Modeling Classroom Discourse: Do Models that Predict Dialogic Instruction Properties Generalize across Populations?

Borhan Samei[1] Andrew M. Olney[1] Sean Kelly[2] Martin Nystrand[3]
Sidney D'Mello[4] Nathan Blanchard[4] Art Graesser[1]
[1] University of Memphis [2] University of Pittsburgh [3] University of Wisconsin [4] University of Notre Dame
bsamei@memphis.edu

ABSTRACT

It has previously been shown that the effective use of dialogic instruction has a positive impact on student achievement. In this study, we investigate whether linguistic features used to classify properties of classroom discourse generalize across different subpopulations. Results showed that the machine learned models perform equally well when trained and validated on different subpopulations. Correlation-Based Feature Subset evaluation revealed an inclusion relationship between different subsets in terms of their most predictive features.

Keywords

Classroom Discourse, Machine Learning, Authenticity, Uptake

1. INTRODUCTION

Previous research on classroom instruction has shown the positive influence of dialogic instruction on student achievement [2]. Dialogic instruction is a classroom discourse strategy based on the free and open exchange of ideas between teachers and students. It is hypothesized that dialogic instruction improves achievement by increasing student engagement in classrooms [3, 5].

Previous efforts to carefully quantify teachers' use of dialogic instruction include three major studies by Nystrand and colleagues [6]. Nystrand et al.'s approach included coding discourse moves with a focus on the nature of question events, which are defined by the discourse context preceding and following a question. Question events include the question along with the response and optional evaluation/follow up. They follow a pattern that mirrors the well-known initiation response, and evaluation sequence (IRE). This coding scheme treats questions as sites of interaction and takes into account the response and evaluation. As a result, the questions alone do not uniquely determine the dialogic properties of the event; instead, they create a context through which dialogic properties may be realized.

In this research, question events were coded with five properties that were hypothesized to relate dialogic instruction to student achievement: authenticity, uptake, level of evaluation, cognitive level, and question source. However, Nystrand and Gamoran found that among these variables, authenticity and uptake were the most strongly related to student achievement [2, 8]. A question is defined as having authenticity when the asker does not have a pre-scripted answer, i.e. an open-ended question, which creates a context for students to contribute to an open ended discussion. Uptake occurs when one asks a question about something that another person has said previously. When teachers exhibit uptake, they incorporate student contributions into the discussion, potentially encouraging additional student contributions.

Question properties were live-coded by observers in Nystrand et al.'s study, a time-consuming and expensive process requiring trained classroom observers. To facilitate research into dialogic instruction, we recently developed a machine learning model to investigate the extent to which question properties can be automatically coded [9]. This previous study showed that machine learned models can predict authenticity and uptake as accurately as human experts in a setting where the questions are presented without the preceding and following context, which was the information available to the machine learned model.

Machine learned models, often referred to as *predictors or classifiers,* are sensitive to the properties of the data set on which they are trained. However, in order to perform large scale analysis, these models must be applicable to new, larger, and more diverse data. An important question in this work is whether the models systematically vary their predictions with different subpopulations in the data (e.g. different demographics). This systematic variation, essentially bias, could lead to incorrect predictions and flawed conclusions when the model is applied to a sample drawn from the same subpopulation as opposed to different subpopulations and indeed any sample where the individuals are spatially or temporally correlated may potentially have problems of generalizability.

Some recent research has focused on examining generalizability of EDM models. For example, Baker and Gowda studied the difference in student behaviors associated with disengagement in urban, suburban, and rural schools and found that urban students went off-task more often and exhibited significantly more careless behaviors than students in the rural and suburban schools [1]. Furthermore, Ocumpaugh et al, found that models trained on a population drawn primarily from one demographic grouping (rural, urban, or suburban) do not always generalize to populations drawn primarily from the other demographic groupings [7]. Generalization can sometimes occur across seemingly distinct contexts. For example, San Pedro et al. (2011) found that their models of detecting student carelessness were generalizable among different tutor interfaces (i.e. with and without an embodied conversational agent), as well as different school settings (i.e. Philippine high school and US middle school) [10].

In this paper we investigated the generalizability of two previously developed models for predicting authenticity and uptake in classroom discourse [9].

2. METHOD

We trained and tested our models using data collected from the Partnership for Literacy Study (Partnership). The data set consists of question events as recorded by the classroom observers. Partnership was a study of professional development, instruction, and literacy outcomes in middle school, in which 120 classrooms in 21 schools were observed twice in the fall and twice in the spring over two years. The Partnership data set consists of observational data which were coded using the CLASS 4.24 computer-based data coding program [9]. Inter-rater agreement was approximately 80% on question properties with observation-level inter-rater correlations averaging approximately .95 [6].

Some of the teachers received special training in the first year and their classes were observed again in the second year. We used teacher training to split the data into Pre-training (N=7082) and Post-training (N=13655) groups. The school location was coded into categories of large and mid-size central city, urban fringe of mid-size city, small town rural outside MSA (metropolitan statistical area), and rural inside MSA. Based on the number of data points in each category, we split the data in two categories: Urban (i.e. Mid-size and Large Central City, N=13126) vs. Non-urban (the rest of categories, N=10911). Table 1 shows the distribution of authenticity and uptake across the different splits.

Table 1. Proportion of Authenticity and Uptake in different subsets and the full data set.

Category	% Authenticity	% Uptake
Non-urban : Urban	54 : 47	23 : 20
Pre-training : Post-training	39 : 52	15 : 24
Full-set	50	21

As seen in Table 1, authentic questions were more frequent than uptake in general, and the Non-urban group had higher rates of both authenticity and uptake than Urban. Overall the distribution of authenticity and uptake was similar among Non-urban, Post-training, and Full-set. Pre-training had the lowest rate of authenticity and uptake compared to others. It is also worth noting that teacher training was apparently quite effective at increasing both authenticity and uptake, as shown by the increase from Pre- to Post-training.

Based on our previous work on automating coding the questions with authenticity and uptake [9], we applied machine learning to train separate classifiers for authenticity and uptake on each of the above subsets. The models use linguistic features utilized in the classification of question types [8], including parts of speech, manually constructed bags of words (e.g., causal antecedent words), and positional information.

Most of the features are binary and indicate the presence/absence of certain keywords or part of speech tags in the question. Other features include attributes that show the position of the target keyword in the question in addition to presence/absence using four values: middle, beginning, end, and none. For example, if a question consisted of four words, e.g. “word1 word2 word3 word4” the position of “word1” is captured as beginning and “word4” as end, furthermore “word2” and “word3” are both captured as middle and if there were only two words in the question, we consider the first one as the beginning and the other as the end.

An example of a feature is causal consequent words, which include “outcomes,” “results,” “effects,” etc. Similarly, procedural words are defined as a set of keywords including “plan,” “scheme,” “design,” etc. Moreover, part of speech tags, such as determiner, noun, pronoun, adjective, adverb, and verb, and certain words such as “What,” “How,” and “Why,” were also included in the feature set. More complete descriptions and justifications of these features for question classification can be found in the mentioned references.

We first trained models on each subset and evaluated their performance using 10-fold cross validation within the subset. Next, we tested generalizability by training on one subset and testing on its dual. For example, a model trained on Urban subset was tested on the Non-urban subset and vice versa. Moreover, the models trained on the full set of data were also tested on each subset. This methodology allows for the following contrasts. First, cross validation within a subset establishes a reasonable upper bound on performance since training and testing instances, while distinct, still come from the same subset. Second, training on one subset and testing on its dual subset establishes a reasonable lower bound on performance, since accuracy would be determined by shared features between the subsets rather than by distinctive properties to each subset. Training on the full data set and testing on subsets (thus training and testing on those subsets) allows similar comparisons of bias. For example, if training on the full set and testing on set A has higher accuracy than testing on set B, we may hypothesize that the features of the full model are better aligned with the features of A, or the prevalence of category distribution in the full set better matches that of A.

3. RESULTS & DISCUSSION

We first trained separate models to predict authenticity and uptake and evaluated the models using on 10-fold cross validation for each subset. For each category (e.g. Urban, Non-urban, etc.) separate decision tree models were trained and evaluated using WEKA [4]. The models for predicting uptake were trained on a random subsample of the data to obtain an even (50-50) distribution. Table 2 shows the performance of the models along with the performance of a model trained on the full set of data.

Table 2. Performance of the decision tree models trained on different data subsets using 10-fold cross validation.

Training Data	Authenticity		Uptake	
	Accuracy	*Kappa*	*Accuracy*	*Kappa*
Non-urban	0.61	0.21	0.59	0.19
Urban	0.62	0.24	0.60	0.20
Pre-training	0.64	0.24	0.61	0.23
Post-training	0.63	0.26	0.61	0.22
Full-set [9]	0.64	0.28	0.62	0.24

As seen in Table 2, the models on different splits show comparable performances, where the maximum difference on their accuracy is 0.03 (3%). To examine performance of these models and their generalizability across different subsets, we trained models on one subset and tested on its dual subset, e.g. Urban – Non-Urban. In Table 3, the performance of each model is tested on its dual. Additionally, the models trained on full set of data are tested on different subsets.

Table 3. Generalizability of models on different splits of data (trained on one tested on other).

Train	Test	Authenticity	Uptake
		Accuracy	*Accuracy*
Non-urban	Urban	0.60	0.63
Urban	Non-urban	0.62	0.62

Full-set	Non-urban	0.70	0.68
Full-set	Urban	0.68	0.68
Pre-training	Post-training	0.59	0.62
Post-training	Pre-training	0.60	0.64
Full-set	Pre-training	0.70	0.68
Full-set	Post-training	0.72	0.67

In Table 3, training on one subset and testing on its dual is never more than 2 percentage points away from the reverse. Thus the results are fairly stable. However there are several patterns of differences of interest. First, accuracy for the authenticity models when trained on Urban and tested on Non-urban is slightly higher than when trained on Non-urban and tested on Urban, however the uptake model performs slightly better when trained on Non-urban and tested on Urban than the reverse. Moreover, uptake and authenticity accuracy were higher for models trained on Post-training and tested on Pre-training compared to the reverse.

These results show that the model's performance when trained on one subset and tested on its dual is comparable to the results presented in Table 2. These results suggest that Pre-training and Non-urban are more likely to be proper subsets of Post-training and Urban respectively than the reverse. In other words, Post-training and Urban models may (by virtue of having better training data for their duals) include features that are effective on Pre-training and Urban, however this could also be due to the base rate or prevalence of authenticity and uptake in these subsets which needs further investigation.

In order to further examine the models, we compared the confusion matrices to illustrate the bias/prevalence of the models. Using the confusion matrices of models presented in Table 2 (i.e., 10-fold cross validated), we subtracted the confusion matrix when training on the Full-set from the others (Figures 1 and 2.) The resulting matrices represent the extent to which the confusion matrix of a model is different from the baseline model (i.e. Full-set). Each of the confusion matrices were separately proportionalized (before subtraction) by size of the corresponding subset to make the values comparable. Positive values in the figures indicate that the associated category occurred more often in the subset than in the Full-set. Likewise negative values mean that the category occurred less often in the subset than the Full-set.

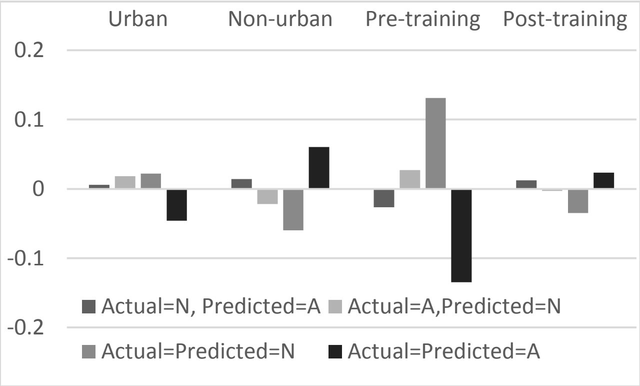

Figure 1. Normalized distance of confusion matrices of Authenticity models on subsets from full-set (A=Authenticity, N= Non-authentic).

It is seen in Figure 1 that the Urban and Post-training authenticity models are the most similar to the Full-set model because their differences with the Full-set are close to zero. This suggests that these models are not biased with respect to the Full-set. However, the Non-urban and Pre-training have larger differences with the Full-set model. Non-urban and Post-training subsets have more true-positives (Actual=Predicted=A) and less true-negatives (Actual=Predicted=N) than the Full-set while the opposite is true for Urban and Pre-training. This contrast in true-positive and true-negatives creates a trade-off in the models which previously appeared to be consistent. Specifically, Figure 1 reveals that Pre-training is more biased towards predicting N (non-authentic instances) than A (authentic instances) which may be due to the fact that there are fewer authentic instances than non-authentic in the Pre-training subset (39% vs. 50%, see Table 1). Conversely, the Non-urban model is biased towards A at the expense of N reflecting the higher distribution of A in the Non-urban subset (54% vs. 50%, see Table 1). Overall, the trade-off between true-positive and true-negative is symmetric which explains why the overall accuracy of the models is not particularly affected despite the differences in error patterns.

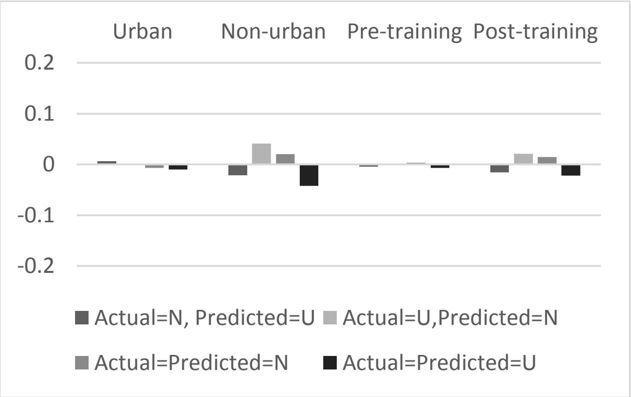

Figure 2. Normalized distance of confusion matrices of Uptake models on subsets from full-set (U=Uptake, N= Non-uptake).

Similar to Figure1, Figure 2 shows the distance between confusion matrices of uptake models. The overall distance of uptake models on subsets compared to Full-set is lower than the distance of authenticity models. Note that the uptake models were trained and 10-fold cross validated on a subsample with an even distribution (50-50) which removes the effect of prevalence on the models. Notably, the Non-urban model sacrifices more true-positives at the expense of false-negatives which explains the lower accuracy of Non-urban in predicting uptake (59% vs. 62%, see Table 2) while the rest of models are very close to the Full-set and hold a balanced tradeoff between true-positive and true-negative.

We examined the models in more detail using Correlation-Based Feature Subset evaluation (CFS). Specifically, we analyzed the frequency of each CFS feature to determine the most important CFS features for each subset. Table 4 shows the CFS results for each model. The features are presented in groups to show whether they were common between the models (shared) or exclusively included in one model only.

Table 4. CFS results, most predictive features of each model grouped based on inclusion.

Models	Authenticity	Uptake
	Urban & Non-Urban	
Shared	Wh, What	Why
Urban only	Be, Judgmental, Enablement	Neg, Pron, Causal_Antecedent
Non-urban only		Disjunction
	Pre-training & Post-training	
Shared	Judgmental, What	Neg, Metacog, Pron, Judgemental, Why
Pre-training only	Comparison	What
Post-training only	Be, Wh, Enablement	Modal, No, Causal_Antecedent

Although the models show similar performance, the most predictive features of each model is different, as seen in Table 4. However there are also marked commonalities among the groups. The features for authenticity on the Non-urban subset, for instance, are fully included in the Urban authenticity subset. Thus this analysis further supports the interpretation of inclusion suggested by the pattern of results in Table 3.

Similarly most of the features of pre-training are included in the post training features, which implies that although teachers' language changed after they received training, the result was that their linguistic behavior broadened with training such that their pre-training behavior was still evident.

4. CONCLUSION

We investigated the generalizability of previously presented models that predict authenticity and uptake in classroom discourse. Overall the results showed that the proposed models' performance is consistent among different subsets of the data set. However, we also found that some subpopulations were potentially more representative of the nature of dialogic instruction than others, making them better for classifier training.

The inclusion relationship between our subsets was investigated by comparing the confusion matrices of our models which revealed that authenticity models of supersets (i.e. Urban and Post-training) were closer to the full-set model than their duals. The consistent accuracy of the models on different subsets was attributed to the tradeoff between true-positive and true-negative predictions which was also explained by the prevalence and bias of the subsets towards one category.

We plan to apply our model to new data which is being collected currently. The proposed models will be applied with the ultimate goal of recording and coding classroom interaction in a fully automatic way and generating statistical reports to show effective instructional strategies. While the models proposed in this paper showed generalizability, another direction of future work is to improve the accuracy by adjusting current features and adding new predictive features to our models.

5. ACKNOWLEDGMENTS

This research was supported by the Institute of Education Sciences (R305A130030) and the National Science Foundation (IIS 1352207). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of these sponsoring agencies.

6. REFERENCES

[1] Baker, R. S., & Gowda, S. M. 2010. An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools: *Proceedings of the 3rd International Conference on Educational Data Mining*, 11-20.

[2] Gamoran, A., & Nystrand, M. 1991. Background and instructional effects on achievement in eighth-grade English and social studies: *Journal of Research on Adolescence, 1*(3), 277-300.

[3] Gamoran, A., & Nystrand, M. 1992. Taking students seriously: *Student engagement and achievement in American secondary schools*, 40-61.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. The WEKA data mining software: an update: *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

[5] Kelly, S. 2007. Classroom discourse and the distribution of student engagement: *Social Psychology of Education, 10*(3), 331-352.

[6] Nystrand, M., & Gamoran, A. 1997. The big picture: Language and learning in hundreds of English lessons: *Opening dialogue*, 30-74.

[7] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. 2014. Population validity for Educational Data Mining models: A case study in affect detection: *British Journal of Educational Technology, 45*(3), 487-501.

[8] Olney, A., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. 2003. Utterance classification in AutoTutor: *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, 1-8.

[9] Samei, B., Olney, A., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., Graesser, A. 2014. Domain Independent Assessment of Dialogic Properties of Classroom Discourse. *Proceedings of the 7th International Conference on Educational Data Mining*, 233-236.

[10] San Pedro, M. O., d Baker, R. S., & Rodrigo, M. M. 2011. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics: *Artificial Intelligence in Education*, 304-311.