Predicting Speech Acts in MOOC Forum Posts

Jaime Arguello
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
jarguello@unc.edu

Kyle Shaffer
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
shafferk@live.unc.edu

Abstract

Students in a Massive Open Online Course (MOOC) interact with each other and the course staff through online discussion forums. While discussion forums play a central role in MOOCs, they also pose a challenge for instructors. The large number of student posts makes it difficult for instructors to know where to intervene to answer questions, resolve issues, and provide feedback. In this work, we focus on automatically predicting *speech acts* in MOOC forum posts. Our speech act categories describe the purpose or function of the post in the ongoing discussion. Specifically, we address three main research questions. First, we investigate whether crowdsourced workers can reliably label MOOC forum posts using our speech act definitions. Second, we investigate whether our speech acts can help predict instructor interventions and assignment completion and performance. Finally, we investigate which types of features (derived from the post content, author, and surrounding context) are most effective for predicting our different speech act categories.

Introduction

A Massive Open Online Course (MOOC) is an online course designed for open, large-scale enrollment. Students progress through the course by watching pre-recorded lectures, completing assignments, and interacting with each other and the course staff via the MOOC's discussion forums. MOOCs are typically free of charge and bring together students from a wide range of countries, cultures, and socio-economic background, providing new opportunities for currently underserved populations and a unique educational environment for students and instructors.

While the MOOC model has enormous potential for narrowing the knowledge gap and facilitating life-long learning, managing a MOOC is a challenging task. The large number of student posts in MOOC discussion forums makes it difficult for an instructor to know where to intervene to answer questions, resolve issues, or provide feedback. Moreover, while MOOC forum posts provide evidence of student learning and motivation (Wen and Yang 2014a; Elouazizi 2014), the large volume of content makes it difficult for an instructor to identify students who may need help or encouragement. In a recent survey, 92 MOOC instructors reported that discussion forum activity is a valuable information source for identifying struggling students, evaluating the effectiveness of the course material and staff, and determining the overall appropriateness of the course difficulty level (Stephens-Martinez, Hearst, and Fox 2014).

In this work, we focus on automatically predicting *speech acts* in MOOC forum posts. In linguistics, a speech act describes the purpose or function of a sentence or utterance in a discourse. Similarly, our speech act categories describe the purpose or function of a post in a MOOC discussion thread. We focus on seven speech act categories: question, answer, issue, issue resolution, positive acknowledgment, negative acknowledgment, and other. Our speech act categories describe whether a post is asking or answering a question about the course material, raising or resolving an issue regarding the course management, or providing positive or negative feedback in response to a previous post. The other category is reserved for posts that serve a different purpose.

Our end goal is to use the speech acts present in a MOOC forum to help identify discussions that may require instructor attention and to help identify students that may need assistance. Specifically, in this paper, we focus on three main research questions (RQ1-RQ3). In our first research question (RQ1), we investigate whether non-expert, crowdsourced workers can reliably label MOOC forum posts using our speech act definitions. MOOCs cover different topics, and therefore predictive models trained on one MOOC may not generalize to another. To this end, we investigate whether gold-standard labels for training a model (possibly for a new MOOC) can be produced reliably, quickly, and inexpensively. Following prior work, we collected redundant labels from crowdsourced workers and combine them into gold-standard labels using a majority vote. We evaluate the quality of our labels by measuring the inter-annotator agreement between the crowdsourced majority vote and the labels from a trained expert on a large subset of our data.

In our second research question (RQ2), we investigate whether our MOOC speech acts can ultimately help instructors with the course management. To this end, we present three analyses. In the first analysis, we consider whether the speech acts in a thread can help predict an instructor intervention. In the second and third analyses, we focus on predicting student performance. In the second, we consider

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

whether the speech acts associated with a student's posts can help predict assignment completion and, in the third, whether they can help predict assignment performance. As a first step, these three analyses were conducted using our gold-standard (crowdsourced) labels, rather than our automatically predicted ones.

Finally, in our third research question (RQ3), we investigate which features are most predictive for different speech acts. To this end, we trained and tested independent binary classifiers to predict the presence of each speech act category in a post. We generated a wide range of features derived from the post content, the author, and its context in the thread. Certain speech acts are likely to appear in sequence (e.g., answers are likely to follow questions). We exploit the sequential correlations between speech acts by incorporating the predicted speech acts from the previous post as features. An extensive feature ablation analysis shows the marginal contribution of our features for different speech acts.

Related Work

Our work is informed by two branches of prior research: (1) predicting speech acts in different domains and (2) using MOOC forum data to predict instructor interventions and to predict student performance and drop-out.

Cohen *et al.* (2004) focused on predicting email speech acts that describe the email sender's intent (e.g., commit, request, deliver, propose) and used simple features such as unigrams, bigrams, POS tags, and the presence of dates and times. In a follow-up study, Carvalho and Cohen (2005) improved upon this simple approach by exploiting the sequential correlations between speech acts in the email thread (e.g., a *deliver* email tends to follow a *request*). Sequential correlations were exploited by including the predicted speech acts from the parent and child emails as input features. Qadir and Riloff (2011) focused on predicting speech acts in on-line discussion board posts in the medical domain. The analysis focused on predicting four speech acts at the sentence level. Interestingly, while their speech act definitions were not tailored to the medical domain, topical features (e.g., whether the post discusses symptoms, drugs, treatments, etc.) were found to be highly predictive. This result suggests that in certain domains, the topic of a message may provide information about its purpose in the conversation. Bhatia *et al.* (2012) focused on predicting speech acts in an online discussion board about database programming and used some of the same speech acts included in the current paper. The proposed model combined features derived from the post (e.g., unigram features), the user (e.g., number of previous posts) and the position of the post in the thread. Positional features were motivated by the observation that questions tend to appear higher in the thread and that answers tend to appear lower. All feature types improved performance.

Within the educational domain, Rus *et al.* (2012) used clustering and lexical similarity to discover speech act categories in logged chat room discussions from three educational online games. Manual annotation confirmed that chatroom exchanges with similar terms serve a similar function. Ravi and Kim (2007) used simple n-gram features to predict questions and answers in educational forum posts and then applied a set of heuristics to the predicted values to detect threads with unanswered questions. In a follow-up study with the same goal, Kim and Kang (2014) trained machine-learned models to predict unresolved threads using features generated from the sequence of predicted speech acts.

Beyond predicting speech acts in different domains, prior work has also analyzed MOOC forum data for different purposes, for example to detect threads that require instructor attention or to predict student performance and drop-out. Chatruvedi *et al.* (2014) focused on predicting instructor interventions retrospectively. The authors combined features from the entire thread (e.g., number of posts, mean time difference between posts) and the previous post (e.g., number of question words) to predict whether the current post was written by an instructor. Wen and Rosé (2014) found a correlation between sequences of student activities in a MOOC and their level of performance at the end of the course (excelled, passed, failed, dropped out). Elouazizi (2014) used linguistic cues to measure cognitive engagement in forum data from three different MOOCs and found evidence of low cognitive engagement.

Finally, prior work has also attempted to model and predict student drop-out. Wen *et al.* (2014b) found a correlation between the ratio of negative-to-positive sentiment words in MOOC forum posts and the rate of student drop-out in a given week. Similarly, Wen *et al.* (2014a) conducted a survival analysis and found that students with MOOC forum posts indicating higher levels of motivation and cognitive engagement were more likely to complete the course.

MOOC Forum Dataset

The MOOC forum dataset used in this paper originates from a MOOC on Metadata offered by our university in the Fall of 2013. The dataset includes all messages written by students and instructors that were not subsequently deleted. The MOOC covered topics such as principles of information organization, metadata schema development and evaluation, and specific metadata schemas used in different domains. The course was delivered using the Coursera platform and spanned a period of eight weeks from August to November 2013. The course had an initial enrollment of just over 27,000 students and an ending enrollment of just under 26,000 in the final week. While an enrollment of 26,000 students seems high, only 1,418 students completed enough course material to earn a Statement of Accomplishment. Prior research suggests that MOOCs have lower completion rates than other educational settings because many (if not most) students enroll without a serious intent to finish (Koller et al. 2013). For example, students may enroll to see what the course is about or to learn about MOOCs in general. That being said, in our view, a close to 5% completion rate seems low and suggests the need for tools to help instructors identify MOOC forum discussions that require their attention and tools to analyze MOOC forum data to measure student motivation and learning.

Students were evaluated on eight weekly assignments that included short-answer and coding segments. Each assign-

ment followed a weeklong learning module dedicated to a specific topic. Each module included video lectures recorded by the instructor along with selected readings. The instructor and one teaching assistant (jointly referred to as "instructors") were responsible for managing the course and responding to students through the online discussion forums.

Before presenting summary statistics, we provide some terminology. Our MOOC forum data consists of two types of messages: *post* and *comments*. Both types of messages are identical except that a comment is structurally tied to a post and, in many cases, is a direct response to the post it is attached to. Posts can have zero or more comments. A *forum* is the coarsest unit of analysis and is comprised of a collection of *threads*. A thread is comprised of a sequence of posts and comments. Since comments are tied to a post, the first message in a thread is always a post. In total, our dataset contains 15 forums, 425 threads, and 2,943 individual messages (2,166 posts and 777 comments). Of these messages, 2,754 were written by students and 189 were written by instructors. The difference between posts and comments is not critical to our work. Thus, we jointly refer to them as "posts".

Speech Act Definitions

Speech act theory arose from research in sociolinguistics and philosophy. Instead of focusing on meaning and structure alone, speech act theory characterizes sentences or utterances in terms of the purpose or function they serve in a discourse. Searle's early taxonomy of speech acts, for example, includes commisives, which commit the speaker to a future action, directives, which request the listener to do something, and expressives, which communicate the speaker's psychological state (Searle 1976).

Prior research has also proposed speech acts that are tailored to a specific domain and can be applied to whole messages, for example emails (Carvalho and Cohen 2005) or discussion forum posts (Kim and Kang 2014). Similarly, our speech acts characterize the purpose of a whole post within a MOOC discussion thread.

We focused on seven speech acts: (1) question, (2) answer, (3) issue, (4) issue resolution, (5) positive acknowledgment, (6) negative acknowledgment, and (7) other. Table 1 provides example statements indicating the presence of each speech act in a post. Our speech acts were defined as follows. A *question* is a request for information related to the course content. While many questions are posed in interrogative form (e.g., "What is the difference between these two concepts?"), they can also be posed as a statement (e.g., "I am unclear about the difference between these two concepts."). An *answer* is a response to a previous question and contributes information that is likely to be useful to the asker (even if it does not definitively answer the question). An *issue* communicates dissatisfaction with the course or a problem with the course management that may require the course staff to take corrective action. For example, an issue may communicate that an assignment question is vague or that the assignment submission system is down. While questions are in regards to the course material, issues are in regards to the course execution and logistics and are likely to be viewed negatively by an instructor. An *issue resolution* is a direct response to a previously raised issue and attempts to resolve it. A *positive acknowledgment* expresses a positive sentiment about a previous post. Positive sentiments include agreement, encouragement, and support. A *negative acknowledgment* expresses a negative sentiment about a previous post. Negative sentiments include disagreement, confusion, or frustration. Finally, the *other* category was reserved for posts that serve a function not captured by any of the other speech acts. These included non-English posts, student introductions, attempts to coordinate in-person study groups, and messages unrelated to the course.

Speech Act	Example
Question	"In Question 8 on the assignment I'm confused about the code formatting. In lectures, the instructor said syntax should be of the form X, but do you have to include Y? Any ideas what I'm doing wrong?"
Answer	"The answer here should follow the form of the practice problems. Hopefully that helps."
Issue	"The wording for Question 6 was confusing and ambiguous. Please consider revising the wording or giving students the points for this question."
Issue Res.	"We are aware of a glitch in our submission form for Homework 2. As a result, the last question has been awarded to each student as a free point."
Positive Ack.	"I'm glad I'm not the only one stuck on this! That was definitely confusing me too!"
Negative Ack.	"The last question may have been difficult, but part of learning new material is working at it. No sense in complaining."
Other	"Hi everyone! I'm a web designer and extremely interested in this course!"

Table 1: Speech Act Examples

Crowdsourced Annotation

Gold-standard speech act labels were collected using Amazon's Mechanical Turk (MTurk). Each MTurk Human Intelligence Task (HIT) asked the MTurk worker to select all the applicable speech act categories for a single post. Posts were displayed within the context of the thread. However, to avoid overwhelming workers, we only included previously written posts. In other words, the post to be labeled was always the last one in the thread and was marked using a 5-pixel-wide gray border. The HIT displayed the post/thread inside a scrollable HTML frame and seven checkboxes displayed next to the frame (one per speech act category). In order to motivate workers to reflect on their choices, we asked them to provide a brief justification for their annotation. Workers were instructed to select at least one checkbox per post and were not allowed to leave the justification field blank.

MTurk workers were given the following instructions. First, we explained what a MOOC is and described the role of discussion forums in a MOOC. Then, we described that the goal in the HIT was to: "Annotate a post from a MOOC discussion thread based on the speech act(s) present in the post." Speech acts were explained as follows: "If you treat a MOOC discussion thread as a conversation, the speech act(s) associated with a post describe its function in the conversation. The speech act(s) associated with the post describe what the author is trying to achieve." We explained our seven speech act categories consistent with the previous section,

except that we included a few additional tips. For the question category, we indicated that questions can be in the form of a statement (e.g., "I need help with HW Question 3."). Furthermore, to help distinguish questions from issues, we explained that asking questions is part of a student's learning process and are not necessarily bad from an instructor's perspective. For the answer category, we indicated that answers may not completely resolve a question, but should provide information that is useful in some way. We also indicated that mere feedback about a previous question (e.g., "I have the same question!") should be labeled as positive or negative acknowledgment. For the issue category, we added that issues may require corrective action by the course staff and are likely to be considered bad from an instructor's perspective. Issues may refer to glitches in the course materials or logistics. For the issue resolution category, we added that issue resolutions may simply indicate that the course staff is aware of the problem and working on a solution. An issue resolution may not completely fix the problem. For the positive acknowledgment category, we added that positive sentiments may include agreement, encouragement, and support. Finally, for the negative acknowledgment category, we added that negative sentiments may include disagreement, confusion, and frustration.

Snow *et al.* (2008) evaluated the quality of crowdsourced labels across several computational linguistics tasks. Results found that combining as few as four redundant crowdsourced labels using a majority vote can produce labels comparable to an expert's. In a similar fashion, we collected five redundant annotations per post and combined them into gold-standard labels using a majority vote. While posts could be associated with multiple speech act categories, we decided to treat each speech act category independently. In this respect, a post was considered a gold-standard *positive* example for a particular speech act if at least 3/5 MTurk workers selected that speech act and was considered a *negative* example otherwise. In total, we collected 14,815 annotations (2,963 posts $\times$ 5 redundant HITs per post), and workers were compensated with $\$0.10$ USD per HIT.

Our HITs were implemented as *external* HITs, meaning that everything besides recruitment and compensation was managed by our own server. Using an external HIT design allowed us to control the assignment of posts to workers, preventing workers from seeing the same post more than once, and to detect and filter careless workers dynamically. MTurk annotation tasks require quality control, and we addressed this in four ways. First, we restricted our HITs to workers with a 95% acceptance rate or greater. Second, to help ensure English language proficiency, we restricted our HITs to workers in the U.S. Third, workers were exposed to several HITs for which an expert assessor (one of the authors) thought that the correct speech act was fairly obvious. Workers who disagreed with the expert on three of these HITs were automatically prevented from completing more HITs. Finally, in order to avoid having a few workers do most of our HITs, workers were not allowed to complete more than 165 HITs (about 1% of the total). Ultimately, we collected annotations from 360 unique workers.

In our first research question (RQ1), we investigate whether crowdsourced workers can reliably label our speech acts in MOOC forum posts. To answer this question, we measured the level of inter-annotator agreement between the MTurk majority vote and an expert assessor. To this end, an expert assessor (one of the authors) labeled a random sample of 1,000 posts (about a third of the full dataset) with respect to each speech act category. Then, for each speech act, we measured the Cohen's Kappa agreement between the MTurk majority vote and the expert. Cohen's Kappa (κ_c) measures the chance-corrected agreement between two annotators on the same set of data. Furthermore, in order to make a full comparison, we also measured the Fleiss' Kappa agreement between MTurk workers across all posts. Fleiss' Kappa (κ_f) measures the chance-corrected agreement between *any* pair of assessors and is therefore appropriate for measuring agreement between MTurk workers who were free to annotate any number of posts (up to a max of 165).

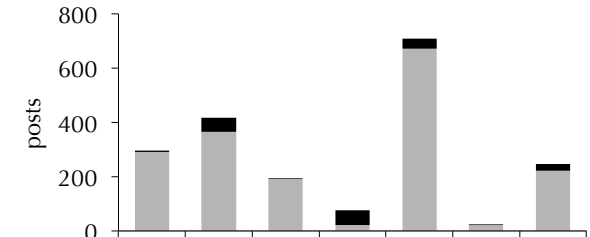
Agreement numbers are provided in Table 2. Two trends are worth noting. First, across all speech acts, the level of agreement between MTurk workers was lower than the level of agreement between the MTurk majority vote and the expert. This result is consistent with previous work (Snow et al. 2008) and suggests that combining redundant crowdsourced labels improves label quality. Second, agreement between the MTurk majority vote and the expert varied across speech acts. Agreement was "almost perfect" for questions ($\kappa_c > 0.80$), close to "almost perfect" for answers ($\kappa_c \approx 0.80$), and "substantial" for the other speech acts ($0.80 \geq \kappa_c > 0.60$) (Landis and Koch 1977). Overall, we view these results as encouraging, but with room for improvement.

The speech acts with the lowest agreement were issue resolution, negative acknowledgment, and other. As described in more detail below, issue resolutions and negative acknowledgments were fairly infrequent. Assessors may need further instructions and examples to reliably recognize these speech acts. The other category occurred more frequently, but was still associated with lower agreement. After examining the data, we found posts where MTurk workers were divided between other and positive acknowledgment. In many of these posts, the author's overall sentiment was positive (e.g., "Hi, I'm **** from ****. Nice to meet you all!"), but the post *did not* directly reference a previous post. Future work may need to provide further instructions to help distinguish between positive acknowledgment and other.

	MTurk Workers	**MV and Expert**
	κ_f	κ_c
Question	0.569	0.893
Answer	0.414	0.790
Issue	0.421	0.669
Issue Resolution	0.286	0.635
Positive Ack.	0.423	0.768
Negative Ack.	0.232	0.633
Other	0.337	0.625

Table 2: Agreement between MTurk workers (κ_f) and between the MTurk majority vote (MV) and the expert (κ_c).

Figure 1 shows the frequencies across speech acts based on our gold-standard, majority vote labels. For each speech act, we indicate the number of posts authored by students and instructors. The sum across all speech acts is less than the total number of posts because a subset of posts did not have a 3/5 majority vote with respect to *any* speech act. As can be seen from Figure 1, questions, answers, and positive acknowledgments were the most common, and instructors intervened most frequently to answer questions, resolve issues, and provide positive acknowledgment.



	Q	A	I	R	P	N	O
instructors	4	53	1	54	37	1	23
students	292	365	193	22	671	22	223

Figure 1: Speech Act Frequencies

Usefulness of Speech Acts

In our second research question (RQ2), we investigate whether our MOOC speech act categories can help guide instructors towards posts or threads that require their attention and whether our speech act categories can help predict student performance. Below, we present three different logistic regression analyses that explore the following questions: (1) Are speech acts helpful for identifying discussions that require instructor attention? (2) Are speech acts helpful for predicting whether a student will complete an assignment? and (3) Are speech acts helpful for predicting a student's grade on an assignment? Our goal was to justify the usefulness of predicting speech acts in MOOC forum posts. As a first step, these analyses were done using gold-standard speech act labels instead of predicted ones.

Predicting Instructor Interventions. This analysis was done retrospectively by trying to predict whether a post was written by an instructor based on the preceding speech acts in the thread. Logistic regression is appropriate for cases where we have a binary dependent variable (1 = post written by instructor, 0 = written by student) and several binary or real-valued independent variables. We considered the following 16 independent variables: 7 binary-valued variables indicating the speech act(s) in the previous post and 7 real-valued variables indicating the frequency of each speech act in all the previous posts in the thread. We also included the number of previous posts in the thread and, in order to model cases where an instructor already intervened, the number of previous posts in the thread written by an instructor.

The model as a whole was statistically significant ($\chi^2(16) = 145.401$, $p < .001$) and explained 12.7% of the variance based on Nagelkerke's R^2. The predictiveness of each independent variable is described in Table 3. The odds ratio measures the change in the probability that the dependent variable = 1 (the post was written by an instructor) per unit increase in the independent variable holding the other variables constant. Four variables (shown in bold) were found to be significant based on Wald's χ^2 test ($p < .05$). A post was 2.278 times more likely to be written by an instructor when the previous post contained an issue and 3.053 times more likely when the previous post contained a negative acknowledgment. Similarly, a post was 1.406 times more likely to be written by an instructor for each additional issue found in the previous posts in the thread. Interestingly, a post was 1.562 times more likely to be written by an instructor for each additional previous post in the thread from an instructor. This contradicts our initial intuition and suggests that certain threads were associated with *repeated* instructor interventions.

Input Variables	Odds Ratio	Wald's $\chi^2(1)$	p-value
Question (Prev)	1.391	1.345	.246
Answer (Prev)	1.566	3.361	.067
Issue (Prev)	**2.278**	**8.782**	**.003**
Resolution (Prev)	0.568	0.894	.345
Positive (Prev)	1.459	3.043	.081
Negative (Prev)	**3.053**	**3.884**	**.049**
Other (Prev)	2.001	2.992	.084
Question (Freq)	0.990	0.006	.938
Answer (Freq)	0.911	1.138	.286
Issue (Freq)	**1.406**	**6.497**	**.011**
Resolution (Freq)	0.876	0.349	.555
Positive (Freq)	0.930	0.768	.381
Negative (Freq)	0.685	2.843	.092
Other (Freq)	0.803	2.362	.124
NumPosts	0.981	0.081	.776
NumPostsInstruct.	**1.562**	**24.545**	**.000**

Table 3: Predicting Instructor Interventions. 'Prev' and 'Freq' denote the variables derived from the previous post and from all the preceding posts in the thread, respectively.

It is not surprising that instructors were more likely to intervene in response to negative acknowledgments and issues. We found several examples of instructors responding to disagreements (e.g., about grading) and to issues related to infrastructure (e.g., browser compatibility issues) or confusion about an assignment. We did not expect that instructors would be more likely to intervene in the presence of more posts in the thread by an instructor. After examining the data, we realized that certain threads were initiated by an instructor as a means to engage the students in discussion. These were threads containing several posts written by an instructor. In retrospect, these are threads where the instructor plays a different role and should be treated differently.

Predicting Assignment Completion. As previously mentioned, students completed a total of eight weekly assignments (denoted as a_1 to a_8). We performed a logistic regression analysis to predict whether a student completed assignment a_i (1 = completed, 0 = not completed). The independent variables were derived from two sets of posts: (1) all the posts written by the student prior to the a_i deadline and (2) only the posts written by the student between assignments $a_{(i-1)}$ and a_i. In this analysis, we only included students who contributed at least one post throughout the

course. Furthermore, we also attempted to filter students who had already dropped by the time of the assignment a_i deadline. To this end, for each assignment a_i, we excluded those students who did not complete assigned $a_{(i-1)}$ and did not complete *any* assignment a_j for $j > i$. The regression analysis considered the following 16 independent variables: 2 real-valued variables indicating the number of posts written by the student prior to a_i and between $a_{(i-1)}$ and a_i, and 14 real-valued variables indicating the frequency of each speech act in those two sets of posts.

Due to space limitations, for the next two analyses, we do not show the full logistic regression output table and focus only on the variables that were significant. As a whole, the model for predicting assignment completion was statistically significant ($\chi^2(16) = 402.206$, $p < .001$), although it only explained 4.2% of the variance (Nagelkerke's R^2). Two variables were found to be significant. Based on the odds ratio, a student was 1.364 times more likely to submit assignment a_i for every post written between $a_{(i-1)}$ and a_i (Wald's $\chi^2 = 8.709$, $p = .003$), and was 1.591 times more likely to submit a_i for every post with an *issue* written between $a_{(i-1)}$ and a_i (Wald's $\chi^2 = 5.243$, $p = .022$). We discuss this result below.

Predicting Assignment Performance. To be consistent with the previous two analyses, we cast assignment performance prediction as a binary task. For those students who completed assignment a_i, we performed a logistic regression analysis to predict whether the student's grade in the assignment was equal to or greater than the median grade or less (1 = greater than or equal to the median, 0 = less). Similar to the previous analysis, we only included students who contributed at least one post throughout the course, and used the same 16 independent variables derived from all of the student's previous posts and from only those written between $a_{(i-1)}$ and a_i.

Again, the model for predicting assignment performance was statistically significant ($\chi^2(16) = 137.940$, $p < .001$), but only explained 1.7% of the variance (Nagelkerke's R^2). Only one variable was found to be significant. Based on the odds ratio, a student who submitted assignment a_i was 1.426 times more likely to perform equal to or above the median for every post with an *issue* written between $a_{(i-1)}$ and a_i (Wald's $\chi^2 = 11.177$, $p = .001$).

While both models for predicting assignment completion and performance were significant, they explained very little of the variance (4.2% and 1.7%, respectively). Predicting assignment completion and performance may require additional measures about the student's background, level of prior knowledge, and motivations for enrolling in the MOOC. That being said, we were interested in why students raising issues were more likely to complete an assignment and perform above the median. After examining the data, we discovered that many of these issues were in regards to the assignment (e.g., finding a question to be vague, or not being able to access a resource needed to answer a question). These issues indicate that the student is committed to completing the assignment and performing well.

Predicting Speech Acts

We trained independent binary classifiers to predict each speech act category in a forum post. In all experiments, we used L2-regularized logistic regression classifiers implemented using the LibLinear toolkit. A logistic regression classifier learns to predict a binary target as a function of a set of features. We grouped features into different categories and evaluated their marginal contribution to performance separately for each speech act category. In total, we investigated 201 features. The numbers in parentheses indicate the number of features in each category.

LIWC Features. Several of our features were generated using Pennebaker's Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al. 2007). LIWC generates a number of different measures for an input span of text. LIWC measures include properties of the text (e.g., number of words, words per sentence) as well as the frequencies associated with words from different lists. LIWC word lists measure phenomena such as sentiment, affect, and references to different cognitive, perceptual, and social processes indicated by the author's language. We used a total of 60 LIWC features. We provide a general description of each LIWC feature category and refer the reader to the complete list of features available in Pennebaker (2007).

- **Affect Features (8):** Measure positive and negative sentiment, as well as more specific emotions such as assent, anger, anxiety, and sadness.
- **Cognitive Features (9):** Measure whether the author is expressing uncertainty, considering a causal relationship, or comparing and contrasting.
- **Personal Concern Features (9):** Measure references to topics of a personal nature (accomplishment, money, home, work, death, religion).
- **Linguistic Features (26):** Measure linguistic properties of the text (number of words, words per sentence, number of words appearing in a dictionary), as well as frequency counts for different types of pronouns (characterized by person and number), verb tenses (past, present, future), and other linguistic expressions (quantifiers, numbers, negations).
- **Perceptual Features (4):** Measure references to things being perceived by the senses (seeing, hearing, feeling).
- **Social Features (4):** Measure references to humans, friends, and family.
- **Spoken Features (3):** Measure nonfluencies ("er", "hm*", "um*") and fillers ("blah", "I mean").

Non-LIWC Features. In addition to LIWC features, we investigated a number of other features.

- **Sentiment Features (4):** The overall positive or negative sentiment in the post is likely to be predictive for positive/negative acknowledgment and issue. In addition to the LIWC affect features described above, we used the positive and negative term lists provided by Liu *et al.* (2005) (over 2,000 positive and 4,500 negative terms) to generate four features: the number and proportion of terms in the post from each list.
- **Unigram Features (100):** Prior work found that certain cue phrases are predictive for certain speech acts (Kim

and Kang 2014), and that in certain domains, the topic of a message can provide evidence of its function in the discourse (Qadir and Riloff 2011). To capture these types of evidence, we included unigram features. Specifically, we used the χ^2 statistic to measure the lack of independence between each unigram and each speech act category and generated 100 unigram features per speech act. Each speech act had its own set of unigram features (those with the highest χ^2 value) and each feature measured the unigram's frequency in the post.

- **Text Similarity Features (6):** Prior work found that thread starters tend to contain questions and that questions and answers share similar terms (Bhatia, Biyani, and Mitra 2012). To exploit this type of evidence, we included six features that measure the text similarity between the post and the surrounding context. We measured text similarity using the cosine similarity with stemming, TF.IDF term-weighting, and stop words included. We included 6 text similarity features: the similarity with the previous post; the similarity with the first post in the thread; and the min, max, mean, and variance of the similarity with all the previous posts in the thread.
- **Temporal Features (3):** Certain speech acts (e.g., questions, answers, and issues) may be more frequent closer to an assignment deadline. To model this type of evidence, we included three features indicating the time (in days, hours, and minutes) to the nearest assignment deadline.
- **Sequential Correlation Features (7):** Prior work found an improvement in performance by exploiting the sequential correlations between email speech acts (Carvalho and Cohen 2005). In our case, answers are likely to follow questions and resolutions are likely to follow issues. To exploit the sequential correlations between speech acts, we included the prediction confidence values for each speech act in the previous post. The confidence values were produce by classifiers that had access to all features except for these ones and were generated using cross-validation.
- **Author (1):** Instructor vs. student.
- **Link (1):** The number of hyperlinks in the post.
- **Modal (2):** The absolute and relative frequencies of modal verbs in the post.
- **Position (2):** The absolute and relative position of the post in the thread.
- **Post_comment(1):** Post vs. comment.
- **Punctuation (13):** The absolute and relative frequencies of different punctuations in the post.
- **Votes (1):** Post up-votes minus down-votes.

Prediction Evaluation Methodology

Training and testing was conducted using 20-fold cross-validation. We report average performance across held-out test sets. We used the same 20 folds in all experiments. Thus, when comparing between two approaches (two sets of features), we measured statistical significance using an approximation of Fischer's randomization test (Smucker, Allan, and Carterette 2007). This is a non-parametric test where the test statistic is the difference in means across a set of *paired* samples—in our case, paired held-out test sets.

Binary classification is often evaluated in terms of precision and recall. Measuring precision and recall requires setting a threshold on the classifier's prediction confidence value. For logistic regression, the default is 0.50. An evaluation based on a single precision/recall operating point shows an incomplete picture. Ultimately, depending on a user's preference, the prediction confidence threshold could be raised or lowered to favor precision over recall or vice-versa. For this reason, we decided to evaluate in terms of *average precision* (AP). AP measures the average of precision values for all prediction confidence threshold values where recall changes.

We explain average precision using the following notation. Let $\mathcal{D}$ denote the test set and $\mathcal{D}_s^+$ denote the subset of instances in $\mathcal{D}$ that are gold-standard positive examples for speech act s. Next, let ϕ_s denote a ranking of $\mathcal{D}$ in descending order of prediction confidence value with respect to s. The prediction confidence values correspond to the output of the logistic regression classifier. Furthermore, let $\phi_s(k)$ denote the instance at rank k in ϕ_s. Finally, let $\mathcal{P}(k)$ denote the precision at rank k—the percentage of top-k instances in ϕ_s that are in $\mathcal{D}_s^+$. Using this notation, average precision is computed as,

$$\text{AP} = \frac{1}{|D_s^+|} \sum_{k=1}^{|\mathcal{D}|} \mathcal{P}(k) \times \mathcal{I}(\phi_s(k) \in \mathcal{D}_s^+),$$

where $\mathcal{I}$ is the indicator function. AP is in the range [0,1] and is proportional to the area on the precision/recall curve.

Regularized logistic regression uses parameter C to control the misclassification cost on the training set. Parameter C was tuned using a second-level of cross-validation. For each first-level training/test set pair, we did 19-fold cross-validation on the training set and used the value of C with the greatest mean AP performance. Parameter C was tuned across values of 2^x where x =-5,-4,-3,-2,-1,0. All feature values were normalized to zero min and unit max. Normalization was done separately for each training/test set pair using the min and max values from the training set.

Speech Act Prediction Results

In our third research question (RQ3), we investigate which features are most predictive for different speech acts.

Feature Ablation Analysis. Results from our feature ablation study are presented in Table 4 in terms of average precision (AP). We were interested in isolating the contribution of our sequential correlation (SC) features in the analysis. Thus, only the model in the last row used SC features. The model in the first row had access to all features except for SC features. The models for the middle rows used all non-SC features except for those in that row's feature group. The model in the last row used all features including SC features. The percentages indicate the percent change in performance compared to the model in the first row. Likewise, the symbols ▼/▲ denote a statistically significant decrease/increase in performance compared to the model in the first row. We used two-tailed tests and set $\alpha = 0.05$. In the middle rows, a large performance *decrease* indicates a large positive contribution for the feature group in that row. In the last row,

	question	answer	issue	issue resolution	positive ack.	negative ack.	other
all features	0.747	0.604	0.643	0.555	0.809	0.160	0.612
-affective	0.747 (0.00%)	0.590 (-2.32%)▼	0.631 (-1.87%)	0.575 (3.60%)▲	0.809 (0.00%)	0.152 (-5.00%)	0.613 (0.16%)
-author	0.746 (-0.13%)	0.605 (0.17%)▲	0.639 (-0.62%)	0.457 (-17.66%)▼	0.807 (-0.25%)	0.147 (-8.13%)▼	0.615 (0.49%)
-cognitive	0.753 (0.80%)	0.604 (0.00%)	0.648 (0.78%)	0.580 (4.50%)	0.811 (0.25%)	**0.171 (6.88%)**▲	0.620 (1.31%)
-cosine	0.737 (-1.34%)	0.589 (-2.48%)	0.634 (-1.40%)	0.570 (2.70%)	0.808 (-0.12%)	0.164 (2.50%)	0.605 (-1.14%)
-linguistic	0.728 (-2.54%)▼	0.594 (-1.66%)	0.644 (0.16%)	0.523 (-5.77%)	0.804 (-0.62%)▼	0.147 (-8.13%)	0.603 (-1.47%)
-links	0.746 (-0.13%)	0.604 (0.00%)	0.643 (0.00%)	0.560 (0.90%)	0.809 (0.00%)	0.161 (0.63%)	0.612 (0.00%)
-modal	0.745 (-0.27%)	0.605 (0.17%)	0.640 (-0.47%)	0.575 (3.60%)▲	0.810 (0.12%)	0.160 (0.00%)	0.613 (0.16%)
-perceptual	0.746 (-0.13%)	0.605 (0.17%)	0.638 (-0.78%)	0.562 (1.26%)	0.810 (0.12%)	0.159 (-0.63%)	0.611 (-0.16%)
-personal_concerns	0.746 (-0.13%)	0.607 (0.50%)	0.640 (-0.47%)	0.550 (-0.90%)	0.810 (0.12%)	0.162 (1.25%)	0.609 (-0.49%)
-temporal	0.748 (0.13%)	0.605 (0.17%)	0.643 (0.00%)	0.592 (6.67%)▲	0.811 (0.25%)	0.166 (3.75%)	0.612 (0.00%)
-position	0.737 (-1.34%)▼	0.604 (0.00%)	0.645 (0.31%)	0.563 (1.44%)	0.811 (0.25%)	0.154 (-3.75%)	0.614 (0.33%)
-post_comment	0.747 (0.00%)	0.600 (-0.66%)	0.642 (-0.16%)	0.550 (-0.90%)	0.807 (-0.25%)	0.138 (-13.75%)	0.608 (-0.65%)
-punctuation	0.561 (-24.90%)▼	0.584 (-3.31%)▼	0.643 (0.00%)	0.555 (0.00%)	0.804 (-0.62%)▼	0.166 (3.75%)	0.613 (0.16%)
-sentiment	0.746 (-0.13%)	0.604 (0.00%)	0.641 (-0.31%)	0.572 (3.06%)	0.806 (-0.37%)▼	0.161 (0.63%)	0.616 (0.65%)
-social	0.748 (0.13%)	0.604 (0.00%)	0.641 (-0.31%)	0.569 (2.52%)	0.809 (0.00%)	0.159 (-0.63%)	0.614 (0.33%)
-spoken	0.745 (-0.27%)	0.605 (0.17%)▲	0.640 (-0.47%)	0.563 (1.44%)	0.811 (0.25%)	0.157 (-1.88%)	0.615 (0.49%)
-unigram	0.685 (-8.30%)▼	0.547 (-9.44%)▼	0.460 (-28.46%)▼	0.426 (-23.24%)▼	0.749 (-7.42%)▼	0.090 (-43.75%)▼	0.506 (-17.32%)▼
-votes	0.746 (-0.13%)	0.604 (0.00%)	0.643 (0.00%)	0.560 (0.90%)	0.809 (0.00%)	0.161 (0.63%)	0.611 (-0.16%)
+sequential corr.	**0.764 (2.28%)**	**0.654 (8.28%)**▲	**0.657 (2.18%)**	**0.596 (7.39%)**	**0.816 (0.87%)**	0.153 (-4.38%)	**0.664 (8.50%)**▲

Table 4: Feature Ablation Study Results.

a large performance *increase* indicates a large positive contribution from our SC features. In several cases, differences in performance were statistically significant, but small. We focus our discussion on differences that were statistically significant *and* had at least a 5% difference in performance (highlighted in gray). Our assumption is that smaller differences are less likely to be noticed, even if significant. The best performing model for each speech act is marked in bold.

The results in Table 4 reveal several important trends. First, we focus on performance across speech acts by comparing the best performing model for each speech act (marked in bold). Performance varied widely across speech acts and was generally consistent with the level of agreement between the MTurk majority vote and the expert assessor (Table 2). Questions had the second highest performance and also the highest level of agreement. This suggests that questions were fairly easy to recognize (by MTurk workers and our models) using a relatively small number of features (e.g., punctuation and question words). On the other hand, issue resolutions and negative acknowledgments had the lowest performance and also the lowest level of agreement. These speech acts were the most infrequent (Figure 1). In total, there were only 76 issue resolutions and 23 negative acknowledgments in our data. More positive examples may be needed to improve performance for these speech acts.

Next, we focus on the contribution from different features. The second important trend is that unigram features were highly effective and improved performance for all speech acts. To gain more insight, Table 5 shows the top 20 unigrams with the greatest χ^2 value for each speech act. The table displays term-stems because we used stemming when generating our unigram features.

Interestingly, unigrams captured different phenomena for different speech acts. For questions, the top unigrams include question words (how, what, where, why), modal verbs (can, would), and terms about confusion or gaps in knowledge (does, know, question, thought, wonder). For answers, the top unigrams include terms related to the course topic of metadata (content, element, format, meta, object, tag, scheme, name), evidence of hyperlinks to external content (href, http, org), and words related to explanation (example, define, describe, depend, mean). For issues, the top unigrams include terms related to problems/errors (error, issue, mistake, omit, problem, typo), infrastructure (browser, chrome, firefox, window), and course material (answer, homework, question, quiz). For issue resolutions, the top unigrams include words about apologizing (apology, sincerely, sorry), fixing problems (fix, remove, resolve, update), explaining errors (apparently, cause, encountered, instead), and grading (credit, extra, score). For positive and negative acknowledgment, the top unigrams generally relate to positive sentiment (agree, great, thank) and negative sentiment (challenge, disagree, disappoint, frustrated, hate). For the other category, many of the top unigrams are Spanish words, as non-English posts were included in the other category.

The third important trend is that, beyond unigram features, different features were highly predictive for different speech acts. The author feature (student vs. instructor) was highly predictive for issue resolution and negative acknowledgment. As shown in Figure 1, most issue resolutions were written by instructors and most negative acknowledgments were written by students. Not surprisingly, punctuation features were highly predictive for questions. Finally, our sequential correlation features were highly predictive for answers and the other category. For answers, this is expected—answers tend to follow questions. For the other category, as it turns out, posts labeled as other tended to follow *other* posts labeled as other. The other category included non-English posts and student introductions. Non-English posts tended to follow each other (students communicating in their native language) and in our data there was a large thread during the first week of the course where students wrote posts introducing themselves to the class.

A few features had a significant negative effect on performance. Cognitive features hurt performance for negative acknowledgments and temporal features hurt performance for issue resolution. We believe this was due to data sparseness. In this analysis, we considered 201 features and both of these speech acts had fewer than 80 positive instances each. We return to this point below.

Feature Ablation Analysis (Excluding Unigrams). Unigram features captured different phenomena for different speech acts (Table 5) and this likely attenuated the contribution of some of our other features. Moreover, a model using unigram features may be less likely to generalize to a new MOOC on a different topic. For example, some of the un-

question	answer	issue	issue_resolution	positive_ack	negative_ack	other
anyon (96.6)	thank (171.8)	strong (350.8)	apolog (161.4)	thank (465.6)	disagre (35.2)	que (300.3)
doe (72.5)	exampl (83.9)	typo (324.6)	resolv (161.2)	cathi (110.1)	frustrat (35.0)	para (230.6)
what (61.9)	href (77.0)	factual (299.7)	fix (104.8)	agre (88.8)	anonym (32.5)	por (163.9)
how (38.5)	depend (71.9)	browser (293.1)	caus (100.2)	http (80.0)	less (27.1)	gracia (143.8)
anybodi (36.3)	describ (68.1)	omit (290.3)	sorri (76.8)	href (69.0)	code (25.8)	con (126.4)
ani (35.5)	http (65.3)	error (266.4)	now (67.3)	great (65.4)	necessarili (24.5)	pero (125.2)
can (32.3)	target (61.6)	problem (261.3)	remov (59.5)	target (58.0)	exercis (23.0)	todo (125.2)
question (29.3)	object (60.3)	detail (253.5)	001 (51.8)	org (46.7)	date (22.2)	curso (125.2)
why (24.2)	meta (58.4)	addit (236.9)	updat (49.3)	can (43.3)	white (21.7)	del (114.4)
would (20.4)	tag (57.1)	window (221.7)	homework (45.9)	not (43.0)	hate (21.6)	the (107.1)
take (20.3)	scheme (55.8)	chrome (210.9)	been (34.6)	pomerantz (42.8)	not (20.0)	como (105.8)
recip (15.9)	org (55.2)	mistak (206.0)	encount (28.5)	which (40.3)	turn (19.3)	mucha (104.8)
note (15.9)	content (53.8)	firefox (178.0)	score (27.5)	page (40.1)	consider (19.3)	est (85.7)
where (15.3)	mean (51.9)	homework (155.3)	appar (26.7)	question (39.3)	disappoint (19.3)	reput (67.6)
wonder (15.2)	element (51.7)	issu (136.6)	issu (25.4)	much (38.1)	accept (18.9)	thank (61.6)
email (14.9)	defin (51.4)	quiz (133.3)	sincer (23.7)	the (37.9)	get (18.5)	that (53.0)
know (14.8)	not (50.0)	messag (116.5)	instead (23.5)	problem (35.6)	challeng (16.5)	com (46.3)
thought (14.3)	format (48.5)	type (111.0)	credit (21.9)	name (35.5)	simpli (16.5)	you (39.0)
regard (14.0)	that (48.0)	question (100.8)	player (21.3)	titl (34.3)	express (16.5)	grei (37.9)
good (12.8)	name (47.9)	answer (99.0)	extra (14.1)	www (33.6)	should (15.8)	http (36.6)

Table 5: Unigrams (term stems) with highest χ^2 value with respect to each speech act category.

igram features for the answer category were related to the topic of metadata. Here, we present a feature ablation analysis that excluded unigram features.

Results are presented in Table 6. We use the same notation and highlighting to mark differences in performance. Again, we focus our discussion on the features that yielded significant differences and at least a 5% change in performance compared to the model in the first row.

Two main trends are worth noting. First, without unigrams, several *additional* feature groups were found to be highly predictive. Affective features were highly predictive for issues, which may convey anxiety or anger. Personal concern features were highly predictive for issue resolutions, which tend to contain "achievement" words related to grading (credit, score). Linguistic features were highly predictive for answers (may contain more numbers, quantifiers, and impersonal pronouns), issues (may contain more negations and past-tense verbs), and other (may contain more non-English "out of dictionary" words). Punctuation features were highly predictive for answers, which may tend to have an absence of question marks. Finally, sequential correlation features were predictive for issue resolutions (tend to follow issues) and for issues (tend to follow other issues).

The second trend is that, without unigrams, no feature group had a significant *negative* effect on performance. Including unigram features doubled the feature space from 101 to 201 features, making it more difficult for the learning algorithm to attenuate the contribution from noisy features (e.g., temporal features with respect to issue resolution). While unigram features were highly effective, they also drastically increase the feature space.

Conclusion

We investigated three research questions. With respect to RQ1, our results show that combining redundant speech act labels from crowdsourced workers can approximate the labels from an expert. Agreement between the crowdsourced majority vote and the expert was at the level of "almost perfect" for questions and answers and "substantial" for the other speech acts. As one might expect, the most infrequent speech acts (issue resolution, negative acknowledgment, and other) had the lowest level of agreement. Improving label quality for these speech acts may require a greater number of redundant annotations. Also, further clarification may help distinguish between confusable speech acts. For example, the other category was often confused with positive acknowledgment. Future instructions may need to emphasize that a positive acknowledgment needs to reference a previous post.

In our second research question (RQ2), we investigated the usefulness of speech acts for predicting instructor intervention, assignment completion, and assignment performance. Our speech acts were the most useful for predicting instructor intervention. A post was significantly more likely to be written by an instructor when the thread contained issues and negative acknowledgments. This is consistent with the fact that posts from instructors mostly contained answers, issue resolutions, and positive acknowledgments. In other words, instructors intervened mostly to answer questions, fix problems, and provide encouragement. Future work might consider using our speech act categories as input features to a model that predicts discussion threads that require an instructor's attention.

Our speech acts were not as useful for predicting assignment completion and performance. While students who raised issues were significantly more likely to complete an assignment and perform above the median, our models explained very little of the variance. MOOC students come from different backgrounds and enroll for different reasons (Koller et al. 2013). A model that predicts student performance based on MOOC forum speech acts may *also* need to consider self-report measures about students' level of prior knowledge and motivations for enrolling. Our analyses for RQ2 suggest that predicting instructor interventions is an easier task. Instructors intervene based on what they see in the forum, and therefore the data contains more of the information necessary to model this behavior.

In terms of RQ3, the most effective features were unigram features and sequential correlation features. Unigram features were highly effective at capturing different phenomena for different speech acts, and sequential correlations were highly effective for speech acts that tend to occur in response to others (e.g., answers and issue resolutions). Other features helped some speech acts, but not others.

Several open questions remain. First, our analyses were conducted on data from a single MOOC. It remains to be

	question	answer	issue	issue resolution	positive ack.	negative ack.	other
all features	0.685	0.547	0.460	0.426	0.749	0.090	0.506
-affective	0.681 (-0.58%)	0.526 (-3.84%)▼	0.392 (-14.78%)▼	0.428 (0.47%)	0.730 (-2.54%)▼	0.087 (-3.33%)	0.505 (-0.20%)
author	0.681 (-0.58%)	0.547 (0.00%)	0.454 (-1.30%)	0.330 (-22.54%)▼	0.746 (-0.40%)	0.095 (5.56%)	0.506 (0.00%)
-cognitive	0.689 (0.58%)	0.546 (-0.18%)	0.444 (-3.48%)	0.432 (1.41%)	0.748 (-0.13%)	0.091 (1.11%)	0.508 (0.40%)
-cosine	0.665 (-2.92%)▼	0.524 (-4.20%)	0.441 (-4.13%)	0.444 (4.23%)	0.747 (-0.27%)	0.082 (-8.89%)	0.505 (-0.20%)
temporal	0.686 (0.15%)	0.546 (-0.18%)	0.462 (0.43%)	0.464 (8.92%)	0.746 (-0.40%)	0.078 (-13.33%)	0.502 (-0.79%)
-linguistic	0.654 (-4.53%)▼	0.510 (-6.76%)▼	0.395 (-14.13%)▼	0.406 (-4.69%)	0.735 (-1.87%)▼	**0.099 (10.00%)**	0.473 (-6.52%)▼
-links	0.685 (0.00%)	0.546 (-0.18%)	0.451 (-1.96%)▼	0.425 (-0.23%)	0.749 (0.00%)	0.087 (-3.33%)	0.512 (1.19%)
-modal	0.686 (0.15%)	0.544 (-0.55%)	0.453 (-1.52%)	0.416 (-2.35%)	0.749 (0.00%)	0.093 (3.33%)	0.515 (1.78%)
-perceptual	0.685 (0.00%)	0.545 (-0.37%)	0.462 (0.43%)	0.411 (-3.52%)	0.749 (0.00%)	0.090 (0.00%)	0.512 (1.19%)
-personal_concerns	0.682 (-0.44%)	0.549 (0.37%)	0.441 (-4.13%)	0.352 (-17.37%)▼	0.750 (0.13%)	0.087 (-3.33%)	0.497 (-1.78%)
-position	0.672 (-1.90%)▼	0.542 (-0.91%)	0.444 (-3.48%)▼	0.419 (-1.64%)	0.749 (0.00%)	0.083 (-7.78%)	0.511 (0.99%)
-post_comment	0.685 (0.00%)	0.546 (-0.18%)	0.456 (-0.87%)	0.426 (0.00%)	0.747 (-0.27%)	0.095 (5.56%)	0.492 (-2.77%)
-punctuation	0.443 (-35.33%)▼	0.516 (-5.67%)▼	0.456 (-0.87%)	0.433 (1.64%)	0.741 (-1.07%)▼	0.084 (-6.67%)	0.483 (-4.55%)▼
-sentiment	0.687 (0.29%)	0.546 (-0.18%)	0.457 (-0.65%)	0.422 (-0.94%)	0.740 (-1.20%)▼	0.089 (-1.11%)	0.506 (0.00%)
-social	0.682 (-0.44%)	0.547 (0.00%)	0.450 (-2.17%)▼	0.420 (-1.41%)	0.749 (0.00%)	0.088 (-2.22%)	0.509 (0.59%)
-spoken	0.685 (0.00%)	0.546 (-0.18%)	0.461 (0.22%)	0.427 (0.23%)	0.749 (0.00%)	0.083 (-7.78%)	0.509 (0.59%)
-votes	0.684 (-0.15%)	0.546 (-0.18%)	0.461 (0.22%)	0.424 (-0.47%)	0.748 (-0.13%)▼	0.088 (-2.22%)	0.505 (-0.20%)
+sequential corr.	**0.715 (4.38%)**▲	**0.609 (11.33%)**▲	**0.492 (6.96%)**▲	**0.490 (15.02%)**▲	**0.759 (1.34%)**	0.082 (-8.89%)	**0.604 (19.37%)**▲

Table 6: Feature ablation results excluding unigram features.

seen whether our results generalize to different MOOCs. Second, most research on predicting instructor interventions has been done retrospectively, by learning to predict places where an instructor already intervened. While this is a convenient framework, future work might consider using speech acts to predict discussions that an instructor may want to be aware of, whether or not they intervene.

References

Bhatia, S.; Biyani, P.; and Mitra, P. 2012. Classifying user messages for managing web forum data. In *Proceedings of the 5th International Workshop in Web and Databases*.

Carvalho, V. R., and Cohen, W. W. 2005. On the collective classification of email "speech acts". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 345–352. ACM.

Chaturvedi, S.; Goldwasser, D.; and Daumé III, H. 2014. Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1501–1511. ACL.

Cohen, W.; R. Carvalho, V.; and M. Mitchell, T. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. ACL.

Elouazizi, N. 2014. Point-of-view mining and cognitive presence in moocs: A (computational) linguistics perspective. *EMNLP 2014* 32.

Kim, J., and Kang, J.-H. 2014. Towards identifying unresolved discussions in student online forums. *Applied Intelligence* 40(4):601–612.

Koller, D.; Ng, A.; Do, C.; and Chen, Z. 2013. Retention and intention in massive open online courses: In depth. *Educause Review* 48(3).

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, 342–351. ACM.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. *The Development and Psychometric Properties of LIWC2007*. LIWC.net.

Qadir, A., and Riloff, E. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 748–758. ACL.

Ravi, S., and Kim, J. 2007. Profiling student interactions in threaded discussions with speech act classifiers. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, 357–364. IOS Press.

Rus, V.; Graesser, A. C.; Moldovan, C.; and Niraula, N. B. 2012. Automatic discovery of speech act categories in educational games. In *Proceedings of the 5th International Conference on Educational Data Mining*, 25–32. www.educationaldatamining.org.

Searle, J. R. 1976. A classification of illocutionary acts. *Language in society* 5(01):1–23.

Smucker, M. D.; Allan, J.; and Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 623–632. ACM.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263. ACL.

Stephens-Martinez, K.; Hearst, M. A.; and Fox, A. 2014. Monitoring moocs: which information sources do instructors value? In *Proceedings of the first ACM conference on Learning@ scale conference*, 79–88. ACM.

Wen, M., and Rose, C. P. 2014. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1983–1986. ACM.

Wen, M., and Yang, Diyi andRose, C. P. 2014a. Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the 8th International Conference in Weblogs and Social Media*, 525–534. AAAI.

Wen, M., and Yang, Diyi andRose, C. P. 2014b. Sentiment analysis in mooc discussion forums: What does it tell us? In *Proceedings of the 7th International Conference on Educational Data Mining*, 130–137. www.educationaldatamining.org.