Domain Independent Assessment of Dialogic Properties of Classroom Discourse

Borhan Samei[1] Andrew M. Olney[1] Sean Kelly[2] Martin Nystrand[3]
Sidney D’Mello[4] Nathan Blanchard[4] Xiaoyi Sun[3] Marcy Glaus[3] Art Graesser[1]

[1] University of Memphis [2] University of Pittsburgh [3] University of Wisconsin [4] University of Notre Dame
bsamei@memphis.edu

ABSTRACT

We present a machine learning model that uses particular attributes of individual questions asked by teachers and students to predict two properties of classroom discourse that have previously been linked to improved student achievement. These properties, uptake and authenticity, have previously been studied by using trained observers to live-code classroom instruction. As a first-step in automating the coding of classroom discourse, we model question properties based on the features of individual questions, without any information about the context or domain. We then compare the machine-coded results to two referents: human-coded individual questions and “gold standard” codes from existing data. The performance achieved by the models is as good as human experts on the comparable task of coding individual questions out of context. Yet ultimately, this study highlights the need to draw on contextualizing information in order to most completely identify question properties associated with individual questions.

Keywords

Classroom Discourse, Machine Learning, Authenticity, Uptake

1. INTRODUCTION

A particular style of classroom discourse, known as dialogic instruction, has been found to improve student achievement [1, 10, 11]. Dialogic instruction involves fewer teacher questions and more conversational turns as teachers and students alike contribute their ideas to a discussion. One way in which dialogic instruction leads to improved learning is by increasing student engagement in classroom instruction [2]. Moreover, when teachers focus on provoking student thought and analysis, and postpone evaluation during question and answer sessions by engaging in dialogic instruction, levels of student effort are more evenly distributed among students [7]. In the first major quantitative study of dialogic instruction, Nystrand and colleagues observed discourse practices in 8th and 9th grade classrooms over two years [9, 11]. Nystrand et al.’s coding approach focused on the nature of question *events*, which include the discourse context preceding and following a given question. Five properties of question events were coded: *authenticity, uptake, level of evaluation, cognitive level, and question source*. Nystrand and Gamoran reported that among these variables, authenticity and uptake are the most important properties affecting student achievment [1, 3, 10].

Within this context of dialogic instruction, authenticity is defined as a question for which the asker does not have a pre-scripted answer, i.e. open-ended questions. Such questions, particularly when asked by the teacher, create a context for students to contribute and develop their understanding to an evolving discussion. For example, “What was your reaction to the end of the story?” is an authentic question which leads to open-ended discussion, whereas questions such as “What was the father's name?” are not authentic.

Uptake in the context of dialogic instruction occurs when one asks a question about something that another person has said previously. Uptake of student ideas by the teacher therefore emphasizes the importance of student contributions. In previous work, these indicators were judged considering the question in context as opposed to just the individual question. Indeed, the very definition of uptake suggests that it is not possible to detect it from an isolated question, though this assumption and the corresponding assumption for authenticity have never been empirically tested.

In previous research, these variables were “live coded” by classroom observers who also recorded the question as an index of the discourse context preceding and following a given question event. Coding of question events, as opposed to isolated questions, are ultimately determined by teacher responses to students. In contrast, we attempt to predict the question event features of uptake and authenticity from the isolated question using machine learning techniques. Our work addresses a previously untested theoretical question of whether it is possible to recover these variables from the question, since the question is only loosely coupled to the event.

Olney et al. proposed a method to classify questions based on part-of-speech tagging, cascaded finite state transducers, and simple disambiguation rules [12]. They used 16 question categories which were defined in previous works on question classification [4, 5]. This classifier was manually designed using expert linguistic knowledge (a rule based system). We believed that this classifier, though designed for a slightly different purpose, used features that might be highly relevant to identifying uptake and authenticity, because we believed that different kinds of questions might lead to different levels of uptake and authenticity. For example, we hypothesize that yes/no questions are less likely to lead to extended discussion containing uptake and authenticity than causal questions about why an event occurred or why someone decided to take a certain course of action.

Based on the definition of uptake and authenticity, we expected to achieve a reasonable performance by using the same features as predictors as Olney et al. The study reported here shows that the performance of a machine learning approach based on features previously used in question classification is as accurate as expert humans on the task of classifying authenticity and uptake in isolated questions.

2. METHOD

Our long-term research goal is to develop cutting-edge classifiers in order to identify dialogic questions properties important to effective classroom discourse. In working towards this goal, in the present study we address two research questions: (a) How well do machine classifiers perform relative to trained human raters in coding individual questions *without supporting contextual information*? and (b) Do property codes ascertained from individual questions, either by human or machine, correspond well to *fully contextualized codes* (i.e., the "gold standard")?

To address these questions, we utilize existing data from a study of classroom instruction where fully contextualized question property codes had previously been generated [6, 7]. In addition, in order to answer the first research question, we collected new human ratings, using only the information available to the machine learning algorithm, i.e., the question out of context. This study represents the first empirical investigation of dialogic question properties at the level of individual questions.

2.1 Dataset

2.1.1 Gold Standard Data

The present study relies on the Partnership for Literacy Study data (Partnership), a study of professional development, instruction, and literacy outcomes in middle school. In Partnership study, 120 classrooms in 23 schools were observed twice in the fall and twice in the spring.

Observational data from Partnership classrooms were coded using CLASS 4.24, a computer-based data collection system [8]. Coding reliability studies using CLASS indicate that raters agree on question properties approximately 80% of the time, with observation-level inter-rater correlations averaging approximately .95 [10]. Importantly, the original Partnership codes were based on the full set of contextualizing information, including preceding discourse and classroom events.

In all, the Partnership data consist of 29,673 teacher and student questions coded using CLASS during question and answer sessions. In the present study, after removing partially incomplete observations where one or more of the question codes were missing, we utilized a subset of 25,711 questions as our training data, a subset of which is excluded from training and used as the "gold standard" for evaluation purposes.

2.1.2 Individual Question Coding

As a baseline for evaluation of our models, we asked four human raters who were experts in classroom discourse to code the questions of separate sample instances selected from the gold standard data (one sample for authenticity and another for uptake). The sample sets contained 100 questions exhibiting each category of the question property and a separate 100 not exhibiting that property. For example, the uptake set contained 100 questions originally rated as non-uptake and 100 as uptake.

All the questions in the samples were represented by plain text and randomly ordered so that human judgments were based on individual questions without any information about the context. The questions for both authenticity and uptake were rated using a binary (Yes/No) scale.

This task was designed to investigate the performance of human experts on rating the questions, using the same information that we use to build our classifier model with. We also calculated the agreement among human raters to address the difficulty of the task of rating questions in isolation. The performance of machine coding was compared to both the original live-coded data (coding in context) and the subset of data re-coded by human experts (coding in isolation).

2.2 Machine Learning

As mentioned earlier, we applied machine learning using the features based on previous work on question classification. The feature set consisted of 30 attributes including part of speech tags and sets of keywords. Most of the attributes are binary representing the presence/absence of certain keywords or part of speech in the question, for example 'NEG' is true if there is a negation keyword in the question or false otherwise. However, for some of the attributes we take into account the position of the keyword in the question by defining four values: middle, beginning, end, and none, in which the first three values show the position of the keyword if present in the question. For example, if a question consisted of four words, e.g. "word1 word2 word3 word4" the position of "word1" and "word4" are captured as beginning and end respectively. "word2" and "word3" are both captured as middle. Moreover, if we only had two words in the question, we consider first one as beginning and the other as end.

Binary attributes

In our feature set we defined binary attributes to represent the presence of particular words in the questions, regardless of position. These words are defined in sets in Olney et al.; therefore we define the attributes as true if any member of the set is present in the sentence. Causal consequent words, for example, were defined by a set of words including "outcomes," "results," "effects," etc. Similarly, procedural words included "plan," "scheme," "design," etc. The rest of binary attributes included feature specification, negation, meta-communication, metacognition, comparison, goal orientation, judgmental, definition, enablement, interpretation, example, quantification, causal antecedent, and disjunction which are also defined as sets of keywords related to them. We also defined some attributes representing certain words such as "happen," "no," and "yes." More complete descriptions of these features and their validation for question classification can be found in [12]. We used the source code from a simplified version of the question classifier released as part of the open-source GnuTutor project [13].

Other attributes

As mentioned above, for some of the attributes we defined values to represent the presence and position of certain words and part of speech tags. These attributes included part of speech tags such as determiner, noun, pronoun, adjective, adverb, and verb along with word lists: Do/Have (e.g. "don't," "having," and etc.), be (am, are, is, etc.), modal (would, might, etc.), and certain words such as "What," "How," and "Why." More complete descriptions and justifications of these features for question classification can be found at the references above. By including features for positional information we hoped to approximate the regular expression patterns of the Olney question classifier. However instead of directly using the patterns discovered previously, we decided to allow new approximate patterns to be discovered during the machine learning process. Although there might be a correspondence between previous work on question categorization and the constructs of authenticity and uptake, a 1-to-1 correspondence assumption appeared to be unwarranted.

The training data was selected from the "live-coded" data set (Partnership) to form a set of coded questions with uniform distribution of the authenticity and uptake variables. In the case of authenticity, the original distribution of data was close to uniform. New sampling to make the distribution completely uniform (base rate of 50%) yielded a set of 25,464 questions.

Uptake originally was defined by three values: test, authentic and no uptake; however we reduced the uptake to a binary scale of uptake and no-uptake. The original test uptake values were taken as no-uptake in the new scale. The argument for collapsing test uptake and no uptake is based on the observation that they have indistinguishable impact on student achievement. Collapsing test and no uptake and normalizing to a uniform distribution yielded a total of 9,579 instances with an even distribution of uptake and no-uptake. The magnitude of this reduction relative to the set of authentic questions reflects the large number of instances that were originally coded as no-uptake.

These selected instances from the original "live coded" data were then separately used as gold standards to train the two classifiers for predicting uptake and authenticity on isolated questions. The subset of instances given to the expert judges was excluded from training data and was used to test the models. We used WEKA [14] to train and test J48 decision tree classifiers to predict authenticity and uptake.

3. RESULTS & DISCUSSION

We evaluate our models' performance by comparing the performance to the expert re-coded sample as our baseline (coding in isolation), using the gold standard data as the reference (coding in context). Thus the baseline performance was measured by evaluating the performance of four experts on the task the machine classifiers faced: coding questions in isolation.

Cohen's kappa was used as a metric to assess reliability between two raters and between the computer and the rater. Results showed low agreement among human raters on the task, which suggests that in most cases human raters could not make strong judgments based only on the features of individual questions in isolation. The minimum kappa among human raters for authenticity was 0.18; however for other pairs the kappa ranged from 0.3-0.5 with a maximum of 0.55 and an average of 0.4. Similarly, the average inter-rater reliability for Uptake was 0.42, with a minimum of 0.31 and maximum of 0.51 kappa.

The overall low agreement among human raters illustrated the difficulty of making judgments based only on the individual questions as opposed to having information about the context and other properties of the classroom discourse around each question.

The machine learning model was trained on the gold standard data that were rated by Partnership observers. We built J48 decision tree models and tested the models on the same samples that were given to human raters—which were excluded from our traning data—and compared the performance of the model with experts in terms of kappa and recognition rate (Table 1).

Table 1. Kappa statistics and recognition rate of human raters and machine leaning model compared to Gold Standard ratings for authenticity (A) and uptake (U).

-	Kappa		Recognition Rate	
	A	U	A	U
R1	0.13	0.22	56%	61%
R2	0.17	0.25	58%	62%
R3	0.25	0.30	62%	65%
R4	0.10	0.23	55%	61%
Model	0.34	0.46	67%	73%

As seen in Table 1, the highest performance of human raters on predicting authenticity yielded an accuracy of 62% and 0.25 kappa. The performance of the model on predicting authenticity was better than human experts with 67% accuracy and 0.34 kappa. Authenticity was better judged in context, which is why human raters (coding in isolation) showed lower performance and agreement than the original raters (coding in context) of ~80%. By outperforming human raters on this task, our model's performance on authenticity implies that the features used in training are as predictive as could be considering the lack of contextual information. The performance of our model on uptake was markedly better than human experts. The highest performance for a human rater is an accuracy of 65% and 0.30 kappa. The performance of the model on predicting uptake is 73% accuracy and 0.46 kappa. A question with uptake, by definition, refers to a previous discourse contribution. However it appears that features of individual questions are indirectly marking uptake, because our feature set has suitability for predicting uptake in the absence of context. We also measure the overall performance of the model on the whole gold standard data using 10-fold cross validation. Table 2 shows the overall performance of the models.

Table 2. Overall performance of models on gold standard data using 10-fold cross validation

Models	Kappa	Accuracy
Authenticity	0.28	64%
Uptake	0.24	62%

The overall performance of the authenticity model on the gold standard data was close to performance on the sample data while the uptake model performed with a lower accuracy; however the results are still close to human raters coding questions in isolation which supports the reasonable performance of our models on this task.

To take a closer look at the models, we ran Correlation-based Feature Subset Selection (CFS) on our feature sets. CFS considers the individual predictive ability of each feature along with the degree of redundancy between them to evaluate the worth of a subset of attributes. The results showed that the highest ranked attributes used in predicting authenticity were: Judgmental keywords, WH words, Enablement keywords, and "what." Similar analysis on the decision tree for uptake yielded the following most useful attributes: negation keywords, Judgmental keywords, and "why." The importance of such features for predicting uptake can be inferred from the definition.

Although the CFS analysis identified Judgmental, Negation, and Enablement keywords as the most predictive keyword sets, the CFS analysis was unable to identify the actual keywords used because these keywords had been replaced by the labels corresponding to the keyword sets. To illustrate the actual words that were coded as these features, for each set of keywords we calculated the frequency of these words in the data set and measured the distribution of each word as a proportion of the frequency of all the words in the keyword set.

The distribution of Judgmental keywords showed that "think" (.83), "should" (.06), and "find" (.05) accounted for 94% of the total Judgmental keywords seen in the data set. Other keywords individually contributed less that 1%.

Similar analysis showed that the most frequent Enablement keyword was “need(ed)” (0.81) while the other enablement keywords were less frequent, e.g. “helpful,” (0.05), and “in order to,” (0.05). Furthermore, “not(n’t)” (0.95), was the most frequent negation word and other negation words such as “never” and “neither” contributed less than 1%.

The following questions, for example, were extracted from our dataset, to illustrate the use of mentioned keywords in the actual questions:

Questions with authenticity:

*“Do you **think** enterprising people always **need** to be audacious?”*
*“Did you **find** it **helpful**?”*
*“Do you **think** it **needed** to go on the next ten lines?”*

Questions with uptake:

*“**Why** do you **think** he wants to **help** the little boy?”*
*“You **think** he ca**n't** get **help**, Can you expand on that?”*
*“Like if I make a connection to my life and **not** to all three of them do you **think** that that might **help**?”*

Considering the size of our training data, these results suggest the coverage of our feature set in classifying questions out of context. Moreover, these features, as used in the models, are consistent with the theoretical definitions of authenticity and uptake.

4. CONCLUSION

We examined the performance of machine learning models compared to human experts in predicting authenticity and uptake on a random set of isolated questions sampled from a previous classroom study. The key aspect of our approach is that we did not use any contextual information regarding the discourse moves in the model, yet we showed that the models perform as well as human experts under the same restrictions.

The original coders (coding in context) achieved approximately 80% agreement, but in the current study the expert re-coders (coding in isolation) achieved only 60% with the original coders. This suggests that, on a coding task with equally probable categories, a roughly 20% gap in agreement could be attributed to missing contextual information. A surprising finding is that isolated questions provide sufficient cues to correctly identify many authentic questions and questions with uptake. Based on this finding it may be the case that authenticity and uptake can be redefined in terms of an adequate window size of context before and after the question. In future studies, we anticipate incorporating both additional preceding context and following context in determining authenticity and uptake codes.

.

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] Gamoran, A., and Nystrand, M. 1991. Background and instructional effects on achievement in eighth-grade English and social studies . *Journal of Research on Adolescence,* 277–300.

[2] Gamoran, A., and Nystrand, M. 1992. Taking students seriously. In F. Newmann, (Ed.), *Student engagement and achievement in American secondary schools*. Teachers College Press, New York.

[3] Gamoran, A., and Kelly, S. 2003. Tracking, instruction, and unequal literacy in secondary school English. In Hallinan, Gamoran, Kubitschek, & Loveless (Eds.). *Stability and change in American education: Structure, process, and outcomes*. Clinton Corners, NY: Eliot Werner.

[4] Graesser, A. C., and Person, N. 1994. Question asking during tutoring. *American Educational Research Journal*, 104-137.

[5] Graesser, A., Person, N., and Huber, J. 1992. Mechanisms that generate questions Erlbaum. *Questions and information systems*.

[6] Kelly, S. 2007. Classroom discourse and the distribution of student engagement within middle school English classrooms. *Social Psychology of Education,* 10, 331–352.

[7] Kelly, S. 2008. Race, social class, and student engagement in middle school English classrooms. *Social Science Research, 37*, 434-448.

[8] Nystrand, M. 1988. CLASS (Classroom language assessment system) 2.0: A Windows laptop computer system for the in-class analysis of classroom discourse. Wisconsin Center for Education Research, Madison.

[9] Nystrand, M. 1997. Opening dialogue: Understanding the dynamics of language and learning in the English classroom . *Teachers College Press*, New York, NY.

[10] Nystrand, M, and Gamoran, A. 1997. The Big Picture: Language and learning in hundreds of English Lessons. In M. Nystrand. *Opening dialogue Understanding the dynamics of language and learning in the English classroom*. Teachers College Press, New York.

[11] Nystrand, M., Wu, L., Gamoran, A., Zeiser, S., and Long, D. A. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes*, 135-198.

[12] Olney, A. M., Louwerse, M., Mathews, E. C., Marineau, J., Mitchell, H. H., and Graesser, A. C. 2003. Utterance classification in AutoTutor. *Human Language Technology - North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 1-8, Philadelphia.

[13] Olney, A. M. 2009. GnuTutor: An open source intelligent tutoring system based on AutoTutor. *Proceedings of the 2009 AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems,* AAAI Press. 70–75

[14] Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., and Cunningham, S. J. 2011. *Weka: Practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann Publishers.