Automating Analysis and Feedback to Improve Mathematics Teachers' Classroom Discourse

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, Wayne Ward
Institute of Cognitive Science
Department of Computer Science
University of Colorado Boulder

Abstract

Our work builds on advances in deep learning for natural language processing to automatically analyze transcribed classroom discourse and reliably generate information about teachers' uses of specific discursive strategies called "talk moves." Talk moves can be used by both teachers and learners to construct conversations in which students share their thinking, actively consider the ideas of others, and engage in sustained reasoning. Currently, providing teachers with detailed feedback about the talk moves in their lessons requires highly trained observers to hand code transcripts of classroom recordings and analyze talk moves and/or one-on-one expert coaching, a time-consuming and expensive process that is unlikely to scale. We created a bidirectional long short-term memory (bi-LSTM) network that can automate the annotation process. We have demonstrated the feasibility of this deep learning approach to reliably identify a set of teacher talk moves at the sentence level with an F1 measure of 65%.

Introduction

Classroom recordings can be understood as a new form of big data: they are now common practice and rapidly becoming a highly utilized resource for teacher learning. These recordings consist of video, audio, and/or transcripts of teaching episodes, including entire lessons or portions of lessons. They may capture whole class discussions, teacher facilitated-group work, or online instruction. For instance, a recent search for "video in teacher education" yielded 1,630,000 results in Google Scholar. Furthermore, an increasing number of websites host large collections of videos of teaching episodes, either collected by research teams or uploaded by teachers themselves; see for example The Teaching Channel, Teacher Tube, The Teaching and Learning Exploratory, and Inside Mathematics. TeacherTube, based on YouTube, contains over 400,000 classroom videos, the majority of which were uploaded by teachers. These sites reflect calls for curated digital video libraries as a means of systematically sharing classroom practice and generating a knowledge base that can support instructional improvement (Hiebert, Gallimore, and Stigler 2002). There is likely to be exponential growth in classroom recordings in the coming decade as more teachers "self-record" using advances in classroom robotics (e.g., SWIVL, Panopto) and other audio/video recording technologies (e.g., improved smart-phone capabilities). Our goal is to utilize this data to provide valuable feedback to teachers on their effective use of discourse practices. Our efforts to develop a deep learning framework for classifying discourse serves as an example of the type of innovative application that can be built using large scale repositories of classroom teaching.

We take our motivation from Vygotsky's theory (Vygotsky 1978) that deliberative social interaction is essential to the development of individuals' mental processes. This theory still resonates within the education community, and has heavily influenced the literature on effective classroom talk (Michaels, O'Connor, and Resnick 2008). Of particular relevance is Vygotsky's argument that instructional moves should be used to support students to shift from being able to learn while assisted to becoming independent, and that these moves should occur at opportune moments within a given student's zone of proximal development. In the mathematics education field, there is widespread agreement that students' understanding should be constructed through the process of interacting within a learning community, with discussions serving as a prominent and normative feature within K-12 classrooms (Evans, Leija, and Falkner 2001). The Common Core State Standards for Mathematical Practice emphasize communication as a means of promoting argumentation and reasoning, and engaging students in the intellectual work of mathematics by vocalizing their thinking and making sense of others' ideas (Franke et al. 2015).

In this paper, we describe the development of a talk moves classification model, based on transcribed audio from teacher and student classroom interactions. This classification model will eventually be incorporated into a larger application (TalkBack) that draws on recorded audio tracks to detect specified discourse moves, and generates personalized feedback through a web-based interface. The TalkBack application and its potential to support instructional improvement is possible due to convergent advances in three areas: the increasingly widespread availability of a new big data source, advances in deep learning for speech and language processing tasks that can be achieved with high levels of reliability, and consensus in mathematics education research on the types of mathematics discussion and teacher supports that promote student learning. In the next section, we discuss related work using deep learning networks to classify classroom speech, followed by educational theory and rationale on talk moves. The Method section describes the data sources and model architecture method, followed by our results, discussion, and conclusion in consecutive sections.

Related work

Our model architecture, as well as its representations, are based on prior research in natural language processing and deep learning. Deep learning models have been extensively used in a variety of natural language processing tasks such as binary and multiclass classification, entailment and semantic relatedness, event extraction, semantic textual similarity, paraphrase detection, machine translation, sequence tagging, and caption image-retrieval tasks, among others (Conneau et al. 2017). In the domain of classroom discourse, several research teams have developed and validated automated systems for discriminating basic discourse structures (e.g. lecture, group work) (Donnelly et al. 2016). For example, (Donnelly et al. 2016) trained supervised machine learning models to classify instructional segments with F1 scores ranging from 0.64 to 0.78 based on data from 76 middle school classes. Additionally, (Donnelly et al. 2017) demonstrated the feasibility of using automatic speech recognition and classification models to automatically segment classroom speech and determine whether or not teachers' utterances contained a question. This work was based on a dataset of 10,080 utterances, and the best performing models achieved an F1 score of 0.69. Other efforts to use automatic speech recognition include segmenting teacher and student classroom speech (D'Mello et al. 2015) and low-level acoustic features (Donnelly et al. 2016). We believe the research presented in this article is the first to use a deep learning model on classroom discourse data to perform a multi-class classification of teachers' sentences.

The boost in performance of deep learning models when compared with traditional models such as Naive Bayes, linear regression and k-nearest neighbours can partially be attributed to recent advances in learning representations such as sentence embedding and continuous word embedding which have made it possible for deep learning models to discriminate between sentences in higher dimensional space while being robust to noise. Recurrent neural networks (RNNs) are more popular than traditional feedforward networks due to their capability to store past inputs in order to produce the current output (Mikolov and Zweig 2012). However, owing to the vanishing gradient problem (Hochreiter 1998), researchers developed the Long Short-Term memory network or LSTM. LSTM replaces the RNN layer units with LSTM units where the activation function is the identity function. Hence, the backpropagated gradient neither explodes or vanishes. This advantage has prompted many researchers to adopt LSTM networks over RNN's. Similar to RNNs, LSTMs have a bidirectional variant where the inputs are provided in both directions (forward and backward). This variant enables the network to preserve information from the past and the future. In some applications where context is important, such as speech and language, bidirectional LSTMs consistently outperformed LSTMs by feeding the inputs in both directions (Conneau et al. 2017). Within our research, which relies heavily on the context of conversations, we incorporated a bidirectional LSTM layer in our deep learning model.

Feature representations are crucial to determining and improving model performance. In our work, we used features common to state-of-the-art language applications, specifically sentence embeddings and word embeddings. For sentence embeddings, we used the ALLNLI corpus (Conneau et al. 2017), which is modelled after the Stanford Natural language inference (SNLI) corpus (Bowman et al. 2015) and MultiNLI corpus (Williams, Nangia, and Bowman 2017), to produce these features for our models. The SNLI corpus is a collection of human written sentence pairs which have been manually labelled to support the task of natural language inference. The labels include neutral entailment and contradiction. Previously, SNLI was used for training sentence embedding models. However, the MultiNLI corpus (Williams, Nangia, and Bowman 2017), which is a multi-genre version of SNLI with 433K sentence pairs, is observed to produce a significant boost in performance when compared to SNLI (Conneau et al. 2017). We retrained the model with 600 dimensions and used the trained model for our sentence embedding representation. In addition to sentence embedding, we have also used the GloVe Bag of words (GloVe BOW) representation as a feature. GloVe or Global vectors for word representation is an unsupervised learning algorithm trained on aggregated word-word co-occurrence statistics from a corpus. In our model, we use the vectors trained on Common Crawl with 840 billion tokens and 300 dimensions. For a given sentence, we identify the GloVe representation for each word in the sentence and compute a mean vector.

Educational Theory and Rationale

Research has demonstrated that implementing accountable talk moves in the classroom has positive links to student learning (Michaels et al. 2010). Yet, many teachers are ill-prepared to routinely create and sustain mathematically-rich and productive discourse in their classrooms (Weiss et al. 2003). These instructional skills are not easily developed and require extensive practice, coupled with timely feedback to support reflection and inform adjustments in instruction (Jacobs et al. 2014).However, missing from the current range of teacher learning tools are those that enable detailed and rapid feedback.

Michaels, O'Connor, Resnick and other colleagues have developed an approach to classroom discourse labeled "accountable" or "academically productive" talk (O'Connor, Michaels, and Chapin 2015). At the heart of accountable talk is the notion that teachers should organize classroom discussions that promote students' equitable participation in a rigorous learning environment. By utilizing specific talk moves, teachers can help ensure that the discussions will be purposeful, coherent, and productive (Michaels et al. 2010). Boston (Boston 2012) explains that accountable talk moves support classroom discourse to go beyond the traditional "Initiate, Response, Evaluate (IRE)" linguistic sequence (Mehan 1979). In particular, accountable talk seeks to replace the act of evaluating with practices that support a collaborative understanding that builds on and extends mathematical ideas (Michaels and O'Connor 2015). In this way, talk moves enable dialogue shifts from teacher-directed recitation to ”true discussions” in which knowledge is informally shared and constructed rather than transmitted (Cazden and others 2003).

Accountable talk moves have been defined and incorporated into a variety of tools for both researchers and practitioners, including the Instructional Quality Assessment toolkit (Boston and Wolf 2006) and the Accountable Talk Sourcebook (Michaels et al. 2010). In our study exploring the feasibility of a deep learning approach to reliably generate information about instructional talk moves, we incorporated 6 teacher talk moves from 3 categories:

1. Accountability to the learning community
   - Keeping everyone together (e.g. ”What did she just say?”)
   - Getting students to relate to another's ideas (e.g. ”Who agrees and who disagrees?”)
2. Ensuring purposeful coherent and productive group discussion
   - Restating (e.g. ”Let me say back what I heard.”)
   - Revoicing (e.g. ”Let me say back what I heard and add on.”)
3. Accountability to rigorous thinking
   - Pressing for accuracy (e.g. ”Can you tell us the steps you used to find the answer?”)
   - Pressing for reasoning (e.g. ”How are these ideas connected?”)

Accountable talk looks ”striking similar to the norms of discourse called for in theories of deliberative democracy” (Michaels, O'Connor, and Resnick 2008) (pg. 285). Specifically, accountable talk supports a discussion-based classroom community with the expectation that all students will have equal access to participation, subject matter content, and developing appropriate habits of mind (Michaels and O'Connor 2015). In a discursive classroom, an environment is constructed such that all students have the potential to contribute to the rational discourse, and that potential is nurtured and socialized (Michaels et al. 2010).

Forming and sustaining such a learning community has the potential to especially support girls and students from home backgrounds where risk-taking and modeling of similar talk patterns may be less common, inculturating them into the norms of democratic discourse that can later be realized in wider civic spheres (Michaels, O'Connor, and Resnick 2008). Shifting away from traditional IRE linguistic patterns towards accountable talk makes space for students' contributions, especially for English Language Learners (ELLs), by encouraging a focus on communicating mathematically and presenting arguments rather than acquiring vocabulary and other low-level linguistic skills (Moschkovich 2002). Furthermore, increased participation by ELLs and students from non-dominant groups can foster dispositions that attend to competencies and resources rather than deficits and obstacles (Moschkovich 2002).

A central premise of our research is that personalized, automated feedback can dramatically enhance teacher learning and support improvements in their instruction. Preliminary evidence indicates that teachers who receive automated feedback regarding the proportion of teacher to student talk in their mathematics lessons significantly increased the relative amount of student talk (Wang, Miller, and Cortina, 2013), suggesting that even basic information about teachers' own classroom discourse patterns can produce changes in the desired directions. Providing educators with longitudinal information about their instructional practices for two years improved students' math achievement after the first year by about four weeks of learning (Wayne et al. 2018).

Method

Data

The data used for training the talk moves classification model were sourced from multiple, existing professional learning resources for math teachers, which include entire lessons and short excerpts from lessons. At present, the data includes 100,683 sentences from 406 lesson transcripts, of which 60,241 are teacher sentences and 40,442 are student sentences. The data (i.e. text transcripts) were coded by two human annotators, applying a set of six, mutually exclusive teacher talk moves, adapted from the Accountable Talk framework (Michaels, O'Connor, and Resnick 2008). The inter-rater agreement/reliability score for each talk move is summarized in table 1. The high levels of inter-rater reliability suggest that machine learning models should be capable of learning to discriminate between these different labels.

Of the teacher sentences, 54.49% contain one of the six talk moves. The number of teacher sentences annotated including a talk move is shown in figure 1. As the figure indicates, the data have an uneven distribution pattern, with more instances of ”Keeping everyone together” and ”Pressing for accuracy” when compared with the other talk moves. The skewed nature of the data suggests there could be significant challenges in developing an accurate automated classification model. Imbalanced learning is still an open challenge in the field of machine learning and deep learning (Krawczyk 2016).

Model architecture

This section describes in detail the deep learning model trained on manually annotated transcripts of classroom lessons. The aim of the model is to automate the annotation and, when incorporated in the TalkBack application, to develop actionable recommendations for changes in teaching practice. The model classifies each teacher sentence, while the student sentences are used as context. If the teacher sentence does not contain a talk move it is classified under the ”None” label.

The text transcripts consist of student-teacher conversations which are segmented into sentences. As the preprocessing step, we converted the sentences into ”turns”. Each turn is comprised of a student utterance followed by

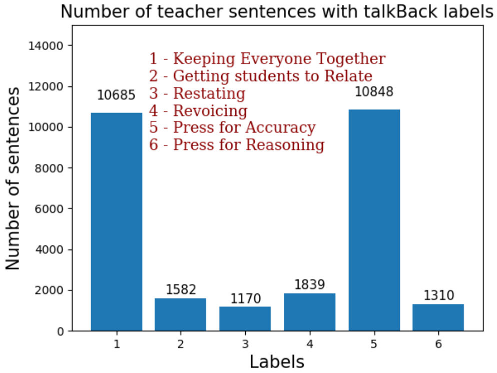

Figure 1: Histogram of teacher sentences containing talk moves.

Table 1: Inter-rater agreement for coding talk moves

Coding Decision	Inter-rater agreement
Sentence containing a talk move (or not)	95%
Talk move label	90%
Keeping everyone together	90%
Getting students to relate	86%
Restating	88%
Revoicing	87%
Press for accuracy	89%
Press for reasoning	91%

a teacher utterance. This organization is important in providing the model with the context required to learn talk moves such as ”Restating” and ”Revoicing”, which have to do with the teacher’s uptake of a student utterance. Moreover, a given teacher or student turn can include multiple sentences. For example, Table 2 provides an example of a turn with 4 sentences. The sentences are stripped of punctuation and case information (uppercase or lower case). We did not remove stop words because we anticipated that certain stop words (such as ”what”) would be crucial in discriminating between the different talk moves.

The model takes a set of feature vectors as inputs and produces a softmax output. The next step involves converting the ”turns” into corresponding feature vectors. Features can be defined as distinctive attributes of a sentence that the model can use to discriminate one type of sentence from another. Some of the successful features used in the domain of natural language processing are word embeddings and sentence embeddings (Palangi et al. 2016). Embedding is the process of representing the word/sentence as a vector in higher dimensional space. Without the trappings of dimensionality, the distance between two vectors in higher dimensional space can be representative of the semantic difference (for example) between those words/sentences. So similar vectors can be grouped together and have similar properties. These embeddings have been incorporated in various tasks such as semantic relatedness task and have been documented to achieve up to 89% accuracy (Conneau et al. 2017). In a nutshell, this step enables the conversion of all sentences into numbers for the model to process.

Table 2: Example Turn of 4 sentences with punctuation removed

student:	so you put the eight on the box
student:	then you get eight
teacher:	oh so were you using this side to help you get that side
teacher:	let me see if i can figure out what you said

At present, the talk moves classifier includes four different features:

- Sentence embedding - We used state-of-the-art sentenced embedding from (Conneau et al. 2017) trained on the ALLNLI corpus. Each sentence is represented as a 600 dimensional vector.
- GloVe BOW embedding - GloVe Bag of word representation. We took the average of the Glove vector normalized over all the words in a given sentence. Each sentence is represented as a 300 dimensional BOW embedding. The GloVe embeddings are not updated during training.
- CountVectorizer - In addition to the above features we implemented a CountVectorizer to keep track of the number of words in each sentence. It is represented as a 100 dimensional vector which includes the top 100 words in the corpus ordered by term frequency. The CountVectorizer includes stop words.
- Role - This is a simple feature that indicates whether the sentence is a student sentence or a teacher sentence. Although we focus on identifying the talk moves for teachers we provide the model with sentences spoken by students as context to help it discriminate between students and teachers.

The deep BI-LSTM model consists of five layers followed by a softmax layer (see figure 2). The first layer merges all of the features into a single vector. Handing the responsibility of merging features at the Graphical Processing unit (GPU) level is significantly better in terms of resources and computation time when compared to Central processing unit (CPU). The second layer is a Bidirectional Long Short-Term memory layer, which is made up of LSTM units. A LSTM unit is composed of a cell, input gate, output gate and forget gate. The gates can be compared with conventional artificial neurons and the cell handles all the ”remembering”. This layer is followed by a time distributed dense layer which is a one to one mapping of the input to output layer. This layer allows for the application of the activation function across every output over time. The time-distributed layer is followed by two dense layers of width ”3015” and ReLU or Rectified linear unit. ReLU is the max function (x,0) with input x. The two major advantages of this unit is sparsity and reduced likelihood of vanishing gradient. All the layers are batch normalized. The final step is a 7-way softmax unit which produces a probability distribution over the talk moves. The output of this layer can be also interpreted as the likelihood of each sentence to be classified as either ”None” or a particular talk move. The output of softmax will enable us to identify the talk move associated with every teacher sentence in the dataset. The model architecture was developed based on previous research on deep learning models to investigate the SNLI corpus (Conneau et al. 2017). The proposed deep learning architecture is a baseline model to classify teachers’ sentences from classroom discourse data.

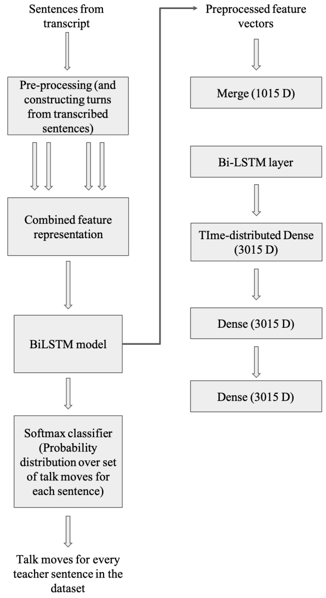
Sentences from transcript
Pre-processing (and constructing turns from transcribed sentences)
Combined feature representation
BiLSTM model
Softmax classifier (Probability distribution over set of talk moves for each sentence)
Talk moves for every teacher sentence in the dataset
Preprocessed feature vectors
Merge (1015 D)
Bi-LSTM layer
TIme-distributed Dense (3015 D)
Dense (3015 D)
Dense (3015 D)

Figure 2: Layer architecture. D represents the width of the layer

Implementation

The model was implemented using the Keras library in Python 2.7. There are different types of parameters involved in the model, and the parameter selection procedure will be discussed in the next section. The maximum number of sentences in a turn was capped at 30 sentences per turn. We used batch training with 128 turns per batch. Each epoch involved multiple batches based on the available data. We ran the model for 30 epochs on the training data before running it on the test data. We used ”RMSProp” optimizer with ”Categorical cross-entropy” as the loss function. The dropout rate was 0.5, the recurrent dropout was 0.5, and learning rate was 0.0001. The width of the intermediate dense layers was 3015. Changing the width of the layer in the range of 1015-4015 did not affect the performance of the model.

Results

After all the pre-processing steps (e.g., converting the transcripts into turns), the dataset was divided into training, validation and testing sets with an 80/10/10 split. i.e. 80% of the data was used for training and rest for validation and test data respectively. The validation set was used for parameter selection i.e. hypertuning.The turns/sentences in the test set were not involved in training to avoid the over-fitting problem. The model performance on the test is a true indicator of what the model has learned.

Metrics

We used an F1 measure as the performance metric. F1 is the harmonic mean of precision and recall and calculated as:

$$F_1 = 2 * \frac{precision*recall}{precision+recall}$$

where precision is the fraction of retrieved talk moves that are relevant to the sentence and recall is the fraction of the talk moves that are successfully retrieved. The F1 score was calculated only across the sentences with talk move labels (micro averaged). Teacher sentences with ”None” labels and student sentences were not considered when calculating the F1 score. On the test set, the model produced an overall F1 score of 65%. The performance graph is shown in figure 3. After a few epochs of training, the performance of the model on the training data (represented by dotted line) reached an F1 measure close to 1.0 while the performance on validation set (represented by solid line) was similar to the performance on the test set.

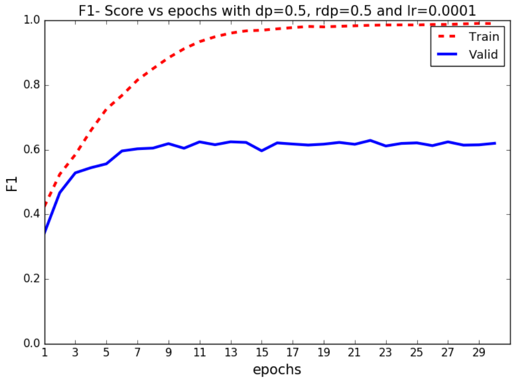

Figure 3: Model performance on test set

Parameters selection

Because GPU-based training benefits from fixed size batches, we needed to decide on a single turn length to use. After examining the statistics of the number of sentences in a turn, we found that although most turns had just two sentences, there were also sentences which were a part of long paragraphs. Based on the analysis, we capped the number of sentences in a turn to be 30. If a turn had fewer sentences than 30 it was pre-padded with zeros and likewise, if it was more than 30, it was truncated. In order to tune the hyperparameters of the model such as loss function, learning rate, dropout rate and recurrent dropout, we performed tests on exhaustive combinations of these parameters. For example, figure 4 represents an example run where the dropout rate was 90% while other parameters were fixed. In addition to training and test, we also plotted the performance of the model on the test set for every epoch to ensure that the valid curve and test curve follow similar performance levels. With 90% dropout, the model had a hard time learning the data.

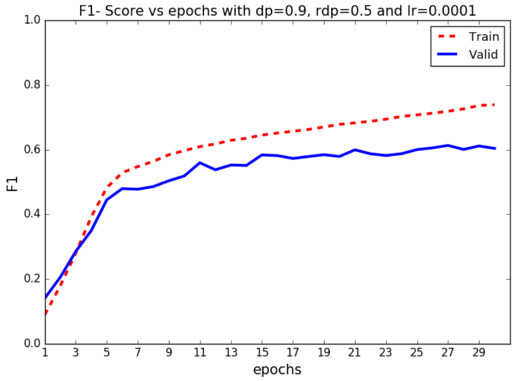

Figure 4: Model performance on test set - Parameter selection

Further justification for the use of BiLSTM is provided in Figure 5, which compares the performance (F1 score) of the model across different types of recurrent layers, including LSTM (Long Short Term Memory Unit), GRU (gated recurrent unit), and RNN (recurrent neural network). Recurrent neural networks are typically used in scenarios which involve sequence processing. In our work, sequential context information provides valuable insights that support the prediction of talk moves associated with the teachers' sentences. To validate BiLSTM as our choice of a recurrent layer, we compared this layer with three other popular recurrent layers. For each type of recurrent layer, the model was retrained and cross-validated. As expected, the BiLSTM model outperformed the others, suggesting it is the optimal recurrent layer of choice for talk move classification.

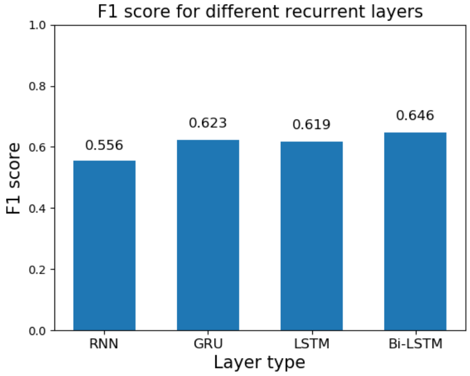

Figure 5: Choice of recurrent layer type

Error Analysis

In addition to the overall F1 score, we calculated the F1 score for individual talk moves. Table 3 shows a confusion matrix that describes the performance of the model on a test set with 4,352 teacher sentences. Numbers 1 though 6 correspond to the various talk moves while number 0 corresponds teacher sentences without a talk move. The columns indicate the number of teacher sentences with predicted talk moves while the rows indicate the actual talk moves. The additional columns show the Precision, Recall and F1-measures of the individual talk moves. As the confusion matrix indicates, Press for Accuracy (number 5) performs well, with 585 sentences correctly identified as this talk move. Meanwhile, Revoicing (number 4) was incorrectly identified as a sentence not containing a talk move for 55 sentences, compared to 43 instances where it was classified correctly.

Next we performed an error analysis to look more closely at the incorrectly classified sentences for each individual talk move. Some of the sentences incorrectly classified as "None" instead of "Keeping Everyone together" were single word sentences such as "okay" and "yes". In addition, we noticed that the model was not performing well for the "Restating" talk move. Most instances of "Restating" occur when the teacher is essentially repeating a prior student utterance. In order to address the above issues, we included the CountVectorizer as one of the input features. We realized that both the word embeddings and sentence embeddings we used did not directly carry information about the number of words in a sentence. Adding the CountVectorizer resulted in a significant boost in performance (3% in overall F1 measure) of the model. Also, we observed that the talk moves "Keeping everyone together" and "Press for accuracy" have a good deal of overlap in some of their language structures, as shown in Table 3. This observation suggests the annotation protocol may need to be revised in order for these talk moves to be more clearly distinguished by the model.

Discussion

The classification performance across each of the talk moves exceeds the majority class baseline (19%) by 19-55%. The majority class baseline is the percentage of time a prediction would be correct if we always chose the most frequently occurring class, which in our case is Keeping Everyone Together. The Press for Accuracy and Press for Reasoning labels are already performing at the level of well-trained human coders (F1s $>$ 0.70). The performance on the Revoicing label is surprisingly poor (F = 0.38), as word and sentence embedding features should be very effective at this type of paraphrase detection task. Overall, however, the level of model performance is very encouraging, particularly considering that the amount of training data (100k sentences) used in this study is relatively modest by deep learning standards and, to date, we included only a few basic features to represent our data.

At present, there is a dearth of research literature using deep learning models on classroom data. However, we believe it is fair to compare the performance of our model with other language models that make use of classroom discourse data such as (Donnelly et al. 2017) and (D'Mello et al. 2015). Performance achieved by our model F1 measure of 65% is very similar to the performance of the language models in (Donnelly et al. 2017) and (D'Mello et al. 2015).Taken

Table 3: Confusion matrix showing predicted and actual numbers of sentences with each talk move.

		Predicted										
		0	1	2	3	4	5	6	Total	Precision	Recall	F1
Actual	0-None	2085	112	2	13	19	85	4	2320			
	1-Keeping Everyone Together	255	446	16	17	17	64	4	819	0.69	0.55	0.61
	2-Getting students to relate	21	17	58	0	0	18	2	116	0.66	0.50	0.57
	3-Restating	17	3	0	38	11	0	0	69	0.46	0.55	0.50
	4-Revoicing	55	19	0	11	43	5	0	133	0.45	0.32	0.38
	5-Pressing for accuracy	145	49	11	4	6	585	4	804	0.75	0.73	0.74
	6-Press for reasoning	12	0	1	0	0	21	57	91	0.80	0.63	0.71
	Total	2590	646	88	83	96	778	71				

together, these studies illustrate the potential along with the challenges of inferring patterns based on classroom data.

In our future work we plan to utilize an expanded set of input data for the model based on human annotations of 1) additional categories of teacher talk moves and 2) student talk moves. We conducted a series of experiments to assess the impact of additional data on model performance, where we trained the model with different proportions of the data to gauge how performance changed with each addition. The results for which is summarized in figure 6. Each bar represents the number of sentences and turns used in a training set while the height of the bar represents the F1 measure calculated on the test set. The test set was the same throughout different trials in order to make a fair comparison on the impact of the number of sentences used for training. The lighter bars indicate the available data and the darker bars indicate the projected data. We can observe a linear increase in performance corresponding to an increase in training data, which suggests that the model performance is clearly expected to improve with additional data. More data is required in order to estimate the performance saturation point of the proposed model.

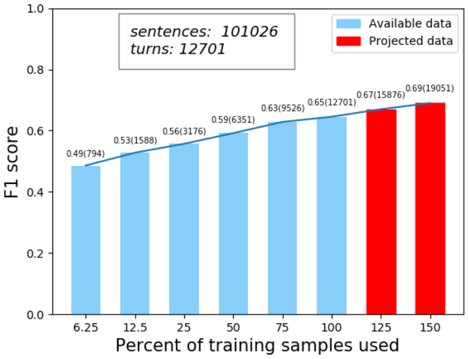

Figure 6: Projected model performance for different proportions of training samples

Conclusion and Future work

Orchestrating instructional conversations is an ”important and universally recognized dimension of teaching”, and prior research has established strong linkages between productive classroom discourse and student achievement (Correnti et al. 2015). Currently, providing teachers with detailed feedback about their discursive strategies requires highly trained observers to hand code transcripts and analyze moves (e.g., (Correnti 2005) and/or one-on-one expert coaching, a time-consuming and expensive process (Robertson, Ford-Connors, and Paratore 2014). Our goal is to develop an application, TalkBack, that will automate this process, and bring feedback on these instructional practices to scale, making them immediately accessible to teachers.

In our study, we trained bi-directional long short-term memory networks incorporating only simple features, such as bag-of-words and sentence embedding, to recognize six types of talk moves used by teachers with an average F1 of 65%, representing up to 74% gain over the class baseline. To train and validate the models, we created a large corpus of annotated transcripts of teacher-student interactions in a wide variety of mathematics learning environments, from small group instruction to K12 classrooms. We achieved an average 90% inter-rater reliability between human experts annotating these data, suggesting that there is an opportunity for significant improvement in model performance.

A central premise of our research is that personalized, automated feedback can dramatically enhance teacher professional learning and support improvements in their instruction. we plan to incorporate the talk moves classifier combining talk move processing, storage, and analytics within a web application. We envision end users (i.e., teachers) uploading or streaming classroom videos through the web application, which will then extract student and teacher sentences from the video and converting them into feature vectors which are readily consumable by the deep learning model. Information from the TalkBack application described above can help teachers understand how their use of talk moves and other patterns of discussion in their classrooms are changing over time. Additionally, the application will allow teachers to view selected video clips, enabling reflection on their own or others’ practice. This type of experience may support teachers to refine their instruction, providing a more inclusive classroom community in which students engage in productive classroom discussions around challenging content.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1600325: Early Stage Research on Automatically Identifying Instructional Moves in Mathematics.

References

Boston, M., and Wolf, M. K. 2006. Assessing academic rigor in mathematics instruction: The development of the instructional quality assessment toolkit. cse technical report 672. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Boston, M. 2012. Assessing instructional quality in mathematics. *The Elementary School Journal* 113(1):76–104.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Cazden, C. B., et al. 2003. Classroom discourse: Courtney b. cazden and sarah w. beck. In *Handbook of discourse processes*. Routledge. 170–202.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Correnti, R.; Stein, M. K.; Smith, M. S.; Scherrer, J.; McKeown, M.; Greeno, J.; and Ashley, K. 2015. Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. *Socializing Intelligence Through Academic Talk and Dialogue. AERA* 284.

Correnti, R. J. 2005. *Literacy instruction in CSR schools: Consequences of design specification on teacher practice.* Ph.D. Dissertation.

D'Mello, S. K.; Olney, A. M.; Blanchard, N.; Samei, B.; Sun, X.; Ward, B.; and Kelly, S. 2015. 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*, 557–566. ACM.

Donnelly, P. J.; Blanchard, N.; Samei, B.; Olney, A. M.; Sun, X.; Ward, B.; Kelly, S.; Nystran, M.; and D'Mello, S. K. 2016. Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 conference on user modeling adaptation and personalization*, 45–53. ACM.

Donnelly, P. J.; Blanchard, N.; Olney, A. M.; Kelly, S.; Nystrand, M.; and D'Mello, S. K. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 218–227. ACM.

Evans, C. W.; Leija, A. J.; and Falkner, T. R. 2001. *Math links: Teaching the NCTM 2000 standards through children's literature*. Libraries Unlimited.

Franke, M. L.; Turrou, A. C.; Webb, N. M.; Ing, M.; Wong, J.; Shin, N.; and Fernandez, C. 2015. Student engagement with others' mathematical ideas: The role of teacher invitation and support moves. *The Elementary School Journal* 116(1):126–148.

Hiebert, J.; Gallimore, R.; and Stigler, J. W. 2002. A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational researcher* 31(5):3–15.

Hochreiter, S. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6(02):107–116.

Jacobs, J.; Koellner, K.; John, T.; and King, C. D. 2014. The process of instructional change: Insights from the problem-solving cycle. In *Transforming Mathematics Instruction*. Springer. 335–354.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5(4):221–232.

Mehan, H. 1979. *Learning lessons*. Harvard University Press Cambridge, MA.

Michaels, S., and O'Connor, C. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue* 347–362.

Michaels, S.; O'Connor, M. C.; Hall, M. W.; and Resnick, L. B. 2010. Accountable talk® sourcebook. *Pittsburg, PA: Institute for Learning University of Pittsburgh. Murphy, PK, Wilkinson, IAG, Soter, AO, Hennessey, MN, & Alexander, JF.*

Michaels, S.; O'Connor, C.; and Resnick, L. B. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education* 27(4):283–297.

Mikolov, T., and Zweig, G. 2012. Context dependent recurrent neural network language model. *SLT* 12:234–239.

Moschkovich, J. 2002. A situated and sociocultural perspective on bilingual mathematics learners. *Mathematical thinking and learning* 4(2-3):189–212.

O'Connor, C.; Michaels, S.; and Chapin, S. 2015. Scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. *Socializing intelligence through academic talk and dialogue* 111–126.

Palangi, H.; Deng, L.; Shen, Y.; Gao, J.; He, X.; Chen, J.; Song, X.; and Ward, R. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24(4):694–707.

Robertson, D. A.; Ford-Connors, E.; and Paratore, J. R. 2014. Coaching teachers' talk during vocabulary and comprehension instruction. *Language Arts* 91(6):416–428.

Vygotsky, L. 1978. Interaction between learning and development. *Readings on the development of children* 23(3):34–41.

Wayne, A.; Garet, M.; Wellington, A.; and Chiang, H. 2018. Promoting educator effectiveness: The effects of two key strategies. ncee 2018-4009. *National Center for Education Evaluation and Regional Assistance*.

Weiss, I. R.; Pasley, J. D.; Smith, P. S.; Banilower, E. R.; and Heck, D. J. 2003. Looking inside the classroom. *Chapel Hill, NC: Horizon Research Inc.*

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.