Utterance-level Modeling of Indicators of Engaging Classroom Discourse

Cathlyn Stone[1], Patrick J. Donnelly[2], Meghan Dale[3], Sarah Capello[3],
Sean Kelly[3], Amanda Godley[3], Sidney K. D'Mello[1]
[1]University of Colorado Boulder; [2]California State University Chico; [3]University of Pittsburgh
430 UCB, Boulder, CO 80309, USA
cathlyn.stone@colorado.edu; pjdonnelly@csuchico.edu; sidney.dmello@colorado.edu

ABSTRACT

We examine the ability of supervised text classification models to identify several discourse properties from teachers' speech with an eye for providing teachers with meaningful automated feedback about the quality of their classroom discourse. We collected audio recordings from 28 teachers from 10 schools in 164 authentic classroom sessions, which we then automatically transcribed into text utterances and then manually coded to identify whether: (1) the utterance contained a question (as opposed to a statement), (2) the question or statement was Instructional vs. Non-Instructional, and (3) the question or statement was Content-Specific. We experimented with Random Forest classifiers and engineered (linguistic, acoustic-prosodic, and contextual) features vs. open-vocabulary n-grams as features to discriminate these discourse variables at the utterance level in a teacher-independent fashion. We achieved AUC scores ranging from 0.71 to 0.77 using open-vocabulary language modeling, which were well above chance (AUC = 0.5), an important step towards our predominant goal of constructing of an automated feedback system for teacher reflection and learning.

Keywords

Automatic Speech Recognition, Natural Language Processing, Dialogic Instruction, Classroom Discourse, Open-vocabulary

1. INTRODUCTION

A teacher's ability to engage students in classroom instruction is of paramount importance in promoting greater student achievement and improving educational outcomes. The level of student engagement is highly dependent upon the ways a teacher interacts with students [1]. The nature of classroom discourse, consisting of the ongoing conversation between the teacher and students, may provide unique insights into a teacher's ability to engage students in the classroom.

Many defining characteristics of classroom discourse have been studied and documented [21]. Traditional methods of classroom instruction are typically presented as monologic discourse, usually in the form of lecture, recitation, and seatwork [20]. However, substantive engagement requires more than just passive listening from students; rather, it requires a degree of student involvement. Not surprisingly, the degree to which classroom discourse is monologic vs. dialogic has been found to greatly influence student engagement, with higher ratios of student talk providing a necessary condition for improved engagement and dialogic interaction [15, 22]. In order to achieve more widespread classroom engagement, students must not only take notes or listen attentively to well-rehearsed lectures from an instructor but must also be engaged in meaningful conversations about a topic. This deep discussion, a hallmark of dialogic instruction [21], is characterized by a symmetrical balance between student and teacher speech, with social interaction shaping the instruction.

In addition to the ratio of teacher speech to student speech, other trends help to characterize dialogic instruction. For example, teachers who ask more questions tend to promote increased student interaction and discussion in classroom discourse [17]. To this note, the Measures of Effective Teaching Study (MET) found question-asking behavior to be a primary factor in the variability of teaching quality [13]. However, questions are not all created equally. Questions associated with classroom management (like attendance-taking or rhetorical questions which require no student response) are not expected to influence student engagement. Compared to informational questions with a known answer, questions which elicit open-ended responses from students are expected to promote increased levels of engagement [34]. These open-ended, or "authentic" questions, draw upon students' ability to put forth independent thought in forming a response, rather than simply perform an affirmation check of the right answer. Authentic questions also serve to initiate discussion in which students can more thoroughly explore an idea and consider different viewpoints [20]. This in turn helps improve their overall understanding and can increase interest in the subject [21].

Additional defining characteristics of dialog associated with increased engagement, and consequently achievement, include higher levels of uptake (teacher questions which incorporate student responses) and cognitive level (the level of cognitive functioning a teacher question seeks to elicit), among other factors [20]. The study by Gamoran and Kelly (2003) demonstrates the benefits of discussion-based approaches to classroom instruction, which contributed the most towards enhanced student performance on complex literacy [12].

Despite the positive correlation of indicators of dialogic instruction (such as authenticity, uptake, and cognitive level) with increased student engagement and achievement [22], the practice has not gained widespread adoption among teachers. Instead, traditional means of instruction and monologic discourse tend to be most prevalent in the classroom [29]. This might be attributed to the challenges encountered by teachers in adopting sustained dialogic discourse into their pedagogical practices. Most importantly, receiving and learning from feedback is essential to assess current abilities and identify areas for improvement. We know that providing teachers with training and data-driven

analysis about their discourse has been shown to positively correlate with student achievement [16]. However, assessment of teachers' instruction via live classroom observations often provides evaluative rather than formative feedback [16]. Moreover, conducting these observations can be expensive, as they require skilled human judges, rubrics, training, and continuous assessment of observers [2]. Therefore, classroom observations occur infrequently, if at all, and must be augmented by additional approaches to further facilitate teacher improvement.

To address this challenge, our study derives from a larger multi-disciplinary project that aims to address a critical lack of quantitative and actionable feedback that teachers receive about the quality of their speech by providing an approach for the automatic analysis of teacher discourse. The present study works towards this overarching goal by automatically classifying teacher utterances from audio recordings of live classroom sessions.

1.1 Related Work

The study of automatic analysis of educational discourse has focused on areas such as online discussions [19], dialog-based intelligent tutoring systems [25], and cognitive models of student learning [5]. In this work we model teachers' classroom discourse through a fully automated process. We examine some of the recent findings in this area.

1.1.1 Modeling classroom discourse

Instructional segment (activity) classification. An instructional segment (or activity) provides coarse-grained information regarding what is occurring at the moment. For example, are students quietly doing seatwork, or is the classroom participating in a discussion or question and answer session? Wang et al. [28] investigated the use of automatic speech recognition of classroom discourse in order to provide feedback for teachers. The authors applied the Language Environment Analysis (LENA) system [11] to segment classroom recordings into broad categories (lecture, discussion, group work). Although the system could provide feedback about the ratio of student to teacher speech, it did not provide any qualitative information about the content or utility of the speech itself. Donnelly et al. [8] also examined automatic identification of instructional activities from classroom recording. Recordings of teacher speech were segmented into individual utterances and transcribed using automatic speech recognition. Using models trained on temporal, natural language, and acoustic features, the authors trained models to identify the dominant (76%) activities (question and answer, procedures and directions, supervised seatwork, group work, and lecture) with accuracies that easily outperformed chance baselines.

Utterance-level classification. At a finer grain than Instructional segment classification, Donnelly et al. [9] built on work by [3] to identify teacher questions within individual utterances. Classroom recordings were segmented, transcribed using Automatic Speech Recognizers (ASRs), and 218 acoustic, linguistic, and contextual features were derived. The acoustic features derived from 384 prosodic, spectral, and voice quality features extracted with the OpenSmile toolkit [10]. Then, a smaller set of 168 acoustic features was obtained by eliminating features with high multicollinearity using tolerance analysis. The transcriptions were analyzed for part-of-speech tags and the presence of specific words (e.g., *why*, *how*) to provide 37 linguistic features. A total of 13 contextual features included timing information, such as the duration of the utterance, the position of the utterance within the class session, and the duration of the pauses preceding and following the utterance. The authors found that combination of all three modalities made no improvement over linguistic features alone in the task of question identification but did yield small improvements in non-question detection.

Session level classification. Given the positive association between the use of authentic questions and student engagement and achievement [1, 20, 21], any system seeking to provide automatic feedback needs to be able to automatically identify this variable. However, the infrequent use of authentic questions compared to other types of dialog leads to highly imbalanced class distributions that make classification tasks difficult [14]. For this reason, Olney et al. [23] aimed to detect the proportion of authentic questions over the course of the class session, rather than seek to classify individual utterances. Using a model trained on word, part-of-speech, syntactic, and discourse features, the prediction of class-level proportions ($r = 0.50$) outperformed aggregated utterance-level classification ($r = 0.27$), and these results were consistent across low and high dialogic classrooms, and on both ASR and human transcripts. In a follow-up study, Cook et al. [7] compared closed- and open-vocabulary techniques (described in Section 1.1.2) for the same task and found that that both approaches were equally predictive of authenticity, but that averaging the models' predictions yielded significant additional improvements.

1.1.2 Computational techniques

We apply techniques from natural language processing and machine learning in the automatic analysis of teacher discourse.

Open-vocabulary language modeling. In contrast to hand-crafted feature sets, open-vocabulary language modeling dynamically generates features for machine learning models by obtaining counts of consecutive words (n-grams) extracted directly from the input text [27]. This "bag-of-n-grams" model then assigns each n-gram feature a value based on its frequency within each utterance. This approach lends itself to human-interpretable analysis of models (e.g., word clouds). Several hyperparameters might also guide the selection of n-grams derived from the training set that should be included in the set of features, as certain n-grams may be unimportant for models, such as n-grams that occur infrequently. These hyperparameters might include whether or not stopwords are removed from text, whether stemming is performed on words, and the types of n-grams considered (unigrams, bigrams, trigrams, etc.). In addition, pointwise mutual information (PMI), described in detail in [6], can be specified as a hyperparamter to filter n-grams by incorporating information about the collocations of words. The PMI for a given n-gram can be defined as $\text{pmi}(n\text{-}gram) = \log\left(p(n\text{-}gram) \,/\, \Pi\, p(word)\right)$ where $p(n\text{-}gram)$ is the probability of an n-gram based on its relative frequency in the training data and $\Pi\, p(word)$ is the product of the probabilities of each word in the n-gram in the training data. This can help ensure only meaningful phrases (such as "*high school*") are used as features. Minimum document frequency might further guide the selection of n-grams (i.e. the n-gram must occur in a specified minimum percentage of all documents to be considered a feature). An overview of these techniques can be found in [6].

1.2 Novelty and Contribution

We apply natural language processing and supervised classification techniques for the *utterance-level* classification of multiple aspects of classroom discourse with an eye for providing

automated feedback to teachers. Our study is novel in multiple respects. First, with the exception of work on detecting individual questions, as noted in Section 1.1.1, existing analyses on the automatic classification of classroom discourse have focused on coarse-grained temporal information ranging from a few minutes to an entire class session. We hypothesize that fine-grained utterance-level information is needed in order to provide meaningful and actionable feedback. Therefore, this study analyzes classroom discourse at the utterance level.

The intrinsic value of a system for automated feedback to teachers is inherently dependent upon the ability to correctly classify *different types* of discourse from automatically segmented recorded speech. Whereas previous work has focused on identifying questions, the present approach considers several more specific discourse variables, which have not previously been studied using automatic recognition methods. In addition to question prediction, we also predict Instructional Questions and Statements as well as Content-Specific Questions and Statements. These discourse variables are described in detail in Section 2.1.2.

2. METHODS

2.1 Dataset

2.1.1 Data collection

Our dataset consists 167 recordings of class sessions, drawn from two sources. One source of data was collected in 2018 and consists of 127 observations from 16 teachers at three schools in western Pennsylvania. Additionally, we newly recoded a subset of the CLASS5 dataset, collected at seven schools in rural Wisconsin over 2014 to 2016. This source of data consists of 40 class observations from 11 teachers [8]. Teachers wore a wireless Samson AirLine 77 vocal headset which transmitted audio to a receiver to then be recorded on a laptop.

2.1.1 Utterance transcription using ASRs

The IBM Watson ASR [26] was used to automatically segment the class recordings into utterances based on hesitations in the audio stream, and to transcribe each resulting utterance. To evaluate the efficacy of the ASR, a sample of 20 utterances per class session was manually transcribed by human coders and these transcriptions were compared to the ASR transcriptions. The average word accuracy (W_{acc}) of the automatic transcriptions across class sessions was 0.602 when considering all utterances. The W_{acc} increased to 0.754 when considering only longer utterances that contained three or more words.

2.1.2 Coding of utterances and coding scheme

To prepare a labeled dataset for training supervised models, a subset of the transcribed utterances was selected for manual annotation by trained human coders. First, to generate this set, any two consecutive utterances were merged together if the pause between them was less than 1.0 seconds. This preprocessing step helped adjoin related phrases together and reduced the number of single word utterances. Next, 200 consecutive merged-utterances were randomly sampled from each class session. If a class session contained less than 200 utterances, then all utterances were sampled for that session. English and language arts content experts trained in the coding schema were given audio excerpts for each utterance in the sampled dataset. The coders manually annotated each utterance with several markers of classroom discourse. Because many of the annotated categories occur only infrequently in the dataset, some markers have been aggregated together to form binary labels, such as Question or Non-Question. Below we describe the variables used in the current work.

Question/Statement/Fragment. Utterances were coded to determine whether they consisted of a question, statement, or fragment. Questions are defined as requests for information, while conversely, statements are utterances which do not request information. Rhetorical questions, such as, *"It's the characteristic of a person, right?*" are not coded as questions because they are not requests for information. Fragments are a single word or a few words that have been separated from a cohesive statement or question in the ASR transcription and appear as an individual utterance. Fragments by themselves are meaningless, and it would not be useful to code their discourse properties. To perform binary classification, we combined statements and fragments to predict whether each utterance was a Question or Non-Question.

Instructional questions. Utterances identified as questions were further coded as Instructional or Non-Instructional Questions. Instructional Questions relate to the lesson and its learning goals, whereas Non-Instructional Questions are irrelevant to the lesson and its learning goals, such as questions about student movement and behaviors. For example, "*Who can tell me what a plot diagram is?*" would be coded as an Instructional Question, while "*Why are you late?*" would be considered Non-Instructional.

Instructional question type. Instructional Questions were further coded as Content-Specific, Generic, or Clarifying. Content-Specific Questions inquire about the content/disciplinary practices of the lesson and its learning goals, such as "*What is the theme of the poem?*", or "*How do you typically revise?*". Generic Instructional Questions are broad questions about organization, materials, behaviors, or checks for understanding connected to the lesson. Examples include "*Where is your paper from yesterday?*" and *"Does that answer your question?*". Clarifying Questions are requests for restatements and repetitions, such as "*Can you say that again?*". We combined the Generic and Clarifying codes to predict the binary classification of Content-Specific Questions vs. Non-Content-Specific Questions.

Instructional statements. Similar to Instructional Questions, utterances identified as Statements were coded as Instructional or Non-Instructional. Instructional Statements relate to the lesson and its learning goals, such as "*A character that moves the action forward but is not central to the story is a minor character*" and *"Today we are going to review literary terms that will be on the quiz on Thursday.*" Non-Instructional Statements are irrelevant to the lesson and its learning goals, such as statements about student movement and behaviors. For instance, "*You shouldn't be walking around the room. Please sit down.*" In addition, short, placeholding utterances that connote continued thinking (e.g., "*hmmm*", "*um*", "*okay*") were coded as Non-Instructional; however, "*okay*" was not automatically coded as Non-Instructional as it can also be an evaluation of a student's response or serve another function, depending on its context.

Instructional statement type. Instructional Statements were further coded as Content-Specific, Generic, or Reading Aloud. Content-Specific Statements are statements about the content/disciplinary practices of the lesson and learning goals. For example, "*The mood of the play contributes to our understanding of the theme of the play.*" Generic statements are broad statements about organization, behaviors, materials, or checks for understanding connected to the lesson, as in *"Take out your journals, and turn to a new page.*" Reading Aloud statements

occur when the teacher or the students are reading aloud from a text verbatim. If the teacher is reading a short Instructional Statement or discussion question out of a textbook, off a PowerPoint slide, off a worksheet, etc., it is not considered Reading Aloud and is coded as Content-Specific. Furthermore, if the teacher stops reading to make a comment or interjects while the students are reading, those utterances are not coded as Reading Aloud. Similar to predictions made for Instructional Question Type, Generic and Clarifying codes were combined to predict Content-Specific Statements vs. Non-Content-Specific Statements.

2.1.3 Prevalence of discourse types

Our dataset contained a total of 24,755 teacher utterances, with 16,977 from the new Spring 2018 data and 7778 from CLASS5. Table 1 provides information about the prevalence of each of these types of discourse variables in this combined dataset.

Table 1: Summary of dataset

	Count	Proportion
Teacher	**24755**	
Question	7792	0.31
Instructional	7267	0.29
Content-Specific	5327	0.22
Non-Content-Specific	1940	0.08
Non-Instructional	525	0.02
Non-Question	16963	0.69
Instructional	12113	0.49
Content-Specific	8369	0.34
Non-Content-Specific	3744	0.15
Non-Instructional	4850	0.20

2.2 Machine learning

Using several modalities of features, we trained Random Forest classifiers implemented using the scikit-learn library [24] to perform a binary (present vs. absent) classification of these discourse features. We constructed models using three representations of the input data: as a set of engineered features computed from the audio and transcribed text of utterances, as a set of features derived via open-vocabulary language modeling, and finally as a combination of both of these sources.

We generated the set of engineered features using the acoustic, context, and linguistic features as described in [9]. Acoustic features were extracted from the audio of utterances using the OpenSmile toolkit [10], using the feature set from the 2009 Interspeech Emotion Challenge. This resulted in 384 acoustic features. Context features describe properties of the utterance such as its duration, its normalized (to unit variance) position in the overall classroom session, and the length of time of the surrounding pauses. In total, we considered 13 context features. Linguistic analyzers parsed the transcribed text and identified the presence of known question words and part of speech tags. These were found using the Brill Tagger [4] to identify certain question words, part-of-speech tags, and other keywords, resulting in 37 total features. The values of all these features were standardized to have a mean of 0 and unit variance. Standardization was computed using the formula $z = (x-u) / s$ such that z is the standardized score, x is the value of an individual sample, u the mean value of all training samples, and s is the standard deviation of the samples.

A bag-of-n-grams representation of input formed the open-vocabulary feature set. N-grams (of which we considered unigrams, bigrams, and trigrams) derived from the texts of transcribed utterances were filtered according to the values of a few hyperparameters. We experimented using minimum document frequencies of 0.01, 0.02, and 0.03; PMI values of 0.2 and 0.4; and either including or excluding stopwords (see Section 1.1.2).

We implemented teacher-level 5-fold cross-validation to determine the best set of hyperparameters for models within each training fold. Specifically, we ensured that all utterances from the same teacher were always kept within the same train/test/validation fold. This helps ensure generalizability of our approach to new data and new teachers. To enable faster training of models, we limited the overall search space of hyperparameters, varying the parameters specified for the open-vocabulary models and leaving other parameters at default values as specified by scikit-learn. To overcome the underlying class imbalance in the dataset (see Table 1), we experimented using the imblearn library [18] to resample the minority class utterances such that both classes were more equally represented in the input dataset. This approach was applied to all models and only performed on the training set; class distributions in the validation and testing sets were unchanged.

3. RESULTS

We examined the ability of different types of models to predict five indicators of teacher discourse using utterances automatically segmented and transcribed by an ASR. We used area under the receiver operating characteristic curve (AUROC or AUC) as our primary outcome metric, which we computed using the pooled predicted probabilities from the five folds of our dataset. An AUC of 0.5 would signify chance performance.

3.1 Predictive language features

Table 2. Top 10 correlated n-grams

Variable	*Top 10 correlated n-grams*	*Example sentences from dataset*
Question	*does, did, think, good, say, mean, yes, guys, kind, make*	"why do you *say* you want to do"
Instructional Question	*does, did, think, good, say, yes, kind, mean, guys, make*	"are you *guys* doing"; "*did* you talk to me on Friday"
Content-Specific Question	*does, think, did, good, kind, say, know, make, mean, people*	"okay why *does* she *think* it's any better for her son"; "what *does* that *mean*"
Instructional Statement	*na, gon, gon na, going, just, like, right, <hesitation>, look, little*	"all *right* now notice what you need to do *look* at this part"
Content-Specific Statement	*like, <hesitation>, going, na, just, gonna, gon, kind, little, right*	"*like* if I'd done that all *right* I have a sample body paragraph here"

Note: <hesitation> expresses a token generated by ASRs to indicate hesitation in speech. Here we treat it as a word.

We analyzed our models to correlate the top 10 n-grams for each discourse variable in order to investigate characteristic language features. We calculated Spearman correlations of n-grams to the class labels (either 0 or 1) of the documents in which they appear. These correlations were averaged across the five folds on which

Random Forest models were trained. These n-grams are listed in Table 2, and we note some expected patterns. For example, one would expect that questions would be characterized by auxiliary verbs such as *does* and *did* as well as action verbs such as *think, say,* and *make*. We also observed considerable overlap between these categories. For example, both Instructional and Content-Specific questions include *does*, *think*, and *did* among their top three n-grams and share eight of ten most common n-grams. Likewise, Instructional and Content-Specific Statements share nine of the top ten most common n-grams. Conversely, we found that Content-Specific Statements and Questions have overlap only in the n-gram *kind*.

3.2 Comparison of feature sets

We constructed Random Forest models using three types of features: (1) engineered acoustic, context, and linguistic features, (2) bag-of-n-grams language features via open-vocabulary language modeling, and (3) a combination of both. Results using the Random Forest model are shown in Figure 1. We found that open-vocabulary language modeling resulted in the highest average AUC scores (average AUC = 0.74) for all discourse variables, followed by the combined set of features (average AUC = 0.72), while the engineered features were the least predictive (average AUC = 0.68). With respect to question detection, for which we have a baseline from previous work [9], we found that the current approach with language features yielded a 11% improvement over engineered features alone. These results demonstrate significant improvement (3-12%) over the previous state of the art.

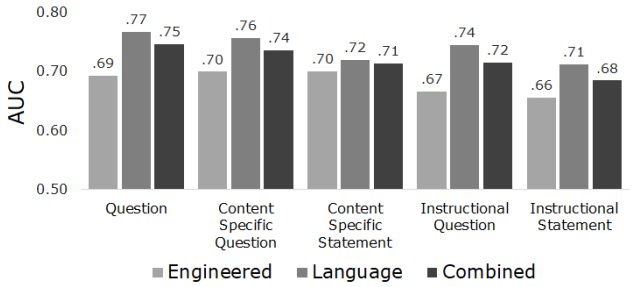

Figure 1. Random Forest: AUROC per feature set

4. DISCUSSION

We investigated the extent to which several characteristics of classrooms discourse could be automatically identified at the utterance level. While prior work focused on predicting questions at the utterance level, in this study we detected several additional discourse characteristics at the utterance level, which would be of paramount importance in a real-time feedback system for teachers. We collected additional data from live classroom sessions during the Spring of 2018 to augment previously collected data. We developed a new coding scheme and manually annotated the dataset. Audio recordings of entire classroom sessions were automatically segmented into utterances which were then manually coded by humans and transcribed into text by the IBM Watson ASR. We then executed machine learning experiments to examine the extent to which these discourse variables could be recognized from the audio signal alone.

4.1 Main findings

First, we observed that open-vocabulary bag-of-n-grams Random Forest models outperformed our previous attempt using models built using only engineered features. These results demonstrate that the specific words a teacher uses, as determined by automatic transcriptions, may be of more utility than acoustic and prosodic cues, timing cues, rates of speech, parts-of-speech analysis, and closed-vocabulary word lists. Moreover, these findings indicate that the words most useful to differentiate between dialogic acts often differ from those anticipated by domain-specific closed-vocabulary lists. For example, closed-vocabulary lists created to predict questions may only look for whose words typically indicative of questions (e.g., *what, where, why, how*), while overlooking other words that may also be useful to distinguish this type of discourse (such as *think, say, mean*).

In summary, our results using open-vocabulary modeling (with AUCs ranging from 0.71 to 0.77) comfortably outperformed chance (AUC = 0.5), and reflect the state of the art performance on automatic modeling of classroom discourse. Further, the fact that the models were trained in a manner that generalizes to new teachers, and that the training data included audio from two different U.S. states across varying grade levels (mainly middle school for CLASS 5 vs. mainly high school for Spring 2018 data), increases our confidence in their generalizability As such, we are optimistic that our present results reflect the feasibility of fully-automated utterance-level classrooms discourse modeling, a key step towards providing actionable feedback for teachers.

4.2 Limitations and future work

Although research indicates that the dialogic indicators of authenticity, uptake, and cognitive level are predictors of enhanced student engagement, this study does not aim to identify these indicators at the utterance level. This is because these variables have extremely low base rates (all under 10%), resulting in severe class imbalance when attempting to identify them from *all* teacher utterances. However, the automatic recognition of the discourse variables in this study serves as a precursor for subsequent approaches to better accurately identify these useful but infrequently occurring dialogic variables. The identification of these key dialogic variables relies on the ability to first correctly differentiate between more generic discourse properties, such as Questions vs. Statements, followed by Content-Specific Questions (of which Authentic Questions are a subset) vs. Instructional Questions.

In addition, we are currently limited by the lack of annotated data to provide sufficient exemplars of these specific dialogic properties (authenticity, uptake, and cognitive level) at the utterance level. Thus, additional collection of data would allow more examples of these more rarely occurring discourse types. Given this new data, we will extend our models to attempt to identify these infrequently occurring dialogic indicators.

Furthermore, continued improvement in the accuracy of our predictions is necessary to ensure the value of the assessment and feedback from our automated system. We plan on exploring several improvements to advance this goal. First, we will incorporate transcription metadata, such as the confidence values of the ASRs, in the models in order to weight individual words in the open-language model based on the quality of the transcription. Since words transcribed with a low confidence may be misidentified, excluding or discounting these words from language model may help to reduce modeling error. Second, we will empirically experiment with varying the pause threshold used for segmentation. Perhaps a slightly longer or shorter gap in speech would provide a better separator of utterances. Third, we

will continue to explore different supervised machine learning models or neural network architectures to further improve our ability to automatically identify these discourse indicators.

4.3 Concluding remarks

We hope that in our continued efforts towards automatic prediction of types of discourse, we can achieve the capability to provide valuable, actionable feedback to teachers about their instructional techniques so that they can better engage students in learning. Certainly, much work remains to be done in this area in order to improve upon our current ability. Nonetheless, this study forms an important step towards our overarching goal and serves as a foundation for future work in this area.

REFERENCES

[1] Applebee, A.N. et al. 2003. Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*. 40, 3 (2003), 685–730.

[2] Archer, J. et al. 2016. *Better feedback for better teaching: A practical guide to improving classroom observations*. John Wiley & Sons.

[3] Blanchard, N. et al. 2016. Automatic detection of teacher questions from audio in live classrooms. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016), International Educational Data Mining Society* (2016), 288-291.

[4] Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language* (1992), 112–116.

[5] Chen, S.Y. and Macredie, R.D. 2004. Cognitive modeling of student learning in web-based instructional programs. *International Journal of Human-Computer Interaction*. 17, 3 (2004), 375–402.

[6] Church, K.W. and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*. 16, 1 (1990), 22–29.

[7] Cook, C. et al. 2018. An Open Vocabulary Approach for Estimating Teacher Use of Authentic Questions in Classroom Discourse. *11th International Conference on Educational Data Mining* (2018), 116-126.

[8] Donnelly, P.J. et al. 2016. Automatic teacher modeling from live classroom audio. *Proceedings of the 24th Conference on User Modeling, Adaptation and Personalization (UMAP 2016)*. (2016), 45-53.

[9] Donnelly, P.J. et al. 2017. Words matter: Automatic detection of questions in classroom discourse using linguistics, paralinguistics, and context. *Lak '17: Proceedings of the seventh international conference on learning analytics & knowledge.* (2017), 218-227.

[10] Eyben, F. et al. 2010. OpenSmile: the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia* (2010), 1459–1462.

[11] Ford, M. et al. 2008. *The LENA Language Environment Analysis System*. Technical Report LTR-03-2. Boulder, CO: LENA Foundation.

[12] Gamoran, A. and Kelly, S. 2003. Tracking, instruction, and unequal literacy in secondary school English. *Stability and change in American education: Structure, process, and outcomes*. (2003), 109–126.

[13] Kane, T. and Cantrell, S. 2010. Learning about teaching: Initial findings from the measures of effective teaching project. *MET Project Research Paper, Bill & Melinda Gates Foundation*. 9, (2010).

[14] Kelly, S. et al. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*. 47, 7 (2018), 451–464.

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

[16] Lai, M.K. and McNaughton, S. 2013. Analysis and discussion of classroom and achievement data to raise student achievement. *Data-based decision making in education*. Springer. 23–47.

[17] Lee, Y. and Kinzie, M.B. 2012. Teacher question and student response with regard to cognition and language use. *Instructional Science*. 40, 6 (2012), 857–874.

[18] Lemaître, G. et al. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*. 18, 1 (2017), 559–563.

[19] Mu, J. et al. 2012. The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*. 7, 2 (2012), 285–305.

[20] Nystrand, M. et al. 1997. *Opening dialogue: Understanding the dynamics of language and learning in the English classroom. Language and Literacy Series*.

[21] Nystrand, M. et al. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes*. 35, 2 (2003), 135–198.

[22] Nystrand, M. and Gamoran, A. 1991. Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*. (1991), 261–290.

[23] Olney, A. et al. 2017. Assessing the Dialogic Properties of Classroom Discourse: Proportion Models for Imbalanced Classes. *Proceedings of the 10th International Conference on Educational Data Mining.* (2017), 162-167.

[24] Pedregosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research*. (2011), 2825-2830.

[25] Rus, V. et al. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine*. 34, 3 (2013), 42–54.

[26] Saon, G. et al. 2015. The IBM 2015 English conversational telephone speech recognition system. In *Proc. Interspeech.* (2015), 3140–3144.

[27] Tan, C.-M. et al. 2002. The use of bigrams to enhance text categorization. *Information processing & management*. 38, 4 (2002), 529–546.

[28] Wang, Z. et al. 2013. Using the LENA in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft*. 4, (2013), 290–305.

[29] Wells, G. and Arauz, R.M. 2006. Dialogue in the classroom. *The journal of the learning sciences*. 15, 3 (2006), 379–428.