Stratified learning for reducing training set size

Peter Hastings[1]*, Simon Hughes[1], Dylan Blaum[2],
Patricia Wallace[2], and M. Anne Britt[2]

[1] DePaul University, Chicago, Illinois
[2] Northern Illinois University, DeKalb, Illinois

Abstract. Educational standards put a renewed focus on strengthening students' abilities to construct scientific explanations and engage in scientific arguments. Evaluating student explanatory writing is extremely time-intensive, so we are developing techniques to automatically analyze the causal structure in student essays so that effective feedback may be provided. These techniques rely on a significant training corpus of annotated essays. Because one of our long-term goals is to make it easier to establish this approach in new subject domains, we are keenly interested in the question of how much training data is enough to support this. This paper describes our analysis of that question, and looks at one mechanism for reducing that data requirement which uses student scores on a related multiple choice test.

1 Introduction

The Next Generation Science Standards (NGSS) provide detailed expectations about engaging students in the practices of constructing scientific explanations and engaging in arguments from evidence about important everyday phenomena using complex literacy and modeling skills [1]. Explaining how or why phenomena occur is a key goal of scientific research [1,13]. However, most students have trouble with explanation and argumentation, particularly in science [6,7,9,12]. In constructing an explanation, students provide an assertion that states how one or more factors lead to the to-be-explained phenomenon through one or more intermediate processes or mechanisms [3,11,13]. Insufficient domain knowledge prevents readers from making the connections required for creating a coherent representation of a scientific explanation [3,10,13].

The high level goal of Project READi is to provide a deeper understanding of how students *read* texts. An important method for assessing that skill is analyzing what they write. In this paper, we are focusing on explanatory essays that students write after reading several short documents. Our long-term goals for this research are to be able to automatically provide analyses of student explanations, and to be able to extend these techniques to other domains. We

* The assessment project described in this article is funded, in part, by the Institute for Education Sciences, U.S. Department of Education (Grant R305F100007). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

are approaching the first goal, but the machine learning mechanism underlying our system is trained on over 1000 essays that have been meticulously annotated by human coders. To achieve the second goal, we need to understand just how much training data we need, and if there are more efficient mechanisms for training. Here we describe one technique.

2 Evaluation context

To describe students' overall skill in constructing causal explanations from reading multiple documents of a variety of types (e.g., descriptive texts, graphs and maps), 10th-grade students in science classes were asked to read a set of documents and write an essay explaining the causes of a scientific phenomenon [2]. Students wrote their essays with the documents available and were given several hints to make sure they understood the task. Then, while they still had the texts, students were given 9 multiple choice questions to assess learning of the causal model with a low-production (high recognition) measure of learning. 1011 students received the coral bleaching assessment ("explain how and why coral bleaching rates vary at different times"). Human coders annotated the essays for concepts mentioned (e.g., increased coral stress) and the connectedness of their causal chains against our causal model (e.g., increased coral stress causes ejection of algae which causes coral bleaching — see [7] for a causal model of coral bleaching). Inter-rater reliability between two human scorers was high (average $\kappa = 0.85$).

For a subset of the essays (440 students; 59.5% female and 33.6% Hispanic, 25.7% African American, 20.0% White, 4.5% Asian) we analyzed their essay quality into four categories [7] based on the completeness and coherence of their explanations. On average, students had difficulty in constructing an explanation from multiple documents with only 30.9% of the essays including an explanation with at least one intervening factor (highest quality). 25.7% of the essays included no target concepts whatsoever (lowest quality), 15.7% included at least one concept but it was not connected to the outcome, and 27.7% made at least one connection to the outcome but with no intervening elements.

The high production essay task and the low production multiple choice measures did significantly converge on assessing student learning. First, there was a significant effect of essay quality category on multiple choice percent correct ($F = 45.12$, $MSE = 2.48$, $p < .001$). The average percents correct on the multiple choice test for the four quality groups were 32%, 47%, 52% and 67%, respectively. Post hoc SNK found that those in the lowest essay category learned less than the middle two groups (which did not differ from each other) which both learned less than the highest quality group. Second, there was a significant correlation between the number of unique core concepts that students mentioned in their causal chains (claims) and their accuracy on the multiple choice measure (Pearson correlation = .43, $p < .001$).

3 Stratified learning

3.1 Identifying concepts and structure

In earlier work, we experimented with a number of different machine learning techniques to detect the core concepts and claims in student essays from several different domains, including history and science [8,7, for example]. We compared the efficacy of a set covering algorithm using frequent multi-word expressions with that of Latent Semantic Analysis (LSA) at detecting how effectively students were using information from different sources when constructing evidence-based history essays [8, for example]. In more recent work, we have focused on detecting causality in scientific explanatory essays. To address this problem, we decomposed the problem of causality detection into two simpler problems, a word-tagging problem and a sentence classification problem. The word tagging problem involves predicting which concept code or codes, if any, are associated with each word in the essay. The sentence classification problem then involves taking the predicted concepts for each word in a sentence, aggregating these predictions across the sentence and then determining what causal relations exist between these identified concepts.

The word tagging problem requires the algorithm to predict varying numbers of concepts per word. This is called a Multi-Label Classification problem (MLC) and presents a challenge as most machine learning algorithms are designed to predict a single class at a time. To solve this problem, we use a problem transformation method called *Binary Relevance* (BR), in which you train a separate binary classifier for each concept code to be predicted. With BR, we moved a fixed-size sliding window of 7 words across the text, using the words within the window and their relative positions as separate features for the classifier [5].

To solve the sentence classification problem, we use stacked generalization [15], feeding the predictions from the word tagging models as features into a 'meta-classifier'. For each sentence, we create a separate feature for each concept code that was predicted for at least one word in the sentence, and also for each unique pair of predicted concepts. Additionally, we compute the minimum and maximum probabilities predicted by each classifier for each concept over all words in the sentence. Using these features, we then train a second ensemble of logistic regression classifiers to detect whether each sentence has a causal relation, and if so between which concepts.

3.2 Meta-evaluation

As described in the introduction, this paper has two important goals: to determine just how much training data is needed to achieve a level of accuracy which is "acceptable" for providing relevant feedback to students, and to determine if there is a more effective training protocol which will enable us to reduce the required amount of training data. The training protocol we explore in this paper was inspired by stratified sampling from the world of statistics[14], so we call it Stratified Learning. One potential problem with machine learning approaches

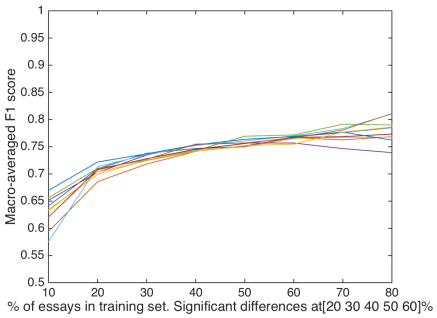

Fig. 1. Averaged F1 scores

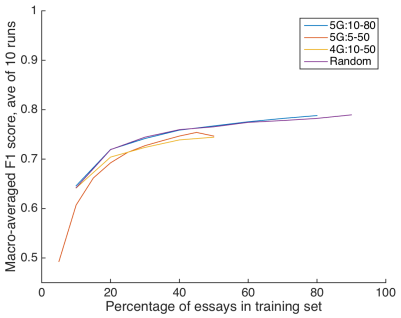

Fig. 2. Coarse-grained F1 scores

occurs when the concepts to be learned have skewed distributions. Stratified learning attempts to avoid this by taking advantage of prior knowledge about the data to be learned. In this case, when we are learning about student essays, we already have their scores on the multiple choice tests, and as mentioned above, we know that there is a moderate correlation between these scores and their essay quality. We hypothesize that by taking advantage of this prior knowledge, and by ensuring a balanced training set, the accuracy of the learned model would be greater than one trained with an equivalent number of essays from an imbalanced set.

With stratified learning, we start with a given, relatively small amount of the training set, and increase the training set size by that same percent, while ensuring that we get a (roughly) even number of items from each stratum or group. In the simple case, we used 5 groups[3] based on the multiple choice scores, and started with 10% of the 1011 essays (in equal groups) in the training data. So there were approximately 22 essays from students with test scores of 0 or 1, 22 essays with scores of 2 or 3, and so on. Fig. 1 shows the F1 scores for this technique for 10 different runs starting with 10% and increasing up to having 80% of the data in the training set. The F1 score for each essay was computed by averaging the accuracy of predictions for all the concept codes and causal connections within each sentence of the essay. Then the Macro-averaged F1 score was calculated based on all the remaining essays in the test set. The divergence that is evident at the 80% level is presumed to be due to the relative scarcity of some groups in the test set at that point. (Note that the Y axis starts here at 0.5 in order to make the distinctions more obvious.)

[3] The choice of group size is significant. As mentioned above, the distribution of multiple choice scores was fairly normal, and the least frequent score, 0, was assigned to 31 students. In order to maintain balanced representation of groups in the training set, some aggregation is necessary otherwise we could only test on a maximum of 31 items from each group. If the aggregation was too broad, however, it would decrease any benefit of balance in the training set.

The most obvious observations from this simulation are that there is consistent increase between the iterations, and that the largest jumps are on the left. As shown in the figure's subtitle, most of the differences between iterations are statistically significant. In terms of the minimum accuracy that is required to provide meaningful feedback to students, we generally find $F1 = 0.7$ to be a useful threshold. The machine learning technique was almost successful at achieving this level with only 10% of the essays (around 100), and it quickly reaches this level when going up to around 200 essays.

We also tested the approach using smaller increments of 5% (labeled "5G:5-50" in Fig. 2), and with 4 groups instead of 5, separated by multiple choice scores of 0–2, 3–4, 5–6, and 7–9 (labeled "4G:10-50"). This tested coarser granularity of group size, while maintaining roughly equal distribution. For this simulation, we took care to avoid the problem that was evident on the right side of Fig. 1, namely increased variance due to greatly diminished size of one or more groups in the *test* set. For this simulation, when we created the initial (balanced) training set at the beginning, we also created a separate, "held out" test set which would not be used as a source for additional training items in later iterations.

Fig. 2 shows the results of these simulations, this time averaged over the 10 runs for clarity. This image makes it very clear that there is a robust increase in the accuracy going from 5 to 10 to 20% of the corpus. This makes us confident of one of the answers which we were after: 100–200 annotated essays should be sufficient to achieve acceptable levels of classification accuracy.

Fig. 2 shows the performance of one additional method: random selection of equivalent increments of essay numbers from the corpus to add to the training set. In other words, this method does not use any balancing at all. Unfortunately for the Stratified Learning approach, the random approach is obviously every bit as effective, without the overhead of matching the scores on the multiple choice test. This provides an answer to the second question. If we are looking for a quicker route to reaching better performance, stratified learning is not it — at least in the case of our stacked learning context, as we will further discuss below.

4 Conclusions and future directions

In this paper, we have addressed two questions related to machine learning approaches for identifying structure in student explanatory essays: how much training data is required, and is training efficiency improved by maintaining a balanced training set. The exploratory goal showed us that a relatively modest training set size of 100–200 annotated essays should be sufficient to achieve adequate classification accuracy with our stacked machine learning mechanism. Our hypothesis about the benefits of stratified learning was not supported, however. There are several possible reasons for this. One is that although the correlation between multiple choice scores and essay quality is moderate, it is not especially high. Alternatively, there may be enough continuity between the lower- and higher-frequency groups that the random sample approach is not significantly disadvantaged relative to the stratified approach.

We have recently begun exploring another sampling mechanism called active learning [4], which shares some similarity with the last "imbalanced" technique we described. With this approach, the system is trained on some subset of items, then attempts to classify the rest. Some of the items that it is least (or most, or some combination) confident in predicting are then added to the training set, and the process repeats. Early simulations show that this technique may actually increase learning efficiency over random selection.

References

1. Achieve, Inc: Next Generation Science Standards: The common core standards for english language arts and literacy in history/social studies and science and technical subjects. Council of Chief State School Officers (2013)
2. Britt, M.A., Wallace, P., Blaum, D., Ko, M., Goldman, S.R., Project READI Science Design Team: Multiple representations in science learning and assessment (April 2015), paper presented in the Multiple Representations and Multimedia: Student Learning and Instruction symposium at AERA conference, Chicago IL
3. Britt, M.A., Richter, T., Rouet, J.F.: Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. Educational Psychologist 49(2), 104–122 (2014), dOI:10.1080/00461520.2014.916217
4. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning 15(2), 201–221 (1994), http://dx.doi.org/10.1007/BF00993277
5. Dietterich, T.G.: Machine learning for sequential data: A review. In: Structural, syntactic, and statistical pattern recognition, pp. 15–30. Springer (2002)
6. Duschl, R., Osborne, J.: Supporting and promoting argumentation discourse in science education. Studies in Science Education 38, 39–72 (2002)
7. Hughes, S., Hastings, P., Britt, A., Wallace, P., Blaum, D.: Machine learning for holistic evaluation of scientific essays. In: Proceedings of Artificial Intelligence in Education 2015 (2015)
8. Hughes, S., Hastings, P., Magliano, J., Goldman, S., Lawless, K.: Automated approaches for detecting integration in student essays. In: Cerri, S., Clancey, W., Papadourakis, G., Panourgia, K. (eds.) Proceedings of Intelligent Tutoring Systems 2012 (2012)
9. Kelly, G.J., Druker, S., Chen, C.: Students' reasoning about electricity: combining performance assessments with argumentation analysis. International Journal of Science Education 20(7), 849–871 (1998)
10. Meyer, B.J., Freedle, R.O.: Effects of discourse type on recall. American Educational Research Journal 22(1), 121–143 (1984)
11. Millis, K.K., Morgan, D., , Graesser, A.C.: The influence of knowledge-based inferences on the reading time of expository text. Psychology of Learning and Motivation 25, 197–212 (1990)
12. Osborne, J., Erduran, S., Simon, S.: Enhancing the quality of argumentation in science classrooms. Journal of Research in Science Teaching 41(10), 994–1020 (2004)
13. Osborne, J., Patterson, A.: Scientific argument and explanation: A necessary distinction? Science Education 95, 627–638 (2011)
14. Shahrokh Esfahani, M., Dougherty, E.R.: Effect of separate sampling on classification accuracy. Bioinformatics 30(2), 242–250 (2014), http://bioinformatics.oxfordjournals.org/content/30/2/242.abstract
15. Wolpert, D.H.: Stacked generalization. Neural networks 5(2), 241–259 (1992)