanding of scientific principles [5]. Such games and simulations can be included in online systems that allow for just-in-time feedback. The dynamic feedback found in online systems affords students the opportunity to confront misconceptions and provides information about areas of struggle or mastery that teachers can use as formative assessments that influence instructional decision making. However, the utility of feedback depends on the ability of an online system to provide an accurate diagnosis of student understanding. Though multiple choice and student behaviors in simulation environments may be readily scored using constraint-based model tutors [6], interpreting and accurately scoring constructed responses in science education has proven much more challenging [2]. These challenges have led researchers to develop content-based automated scoring systems that demonstrate medium to high agreement with human scores. These systems show promise for a number of domains (e.g., math, reading, psychology, biology) and a number of student levels (i.e., middle school, high school, college) [7, 8. 9].

1.2 Current Study

The goal of this study is to introduce CRAT and examine its potential to automatically assign accuracy scores to student constructed responses from an on-line tutor. Constructed responses were collected in the ChemVLab+ tutoring system (chemvlab.org) and scored by expert raters. We used the Constructed Response Analysis Tool (CRAT) to calculate linguistic features related to text content, text summarization, and lexical sophistication and used these linguistic features to predict the human scores.

2. METHOD

2.1 ChemVLab+

The ChemVLab+ is an on-line tutoring system that provides students with opportunities to apply chemistry knowledge to meaningful contexts and to receive immediate, individualized tutoring. Of interest in the current study are the four stoichiometry activities contained within ChemVLab+. The activities engage students in a variety of problem-solving tasks using interactive simulations including a virtual chemistry lab. At the end of each activity, students respond to one to three open-ended questions (i.e., constructed responses) designed to evaluate their ability to synthesize the information they had learned. The four stoichiometry activities included a total of 10 questions.

2.2 Participants

A total of 1392 high school chemistry students from the classes of thirteen teachers in the California bay area used the Stoichiometry module. Students used the online activities as part of their normal coursework.

2.3 Human Scores of Constructed Responses

All constructed responses were coded by two independent raters familiar with the chemistry content. Coders used an annotated rubric that described criteria for each score and provided examples of responses receiving those scores. Reliability of scoring varied across the questions, and interrater reliability ranged from Cohen's κ = 0.55 to .92. Each question had three possible scores, except for the two lowest reliability questions, (items 1 and 2.1), which had four possible scores. When the highest two scores in these questions were collapsed, interrater reliability increased from 0.56 to 0.68 for item 1 and from 0.59 to 0.69 for item 2.2.

2.4 Selection of Constructed Responses

We selected student constructed responses from question 1 in the stoichiometry lab to test CRAT. The question had the greatest number of student answers (n = 1374). The question asked students to explain the relationship between the amount of sugar, the volume of the drink, and concentration of the sports drink.

2.5 CRAT

CRAT is an easy to use constructed response analysis engine that calculates indices related to a) the linguistic and semantic similarities between a source text and a constructed response, b) the linguistic sophistication of a constructed response, and c) text properties (e.g., length and syntactic categories). It is freely available, cross-platform, and is accessed via a graphic user interface (GUI). The similarity indices include lexical similarity calculated using key word overlap, synonym overlap, and latent semantic analysis (LSA) similarity [10] and phrasal similarity calculated using key bigram and trigram overlap and key part of speech sensitive slot-grams (e.g., a trigram with an open slot such as *into the* ____). The constructed response sophistication indices include psycholinguistic word information indices (e.g., concreteness and familiarity [11, 12]), lexical frequency and range (words that occur in a wider range of texts) indices based on the British National corpus (BNC [13]) and the Corpus of Contemporary American English (COCA [14]), and syntactic categories (e.g., number of adjectives and nouns). For COCA, CRAT reports on frequency and range indices for a number of different genres including academic, newspaper, and fiction genres. Selected index features are outlined below. See http://www.soletlab.com to download the tool and to access the complete list of indices.

2.5.1 Function and content word only indices

CRAT indices generally consider all words in a text. CRAT also includes index variants that include only the content words (e.g., nouns, verbs, adjectives, adverbs) and only the function words (e.g., determiners, prepositions, etc.). Content word indices and function word indices are designed to provide more fine-grained analyses, and have been shown to be more predictive, in some cases, than when all words are considered in an index [15].

2.5.2 Text and sentence minimum indices

CRAT indices generally comprise the average score for all instances of a feature across an entire text. Additionally, CRAT calculates index variants that comprise average minimum scores for each sentence in a text in order to assess smaller texts that may be a single sentence in length.

2.5.3 Key word exclusion indices

In addition to the index variants outlined above, constructed response sophistication indices include variants that exclude words that occur more frequently in the source text than would be expected (i.e., words that are "key"). The key word exclusion index variants were included to minimize interference from sophisticated language in the source text on the constructed response produced.

2.5.4 Latent Semantic Analysis Weighting

One variable that can affect LSA similarity scores is the weighting scheme employed. CRAT includes LSA variants calculated from the TASA corpus using normalized weighting, rare words dominated weighting, and frequent words dominated weighting. Normalized weighting considers all words in a reference corpus equally. Rare words dominated weighting assign higher scores to words that occur infrequently in the reference corpus. Frequent words dominated weighting assigns higher scores to words that frequently occur in the reference corpus [16].

2.6 Summary Input

CRAT is a domain specific tool and uses system input (i.e., source texts) to develop knowledge spaces for the domain of interest. The source texts used to develop knowledge spaces can be textbooks, lecture notes, presentations, or any type of text that generalizes expected knowledge on the part of the student. For this analysis, we used the hints provided to the students during specific activities within the ChemVLab+ system. These hints provide an overview of the input the student received and are designed to provide informational hints to students if they are unable to generate the information individually. The hints available to students in question 1 of the stoichiometry lab comprised over 5,000 words and focused specifically on the relationship between sugar, volume, and concentration in a sports drink.

2.7 Statistical Analysis

The indices reported by CRAT that yielded non-normal distributions were removed. A multivariate analysis of variance (MANOVA) was conducted to examine which indices reported differences between the three levels of scores for each student response (incomplete or incorrect, partially correct, and correct responses). The MANOVA was followed by stepwise discriminant function analysis (DFA) using the selected normally distributed indices from CRAT that demonstrated significant differences between responses that were incorrect or incomplete, partially correct, and correct and did not exhibit multicollinearity (r > .90) with other CRAT indices. In the case of multicollinearity between indices, the index demonstrating the largest effect size was retained in the analysis. The DFA was used to develop an algorithm to predict group membership through a discriminant function co-efficient. A DFA model was first developed for the entire corpus of constructed responses. This model was then used to predict group membership of the constructed responses using leave-one-out-cross-validation (LOOCV) in order to ensure that the model was stable across the dataset.

3. RESULTS

3.1 MANOVA

A MANOVA was conducted using the NLP indices calculated by CRAT as the dependent variables and the human scores of the student responses as the independent variables. Of the 759 indices

Table 1: Descriptive statistics and MANOVA results for CRAT variables

Index	Incomplete/incorrect Mean (SD)	Partially correct Mean (SD)	Correct Mean (SD)	*F*	*η2*
Semantic similarity (LSA) response and input (rare word dominated)	0.362 (0.159)	0.458 (0.111)	0.499 (0.079)	102.799**	0.131
Semantic similarity (LSA) response and input (frequent word dominated)	0.403 (0.155)	0.5 (0.113)	0.531 (0.096)	95.432**	0.122
Academic frequency COCA function words	24524.248 (16585.406)	36788.308 (13168.904)	34324.442 (11401.743)	76.716**	0.101
Written frequency (BNC) function words	1.000 (0.441)	1.227 (0.291)	1.25 (0.256)	53.237**	0.072
Percentage of adjectives	0.086 (0.082)	0.112 (0.069)	0.135 (0.074)	38.42**	0.053
Academic range (COCA) all words	-0.494 (0.254)	-0.401 (0.114)	-0.411 (0.096)	24.093**	0.034
Number of words	24.417 (29.134)	33.476 (53.923)	38.618 (39.975)	16.736**	0.024
Range (SUBTLEXus) content words (no key words)	3737.317 (1693.106)	3227.84 (1437.09)	3213.191 (1105.223)	15.819**	0.023
Academic frequency (COCA) content words sentence minimum	0.743 (0.705)	0.941 (0.532)	0.922 (0.487)	12.386**	0.018
Word familiarity (MRC) sentence minimum	497.207 (206.379)	560.031 (126.208)	529.915 (165.372)	10.534**	0.015
Percent content words	0.635 (0.147)	0.597 (0.085)	0.606 (0.091)	9.621**	0.014
Word familiarity (MRC) content words (no key words)	465.777 (132.451)	483.335 (87.545)	495.526 (77.668)	6.393*	0.009
Range (COCA all words sentence minimum)	-1.937 (0.143)	-1.96 (0.083)	-1.956 (0.08)	4.063*	0.006
Academic range (COCA; no key words)	0.712 (0.081)	0.693 (0.076)	0.689 (0.137)	3.865*	0.006

** p < .05,* ** p < .001

Table 2. Confusion matrix for DFA results for classifying scored responses

		Incomplete/incorrect	Partially correct	Correct	F_1 score
Whole set	Incomplete/incorrect	**605**	202	138	0.755
	Partially correct	31	**119**	60	0.400
	Correct	21	67	**129**	0.474

		Incomplete/incorrect	Partially correct	Correct	F_1 score
LOOCV	Incomplete/incorrect	**603**	203	139	0.752
	Partially correct	33	**113**	64	0.379
	Correct	22	70	**125**	0.459

reported by CRAT, 96 of these indices were normally distributed and not multi-collinear with one another. Of these 96 indices, 85 of the indices reported significant differences in the MANOVA analysis. These indices were related to overlap between the constructed response and the input received in the tutor, lexical sophistication, response length, response descriptiveness, and percentage of content words in the response. These indices were used in the subsequent DFA.

3.2 Discriminant Function Analysis

A stepwise DFA using the 85 indices selected through the MANOVA retained 14 variables related to semantic overlap between response and input, text descriptiveness, lexical sophistication, response length, and the use of content words. The indices retained in the DFA along with their means, standard deviations, *F* scores, *p* values, and effect sizes are reported in Table 1.

The results demonstrate that the DFA using these 14 indices correctly allocated 853 of the 1372 student responses in the total set, $\chi 2$ (df=4) = 393.169 p < .001, for an accuracy of 62.2%. For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 841 of the 1372 texts for an accuracy of 61.3% (see the confusion matrix reported in Table 2 for results and F_1 scores). The Cohen's Kappa measure of agreement between the predicted and actual class label was 0.404, demonstrating moderate agreement.

4. DISCUSSION

This analysis provides an initial assessment of the extent to which the linguistic indices reported by the Constructed Response Analysis Tool (CRAT) are predictive of constructed responses. We examined student constructed responses to a single question in the ChemVLab+ system related to stoichiometiry. We found that 86 CRAT indices demonstrated differences between the three levels of human ratings (incomplete/incorrect, partially correct, and correct) and 14 of these variables were significant predictors of human scores in a DFA with a reported accuracy of 62%. The results suggest that the CRAT tool can be used to automatically classify student constructed responses based on human ratings of response accuracy. While preliminary, the results support the use of NLP tools in constructed response scoring and point toward specific linguistic features that can be used to predict human ratings of accuracy for student constructed responses.

The discriminant function analysis indicated that the strongest predictors of human accuracy scores were related to semantic similarity between the constructed response and the knowledge space provided (i.e., the available student hints in the ChemVLab+). The results indicated that student responses that had a higher semantic overlap with the hints were more likely to be correct or partially correct. These results held for rare word and frequent word LSA overlap. This suggests that students whose responses better represent the semantic space of the domain are more likely to produce correct responses.

Beyond semantic overlap with the hints, the next strongest predictors of human scores of student responses were related to the frequency of function words. These indices indicated that students who used more frequent function words were rated as having higher response scores (for both academic and written frequency). This likely indicates that students who used function words that occur more frequently in written contexts (i.e., academic writing and writing in general) construct more accurate responses. Thus, more successful students were those who were more likely to use writing styles frequent in academic English.

More successful answers also differed in the properties of the words they contained. More accurate answers were more descriptive in that they contained a greater number of adjectives. Though longer, successful answers contained fewer content words (i.e., they contained more function words). Successful answers contained more specific words (i.e., words that demonstrated a lower range score) and also contained more familiar and frequent words.

The model developed in this pilot study reports a level of accuracy that is appropriate to provide automated feedback to users in a tutoring system such as ChemVLab+. This feedback could include a summative score to provide users with an overall assessment of the quality of the constructed response. In addition, the model could be used to provide formative feedback to users in terms of language use (i.e., the use of academic language) and appropriate content (i.e., is writer covering the content of the question appropriately). Such feedback could be used by students to revise their responses and engage more deeply with the system. However, we would caution against using the reported model in high stakes assessments where accuracy is at a premium, although this advice should be empirically tested on a number of high stakes test corpora.

CRAT differs from many other scoring systems in that it is domain specific. Domain specificity has advantages as many of the key word and semantic indices can be trained on targeted content that increases construct validity and ensures that topic adherence on the part of the student remains an important component of constructed response scoring. Training the system, however, requires source texts that provide background about the topic. In some cases, these texts may be difficult to transfer to text files (in the case of lectures) or they may not exist within a system, limiting the generalizability of CRAT across a number of system.

Lastly, it remains an open question if a model trained on one area of chemistry will transfer to another area of chemistry or to domains outside of chemistry. For instance, the model developed here needs to be tested on similar but not overlapping chemistry topics and questions to test the model's generalizability within a macro-domain (e.g., with chemistry questions that address molecular equilibrium and acid bases). In addition, the model should be tested on domains outside of chemistry to assess whether constructed responses in various domains can be accurately scored based on a combination of semantic and keyword overlap between the response and the source and the use of academic language by system users.

5. CONCLUSION

This study introduces a freely available tool for constructed response scoring and tests the tool on a dataset of chemistry responses collected in the ChemVLab+. The findings indicate that the Constructed Response Analysis Tool (CRAT) can differentiate and classify student responses based on semantic overlap with text input, syntactic categories, text length, and lexical sophistication indices. Overall, the findings suggest that successful student responses contain greater overlap with the content learned and use more academic function words, more words in general, more descriptive words, and more familiar and frequent words that are also more specific.

Additional studies will be conducted to refine and continue to develop CRAT. For example, a future direction includes assessing the value of including indices of semantic overlap that use Latent Dirichlet allocation (LDA) spaces, allowing for topic modeling along with semantic graph analyses. CRAT also needs to be tested on additional constructed responses, including responses from a variety of domains. Lastly, the models developed using the CRAT tool should be assessed for application in providing feedback to users in instructional systems. Such follow up studies will provide additional information about the reliability of CRAT and the linguistic features within CRAT that are predictive of human ratings of constructed responses within different domains and on-line learning environments.

6. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences and National Science Foundation (IES R305A080589, IES R305A100069, IES R305G20018-02, DRL -1418072, and DRL- 1418378). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES or the NSF.

7. REFERENCES

[1] NGSS Lead States. 2013. *Next Generation Science Standards: For States, By State*s. Washington, DC: The National Academies Press.

[2] Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19-28.

[3] Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103–118.

[4] Williamson, D., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.

[5] Honey, M. A., & Hilton, M. (Eds.). (2010). *Learning science through computer games and simulations*. Washington, DC: National Academies Press.

[6] Mitrovic, A. (2012). Fifteen years of constraint-based tutors: what we have achieved and where we are

going. *User Modeling and User-Adapted Interaction*, *22*(1-2), 39-72.

[7] Attali, Y., & Powers, D. (2008). Effect of immediate feedback and revision on psychometric properties of open-ended GRE subject test items. GRE Board Research Rep. No. 04-05; ETS RR-08-21. Princeton, NJ: Educational Testing Service.

[8] Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education*, 9, 133–150.

[9] Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2005). Automated scoring for creative problem-solving ability with ideation-explanation modeling. *In Proceedings of the Thirteenth International Conference on Computers in Education* (pp. 522–529). Singapore: IOS Press.

[10] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

[11] Brysbaert, M., Warriner, A.B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*. doi:10.3758/s13428-013-0403-5

[12] Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505. doi:10.1080/14640748108400805

[13] British National Corpus, version 3 (BNC XML ed.). (2007). Retrieved from http://www.natcorp.ox.ac.uk

[14] Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, *25*(4), 447-464.

[15] Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49*(4), pp. 757-786. doi: 10.1002/tesq.194

[16] McNamara, D. S., Cai, Z., & Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. *Handbook of latent semantic analysis*, 379-400.