Identifying Teacher Questions Using Automatic Speech Recognition in Classrooms

Nathaniel Blanchard[1], Patrick J. Donnelly[1], Andrew M. Olney[2], Borhan Samei[2], Brooke Ward[3], Xiaoyi Sun[3], Sean Kelly[4], Martin Nystrand[3], & Sidney K. D'Mello[1]
[1]University of Notre Dame; [2]University of Memphis;
[3]University of Wisconsin-Madison; [4]University of Pittsburgh
384 Fitzpatrick Hall
Notre Dame, IN 46646, USA
nblancha|sdmello@nd.edu

Abstract

We investigate automatic question detection from recordings of teacher speech collected in live classrooms. Our corpus contains audio recordings of 37 class sessions taught by 11 teachers. We automatically segment teacher speech into utterances using an amplitude envelope thresholding approach followed by filtering non-speech via automatic speech recognition (ASR). We manually code the segmented utterances as containing a teacher question or not based on an empirically-validated scheme for coding classroom discourse. We compute domain-independent natural language processing (NLP) features from transcripts generated by three ASR engines (AT&T, Bing Speech, and Azure Speech). Our teacher-independent supervised machine learning model detects questions with an overall weighted F_1 score of 0.59, a 51% improvement over chance. Furthermore, the proportion of automatically-detected questions per class session strongly correlates (Pearson's r = 0.85) with human-coded question rates. We consider our results to reflect a substantial (37%) improvement over the state-of-the-art in automatic question detection from naturalistic audio. We conclude by discussing applications of our work for teachers, researchers, and other stakeholders.

1 Introduction

Questions are powerful tools that can inspire thought and inquiry at deeper levels of comprehension (Graesser and Person, 1994; Beck et al., 1996). There is a large body of work supporting a positive relationship between the use of *certain types of questions* with increased student engagement and achievement (Applebee et al., 2003; Kelly, 2007). But not all questions are the same. Questions that solicit surface-level facts (called test questions) are far less predictive of achievement compared to more open-ended (or dialogic) questions (Nystrand and Gamoran, 1991; Gamoran and Nystrand, 1991; Applebee et al., 2003; Nystrand, 2006).

Fortunately, providing teachers with training and feedback on their use of instructional practices (including question-asking) can help them adopt techniques known to be associated with student achievement (Juzwik et al., 2013). However, automatic computational methods are required to analyze classroom instruction on a large scale. Although there are well-known coding schemes for manual coding of questions in classroom environments (Nystrand et al., 2003; Stivers and Enfield, 2010) research on *automatically* identifying these questions in live classrooms is in its infancy and is the focus of this work.

1.1 Related Work

To keep scope manageable, we limit our review of previous work to question detection from automatic speech recognition (ASR) since the use of ASR transcriptions, rather than human transcriptions, is germane to the present problem.

Boakye et al. (2009) trained models to detect questions in office meetings. The authors used the ICSI Meeting Recorder Dialog Act (MRDA) corpus, a set of 75 hour-long meetings recorded with headset and lapel microphones. Their ASR system achieved a word error rate (WER), a measure of edit distance comparing the hypothesis to the original transcript, of 0.38 on the corpus. They trained an AdaBoost classifier to detect questions from word, part-of-speech, and parse tree features derived from the ASR transcriptions, achieving F_1 scores of 0.52, 0.35, and 0.50, respectively, and 0.50 combined. Adding contextual and acoustic features slightly improved the F_1 score to 0.54,

Proceedings of the SIGDIAL 2016 Conference, pages 191–201,
Los Angeles, USA, 13-15 September 2016. ©2016 Association for Computational Linguistics

suggesting the importance of linguistic (as opposed to contextual or acoustic) information for question detection.

Stolcke et al. (2000) built a dialogue act tagger on the conversational Switchboard corpus using ASR transcripts (WER 0.41). A Bayesian model, trained on likelihoods of word trigrams from ASR transcriptions, detected 42 dialogue acts with an accuracy of 65% (chance level 35%; human agreement 84%). Dialogue acts such as statements, questions, apologies, or agreement were among those tagged. Limiting the models to consider only the highest-confidence transcription (the 1-best ASR transcript) resulted in a 62% accuracy with the bigram discourse model. Additionally, the authors noted a 21% decrease in classification error when human transcripts were used instead.

Stolcke et al. (2000) also attempted to leverage prosody to distinguish yes-no questions from statements, dialogue acts which may be ambiguous based on transcripts alone. On a selected subset of their corpus containing an equal proportion of questions and statements they achieved an accuracy of 75% using transcripts (chance 50%). Adding prosodic features increased their accuracy to 80%.

Orosanu and Jouvet (2015) investigated discrimination between statements and questions from ASR transcriptions from three French-language corpora. Their training set consisted of 10,077 statements and 10,077 questions, and their testing set consisted of 7,005 statements and 831 questions. Using human transcriptions, the models classified 73% of questions and 78% of statements correctly. When the authors tested the same model against ASR transcriptions, they observed a 3% reduction in classification accuracy. The authors also compared their datasets based on differences in speaking styles. One corpus consisted of unscripted, spontaneous speech from news broadcasts (classification accuracy 70%; WER 22%), while the other contained scripted dialogue from radio and TV channels (classification accuracy 73%; WER 28%).

All the aforementioned studies have used manually-defined sentence boundaries. However, a fully-automatic system for question detection would need to detect sentence boundaries without manual input. Orosanu and Jouvet (2015) simulated imperfect sentence boundary detection using a semi-automatic method. They substituted sentence boundaries defined by human-annotated punctuation with boundaries based on silence in the audio. When punctuation aligned with silence, the boundaries were left unchanged from the manually-defined boundaries. This semi-automatic approach to segmentation resulted in a 3% increase in classification errors.

Finally, in preliminary precursor to this work, we explored the potential for question detection in classrooms from automatically-segmented utterances that were transcribed by humans (Blanchard et al., 2016). We used 1,000 random utterances from our current corpus which we manually transcribed and coded as containing a question or not (see Section 2.3). Using leave-one-speaker-out cross-validation, we achieved an overall-weighted F_1 score of 0.66, with an F_1 of 0.53 for the question class. That work showed that question detection was possible from noisy classroom audio, albeit with human transcriptions.

1.2 Challenges, Contributions, and Novelty

We describe a novel question detection scenario in which we automatically identify teacher questions using ASR transcriptions of teacher speech in a real-world classroom environment. We have previously identified numerous constraints that need to be satisfied in order to facilitate question detection at scale. Such a system must be affordable, cannot be disruptive to either the teacher or the students, and must maintain student privacy, which precludes recording or filming individual students. Therefore, we primarily rely on a low-cost, wireless headset microphone for recording teachers as they move about the classroom freely. This approach accommodates various seating arrangements, classroom sizes, and room layouts, and attempts to minimize ambient classroom noise, muffled speech, or classroom interruptions, all factors that reflect the reality of real-world environments.

There are a number of challenges with this work. For one, teacher questions in a classroom differ from traditional question-asking scenarios (e.g., meetings, informal conversations) where the goal of a question is to elicit information and the questioner usually does not know the answer ahead of time. In contrast, rather than information-seeking, the key goal of teacher questions is to assess knowledge and to prime thought and discussion (Nystrand et al., 2003), thereby introducing difficulties in coding questions themselves.

We note that ASR on classroom speech is particularly challenging given the noisy environment that includes classroom disruptions, accidental microphone contact, and sounds from students, chairs, and desks. Previous work on this data

yielded WERs ranging from 0.34 to 0.60 (D'Mello et al., 2015), suggesting that we have to contend with rather inaccurate transcripts.

In addition, previous work reviewed in Section 1.1 has focused on human-segmented speech, which is untenable for a fully-automated system. Therefore, our approach uses an automated approach to segment speech, which itself is an imperfect process.

This imperfect pipeline ranging from question coding to ASR to utterance segmentation accurately illustrates the difficulties of detecting questions in real-world environments. Nevertheless, we make several novel contributions while addressing these challenges. First, we implement fully automated methods to process teacher audio into segmented utterances from which we obtain ASR transcriptions. Second, we combine transcriptions from multiple ASR engines to offset the inevitable errors associated with automatically segmenting and transcribing teacher audio. Third, we restrict our feature set to domain-independent natural language features that are more likely to generalize across different school subjects. Finally, we use leave-one-teacher-out cross-validation so that our models generalize across teachers rather than optimizing for individual teachers.

The remainder of the paper is organized as follows. First, we discuss our data collection methods, data pre-processing, feature extraction approach, and our classification models in Section 2. In Section 3, we present our experiments and review key results. We next discuss the implications of our findings and conclude with our future research directions in Section 4.

2 Method

2.1 Data Collection

Data was collected at six rural Wisconsin middle schools during literature, language arts, and civics classes taught by 11 different teachers (three male; eight female). Class sessions lasted between 30 and 90 minutes, depending on the school. A total of 37 classroom sessions were recorded and live-coded on 17 separate days over a period of a year, totaling 32:05 hours of audio.

Each teacher wore a wireless microphone to capture their speech. Based on previous work (D'Mello et al., 2015), a Samson 77 Airline wireless microphone was chosen for its portability, noise-canceling properties, and low-cost. The teacher's speech was captured and saved as a 16 kHz, 16-bit single channel audio file.

2.2 Teacher Utterance Extraction

Teacher speech was segmented into utterances using a two-step voice activity detection (VAD) algorithm (Blanchard et al., 2015). First, the amplitude envelope of the teacher's low-pass filtered speech was passed through a threshold function in 20-millisecond increments. Where the amplitude envelope was above threshold, the teacher was considered to be speaking. Any time speech was detected, that speech was considered part of a *potential utterance*, meaning there was no minimum threshold for how short a potential utterance could be. Potential utterances were coded as complete when no speech was detected for 1,000 milliseconds (1 second).

The thresholds were set low to ensure capture of all speech, but this also caused a high rate of false alarms in the form of non-speech utterances. These false alarms were filtered from the set of potential utterances with the Bing ASR engine (Microsoft, 2014). If the ASR engine rejected a potential utterance then it was determined to not contain any speech. Additionally, any utterances less than 125 milliseconds was removed, as this speech was not considered meaningful.

We empirically validated the effectiveness of this utterance detection approach by manual coding a random subset of 1,000 potential utterances as either containing speech or not. We achieved high levels of both precision (96.3%) and recall (98.6%) and an F_1 score of 0.97. We applied this approach to the full corpus to extract 10,080 utterances from the 37 classroom recordings.

2.3 Question Coding

One limitation of automatically segmented speech is that each utterance may contain multiple questions, or conversely, a question may be spread across multiple utterances (Komatani et al., 2015). This occurs partly because we use both a static amplitude envelope threshold and a constant pause length to segment utterances rather than learning specific thresholds for each teacher. However, the use of a single threshold increases generalizability to new teachers. Regardless of method, voice activity detection is not a fully-solved problem and any method is expected to yield some errors.

To address this, we manually coded the 10,080 extracted utterances as "containing a question" or "not containing a question" rather than "question" or "statement." The distinction, though subtle, indicated that a question phrase that is embedded

within a large utterance would be coded as "containing a question." Conversely, we also ensured that if a question spans adjacent utterances then each utterance would be coded as "containing a question." We also do not distinguish among different questions types in this initial work.

Our definition of "question" follows coding schemes that are uniquely designed to analyze questions in classroom discourse (Nystrand et al., 2003). Questions are utterances in which the teacher solicits information from a student either procedurally (e.g., "*Is everyone ready?*"), rhetorically (e.g., "*Oh good idea James why don't we just have recess instead of class today*"), or for knowledge assessment/information solicitation purposes (e.g., "*What is the capital of Indiana, Michael?*"). Likewise, the teacher calling on a different student to answer the same question (e.g., "*Nope. Shelby?*") would also be considered a question, although in some coding schemes, the previous example would be classified as "Turn Eliciting" (Allwood et al., 2007). We do not consider certain cases questions, such as when the teacher calls on a student for other reasons (e.g., to discipline them) or when the teacher reads from a novel in which a character asked a question.

The coders were seven research assistants and researchers whose native language was English. The coders first engaged in a training task by labeling a common evaluation set of 100 utterances. These 100 utterances were manually selected to exemplify difficult cases. Once coding of the evaluation set was completed, the expert coder, who had considerable expertise with classroom discourse and who initially selected and coded the evaluation set, reviewed the codes. Coders were required to achieve a minimal level of agreement with the expert coder (Cohen's kappa, $\kappa = 0.80$). If the agreement was lower than 0.80, then errors were discussed with the coders.

After this training task was completed, the coders coded a subset of utterances from the complete dataset. Coders listened to the utterances in temporal order and assigned a code (question or not) to each based on the words spoken by the teacher, the teachers' tone (e.g., prosody, inflection), and the context of the previous utterance. Coders could also flag an utterance for review by a primary coder, although this was rare. In all, 36% of the 10,080 utterances were coded as containing questions. A random subset of 117 utterances from the full dataset were selected and coded by the expert coder. Overall the coders and the primary coder obtained an agreement of $\kappa = 0.85$.

2.4 Automatic Speech Recognition (ASR)

We used the Bing and AT&T Watson ASR systems (Microsoft, 2014; Goffin et al., 2005), based on evaluation in previous work (Blanchard, 2015; D'Mello et al., 2015). For both of these systems, individual utterances were submitted to the engine for transcription. We also considered the Azure Speech API (Microsoft, 2016) which processes a full-length classroom recording to produce a set of time-stamped words, from which we reconstructed the individual utterances.

We evaluated the performance of the ASR engines on a random subset of 1,000 utterances. We considered two metrics: word error rate (WER), which accounts for word order between ASR and human transcripts, and simple word overlap (SWO), a metric that does not consider word order. WER was computed by summing the number of substitutions, deletions, and insertions required to transform the human transcript into the computer transcript, divided by the number of words in the human transcript. SWO was computed by dividing the number of words that appear in both the human and computer transcripts by the number of words in the human transcript. Table 1 presents the WER and SWO for the three ASR systems, where we note moderate accuracy given the complexity of the task in that we are processing conversational speech recorded in a noisy naturalistic environment.

Table 1. ASR word error rate and simple word overlap averaged by teacher for 1,000 utterances, with standard deviations shown in parentheses.

ASR	WER	SWO
Bing Speech	0.45 (0.10)	0.55 (0.06)
AT&T Watson	0.63 (0.11)	0.42 (0.11)
Azure Speech	0.49 (0.07)	0.64 (0.16)

2.5 Model Building

We trained supervised classification models to predict if utterances contained a question or not (as defined in Section 2.3).

Feature extraction. In this work we focused on a small set of domain-general features rather than word specific models, such as n-grams or parse trees. Because we sampled many different teachers and classes, the topics covered vary significantly between class sessions, and a content-heavy approach would likely overfit to specific topics. This decision helps emphasize generalizability across topics as our models are intended to

be applicable to class sessions that discuss topics not covered in the training set.

Features (N = 37) were generated using the ASR transcripts for each utterance obtained from Bing Speech, AT&T Watson, and Azure Speech engines. Of these, 34 features were obtained by processing each utterance with the Brill Tagger (Brill, 1992) and analyzing each token (Olney et al., 2003). Features included the presence or absence of certain words (e.g., *what*, *why, how),* categories of words (e.g., definition, comparison), or part-of-speech tags (e.g., presence of nouns, presence of adjectives). These features were previously used to detect domain-independent question properties from human-transcribed questions (Samei et al., 2014). We supplemented these features with three additional features: proper nouns (e.g., student names), pronouns associated with uptake (teacher questions that incorporate student responses), and pronouns not associated with uptake, as recommended by a domain expert on teacher questions.

We extracted all 37 NLP features for each ASR transcription, yielding three feature sets. We also created a fourth set of NLP features that combined the features from the individual ASRs. For this set, each feature value was taken as the proportion of each features' appearances in the three ASR outputs. For example, if a feature was present in an utterance as transcribed by Bing and AT&T, but not Azure, then the feature's value would be 0.67.

Oversampling. We supplemented our imbalanced training data with synthetic instances (for the minority question class) generated with the Synthetic Minority Over-sampling Technique (SMOTE) algorithm (Chawla et al., 2011). Class distributions in the testing set were preserved.

Classification and validation. We considered the following classifiers: logistic regression, random forest, J48 decision tree, J48 with Bagging, Bayesian network, *k*-nearest neighbor (k = 7, 9, and 11), and J48 decision tree, using implementations from the WEKA toolkit (Hall et al., 2009). For each classifier, we tested with and without wrapping the classifiers with MetaCost, a cost-sensitive procedure for imbalanced datasets that assigned a higher penalty (weights of 2 or 4) to misclassification of the question class.

Classification models were validated using a leave-one-teacher-out cross-validation technique, in which models were built on data from 10 teachers (training set) and validated on the held-out teacher (testing set). The process was repeated until each teacher was included in the testing set. This cross-validation technique tests the potential of our models to generalize to unseen teachers both in terms of acoustic variability and in terms of variability in question asking.

3 Results

3.1 Classification Accuracy

In Table 2 we present the best performing classification model for each ASR and their combination based on the F_1 score for the question class (target metric). Table 2 includes the F_1 score for the question class, the F_1 score for the non-question class, and the overall weighted F_1 score. The best-performing individual ASR models were each Bayesian networks. The combined model was built with J48 with Bagging and with MetaCost (miss weight of 2). We show the confusion matrix for this model in Table 3.

Table 2. Results of best models for question detection.

Model	F_1 Question	F_1 Not-Question	F_1 Overall
AT&T	0.52	0.68	0.63
Azure	0.53	0.67	0.63
Bing	0.54	0.67	0.63
Combined	**0.59**	**0.74**	**0.69**

Table 3. Confusion matrix of combined ASR model for Question (Q) and Utterances (U).

n	Actual	Predicted	
		Q	*U*
3586	*Q*	2273	1313
6494	*U*	1946	4548

Overall, these results show a general consistency between the models using individual ASR transcriptions, which imply the relative success of each despite the differences in WER. Furthermore, we note that the combination of three ASR transcriptions resulted in improved performance compared to models built using individual ASR transcriptions. Using the combined model, we achieved slightly higher recall (0.63) than precision (0.57) for identifying questions.

We also compared our results to a chance model that assigned the question label at the same rate (42%) as our model, but did so randomly across 10,000 iterations. We consider this approach to computing chance to be more informative than a naïve minority baseline model (as the class of interest is the minority class) that would

yield perfect recall but negligible precision. The chance model had a mean recall of 0.42 and precision of 0.36 for the question class. From these averages, we calculated the chance F_1 score for questions (0.39). Our combined model achieved an F_1 score of 0.59 for the question class, which represents a 51% improvement over chance.

3.2 Feature Analysis

We explored the utility of the individual features using forward stepwise feature selection (Draper et al., 1966). For each individual ASR engine we identified the features selected in all folds of the teacher-level cross-validation procedure. We found four of the features were used in all three of the ASR models: *how*, *what*, *why*, and *wh-* (any word that starts with "*wh-*", including *who* and *where*). The selection of these features across the different ASR feature sets is perhaps unsurprising, but these results confirm that identifying question words are paramount for detecting questions regardless of the specific ASR engine.

3.3 Consistency Across Class-Sessions

The models were trained using leave-one-teacher-out cross-validation, but we perform additional post-hoc analyses exploring the model's accuracy across the 37 individual class sessions. This analysis allows an investigation of the stability of our model for individual class sessions, which will be essential for generalizability to future class sessions and topics.

Question Rate Analysis. Some applications only require an overall indication of the rate of question asking rather than identifying individual questions. To analyze the use of our model to these applications, we compared the proportion of predicted to actual questions for each class session (see Figure 1). There was a mean absolute difference of 0.08 (SD = 0.06) in the predicted proportion of questions compared to the true proportion (Pearson's r = 0.85). This small difference and strong correlation indicates that even though there are misclassifications at the level of individual utterances, the error rate is ameliorated at the session level, indicating the model performs well at correctly predicting the proportion of questions in a class session.

Performance Across Class-Sessions. Figure 2 presents a histogram of F_1 scores for the question class by class session. We note that model accuracy was evenly spread across class sessions rather than being concentrated on the tails (which would indicate a skewed distribution). In particular, 25% of the class sessions scored below 0.47 and 25% of the sessions scored above 0.66, yielding an interquartile range of 0.47 to 0.66. Encouragingly, the poorest performing class session still yielded an F_1 score of 0.33 while the best class session had a score of 0.84.

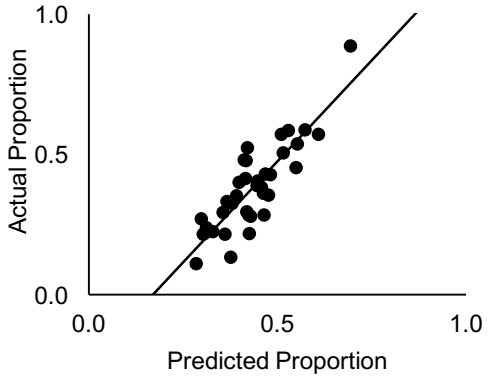


Figure 1. Proportion of predicted to actual questions in each class session.

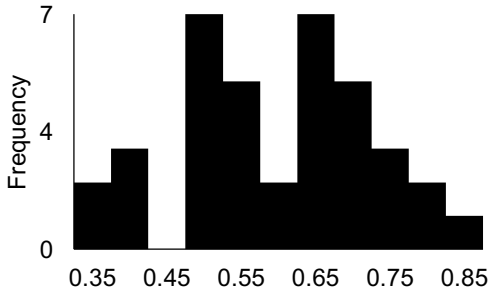


Figure 2. Histogram of F_1 scores for the question class by class-session.

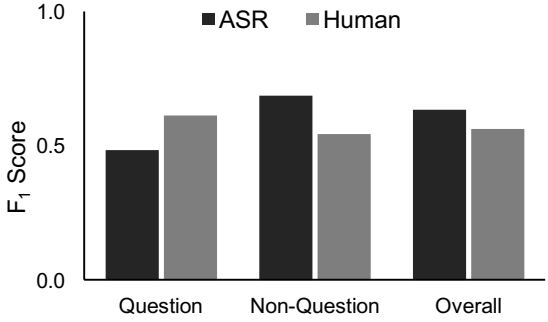


Figure 3. Models (ASR vs. human) built on 1,000 utterance subset.

3.4 Effects of ASR Errors

We explored how our models were affected by ASR errors. We built models on the subset of 1,000 utterances that we manually transcribed to evaluate WER and SWO of the ASRs in Section 2.4. Specifically, we retrained the J48 classifier reported in Section 3.1 on this data subset, using the combination of features from the three individual ASRs, comparing it to the same model built

Table 4. Confusion matrix showing a comparison of the ASR and Human models.

Actual		Predicted			
		Human Q	**Human Q**	**Human NQ**	**Human NQ**
Priors		**ASR Q**	**ASR NQ**	**ASR Q**	**ASR NQ**
0.30	**Q**	**0.15**	*0.07*	*0.03*	**0.05**
0.70	**NQ**	**0.18**	*0.16*	*0.09*	**0.28**

Note: Q indicates a question and NQ indicates a non-question. Bolded numbers indicate both models agreed while italicized numbers indicate disagreement.

using features extracted from the human transcriptions. The results of leave-one-teacher-out cross-validation are shown in Figure 3.

We must note that direct comparisons of models built on this subset of 1,000 instances with those built on the full data set (Section 3.1) are improper due to significantly fewer training instances in the former. In general, the human model achieved a higher F_1 for the question class compared to the combined ASR model, while the ASR model has a higher F_1 for the non-question class. We also note the tendency of the human model to over-predict questions, potentially resulting from the use of the MetaCost wrapper.

We further compared the predictions of the human and ASR models and observed that both models agreed in classifying utterances, either correctly or incorrectly, as questions and non-questions 65% of the time (see Table 4). They differed 35% of the time, disagreeing 25% of the time for non-questions and 10% of the time for questions. We note that, when the models disagreed, the human model was more likely to classify a non-question as a question (16%) compared to the ASR (9%), presumably due to its tendency to over-predict questions as noted above.

3.5 Analysis of Classification Errors

We selected a random sample of 100 incorrectly classified utterances using the human transcription model (so as to eliminate ASR errors as a potential explanation) to study possible causes of errors. We identified 44 utterances with common error patterns, whereas the cause of the error could not be easy discerned for the remaining 56 incorrectly classified utterances.

Out of the 44 errors, 24 were misses (questions predicted as non-questions). In 5 of these 24 misses, the question was only one part of the utterance (e.g., "*If I could just get this thing to open I'd be fine. Can you do it*?"). The remaining 19 errors yielded examples of question types that may be problematic for our model. These include calling on individual students (e.g., "*Sam?*"), rhetorical questions (e.g., "*musical practice, right?*"), implicit questions requiring clues from previous context (e.g., "*why did she say that?*"), fill-in-the-blank questions (e.g., "*Madagascar and _____?*"), and students being directed to speak, rather than being asked a traditional question (e.g., "*tell us about it*").

Additionally, there were 20 false alarms (non-questions incorrectly classified as questions). Nine of these non-questions were offhand/casual statements made by teachers ("*I don't know if you guys should call him that or not*" said jokingly) while interacting with student, indicative of the difficulty of classifying questions in contexts with informal dialogue. Five short utterances may have been classified incorrectly because of limited context (e.g., "*good.*" vs. "*good?*", "*okay.*" vs. "*okay?*"). Three misclassifications involved teachers reading directly from a book, (e.g., quoting a passage from a novel in which a character asks a question). Additionally, there was one aborted statement and one aborted question, in which the teacher started to say something but changed course mid-sentence (e.g., "*No wh- ... put that away!*"). Finally, in another case, the teacher paused midsentence, resulting in a very short utterance that left the full intent of the statement to the next utterance (e.g., "*Juliet reversed course, the nurse…*"). This last example highlights the difficulties of classifying questions with imperfect sentence boundaries (see Section 2.3) as is the case with our data. In general, 15 of the 20 false alarms were associated with changes in speaking style from traditional teacher speech in classrooms.

4 General Discussion

The importance of teacher questions in classrooms is widely acknowledged in both policy (e.g., Common Core State Standards for Speaking and Listening (2010)) and research (Nystrand and Gamoran, 1991; Applebee et al., 2003; Nystrand et al., 2003). Teacher questions play a central role in student engagement and achievement, suggesting that automating the detection of questions

might have important consequences for both research on effective instructional strategies and on teacher professional development. Thus, our current work centers on a fully-automated process for predicting teacher questions in a noisy real-world classroom environment, using only a full-length audio recording of teacher speech.

4.1 Main Findings

We present encouraging results with our automated processes, consisting of VAD to automatically segment teacher speech, ASR transcriptions, NLP features, and machine learning. In particular, our question detection models excel in aggregation of utterances: the detected proportion of questions per class strongly correlates with the proportion of actual questions in the classroom (Pearson's r = 0.85). In addition, our models provided promising results in the detection of individual questions, although further refinement is needed. Both types of analysis are useful in providing formative feedback to teachers, at coarse- and fine-grained levels, respectively.

A key contribution of our work over previous research is that our models were trained and tested on automatically-, and thus imperfectly-, segmented utterances. This extends the work of (Orosanu and Jouvet, 2015) which artificially explored perturbations of a subset of utterance boundaries using the automatic detection of silence within human-segmented spoken sentences. To our knowledge, our work is the first to detect spoken questions using a fully automated process. Our best model achieved an overall F_1 score of 0.69 and an F_1 score of 0.59 for the question class. This represents a substantial 37% improvement in question detection accuracy over a recent state-of-the-art model (Boakye et al., 2009) that reported an overall F_1 of 0.50; the authors do not report F_1 for the question class so the comparison is based on the overall F_1.

We validated our models using leave-one-teacher-out cross-validation, demonstrating generalizability of our approach across teachers in this dataset. Furthermore, we analyzed model performance by class session, finding our model was consistent across class sessions, an encouraging result supporting our goals of domain-independent question detection.

We also explored the differences between models using ASR transcriptions and using human transcriptions. Overall, the results were quite comparable suggesting that imperfect ASR need not be a barrier against automated question detection in live classrooms.

4.2 Limitations and Future Work

This study is not without limitations. We designed our approach to avoid overfitting to specific classes, teachers, or schools. However, all of our recordings were collected in Wisconsin, a state that uses the Common Core standard. It is possible that the Common Core may impose aspects of a particular style of teaching that our models may overfit. Similarly, although we used speaker-independent ASR and teacher-independent validation techniques to improve generalizability to new teachers, our sample of teachers are from a single region with traditional Midwestern accents and dialects. Therefore, broader generalizability across the U.S. and beyond remains to be seen.

We acknowledge that our method for teacher utterance segmentation may potentially be improved using proposed techniques in related works. Komatani et al. (2015) has explored detecting and merging utterances segmented mid-sentence, allowing analysis to take place on a full sentence, rather than a fragment, which may improve question detection by merging instances in which questions were split. An alternative approach would be to automatically detect sentence boundaries within utterances, and extract features from each detected sentence.

Our analysis of errors in Section 3.5 suggests that acoustic and contextual features may be needed to capture difficulty to classify questions. Additionally, related work on question detection (see Section 1.1) suggested that acoustic, contextual, and temporal features (Boakye et al., 2009) may aid in the detection of questions. We will explore this in future work to determine if features capturing these properties will help improve our models for this task. Likewise, we will also explore temporal models, such as conditional random fields and bi-directional long-short-term neural networks, which might better capture questions in the larger context of the classroom dialogue. This temporal analysis may help find sequences of consecutive questions, such as those present in question-and-answer sessions or in classroom discussions.

Further, Raghu et al. (2015) has explored using context to identify non-sentential utterances (NSUs), defined as utterances that are not full sentences but convey complete meaning in context. The identification of NSUs may improve our model's ability to differentiate between difficult cases (e.g., calling on students, saying a student's name for discipline).

In addition to addressing these limitations by collecting a more representative corpus and computing additional features, there are several other directions for future work. Specifically, we will focus on classifying question properties defined by Nystrand and Gameron (2003). While we have explored these properties in previous work (Samei et al., 2014; Samei et al., 2015), that work used perfectly-segmented and human-transcribed question text. We will continue this work using our fully-automatic approach that employs automatic segmentation and ASR transcriptions.

4.3 Concluding Remarks

We took steps towards fully-automated detection of teacher questions in noisy real-world classroom environments. The present contribution is one component of a broader effort to automate the collection and coding of classroom discourse. The automated system is intended to catalyze research in this area and to generate personalized formative feedback to teachers, which enables reflection and improvement of their pedagogy, ultimately leading to increased student engagement and achievement.

5 Acknowledgements

This research was supported by the Institute of Education Sciences (IES) (R305A130030). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the IES.

References

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3–4):273–287.

Arthur N Applebee, Judith A Langer, Martin Nystrand, and Adam Gamoran. 2003. Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, 40(3):685–730.

Isabel L. Beck, Margaret G. McKeown, Cheryl Sandora, Linda Kucan, and Jo Worthy. 1996. Questioning the author: A yearlong classroom implementation to engage students with text. *The Elementary School Journal*:385–414.

Nathaniel Blanchard, Patrick J Donnelly, Andrew M Olney, Borhan Samei, Brooke Ward, Xiaoyi Sun, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. 2016. Automatic detection of teacher questions from audio in live classrooms. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016),* pages 288–291. International Educational Data Mining Society.

Nathaniel Blanchard, Michael Brady, Andrew Olney, Marci Glaus, Xiaoyi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney K. D'Mello. 2015. A Study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, pages 23-33. International Educational Data Mining Society.

Nathaniel Blanchard, Sidney D'Mello, Martin Nystrand, and Andrew M. Olney. 2015. Automatic classification of question & answer discourse segments from teacher's speech in classrooms. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015),* pages 282–288. International Educational Data Mining Society.

Kofi Boakye, Benoit Favre, and Dilek Hakkani-Tür. 2009. Any questions? Automatic question detection in meetings. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding, (ASRU),* pages 485–489. IEEE.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Workshop on Speech and Natural Language*, pages 112–116. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research,* 16:321–357.

Sidney K D'Mello, Andrew M Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly. 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings*

of the 2015 International Conference on Multimodal Interaction, pages 557–566. ACM.

Norman Richard Draper, Harry Smith, and Elizabeth Pownell. 1966. *Applied regression analysis*. Wiley New York.

Adam Gamoran and Martin Nystrand. 1991. Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence*, 1(3):277–300.

Vincent Goffin, Cyril Allauzen, Enrico Bocchieri, Dilek Hakkani-Tür, Andrej Ljolje, Sarangarajan Parthasarathy, Mazin G. Rahim, Giuseppe Riccardi, and Murat Saraclar. 2005. The AT&T WATSON Speech Recognizer. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* pages 1033–1036. IEEE.

Arthur C. Graesser and Natalie K. Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Mary M Juzwik, Carlin Borsheim-Black, Samantha Caughlan, and Anne Heintz. 2013. *Inspiring dialogue: Talking to learn in the English classroom*. Teachers College Press.

Sean Kelly. 2007. Classroom discourse and the distribution of student engagement. *Social Psychology of Education*, 10(3):331–352.

Kazunori Komatani, Naoki Hotta, Satoshi Sato, and Mikio Nakano. 2015. User adaptive restoration for incorrectly segmented utterances in spoken dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 393.

Microsoft. 2014. The Bing Speech Recognition Control. Technical report.

Microsoft. 2016. Azure Speech API. Technical report.

Martin Nystrand. 2006. Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*:392–412.

Martin Nystrand and Adam Gamoran. 1991. Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*:261–290.

Martin Nystrand, Lawrence L Wu, Adam Gamoran, Susie Zeiser, and Daniel A Long. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes*, 35(2):135–198.

Andrew Olney, Max Louwerse, Eric Matthews, Johanna Marineau, Heather Hite-Mitchell, and Arthur Graesser. 2003. Utterance classification in AutoTutor. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing-Volume 2*, pages 1–8. Association for Computational Linguistics.

Luiza Orosanu and Denis Jouvet. 2015. Detection of sentence modality on French automatic speech-to-text transcriptions. In *Proceedings of the International Conference on Natural Language and Speech Processing*.

Dinesh Raghu, Sathish Indurthi, Jitendra Ajmera, and Sachindra Joshi. 2015. A statistical approach for non-sentential utterance resolution for interactive QA system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 335.

Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier. 2013. An open-source state-of-the-art toolbox for broadcast news diarization. Technical report.

Borhan Samei, Andrew Olney, Sean Kelly, Martin Nystrand, Sidney D'Mello, Nathan Blanchard, Xiaoyi Sun, Marci Glaus, and Art Graesser. 2014. Domain independent assessment of dialogic properties of classroom discourse. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)* pages 233-236. International Educational Data Mining Society.

Borhan Samei, Andrew M Olney, Sean Kelly, Martin Nystrand, Sidney D'Mello, Nathan Blanchard, and Art Graesser. 2015. Modeling

classroom discourse: Do models that predict dialogic instruction properties generalize across populations? *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015),* pages 444-447. International Educational Data Mining Society.

Tanya Stivers and Nick J. Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.