Towards Domain-Independent Assessment of Elementary Students' Science Competency using Soft Cardinality

Samuel P. Leeman-Munk, Angela Shelton, Eric N. Wiebe, James C. Lester
North Carolina State University
Raleigh, North Carolina 27695
{ spleeman, anshelto, wiebe, lester } @ ncsu.edu

Abstract

Automated assessment of student learning has become the subject of increasing attention. Students' textual responses to short answer questions offer a rich source of data for assessment. However, automatically analyzing textual constructed responses poses significant computational challenges, exacerbated by the disfluencies that occur prominently in elementary students' writing. With robust text analytics, there is the potential to analyze a student's text responses and accurately predict his or her future success. In this paper, we propose applying soft cardinality, a technique that has shown success grading less disfluent student answers, on a corpus of fourth-grade responses to constructed response questions. Based on decomposition of words into their constituent character substrings, soft cardinality's evaluations of responses written by fourth graders correlates with summative analyses of their content knowledge.

1 Introduction

As a tool for automated assessment, short answer questions reveal cognitive processes and states in students that are difficult to uncover in multiple-choice equivalents (Nicol, 2007). Even when it seems that items could be designed to address the same cognitive construct, success in devising multiple-choice and short answer items that behave with psychometric equivalence has proven to be limited (Kuechler & Simkin, 2010). Because standards-based STEM education in the United States explicitly promotes the development of writing skills for which constructed response items are ideally suited (NGSS Lead States, 2013; Porter, McMaken, Hwang, & Yang, 2011; Southavilay, Yacef, Reimann, & Calvo, 2013), the prospect of designing text analytics techniques for automatically assessing students' textual responses has become even more appealing (Graesser, 2000; Jordan & Butcher, 2013; Labeke, Whitelock, & Field, 2013).

An important family of short answer questions is the constructed response question. A constructed response question is designed to elicit a response of no more than a few sentences and features a relatively clear distinction between incorrect, partially correct, and correct answers. Ideally, a system designed for *constructed response analysis* (CRA) would be machine-learned from examples that include both graded student answers and expert-constructed "reference" answers (Dzikovska, Nielsen, & Brew, 2012).

The challenges of creating an accurate machine-learning-based CRA system stem from the variety of ways in which a student can express a given concept. In addition to lexical and syntactic variety, students often compose ill-formed text replete with ungrammatical phrasings and misspellings, which significantly complicate analysis. The task of automated grading also becomes increasingly difficult as the material graded comes from questions and domains more and more distant from that of human graded responses on which the system is trained, leading to interest in domain-independent CRA systems designed to deal with this challenge (Dzikovska et al., 2013).

In this paper we explore the applications of soft cardinality (Jimenez, Becerra, & Gelbukh, 2013), an approach to constructed response analysis that has shown prior success in domain-independent CRA. We investigate whether soft cardinality is robust to the disfluency common among elementary students and whether its analyses of a student's work as she progresses through a problem-solving session can be used to roughly predict the content knowledge she will have at the end.

Because like other bag of words techniques, soft cardinality is independent of word order, it is robust to grammatical disfluencies. What distinguishes soft cardinality, however, is its character-overlap technique, which allows it to evaluate word similarity across misspellings. We evaluate soft cardinality on a dataset of textual responses to short-text science questions collected in a

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 61–67,
Baltimore, Maryland USA, June 26, 2014.

study conducted at elementary schools in two states. Responders were in fourth grade and generally aged between nine and ten. We train our system on student responses to circuits questions and test it on two domains in the physical sciences—circuits and magnetism. The results indicate that, soft cardinality shows promise as a first step for predicting a student's future success with similar content even grading unseen domains in the presence of high disfluency.

This paper is structured as follows. Section 2 provides related work as a context for our research. Section 3 introduces the corpus, collected on tablet-based digital science notebook software from elementary students. Section 4 describes soft cardinality and an evaluation thereof. Section 6 discusses the findings and explores how soft cardinality may serve as the basis for future approaches to real-time formative assessment.

2 Related Work

Short answer assessment is a much-studied area that has received increased attention in recent years. Disfluency and domain-independence have been the beneficiaries of some of this attention, but cutting edge systems seem to be designed first for correctly spelled in-domain text, and then have domain-independence and disfluency management added afterwards.

For example, one system from Educational Testing Services (ETS) uses an approach to domain independence called "domain adaptation" (Heilman & Madnani, 2013). Domain adaptation generates a copy of a given feature for grading answers to seen questions, answers to unseen questions in seen domain, and answers to questions in unseen domains, and each of these has a separate weight. An item represented in the training data uses all three of these feature copies, and an item from another domain will only use the latter, "generic" feature copy.

Spell correction is also often treated as a separate issue, handled in the data-cleaning step of a CRA system. The common approach at this step is to mark words as misspelled if they do not appear in a dictionary and replace them with their most likely alternative. This technique only corrects non-word spelling errors (Leacock & Chodorow, 2003). Another approach is to use *Soundex hashes* that translate every word into a normalized form based on its pronunciation (Ott, Ziai, Hahn, & Meurers, 2013). This second approach is generally featured alongside a more traditional direct comparison.

The primary limitation of CRA for elementary school education is that evaluations of state-of-the-art systems on raw elementary student response data are limited. C-rater provides a small evaluation on fourth-grade student math responses, but most evaluation is on seventh, eighth and eleventh grade students (Leacock & Chodorow, 2003; Sukkarieh & Blackmore, 2009). Furthermore, the two datasets presented in SemEval's shared task (Dzikovska et al., 2013) for testing and training featured relatively few spelling errors. The BEETLE corpus was drawn from undergraduate volunteers with a relatively strong command of the English language, and the SciEntsBank corpus, which was drawn from 3-6th graders, was originally intended for speech and as such was manually spell-corrected. The Hewlett Foundation's automated student assessment prize (ASAP) shared task for short answer scoring was drawn entirely from tenth grade students (Hewlett, 2012).

3 Corpus

We have been exploring constructed response assessment in the context of science education for upper elementary students with the LEONARDO CYBERPAD (Leeman-Munk, Wiebe, & Lester, 2014). Under development in our laboratory for three years, the CYBERPAD is a digital science notebook that runs on tablet and web based computing platforms. The CYBERPAD integrates intelligent tutoring systems technologies into a digital science notebook that enables students to model science phenomena graphically. With a focus on the physical and earth sciences, the LEONARDO PADMATE, a pedagogical agent, supports students' learning with real-time problem-solving advice. The CYBERPAD's curriculum is based on that of the Full Option Science System (Foss Project, 2013). As students progress through the curriculum, they utilize LEONARDO's virtual notebook, complete virtual labs, and write responses to constructed response questions. To date, the LEONARDO CYBERPAD has been implemented in over 60 classrooms around the United States.

The short answer and pre/post-test data used in this investigation were gathered from fourth grade students during implementations of The CYBERPAD in public schools in California and North Carolina. The data collection for each class took place over a minimum of five class periods with students completing one or more new investigations each day. Students completed

investigations in one or both of two modules, "Energy and Circuits," and "Magnetism." Most questions included "starter text" that students were expected to complete. Students were able to modify the starter text in any way including deleting or replacing it entirely, although most students simply added to the starter text. Example answers can be found in a previous work on the same dataset (Leeman-Munk et al., 2014).

Two human graders scored students' responses from the circuits module on a science score rubric with three categories: *incorrect*, *partially correct*, and *correct*. The graders graded one class of data and then conferred on disagreeing results. They then graded other classes. On a sample of 10% of the responses of the classes graded after conferring, graders achieved a Cohen's Kappa of 0.72.

The graders dealt with considerable disfluency in the student responses in the LEONARDO corpus. An analysis of constructed responses in the Energy and Circuits module reveals that 4.7% of tokens in all of student answers combined are not found in a dictionary. This number is higher in the Magnetism module, 7.8%. This is in contrast to other similar datasets, such as the BEETLE corpus of undergraduate text answers to science questions, which features a 0.8% rate of out-of-dictionary words (Dzikovska, Nielsen, & Brew, 2012). In each case, the numbers underestimate overall spelling errors. Misspellings such as 'batter' for 'battery', are not counted as missing in a dictionary test. These *real-word spelling errors* nevertheless misrepresent a student's meaning and complicate analysis. We describe how soft cardinality addresses these issues in Section 4.

4 Methodology and Evaluation

Soft cardinality (Jimenez, Becerra, & Gelbukh, 2013) uses decompositions of words into character sequences, known as *q-grams*, to gauge similarity between two words. We use it here to bridge the gap between misspellings of the same word. Considering "dcells" in an example answer, "mor dcells," and "D-cells" in the reference answer, we can find overlaps in "ce," "el," "ll," "ls," "ell," "lls," and so on up to and including "cells." This technique functions equally well for real-word spelling errors such as if the student had forgotten the "d" and typed only "cells." Such overlaps signify a close match for both of these words. We evaluated the soft cardinality implementation of a generic short answer grading framework that we developed, WRITEEVAL, based on an answer grading system described in an earlier work (Leeman-Munk et al., 2014). We used 100-fold cross-validation on the "Energy and Circuits" module. We compare WRITEEVAL using soft cardinality to the majority class baseline and to WRITEEVAL using Precedent Feature Collection (PFC), a latent semantic analysis technique that performs competitively with the second highest-scoring system in Semeval Task 7 on unseen answers on the SciEntsBank corpus (Dzikovska et al., 2013). Using a Kruskal-Wallis test over one hundred folds, both systems significantly outperform the baseline ($p<.001$), which achieved an accuracy score of .61. We could not evaluate the scores directly on the Magnetism dataset as we did not have any human-graded gold standard for comparison.

To evaluate soft cardinality's robustness to disfluency, we created a duplicate of the Energy and Circuits dataset and manually spell-corrected it. Table 1 and Figures 1 and 2 show our results. Using the Kruskal-Wallis Test, on the uncorrected data PFC's accuracy suffered with marginal significance ($p = .054$) while macro-averaged precision and recall both suffered significantly ($p < .01$). Soft cardinality suffered much less, with a marginally significant decrease in performance ($p=.075$) only in recall. The decreases in accuracy and precision had $p=.88$ and $p=.25$ respectively.

To determine the usefulness of automatic grading of science content in predicting the overall trajectory of a student's performance, we computed a running average of the grades given by soft cardinality (converted to '1', '2', and '3' for incorrect, partially correct, correct) on students' answers as they progressed through the Energy and Circuits module and the Magnetism module. Because we would intend to be able to use this technique in a classroom on entirely new questions and student answers, we use running average instead of a regression, which would require prior data on the questions to determine the weights.

Students completed a multiple-choice test before and after their interaction with the CYBERPAD. The Energy and Circuits module and the Magnetism module each had different tests – there were ten questions on the Energy and Circuits test and twenty on the Magnetism test. We calculated the correlation of our running average of formative assessments against the student's score on the final test.

A critical assumption underlying the running average is that students answered each question

in order. Although WRITEEVAL does not prevent students from answering questions out of order, it is organized to strongly encourage linear progression.

We excluded empty responses from the running average because we did not want an artificial boost from simply noting what questions students did and did not answer. Data from students who did not take the pre or post-test was excluded, and students missing responses to more than twenty out of twenty-nine questions in Magnetism or fifteen out of twenty questions in Energy and Circuits were excluded from consideration. After cleaning, our results include 85 students in Energy and Circuits and 61 in Magnetism.

Sp.Cr.	System	Accuracy	Precision	Recall
Yes	SoftCr	.68	.55	.54
No	SoftCr	.68	.52	.50*
Yes	PFC	.78	.61	.58
No	PFC	.74*	.54**	.52**

Table 1. Accuracy and Macro-Averaged Precision and Recall for Soft-Cardinality and PFC on spell-corrected and uncorrected versions of the LEONARDO Energy and Circuits module.
*marginally significant decrease from spell-checked
**significant decrease from spell-checked

Figure 1 depicts the correlation between the running average of automatic scoring by WRITEEVAL soft cardinality, PFC, and human scores with post-test score on the responses in the Energy and Circuits module. When spell-corrected, the correlation, as shown in Figure 2, surprisingly becomes worse. We discuss a possible reason for this in the discussion section.

Figure 3 shows correlation of the running average of Magnetism's automatic scores with post-test. For soft cardinality, significant correlation starts five questions in and stays for the rest of the 29. As it relies heavily on relevant training data, PFC is less stable and does not achieve nearly as high a correlation.

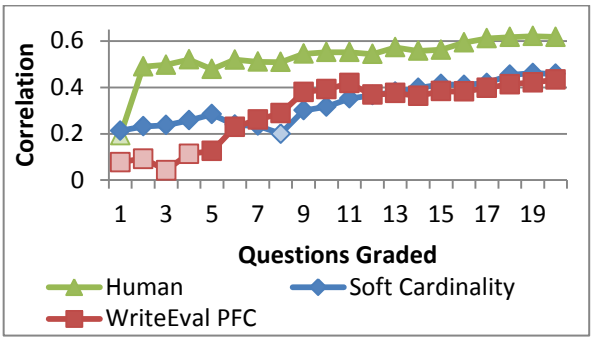

Figure 1. Correlation of grading systems on Energy and Circuits with post-test score. Dark-colored points indicate significant correlation (p<.05)

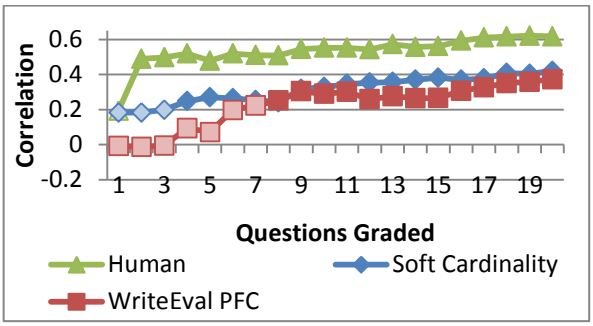

Figure 2. Correlation of grading systems on spell-corrected Energy and Circuits with post-test score. Dark-colored points indicate significant correlation (p<.05)

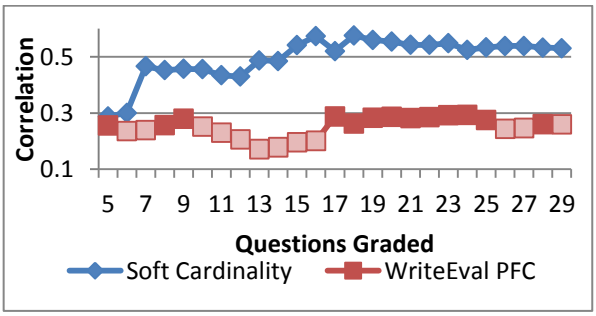

Figure 3. Correlation of the Running Average of WRITEEVAL with soft cardinality with post-test Scores on the Magnetism module of the LEONARDO corpus. Dark-colored points indicate significant correlation (p<.05)

5 Discussion

The evaluation suggests that a relatively simple technique such as soft cardinality, despite performing less well than a domain specific technique in the presence of relevant training data, is more robust to spelling errors and can be far more effective at grading questions and domains not present in the training data.

Soft cardinality is representative of the potential of domain independent, disfluency-robust CRA systems.

The improvement against the gold standard on spell-corrected data but loss of correlation against the post-test scores suggests that poor spelling is a predictor of poor post-test

knowledge at the end of a task. This could be because the students were less able to learn the material due to their poor language skills, they were less able to complete the test effectively despite knowing the material again due to poor language skills, or it could be a latent factor that affects both the students use of language and their eventual circuits knowledge such as engagement. This result shows the challenge of separating different skills in evaluating students.

The significance of soft cardinality's correlation over the running average for all but the eighth question as well as the generally high significant correlation achieved in the magnetism evaluation indicates the predictive potential of soft cardinality. Soft cardinality's performance in Magnetism suggests that with only a relatively limited breadth of training examples it can effectively evaluate answers to questions in some unseen domains. It is important to note that Energy and Circuits and Magnetism are both subjects in the physical sciences, and the questions and reference answers themselves were authored by the same individuals. As such this result should not be overstated, but is still a promising first step towards the goal of domain-independence in constructed response analysis.

6 Conclusion

This paper presents a novel application of the soft cardinality text analytics method to support assessment of highly disfluent elementary school text. Using q-gram overlap to evaluate word similarity across nonstandard spellings, soft cardinality was evaluated on highly disfluent constructed response texts composed by fourth grade students interacting with a tablet-based digital science notebook. The evaluation included an in-domain training corpus and another out-of-domain corpus. The results of the evaluation suggest that soft cardinality generates assessments that are predictive of students' post-test performance even in highly disfluent out-of-domain corpora. It offers the potential to produce assessments in real-time that may serve as early warning indicators to help teachers support student learning.

Soft cardinality's current performance levels suggest several promising directions for future work. First, it will be important to develop techniques to deal with widely varying student responses without relying directly on training data. These techniques will take inspiration in part from bag-of-words techniques such as soft cardinality and Precedent Feature Collection, but will themselves likely take word order into account as there is a sizeable subset of answers whose meaning is dependent on word order. The use of distributional semantics will also be of help in resolving similarities between different words. Secondly, work should be done to consider answers in more detail than simple assessment of correctness. More detailed rubrics such as Task 7's 5-way rubric (Dzikovska et al., 2013) would allow for more detailed feedback from tutors. Further, detailed analysis of individual understandings and misconceptions within answers would be even more helpful, and will be the focus of future work. Third, it will be instructive to incorporate the WRITEEVAL framework into the LEONARDO CYBERPAD digital science notebook to investigate techniques for classroom-based formative assessment that artfully utilize both intelligent support by the PADMATE onboard intelligent tutor and personalized support by the teacher. Finally, it will be important to to investigate additional techniques to evaluate student answers more accurately using less training data from more distant domains.

Reliable analysis of constructed response items not only provides additional summative analysis of writing ability in science, but also gives the teacher a powerful formative assessment tool that can be used to guide instructional strategies at either the individual student or whole class level. Given that time for science instruction is limited at the elementary level, the use of real-time assessment to address student misconceptions or missing knowledge immediately can be an invaluable classroom tool.

7 Acknowledgements

The authors wish to thank our colleagues on the LEONARDO team for their contributions to the design, development, and classroom implementations of LEONARDO: Courtney Behrle, Mike Carter, Bradford Mott, Peter Andrew Smith, and Robert Taylor. This material is based upon work supported by the National Science Foundation under Grant No. DRL1020229. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Dzikovska, M., Brew, C., Clark, P., Nielsen, R. D., Leacock, C., Mcgraw-hill, C. T. B., & Bentivogli, L. (2013). SemEval-2013 Task 7 : The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 263–274).

Dzikovska, M., Nielsen, R., & Brew, C. (2012). Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 200–210). Montreal, Canada. Retrieved from http://dl.acm.org/citation.cfm?id=2382057

Foss Project. (2013). Welcome to FossWeb. Retrieved October 20, 2013, from http://www.fossweb.com/

Graesser, A. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*(2), 1–33. Retrieved from http://www.tandfonline.com/doi/full/10.1076/1049-4820(200008)8%3A2%3B1-B%3BFT129

Heilman, M., & Madnani, N. (2013). ETS : Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 1, pp. 96–102).

Hewlett, W. (2012). The Hewlett Foundation: Short answer scoring. Retrieved March 16, 2014, from https://www.kaggle.com/c/asap-sas/data?Data_Set_Descriptions.zip

Jimenez, S., Becerra, C., & Gelbukh, A. (2013). SOFTCARDINALITY: hierarchical text overlap for student response analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 280–284). Retrieved from http://www.gelbukh.com/CV/Publications/2013/SOFTCARDINALITY Hierarchical Text Overlap for Student Response Analysis.pdf

Jordan, S., & Butcher, P. (2013). Does the Sun orbit the Earth ? Challenges in using short free-text computer-marked questions . In *Proceedings of HEA STEM Annual Learning and Teaching Conference 2013: Where Practice and Pedagogy Meet*. Birmingham, UK.

Kuechler, W., & Simkin, M. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55–73. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4609.2009.00243.x/full

Labeke, N. Van, Whitelock, D., & Field, D. (2013). OpenEssayist: extractive summarisation and formative assessment of free-text essays. In *First International Workshop on Discourse-Centric Learning Analytics*. Leuven, Belgium. Retrieved from http://oro.open.ac.uk/37548/

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405. Retrieved from http://link.springer.com/article/10.1023/A%3A1025779619903

Leeman-Munk, S. P., Wiebe, E. N., & Lester, J. C. (2014). Assessing Elementary Students' Science Competency with Text Analytics. In *Proceedings of the Fourth International Conference on Learning Analytics & Knowledge*. Indianapolis, Indiana.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States.* Washington DC: National Academic Press.

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03098770601167922

Ott, N., Ziai, R., Hahn, M., & Meurers, D. (2013). CoMeT : Integrating different levels of linguistic modeling for meaning assessment. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 608–616).

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards the new US

intended curriculum. *Educational Researcher*, *40*(3), 103–116. Retrieved from http://edr.sagepub.com/content/40/3/103.short

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (pp. 38–47). New York, New York, USA: ACM Press. doi:10.1145/2460296.2460307

Sukkarieh, J., & Blackmore, J. (2009). C-rater: Automatic content scoring for short constructed responses. *Proceedings of the 22nd International FLAIRS Conference*, 290–295. Retrieved from http://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/download/122/302