Semi-Automatic Detection of Teacher Questions from Human-Transcripts of Audio in Live 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@nd.edu; sdmello@nd.edu

ABSTRACT

We investigate automatic detection of teacher questions from automatically segmented human-transcripts of teacher audio recordings collected in live classrooms. Using a dataset of audio recordings from 11 teachers across 37 class sessions, we automatically segment teacher speech into individual teacher utterances and code each as containing a teacher question or not. We trained supervised machine learning models to detect questions using high-level natural language features extracted from human transcriptions of a random subset of 1,000 segmented utterances. The models were trained and validated independently of the teacher to ensure generalization to new teachers. We are able to detect questions with a weighted F_1 score of 0.66, suggesting the feasibility of question detection on automatically segmented audio from noisy classrooms. We discuss the possibility of using automatic speech recognition to replace the human transcripts with an eye towards providing automatic feedback to teachers.

Keywords

Automatic Speech Recognition, Natural Language Processing, Classroom Environments, Question Detection

1. INTRODUCTION

Teachers employ a wide array of instructional strategies in their classrooms due to individual teaching styles, requirements of the curricula, and other constraints. These strategies may include lectures, asking questions and evaluating student responses, or assigning small-group work, among many others. However, these approaches are not equally effective at promoting student achievement. Certain techniques, such as asking particular types of questions or facilitating a classroom-wide discussion, have been shown to predict student engagement and achievement growth above others [1], [2].

Research also indicates that providing teachers with feedback on their instructional practices can lead to improved student achievement [3]. But where does the feedback come from? Currently, the onus is on trained human judges who analyze teacher instruction by observing live classrooms. For example, the Nystrand and Gamoran coding scheme [4], [5] provides a general template for observers to document and analyze teacher instructional practices. This scheme has been empirically validated in numerous studies across hundreds of middle school and high school classrooms [6]–[8]. Unfortunately, this is an expensive and labor intensive process that hinders the ability to analyze classroom instruction at scale. Instead, computational methods that can automatically analyze classroom instruction at scale are needed. We take a step in this direction by considering the possibility of detecting teacher questions in live classrooms. We focus on questions because they are a central component of dialogic instruction, often serving as a catalyst for in-depth classroom discussions and so called 'dialogic spells' [9].

The classroom environment provides a unique set of challenges for the automatic analysis of questions. There are also numerous constraints as discussed in detail by D'Mello et al. [10]. Briefly, the analytic approach should not be disruptive to either the teacher or the students. Secondly, it must be affordable to enable wide-spread adoption across classrooms. Finally, for privacy concerns, video recordings are not possible unless students can be de-identified.

We attempted to overcome these challenges by designing a system that includes a low cost, wireless headset microphone to record teachers as they move about the classroom freely. Our system accommodates various seating arrangements, classroom sizes, and room layouts, but attempts to mitigate complications due to ambient classroom noise, muffled speech, or classroom interruptions, factors that reflect the reality of real-world environments.

There is the open question as to whether the data collected in this fashion can be of sufficient quality for automatic question detection. As an initial step, we consider semi-automated question detection from human-transcripts of automatically-segmented teacher audio. If successful, the next step would be to apply our basic approach by using automatic speech recognition (ASR) in lieu of human transcriptions.

1.1 Related Work

Our work is related to previous attempts at automatic detection of questions from transcriptions of audio albeit outside of the noisy classroom interaction context we consider here. We limit our review to experiments that include ASR, as our ultimate goal is in full automation of question detection.

In a study attempting to detect questions in office meetings, Boakye et al. [11] trained models using the ICSI Meeting Recorder Dialog Act (MRDA) corpus, a dataset of 75 hour-long meetings recorded with headset and lapel microphones. Using an AdaBoost classifier to detect questions from human transcriptions, the authors obtained an F_1 score of 67.6 by combining various NLP features.

Stolcke et al. [12] built a dialogic act tagger on the conversational switchboard database. A Bayesian network modeling word and trigrams discourse grammars, from human transcriptions achieved a recognition rate of 71% to detect a set of dialogic acts, such as statements, questions, apologies, or agreement (chance level 35%; human agreement 84%). The authors further attempted to distinguish questions from statements, two speech acts often confused by their model. They obtained an accuracy of 86% on a subset of their dataset containing equal proportions of questions and statements using only word features (chance accuracy 50%). This result, while promising, is based on an artificially balanced dataset of statements and questions.

Most recently, Orosanu and Jouvet [13] investigated classification of sentences labeled as either statements or questions in three French language corpora, testing on a set of 7,005 statements and 831 questions. The models accurately classified 75.5% of questions and 72.0% of statements using human transcripts. The authors compared the results of using human-annotated sentence boundaries against a semi-automatic method for boundary detection. A subset of sentences, those without prior and proceeding silences of an undefined length, were split once on the longest silence in the sentence; the remainder of the sentences were left unchanged. Semi-automatic splitting led to a 3% increase in classification errors. Although only a subset of sentences were split and there were no cases where sentences were combined, the results suggest that detecting questions from imperfect boundaries may be possible.

1.2 Contributions and Novelty

We describe an approach to automatically identify teacher questions from human-transcriptions of teacher audio recorded in live classrooms. We make several contributions while addressing these challenges. First, we examine a dataset of full length recordings of real world class sessions, drawn from multiple teachers and schools. Second, we only use teacher audio because it is the most scalable and practical option. Third, we automatically segment audio recordings into individual teacher utterances in a fully automated fashion and manually transcribe a subset of these utterances for use in our classification models. Fourth, we restrict our feature set to high-level natural language features that are more likely to generalize to classes on different topics rather than low-level domain-specific words. Finally, we design our models to generalize across teachers rather than optimizing to the speech patterns of individual teachers.

2. METHOD

2.1 Recording Teacher Audio

Data was collected at six rural Wisconsin middle schools during literature, language arts, and civics classes. Class sessions were taught by 11 teachers (three male; eight female) and lasted between 30 and 90 minutes. The teachers carried out their normal lesson plan, allowing the collection of a corpus of real-world samples of classrooms. Based on previous work [10], a Samson 77 Airline wireless microphone was chosen for teachers to wear while teaching. Teacher speech was captured and saved as a 16 kHz, 16-bit single channel audio file. A total of 37 class sessions were recorded on 17 separate days over a period of a year. The recordings contain a total of 32 hours and five minutes of audio.

2.2 Teacher Utterance Detection

Teacher speech was segmented into utterances using a voice activity detection (VAD) technique described in [14] and briefly reviewed here. Audio from the teacher's microphone was automatically split into potential utterances, consisting of either teacher speech or other sounds (e.g., accidental microphone contact, classroom noise), based on pauses (i.e., periods of silence) between speech. The beginning of a potential utterance was automatically identified when the amplitude envelope rose above a preset threshold. The end point of the utterance was automatically identified when the amplitude envelope dropped below this threshold for at least 1000 milliseconds, a pause of one second. The threshold was set to be sufficiently low so as to capture all instances of speech, also causing a high rate of false-alarms. False alarms were eliminated by filtering all potential utterances with Bing ASR [15]. If the ASR rejected a potential utterance, then it was discarded as a non-speech segment.

We validated the effectiveness of our VAD approach in an experiment by hand coding a random subset of 1,000 potential utterances as either containing speech or not containing speech [11]. We achieved an F_1 score of 0.97, which we deemed sufficiently accurate for the purposes of this study. Therefore, we applied our approach for VAD to the full dataset of 37 classroom recordings and extracted 10,080 utterances.

2.3 Question Coding and Transcription

We manually coded the complete set of automatically extracted utterances as containing a question or not. It should be noted that a known limitation of annotating automatically segmented speech is that each utterance may contain multiple tags (questions in this case), or conversely, a tag may be spread across over multiple utterances. This occurs because we use both a fixed amplitude envelope threshold and pause length to segment utterances, rather than creating specific thresholds for each teacher or class-session. This fully automates the VAD detection process, and allows us to test generalizability to new teachers. For this work, we allow question tags to span multiple utterances, since the entire content of question is likely to be essential to future work aimed at providing feedback to teachers.

We define a question after the question coding scheme developed by Nystrand and Gameron [4], [5], which is specific to classrooms. For example, calling on students in class (e.g., "What is the capital of Iowa [pause] Michael") is considered a question. Likewise, the teacher calling on a different student to answer the same question after evaluating the previous response (e.g., "Nope [pause] Shelby") is also considered a question. Calling a student name for other reasons, such as to discipline them, is not a question (e.g., "Steven"). Thus, question coding involves ascertaining both the context and intentionality of the utterance.

The coders were seven research assistants and researchers whose native language was English. Coders listened to the utterances in temporal order and assigned a label (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 occurred rarely.

As training, the coders first engaged in a task of labeling a common evaluation set of 100 utterances. These 100 utterances were selected to exemplify difficult cases. Once coding of the evaluation set was completed, the primary 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 primary coder (Cohen's kappa, κ = 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. 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 $\kappa = 0.85$ on this evaluation set.

From the full dataset of 10,080 labeled utterances, we selected a random (without replacement) subset of 1,000 utterances for manual transcription by humans. 30% of the utterances in this subset contained a question, which is slightly lower than the 36% question rate on the entire dataset.

2.4 Model Building

We trained and tested supervised classification models to predict if utterances contained part (or all) of a question, or did not contain a question. The model building process involved the following steps.

Features. Features were generated using the human transcripts for each utterance. We limited our feature set to a set of 37 generalizable NLP features to limit overfitting to teacher dialect or classroom subject/domain. These 34 features were obtained by processing each utterance with the Brill Tagger [16]. Each tagged token was examined for features (see [17] for further details) based on the semantics of various question types (e.g., causal, interpretation, disjunction) or the syntax of questions (e.g., WH-words and modal verbs). These 34 features capture key word (e.g., *why, how),* word categories (e.g., procedural), and parts of speech (e.g., noun, verb), and have previously been used to detect domain independent question properties associated with learning from human-transcribed questions [18]. Three additional features include proper nouns (e.g., student names), pronouns associated with teacher questions incorporating student responses (a type of question known as uptake), and pronouns not associated with uptake.

Minority oversampling. We supplemented *training* data with additional synthetic instances generated by the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [19] in order to eliminate skew in the training set. Importantly, SMOTE was only applied to the training set and the original distributions in the testing set were not altered.

Classification and validation. We explored a number of classifiers: Naïve Bayes, 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 [20]. We also combined the classifiers with MetaCost, which penalized misclassifications of the minority class (weights of 2 and 4). All 37 features were used in the models.

We validated the classification models with leave-one-teacher-out cross-validation, in which models were built on data from 10 teachers (the training set) and validated on the held-out teacher (the testing set). The process was repeated for 11 folds so that each teacher appeared once in the testing set. This cross validation technique tests the potential of our models to generalize to new teachers in terms of variability in question asking and language.

3. RESULTS

The best performing model was Naïve Bayes, which achieved the overall highest F_1 score (0.53) for detecting utterances containing questions (the minority class). This model achieved an overall weighted F_1 score of 0.66 (see Table 1 for the confusion matrix). Additionally, we also compared our results to a chance-model that assigned the question label at the same rate as our model, but did so randomly. We calculated the chance recall and precision for the question label as the average value per teacher over 10,000 iterations. We consider this approach to computing chance to be more informative than a naïve minority baseline model that would yield perfect recall but negligible precision. We observed an encouraging level of recall (0.61) for the question class, which reflects the model's ability to detect questions from utterances well above both chance precision (0.32) and recall (0.42). However, we note that further refinement is needed to improve the model's precision (0.47), which is hindered by the frequent misclassification of utterances as questions.

Table 1. Confusion matrix of 1,000 utterance subset, showing the count and the proportion in parenthesis.

Instances	Actual	Predicted	
		Question	*Utterance*
320	*Question*	195 (0.61)	125 (0.39)
680	*Utterance*	224 (0.33)	456 (0.67)

4. GENERAL DISCUSSION

Questions play a central role in dialogic instruction in classrooms. The importance of dialogue and discussion is widely acknowledged in research [6], [9], [20] and public policy (e.g., Common Core State Standards for Speaking and Listening). The ability to automatically detect questions for both research and teacher professional development might have important consequences in improving student engagement. Towards this goal, our current work focuses on semi-automatic prediction of individual teacher questions teacher audio recorded in live classrooms.

We demonstrated promising results with our approach, consisting of manually transcribed automatically segmented teacher speech, high-level language features, and machine learning. Our best model, validated independently of the teacher, achieved an overall F_1 score of 0.66 and a F_1 score for the question class of 0.53. This reflects a modest improvement in overall classification (F_1 of 0.63) and a significant improvement in question detection accuracy (F_1 of 0.40) over a recent state of the art model [13].

A major contribution of our work is that our models were trained and tested only on automatically, and thus imperfectly, segmented utterances. This confirms that question detection on imperfect sentence boundaries is possible, a result that furthers the work of [13], in which the authors split a subset of manually defined sentences on the longest silence in the sentence (see Section 1.1).

Despite these encouraging results, this study is not without limitations. Most importantly, we only considered manually transcribed speech in order to examine the feasibility of the automatic identification of questions derived from noisy classroom environments. To fully automate our approach we will need to incorporate ASR engines. We expect that the incorporation of noisy ASR will contribute to additional errors in classification, a possibility we are studying in ongoing work that applies automatic speech recognition (ASR) on our full dataset of 10,080 utterances.

Research [11]–[13] indicates that acoustic and contextual features may be important to capture certain difficult types of questions and we will explore the use of these features in future work. Furthermore, additional data collection which includes a second microphone that captures general classroom activity is ongoing. This second channel of audio, when combined with the recording of the teacher, will allow modelling patterns of teacher-student interactions, potentially revealing question-response patterns between teachers and students. Finally, we will extend our approach to classify the question properties defined by Nystrand and Gameron [9]. We have previously explored this task using human transcriptions of manually segmented questions [18], [21], but will extend this work using our approach that employs automatic segmentation and subsequently ASR transcriptions.

In summary, we took steps towards fully automating the detection of teacher questions from audio recordings of live classrooms. We will continue to refine and improve these models as we extend our approach to use ASR transcriptions of the utterances. The present contribution is one component of a broader effort to automate the collection and coding of classroom discourse to improve learning. The automated system is intended to generate personalized formative feedback to teachers, enabling reflection and improvement of their pedagogy, with the ultimate goal of increasing student engagement and achievement.

5. ACKNOWLEDGMENTS

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.

6. REFERENCES

[1] S. Kelly, "Classroom discourse and the distribution of student engagement," *Soc. Psychol. Educ.*, vol. 10, no. 3, pp. 331–352, 2007.

[2] W. Sweigart, "Classroom talk, knowledge development, and writing," *Res. Teach. Engl.*, vol. 25, no. 4, pp. 469–496, Dec. 1991.

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

[4] M. Nystrand, A. Gamoran, R. Kachur, and C. Prendergast, "Opening dialogue," *Teach. Coll. Columbia Univ. N. Y. Lond.*, 1997.

[5] A. Gamoran and S. Kelly, "Tracking, instruction, and unequal literacy in secondary school english," *Stab. Change Am. Educ. Struct. Process Outcomes*, pp. 109–126, 2003.

[6] A. N. Applebee, J. A. Langer, M. Nystrand, and A. Gamoran, "Discussion-Based Approaches to Developing Understanding: Classroom Instruction and Student Performance in Middle and High School English," *Am. Educ. Res. J.*, vol. 40, no. 3, pp. 685–730, 2003.

[7] M. Nystrand, "Research on the role of classroom discourse as it affects reading comprehension," *Res. Teach. Engl.*, vol. 40, no. 4, pp. 392–412, May 2006.

[8] M. Nystrand and A. Gamoran, "Instructional discourse, student engagement, and literature achievement," *Res. Teach. Engl.*, pp. 261–290, 1991.

[9] M. Nystrand, L. L. Wu, A. Gamoran, S. Zeiser, and D. A. Long, "Questions in time: Investigating the structure and dynamics of unfolding classroom discourse," *Discourse Process.*, vol. 35, no. 2, pp. 135–198, 2003.

[10] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly, "Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2015, pp. 557–566.

[11] K. Boakye, B. Favre, and D. Hakkani-Tur, "Any questions? Automatic question detection in meetings," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 485–489.

[12] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguist.*, vol. 26, no. 3, pp. 339–373, 2000.

[13] L. Orosanu and D. Jouvet, "Detection of sentence modality on French automatic speech-to-text transcriptions," in *International Conference on Natural Language and Speech Processing*, Alger, Algeria, 2015.

[14] N. Blanchard, M. Brady, A. M. Olney, M. Glaus, X. Sun, M. Nystrand, B. Samei, S. Kelly, and S. D'Mello, "A study of automatic speech recognition in noisy classroom environments for automated dialog analysis," in *Artificial Intelligence in Education*, 2015, pp. 23–33.

[15] Microsoft, "The Bing Speech Recognition Control," May 2014.

[16] E. Brill, "A simple rule-based part of speech tagger," in *Proceedings of the Workshop on Speech and Natural Language*, 1992, pp. 112–116.

[17] A. Olney, M. Louwerse, E. Matthews, J. Marineau, H. Hite-Mitchell, and A. Graesser, "Utterance classification in AutoTutor," in *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing Volume 2*, 2003, pp. 1–8.

[18] Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., Sun, X., Glaus, M. & Graesser, A., "Domain independent assessment of dialogic properties of classroom discourse," in *7th International Conference on Educational Data Mining*, 2014, pp. 233–236.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, pp. 321–357, 2011.

[20] National Governors Association Center for Best Practices and Council of Chief State School Officers, "Common Core State Standards Speaking & Listening." National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C., 2010.

[21] Samei, Borhan, Olney, Andrew M., Kelly, Sean, Nystrand, Martin, D'Mello, Sidney, Blanchard, Nathaniel, and Graesser, Art, "Modeling classroom discourse: Do models of predicting dialogic instruction properties generalize across populations?," in *Proceedings of the Eighth International Conference on Educational Data Mining*, Madrid, Spain, 2015, pp. 444–447.