I designed and performed following four experiments in order to evaluate output of LDA, M4 and Block HMM models.

1. **Using topic distribution per turn as features to predict annotations of the dialogue**
2. **Evaluating topic clusters (at the level of whole dataset) based on word rankings in them**
3. **Evaluating topic clusters (at the level of section, condition and role) based on word rankings in them**
4. **Using topic clusters to induce topic for a given dialogue**

Each of the above experiments and their results have been explained in detail below.

Using topic distribution as features per turn to predict annotations of the dialogue

I used the topic distribution per turn as features to predict columns of annotations using two approaches.

- **Multivariate linear regression**
- **Naive Bayes**

Multivariate linear regression

Experiment

I implemented the most widely used predictive model - Multivariate linear regression - to predict annotations for each turn of given data using topic distribution per turn as features. As the topic models were built with 5 set as the number of topics, I had to implement linear regression for multiple variables (5 in this case). The task of predicting annotations for each turn here is a classification problem and one may argue that it is not a good idea to approach a classification problem through regression but I think it's always possible to do a continuous regression to fit the data and truncate the continuous prediction to yield discrete classifications. The reason I tried this approach is so that I can see if I can relate a topic with an annotation by looking at the coefficients that multiply the predictor variables (topics). We will see what I found in this regard in the results section.

Note that for now I tried to predict annotations ("AT" and "Reas") having exactly two classes with this approach. Also, in order to truncate the continuous prediction to yield discrete classification, my code automatically calculates the threshold to divide the continuous prediction into two classes such that accuracy is the highest.

One more thing should be noted that for block HMM model, each turn can be assigned only one topic (for other models each word is assigned a topic) and hence topic number is the only feature to predict an annotation.

Whole corpus was splitted into the ratio of .8 : .2 for having training and test data set.

Results

Model	Annotation - “AT”	Annotation - “Reas”
LDA	Accuracy: 95.5 % Kappa: .91 F measure: .24 (.24, .22)	Accuracy: **97.1 %** Kappa: **.94** F measure: **.5 (.33, 1.0)**
Block HMM	Accuracy: 95.5 % Kappa: .91 F measure: .11 (1.0, .05)	Accuracy: 95.5 % Kappa: .91 F measure: .3 (.33, .27)
M4	Accuracy: 95.5 % Kappa: .91 F measure: .19 (.55, .12)	Accuracy: 96.6 % Kappa: .93 F measure: .46 (.33, .75)

Note: bracketed data denote recall and precision respectively for f-measure

The suspiciously high numbers for accuracy in above table is attributed to the fact that one of the classes is highly dominant like there are very few *reasoning* dialogues in the whole corpus and hence, predicting dominant class for each turn will still give a high accuracy. F measure is affected less by this issue and hence we can see lower numbers for f-measure in above table.

The numbers in the above table indicate LDA to be the best model followed by M4 and Block HMM. It is interesting to see that LDA gives 100% precision while predicting “Reasoning” annotation and Block HMM model gives 100% recall while predicting “Accountable Talk” annotation.

In addition to this, I looked at the coefficients that multiply the predictor variables (topics) for LDA and M4 model for predicting “Reasoning” annotation.

Predictive variable	multiplicative coefficient	
	LDA	M4
Topic1	[-0.01472628]	*[0.02858743]*
Topic2	[-0.01079976]	[-0.00084913]
Topic3	[-0.02676404]	[-0.01120302]
Topic4	**[0.10396077]**	[0.00977066]
Topic5	[-0.01965642]	*[0.01917388]*

As can be seen from above table, Topic4 is clearly indicative of dialogues with “Reasoning” annotation and as we will see later Topic4 is about details of experiment being conducted in class. In case of M4 model there is no such effect.

Naive Bayes

Experiment

In this approach, I still used topic distribution per turn as features to predict annotations but probabilistically. The assumption here is that the probability of each predictive variable (topic) belonging to each class (annotation classes) is independent of all other attributes.

Five features (one in case of Block HMM) were used to predict classes for a type of annotation. With Naive Bayes approach, all type of annotations were being predicted irrespective of the number of classes.

Whole corpus was splitted into the ratio of .8 : .2 for having training and test data set.

Result

Model	offtask	cheating	at	reas	trans	neg	het
LDA	**41.6 %**	NA	**72.8 %**	**85.15 %**	**80.02 %**	**43.08 %**	NA
Block HMM	32.3 %	7.7 %	30.26 %	55.9 %	45.13 %	32.89 %	50.0 %
M4	35.9 %	NA	41.54 %	34.88 %	20.05 %	35.90 %	NA

Similar to results obtained using regression predictive modeling technique, here also LDA seems to be the best model.

One observation here was that this experiment was giving different results on different runs due to probabilistic nature of the technique may be. So above numbers can be different by even 10% on a rerun.

Evaluating topic clusters (at the level of whole dataset) based on word rankings in them

I wrote a script to process data in “TopicWordWeights” file for each model to get word rankings for each topic. I generated two versions of rankings - one with and one without stopword removal in the word lists.

Note : I imported stopwords from nltk.corpus and so, nltk must be downloaded on the machine in order to run the script.

I stored two versions of word rankings of topics for each model to qualitatively analyze if the generated topic clusters make any sense to me. Relevant files (having topic word rankings) have been submitted for review.

Below is the list of **top 25 words** in each topic for each model. Note that stop words have been removed from top 25 words from below lists though analysis was done including stop words also separately.

LDA

##Topic 1:##############################
write worksheet. **explanations observations** talking happen **predictions** discuss **predict** looking conditions specific
##Topic 2:##############################
one would team mates example, encourage see please so, benefit asking said
##Topic 3:##############################
watch video nice going all. sure ok, make **videos** role gotta **folder** please **cell**
##Topic 4:##############################
water condition **glucose** think **distilled solution** replace **cell starch** inside **weight model iodine water**. **suspension**. **starch**, **solution**. presence
##Topic 5:##############################
observed get condition video discussing move predictions. compare lets good discuss

Topic 4 is the most salient one and can be attributed to dialogues having experiment details.

Topic 1 either contains verbs or their nominalizations e.g. predict and predictions. Also looking at the dialogues in the corpus, dialogues by Tutor tend to have words from this topic. May be thats why this topic contains words like explanation, observation, prediction etc. a lot or verbs asking students to do stuff

Topic 3 seems to be containing words from dialogues by Tutor talking about videos on cell condition. It also contains words from Tutor dialogues on off task topics like social messages e.g. “nice talking to you”, “please … “, “sure ok”

It can be observed that words spoken by Tutor are highly ranked words in most of the topics. That can be explained by that fact that Tutor (Computer agent’s) were mostly the same across class sections and thus, same words occurred multiple times thereby moving them higher on topic lists. On the other hand, students were different across different sections and hence different dialogues having different words many of which were typos unlike tutor which is a computer agent and spoke well formed and spelling correct sentences.

Block HMM

##Topic 1:##########################
condition video discussing move observed c
##Topic 2:##########################
water think *glucose* happen *weight* nice talking predict get discuss b would specific
##Topic 3:##########################
would team one mates example encourage explanation said please see benefit
##Topic 4:##########################
video b conditions observed condition watch write worksheet. cell explanations sure make discuss observations
##Topic 5:##########################
condition starch replace **glucose water** watch videos video folder desktop there. **suspension. solution. solution** presence order modification made lugol's **iodine**

Here Topic 5 is the most salient one again referring to words about experiment details but is not as good as Topic 4 of LDA as Topic 5 here also contains words about videos of cell conditions which seems like a different topic to me.

First three topics generated by Block HMM had many stop words and thats why these topics have fewer words after stop word removal. First topic has only 6 words left after stop word removal. Clearly it’s not a coherent and distinct topic cluster to be made.

Topic 2 again contains words relating to experiments e.g. water, glucose, weight etc.

One striking observation is that word “video” was one of the top 25 words in 3 of the 5 clusters above and that too when it has been used only in one sense throughout the corpus.

Clearly, clusters formed by Block HMM model are inferior to those formed by LDA model.

M4

##Topic 1:##############################
video water think watch explanation ok
##Topic 2:##############################
condition **glucose** b c cell happen **starch** write **solution distilled** replace conditions worksheet. please explanations watch
##Topic 3:##############################
observed condition predictions. compare move different back discuss happened conditions discussing video watching put mix
##Topic 4:##############################
would one team mates example encourage get please whether benefit asking said see right
##Topic 5:##############################
nice all. responsible water sure talking make looking specific :- used today student. strategy opportunities instructor discussion conversation. all!

Here Topic 2 is the most salient one pointing towards experiment details. Without giving much details, I feel that clusters formed by M4 comes in between LDA and Block HMM in terms of quality.

Evaluating topic clusters (at the level of section, condition and role) based on word rankings in them

For this experiment, I built a hierarchy of topic clusters at the level of **section** (A, B, C, D), **condition** (no, direct, indirect) and **role** (teacher, student). Hierarchical topic clusters can be seen in wordrankings<model>.txt. I analyzed then qualitatively to see if these hierarchical clusters seem coherent and distinct.

LDA

At this fine grained level, topic clusters are more rich and informative. Lets see following topic cluster at level of

Section:A Condition:indirect Role:Tutor

condition solution In water starch replace glucose distilled A suspension presence placed order opposite modification model made inside immerse detect cell We Lugol Iodine C B

This cluster clearly indicates experiment details in above cluster. On a side note, I made an observation that topic clusters at level of direct condition and student role are much smaller than clusters at level of other conditions and student role which says that students say more in absence of any support or having less support. Clusters at level of tutor role are more rich qualitatively than clusters at level of student role though for experiment detail topic even clusters at student role level are good enough.

Block HMM and M4

Like before, degradation of quality can be seen when we move from LDA model to Block HMM and M4 models. As you can see below, experiment topic cluster for both models though capture relevant words like that of LDA model but in addition it an amount of noise has also been added.

Block HMM: A ind Tutor Topic4

starch replace condition water watch video there. suspension. solution. solution presence order modification made glucose folder distilled detect We Videos Lugol's Iodine In Go Desktop C A.

M4: A ind Tutor Topic1

condition In C B A {s003} write worksheet. watch starch solution s011 replace happen glucose distilled cell build You Please Conditions water. video two suspension. solution. smiley showing s007 presence predictions predict placed pages order opposite. observations move modification model meet made inside immerse going folder explanations discussing discuss detect conditions. change. agreement When We S007 Lugol's Iodine Condition C. As A.

Using topic clusters to induce topic for a given dialogue

This experiment was conducted to check how random dialogues will fit into these topic clusters formed by these models. I used the topic clusters (with stop words) generated in previous experiment and calculated cosine similarity of a new dialogue with each of the five clusters to find which topic is the best fit for the random dialogue. I know it's a very rudimentary test I have performed and may be improved if stop word removal and stemming is applied to both topic cluster and test dialogue. Results with 16 random dialogues can be seen below.

	LDA		Block HMM	
	Expected topic	Output topic	Expected Topic	Output topic
Hi all! Today, each of you will be both the instructor and the student. Each of you will be responsible for looking for opportunities for a specific discussion strategy to be used in the conversation.	1	1	1	1
Lets get started by introducing ourselves. I am Alex.	5	5	3	1
i am {s001}	2	0	1	-1
Hi {s001} you are the Revoicer. When an explanation or idea is given which would benefit from revoicing, please encourage one of your team mates to do so, for example, by asking them to explain what was said in their own words.	2	2	2	2
{s005} you are the Challenger. When you see a statement being made (whether it is right or wrong) which would benefit from being challenged, please encourage one of your team mates to do so, for example, by asking them whether they agree or disagree with what was said, and why.	2	2	2	2
Its nice to meet you all. :) Please make sure you have understood what you are responsible for by looking over pages 7-9 in your book and remembering what role you have been assigned.	3	3	3	3
In condition A, we placed a glucose solution inside the cell model and immerse it in distilled water. In condition B, we did the opposite.	4	5	3	3
Nice to meet u 2.	3	0	1	-1
Please discuss what you predict will happen in these two conditions.	1	5	1	3
In condition C, we made a modification to condition A. We replace the glucose solution with a starch suspension. In order to detect the presence of starch, we replace the distilled water with Lugol's Iodine solution.	4	5	4	0
You should now move on to discussing what will happen in Condition C and your explanation for this change.	5	5	0	2
When you are in agreement, write down your predictions and explanations for Conditions A, B and C on your worksheet.	1	1	3	3
You are now going to watch a video showing the cell in Conditions A, B and C.	3	3	3	3
As you watch the video, write down your observations on your worksheet.	1	3	3	4
Go to the Videos folder on the Desktop, and watch the video which is there.	3	5	4	3
I predict that in Condition A the cell model would gain more water and Codition b would loose the water in the cell model	4	2	1	2
Accuracy	**50 %**		**43.75 %**	

The experiment results here show LDA to be again better than Block HMM though only slightly.

****

Instructions about running scripts and input and output files details

Using topic distribution per turn as features to predict annotations of the dialogue

Regression approach

Run : python mlr.py <model> <attribute>

where model can be *LDA*, *blockHMM* or *M4*
and attribute can be *at* or *reas*

Sample output:
python mlr.py LDA reas
[[0.05566939]
 [-0.01472628]
 [-0.01079976]
 [-0.02676404]
 [0.10396077]
 [-0.01965642]]
Accuracy : 0.970588235294
Kappa : 0.941176470588
F measure : 0.5
Recall : 0.333333333333
Precision : 1.0

where array in the beginning is multiplicative coefficients of predictive variables (topic)

Naive Bayes approach

Run : python naivebayesian.py <model> <attribute>

where model can be *LDA*, *blockHMM* or *M4*
and attribute can be *offtask*, *cheating*, *at*, *reas, trans, neg or het*

Sample output:

python naivebayesian.py LDA reas
Split 965 rows into train=772 and test=193 rows
Accuracy: 84.1243523316%
Kappa: 0.682487046632%

These scripts (mlr.py and naivebayesian.py) use output files of models
LDA - Output.txt
blockHMM - Output1.txt
M4 - Output2.txt

Evaluating topic clusters (at the level of whole dataset) based on word rankings in them

Run python wordrankings.py <model> <attribute>

where model can be *LDA*, *blockHMM* or *M4*

This script uses TopicWordWeights file of each model
LDA - TopicWordWeights.txt
blockHMM - TopicWordWeights1.txt
M4 - TopicWordWeights2.txt

Output of the script can be seen in TopicWordRankings<model>.txt and TopicWordRankings<model>WithoutStopWords.txt

Evaluating topic clusters (at the level of section, condition and role) based on word rankings in them

Running mlr.py script for first experiment automatically create hierarchichal topic clusters files named wordranking<model>.txt

Using topic clusters to induce topic for a given dialogue

Running wordrankings.py for second experiment automatically creates files having new topics for input dialogues. Check newdialogueTopics_<model>.txt files

Test dialogues are in input.txt