Classifying Student Dialogue Acts with Multimodal Learning Analytics

Aysu Ezen-Can, Joseph F. Grafsgaard, James C. Lester, Kristy Elizabeth Boyer
Department of Computer Science
North Carolina State University
aezen, jfgrafsg, lester, keboyer@ncsu.edu

ABSTRACT

Supporting learning with rich natural language dialogue has been the focus of increasing attention in recent years. Many adaptive learning environments model students' natural language input, and there is growing recognition that these systems can be improved by leveraging multimodal cues to understand learners better. This paper investigates multimodal features related to posture and gesture for the task of classifying students' dialogue acts within tutorial dialogue. In order to accelerate the modeling process by eliminating the manual annotation bottleneck, a fully unsupervised machine learning approach is utilized for this task. The results indicate that these unsupervised models are significantly improved with the addition of automatically extracted posture and gesture information. Further, even in the absence of any linguistic features, a model that utilizes posture and gesture features alone performed significantly better than a majority class baseline. This work represents a step toward achieving better understanding of student utterances by incorporating multimodal features within adaptive learning environments. Additionally, the technique presented here is scalable to very large student datasets.

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms

Human Factors

Keywords

Text-based learning analytics, Multimodal learning analytics, Tutorial dialogue, Dialogue act modeling

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3417-4/15/03
http://dx.doi.org/10.1145/2723576.2723588 ...$15.00.

1. INTRODUCTION

The research community has endeavored for several decades to build effective systems that support learners [44]. Within the learning analytics community, there has been significant work on analyzing students' clicking behavior [46], engagement [9], interactions with the learning environment [1], and textual analysis [26, 38] to understand how students learn and how best to support them. Textual natural language data is a rich source of information that can support these goals.

Understanding students through their natural language has been the focus of researchers for a broad range of goals including assessing students' science competency [26], identifying exploratory dialogue [16], identifying idea distribution [10] and relating student posts to knowledge creation [8]. For this paper's goal of improving natural language interaction in a learning environment, the focus is specifically on identifying dialogue acts, which represent the communicative intentions of each utterance [37, 40]. These dialogue acts have been shown to be correlated with learning [42]. For example, "[the task] just means to assign the name right?" and "spaces or no spaces?" are both questions, and the goal of dialogue act classification is to automatically detect their types. Automatically identifying dialogue acts has long been a goal for dialogue researchers [39], yet only very recently have multimodal features been considered for this task.

Dialogue act classification is a very important area of research not only for automated systems that adapt to learners, but also for understanding the processes that underlie learning. For example, dialogue act analysis can reveal patterns that are particularly effective for supporting students with varying levels of self-efficacy [45], different personality profiles [41], and different learner characteristics [29]. For adaptive systems, dialogue act classification is a crucial step toward providing rich natural language within tutoring systems that can bridge the gap between one-on-one tutoring and automated learning environments [17]. Moreover, scalable dialogue act modeling techniques can be applied across massive student data, lending insights into how people learn at scale.

Multimodal learning analytics incorporate features of different categories into the learning analytics tasks [3]. Some categories of multimodal features, including posture [21] and facial expressions indicating confusion [5], have been found useful for dialogue act classification. However, these prior approaches have relied upon supervised machine learning techniques, which suffer from a manual annotation bottleneck that is problematic for scaling these models across domains or even across corpora. Therefore, how to utilize multimodal features within unsupervised dialogue act models is an important open research question. Unsupervised machine learning approaches hold great promise for addressing this shortcoming by eliminating the need for dialogue act taxonomy engineering and manual labeling of utterances [11, 15]. In addition, the groupings produced by unsupervised models are fully data driven, which may differ from manual labels and provide important insights into the data.

This paper presents the first model to incorporate multimodal features into an *unsupervised* dialogue act classifier for the learning analytics domain. Motivated by the importance of analyzing the *process* of learning rather than the end *product* only [2], we analyze posture and gesture features of students in the course of tutoring and utilize this information to enhance our understanding of student dialogue. The results demonstrate that information about students' posture and gesture significantly improves dialogue act classification performance when judged against gold standard dialogue act labels. Furthermore, analyses show that utilizing solely posture and gesture lead to better performance than majority baseline chance, even in the absence of any linguistic information about the utterances being classified. This finding highlights the importance of multimodal features for building rich understanding models of student utterances. This work is a step toward developing tutorial dialogue systems that rely on unsupervised models to provide flexible and effective dialogue to support learning. Moreover, the techniques investigated here have broad application for modeling natural language interactions that support learning because the modeling does not utilize manual labels: the clustering is fully data-driven and the multimodal features are automatically extracted, making it scalable for massive student data across domains.

2. RELATED WORK

A growing body of findings indicates that multimodal features play an important role in learning analytics [3]. Empirical studies suggest that multimodal approaches are promising for assessing learners' interaction experience [24]. Research has proceeded to identify relationships between multimodal cues and cognitive-affective states [19] and learning itself [48, 49].

The importance of multimodal features in dialogue has been widely observed. For example, gaze and gesture help with modeling the flow of conversation while showing the relation of dialogue acts to turn taking [4]. The importance of nonverbal cues is also widely observed for discovering conversational patterns [23] and for determining the addressee of an utterance [43].

Specifically, posture has been shown to be promising for recognizing affective states such as boredom, frustration [47] and disengagement [12, 33, 36]. Automatic tracking of these posture features has improved substantially, allowing extraction of these features both from two-dimensional [12, 36] and three-dimensional [19] video using computer vision techniques. Following the line of investigation into postural features, gestural features have also gained attention from the community. Motivated by the cultural influence of gestures [28], they have also been studied in the intelligent tutoring systems community [18, 33, 47]. Gestures are related to student affective states: one-hand-to-face gestures are associated with less negative affect whereas two-hands-to-face gestures are associated with reduced focus [19]. The growing body of work in posture and gesture motivates research on dialogue incorporating these multimodal features.

From a dialogue perspective, dialogue act classification, the task of inferring the action and intention underlying utterances, has been extensively studied [32, 39]. There is a rich body of work on supervised dialogue act classification techniques where a machine learner is trained on manual dialogue act tags. However, the work utilizing nonverbal cues constitutes a very small subset of that larger body of work, with acoustic and prosodic cues [25, 32], facial expressions [5], pointing gestures [7] and body posture [21] among the modalities that are found promising. There is a much smaller body of work on unsupervised dialogue act modeling, most outside of the educational domain [13, 34]. It is crucial to utilize fully data-driven methods for modeling student utterances for rapid development of adaptive systems. This paper is the first to consider these multimodal cues for inclusion within unsupervised dialogue act classifiers for learning analytics with the overarching goal of understanding students better [11, 15].

3. CORPUS

Tutorial dialogue is one of the most effective means of supporting human learning and is an important source for textual learning analytics. The work reported in this paper uses a tutorial dialogue corpus collected in a computer-mediated textual environment for task-oriented tutoring of introductory computer science. The corpus consists of student-tutor interactions while collaborating on computer programming problems in the Java programming language (see Table 2 for an excerpt). This corpus reflects effective tutoring, with students correctly answering 49% of missed pretest questions on the posttest and demonstrating positive overall learning gain (p<0.001).

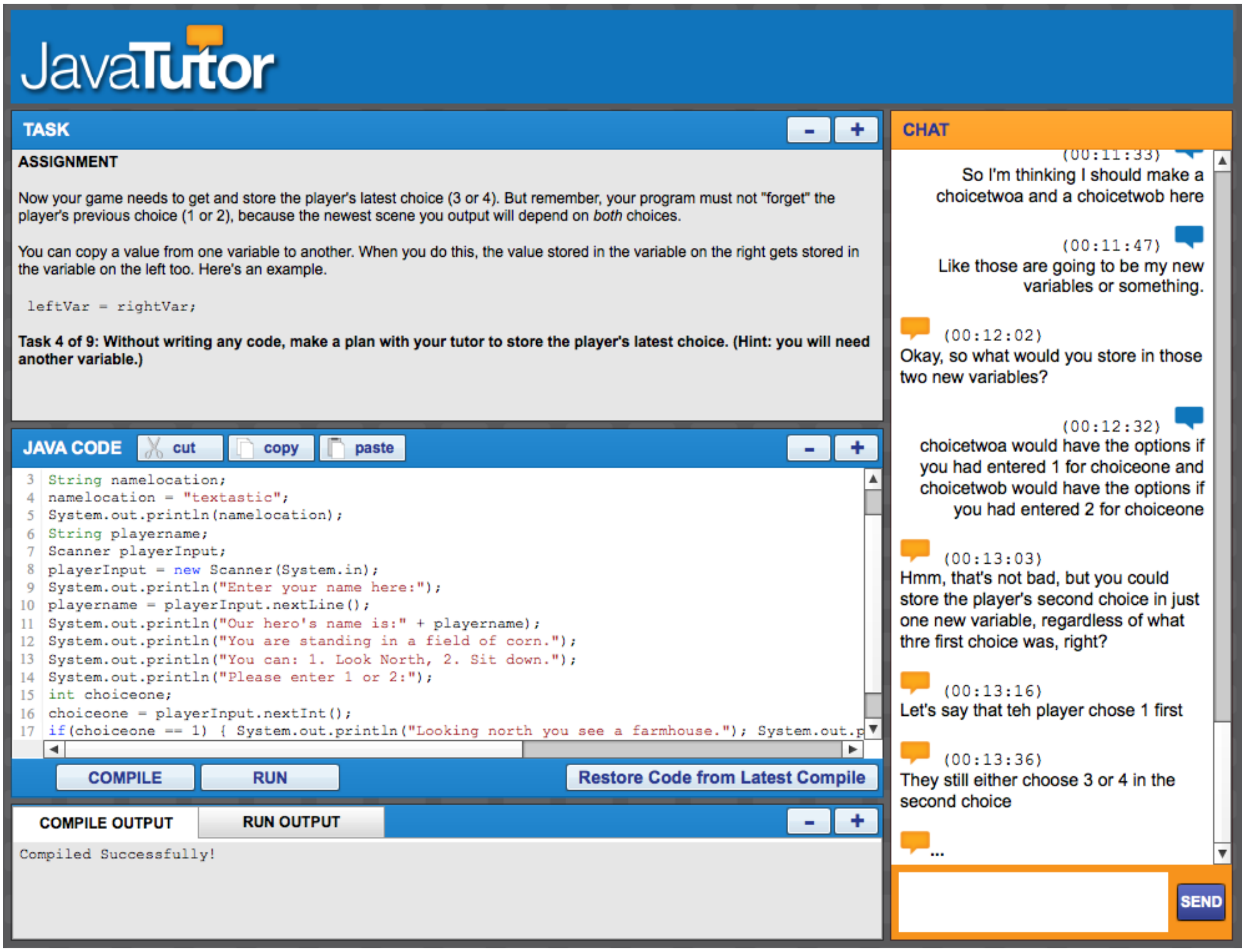
As shown in Figure 1, the interface consists of four panels: the task pane where the tasks to be completed are explained, the code pane in which the students implement their solutions, the output pane where students could see the output of compiling/running their programs, and the dialogue pane which allowed tutor-student textual dialogue.

The multimodal corpus includes 1,443 student dialogue utterances which were manually annotated in prior work (see Table 1) [21]. There are 7 manually labeled dialogue act tags in the corpus: Answer ("A", 43.28% —the majority baseline), statement ("S", 20.46%), acknowledgment ("ACK", 20.2%), question ("Q", 14.16%), clarification ("C", 0.9%), request for feedback ("RF", 0.5%) and other utterances ("O", 0.5%). Because the focus of this paper is on unsupervised dialogue act classification, these dialogue act tags are only used for evaluation purposes with held-out cross-validation test data, while the models are built on unlabeled data.

The students (n=37) were recorded with Kinect cameras (Figure 2), and the videos were processed to extract posture and gesture features. In prior work, the posture features were calculated from head and torso distances and gesture features include one-hand-to-face and two-hands-to-face and the performance of these algorithms compared to manual tags was 92.4%, indicating high reliability [19].

4. FEATURES

The goal of this work is to investigate the extent to which posture and gesture features improve unsupervised dialogue



ASSIGNMENT

Now your game needs to get and store the player's latest choice (3 or 4). But remember, your program must not "forget" the player's previous choice (1 or 2), because the newest scene you output will depend on *both* choices.

You can copy a value from one variable to another. When you do this, the value stored in the variable on the right gets stored in the variable on the left too. Here's an example.

```
leftVar = rightVar;
```

Task 4 of 9: Without writing any code, make a plan with your tutor to store the player's latest choice. (Hint: you will need another variable.)

```
String namelocation;
namelocation = "textastic";
System.out.println(namelocation);
String playername;
Scanner playerInput;
playerInput = new Scanner(System.in);
System.out.println("Enter your name here:");
playername = playerInput.nextLine();
System.out.println("Our hero's name is:" + playername);
System.out.println("You are standing in a field of corn.");
System.out.println("You can: 1. Look North, 2. Sit down.");
System.out.println("Please enter 1 or 2:");
int choiceone;
choiceone = playerInput.nextInt();
if(choiceone == 1) { System.out.println("Looking north you see a farmhouse."); System.out.p
```

```
Compiled Successfully!
```

(00:11:33)
So I'm thinking I should make a choicetwoa and a choicetwob here

(00:11:47)
Like those are going to be my new variables or something.

(00:12:02)
Okay, so what would you store in those two new variables?

(00:12:32)
choicetwoa would have the options if you had entered 1 for choiceone and choicetwob would have the options if you had entered 2 for choiceone

(00:13:03)
Hmm, that's not bad, but you could store the player's second choice in just one new variable, regardless of what thre first choice was, right?

(00:13:16)
Let's say that teh player chose 1 first

(00:13:36)
They still either choose 3 or 4 in the second choice

...

Figure 1: The JavaTutor tutorial dialogue interface with four panels.

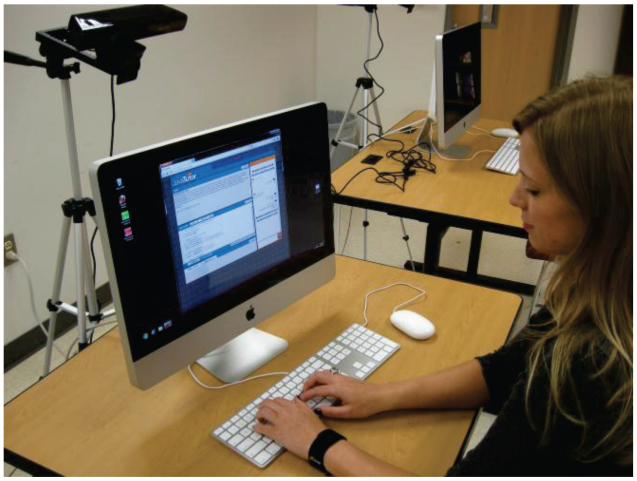


Figure 2: The workstation with Kinect depth-sensor, webcam and tutorial dialogue interface.

act models. Because the parallel streams (student coding activities, multimodal features, and dialogue) offer rich sources of information, we hypothesize that models of student utterances highly benefit from utilizing these features. Four sets of features were considered within the experiments. Three of these sets—lexical, dialogue-context and task features—have been shown to improve unsupervised dialogue act classification in prior work [14]. The fourth set consists of multimodal features of posture and gesture.

Lexical Features. Words and punctuation of each student utterance are provided to the model. Because the overarching goal of dialogue act classification is to understand learners effectively in real-time systems, features such as part-of-speech tags which are time-consuming to extract and have been observed not to improve the accuracy of some dialogue act models [6] are omitted, leaving only unigrams and word-orderings for consideration.

Dialogue-Context Features. Four dialogue-context features shown useful in prior work [21] are included in the model: utterance position in relation to the beginning of the dialogue, utterance length, author of the previous dialogue message (tutor or student), and previous tutor dialogue act.

Student Dialogue Act	**Example**	**Distr. (%)**
Answer (A)	*pretty good, just a lot of homework*	43.28
Statement (S)	*it's very interesting to me*	20.46
Acknowledgement (ACK)	*alright*	20.20
Question (Q)	*how can the errors be fixed?*	14.16
Clarification (C)	**html messing*	0.90
Request for Feedback (RF)	*better?*	0.50
Other (O)	*haha*	0.50

Table 1: Student dialogue act tags, sample utterances and their frequencies.

Because in a tutorial dialogue system the tutor moves are system-generated, their dialogue acts are known. We use the previous tutor dialogue act as feature in our models. This type of dialogue history has been shown effective for dialogue act classification [7,14,27,35].

Task Features. The parallel task stream present in tutorial dialogue is a rich information source that may not be directly represented in the dialogue. This stream consists of task actions, in our case compiling, running of code, changing code and sending messages. Utilizing these features can help capture the whole dialogue in a more comprehensive manner. To do this, we use interaction traces between tutors and students to obtain task features that can help the dialogue act classification task [14]. The programming activities logged throughout the course of tutoring include the previous task action preceding each student utterance (composing an utterance, writing/compiling/running code), the status of the most recent coding action (begin, success, error, stop, input_sent), number of messages sent since the beginning of the task, and number of errors present in the student code.

Posture Features. Four posture features are utilized: head distance (distance between camera and head), mid torso, lower torso, and the average of these three features [19] as shown in Figure 3. Approximately eight frames per second were recorded from a Kinect depth camera. However, utterances occur less frequently. Because the granularity of posture features are different from granularity of utterances, representation constitutes a challenge. We take the average of the feature values ten seconds before an utterance and ten seconds after the previous utterance, which is the minimum granularity that allows us to observe change in the features.

Gesture Features. The gesture features include two different hand-to-face features: one-hand-to-face (see Figure 4) and two-hands-to-face (see Figure 5) indicating the hand positions of students [20]. For matching the gesture features to utterances, we count the number of values detected between two utterances within a ten-second frame. For instance, for a particular utterance, the number of times the one-hand-to-face feature gets detected after the previous utterance of that particular utterance is counted which allows us to match gesture features to each utterance.

Student modifies code.
Student receives a compile error.
One-hand-to-face gesture recognized.
Student: which do i put first? [*Question*]
Tutor: try it. [*Statement*]
Change in head depth detected.
Student receives a compile error.
Tutor: what you had was close. [*Statement*]
Tutor: go back to that [*Statement*]
Student modifies code.
Student compiles code successfully.
Student: is the order wrong? [*Question*]
Tutor: no, the literal is just [*Statement*]
Tutor: Player's name is [*Statement*]
Student modifies code.
Tutor: dont put your name [*Hint*]
Student runs the code successfully.
Tutor: that is excellent. [*Positive Feedback*]
Tutor: i could tell a lot of learning was going on [*Statement*]
Change in mid-depth detected.
Student: it's very interesting to me [*Statement*]
Tutor: good. you are good at it. [*Statement*]
Tutor: try things. make mistakes. learn. [*Statement*]
Tutor: one more screen. [*Statement*]

Table 2: Excerpt of dialogue from the corpus and the corresponding dialogue act tags.

5. METHODOLOGY

For unsupervised classification of dialogue acts, we use a framework that calculates similarities between utterances using their longest-common-subsequences (explained later in this section) and then utilize those similarities within k-medoids clustering [14]. For features other than the lexical features (task, dialogue-context and multimodal features) we use Cosine similarity, which captures similarity independent of the length of utterances. K-medoids is a widely-used clustering algorithm that groups utterances according to their closest centroids within clusters [30]. For this algorithm, the number of clusters needs to be selected. k=5 was found to be the optimal number of clusters in prior work by using the Bayesian Information Criterion, which penalizes the number of parameters the model uses [14].

In addition, our prior work for representing dialogue history, which was shown to significantly improve upon the prior performance of unsupervised dialogue act models, is adopted in this work [14]. Specifically, we branch the clustering model by student utterances according to the previous tutor dialogue acts. Nine branches of student utterances are formed, one for each tutor dialogue act. In this way, the student utterances in the training set are clustered while taking the previous tutor move into account. Each branch has student utterances that share the previous tutor dialogue act and therefore are more granular for clustering. Then, clustering is performed within each branch. Note that each student utterance is clustered only with utterances that had the same previous tutor dialogue act.

Classifying test utterances. Once we have the clusters that are produced using the branching and clustering technique in the training set, each unseen utterance from the test

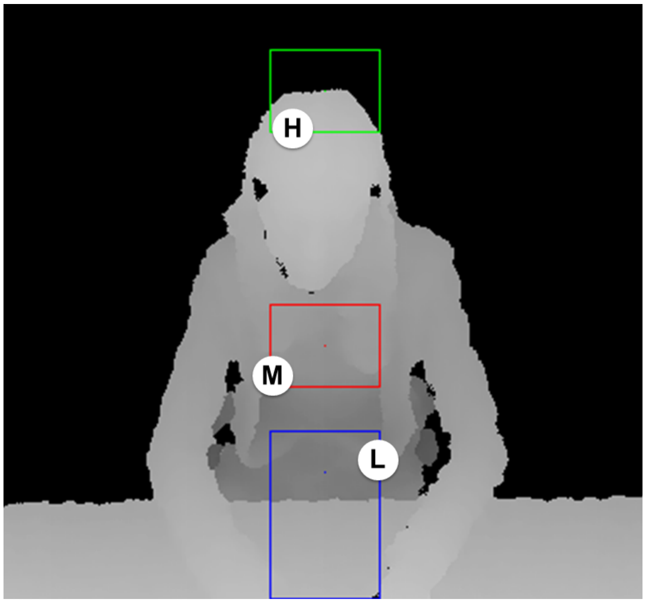


Figure 3: Output of the posture algorithm.

set is classified using the model created in training. For each utterance in the test set, we choose the branching that it should follow in the existing model according to its previous tutor dialogue act. Having chosen the branching, the average distance between the target utterance and each cluster in the clustering group is calculated, where the clustering group represents all clusters in that particular branch. The distance from the target utterance to utterances in each cluster is calculated and divided by the number of utterances in each cluster, producing one average distance to each cluster. The closest cluster which has the smallest average distance determines the target utterance's dialogue act. For instance, if the previous tutor dialogue act of the test utterance was a statement, then the utterance is modeled within clusters that shared the same previous tutor dialogue act in the training set. The process is depicted in Figure 6 where the student utterance to be classified is u_i with its posture and gesture features p_i and g_i respectively. The branching is done based on the previous tutor utterance of u_i (u_{t-1}) and the chosen branch is used for clustering u_i. Using this framework which has been shown to outperform previous state-of-the-art unsupervised dialogue act classifiers [14], the experimental results (Section 6) will demonstrate the additional benefit of using multimodal features for dialogue act classification.

Distance metric. For calculating similarities between utterances, we take word ordering into account to better capture the underlying intentions of each utterance. As an example, consider two utterances 'I should declare a variable' and 'should I declare a variable'. These two utterances have the same set of words when compared with a bag-of-words approach that does not take the order of words into account. However, the first utterance is a statement whereas the latter is a question. To distinguish them, it is necessary to take the word ordering into account. We utilize *longest common subsequence* [22], shared subsequences of not-necessarily contiguous words between utterances, to calculate the similarity between two utterances considering word ordering [14]. Unlike any distance metric that does not exploit utterance ordering information (Cosine, Euclidean, Manhattan, Jaccard), these two sentences are considered different by longest common subsequence. This discriminative power is desirable.

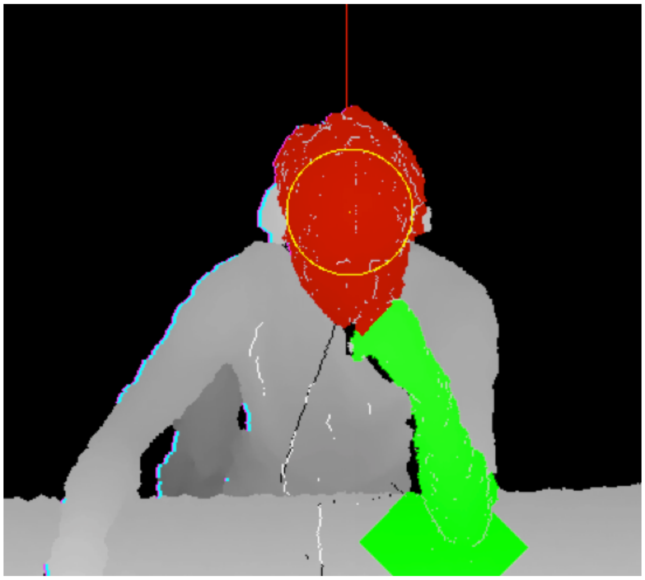


Figure 4: Output of the gesture algorithm showing the one-hand-to-face feature.

6. EXPERIMENTS

The goal of unsupervised dialogue act classification is to group together utterances with the same dialogue act. There are different techniques to accomplish this in an unsupervised way including k-means clustering [34], Dirichlet process mixture model [11] and query-likelihood clustering [13]. For this work we utilized k-medoids clustering because our prior work has established that this technique outperforms its counterparts for our corpus [14].

We hypothesized that including posture and gesture information would improve dialogue act classification performance significantly. Therefore, we conducted experiments with and without these features. We created unsupervised dialogue act classifiers that utilized posture and gesture features as well as models that did not use these features. To investigate how these models compared to each other, we compared the performance of models with the same test sets. For instance, we compared how well the utterances of a student in the test set were classified using the model having access to multimodal features and using the model that did not take this information into account. In this way, we aim to draw conclusions on the importance of multimodal features for dialogue act classification.

For testing, leave-one-student-out cross-validation was performed: for each fold, each student's utterances were either all in the test set or all in the training set, but not in both. To evaluate how well the model performed for each unseen utterance, we computed test set accuracy. Test set accuracy calculates how well the clustering model classifies the label of unseen utterances. Accepting the closest cluster as the cluster of the test utterance (as described in Section 5), the majority vote of the cluster was given as the label to the test instance. The average accuracy for the test set was computed as the number of correct classifications divided

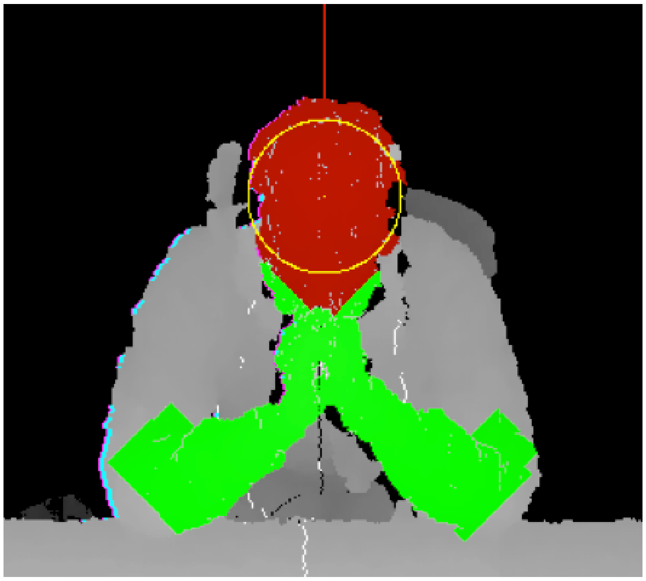


Figure 5: Output of the gesture algorithm showing the two-hands-to-face feature.

by the number of utterances in the test set. The formula for test set accuracy is as follows where n is the number of utterances in each fold of the test set and c_i is the cluster of utterance i (u_i):

$$\frac{\sum_{i=1}^{n} majority\ label\ of\ c_i = label\ of\ u_i}{n}$$

Because we applied leave-one-student-out cross-validation, we took an average of all students (folds) to report average test set accuracy. The t-tests were conducted comparing each classifiers' performance (with and without multimodal features) for each student.

7. RESULTS AND DISCUSSION

This section presents experimental results for unsupervised dialogue act classification based on multimodal features. We compared models built separately using posture and gesture features to models that did not have access to this information. Each comparison in this section was conducted with a one-tailed t-test for $n = 37$ students. The threshold for statistical reliability was taken as $p = 0.05$.

The leave-one-student-out cross-validation accuracies with respect to manual dialogue act labels were statistically significantly better with the addition of posture and gesture features ($p < 0.05$). The average accuracy for the model without using multimodal features was 61.8% ($\sigma = 2$) and this number increased to 67% ($\sigma = 1.9$) with the inclusion of multimodal features, 8% improvement. The confusion matrices for both cases are shown in Figures 7 and 8. Less frequent dialogue acts were eliminated from the confusion matrix because the model never predicted those acts ("Request for Feedback" and "Other").

The experimental results show that including posture and gesture features improved unsupervised dialogue act classification performance significantly. For the most frequent dialogue acts (statements, answers, questions and acknowledgments), only statements' classification accuracy decreased

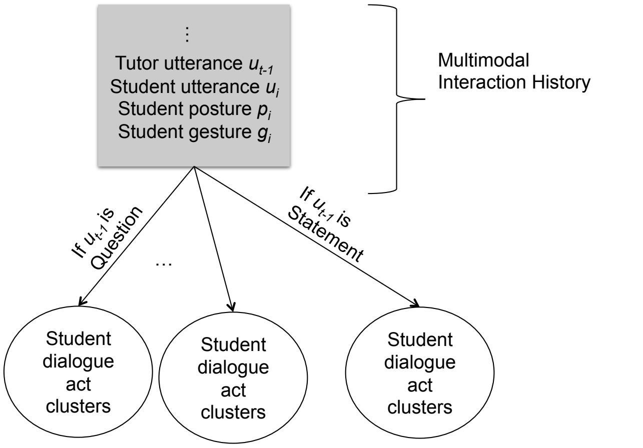

⋮
Tutor utterance u_{t-1}
Student utterance u_i
Student posture p_i
Student gesture g_i
Multimodal Interaction History
If u_{t-1} is Question
...
If u_{t-1} is Statement
Student dialogue act clusters
Student dialogue act clusters
Student dialogue act clusters

Figure 6: Branching student utterances according to previous tutor dialogue act and choosing which clustering group to use for unseen utterances.

Predicted

Manual Label

Dialogue acts	S	A	Q	ACK	C
S	161	38	24	63	2
A	22	594	2	4	0
Q	59	70	35	43	0
ACK	114	43	28	103	0
C	8	3	1	1	0

Figure 7: Confusion matrix for the model *without* posture and gesture (61.8% accuracy).

Predicted

Manual Label

Dialogue acts	S	A	Q	ACK	C
S	139	38	38	71	**2**
A	15	594	6	7	0
Q	38	69	68	32	0
ACK	58	43	34	153	0
C	7	3	0	3	0

Figure 8: Confusion matrix for the model *with* posture and gesture (67.05% accuracy).

with multimodal features. In order to gain better insights and understand which dialogue acts benefit more from multimodal features, we compared the two models qualitatively.

We observed that for distinguishing questions that are very similar to statements in structure, multimodal features are highly beneficial. In contrast, when multimodal sensors detect features that might be indicative of confusion, although students may utter statements, the models decided that they asked questions requesting help. The nature of the corpus is highly influential here: because the students are engaging in dialogue while completing a learning task, nonverbally expressed confusion may relate to the learning task and not necessarily be indicative that the student is expressing a question dialogue act. Table 3 (shown on the next page) depicts sample utterances which were incorrectly classified with the model that did not utilize posture and gesture features and were corrected with the help of multimodal features. We provide five types of corrections in the table, two of which were more frequently seen: questions misclassified as statements and questions misclassified as acknowledgements. These results show that utilizing posture and gesture features, the dialogue act classifier became more successful in distinguishing questions. Especially for utterances that were not syntactically questions such as “so the computer reads it from right to left?”, multimodal features helped enrich the information present in the utterance by incorporating information about students’ posture and gesture. For acknowledgements that were corrected from statements with the help of multimodal features, the students were closer to the workstation (according to lower torso distance) and both one-hand-to-face and two-hand-to-face gestures were present.

Table 4 shows sample utterances that caused the dialogue act classification model to be confused with the addition of multimodal features, i.e. utterances that were classified correctly with the model that did not have access to multimodal features but were incorrectly classified with the addition of these features. Most of the misclassifications caused by multimodal features were on questions. Comparison between two models showed that increase in student’s mid or lower torso depth i.e., students moving farther from the camera, or one-hand-to-face gesture detection increases the chances of the model classifying the utterance as a question because this pattern is seen in other questions as well. Therefore, even though an utterance may have a statement label, observing students moving farther from the computer triggered a question classification. That may be one of the reasons why a decrease in the accuracy of statements was observed when the multimodal features were incorporated.

Another important finding of the experiments is that when posture and gesture were used with no other features, the average cross-validated accuracy was 53.2%, whereas the majority chance baseline was 43%. This finding suggests that, even before knowing the content of an utterance, it is possible to predict the dialogue acts by analyzing multimodal features of students. This information can be especially helpful for systems that aim to provide remedial support without an explicit request from students. Being able to predict what the next dialogue act would be even before the student utters words can be a significant advantage for understanding students.

A notable limitation of the current approach is that collection of posture and gesture data is not yet a fully scalable approach. However, given the continued decrease in the cost of high-resolution video equipment, these approaches are expected to become more scalable. Additionally, work in building affect detectors suggest that it might be possible to infer these multimodal events based on streams that do not require expensive sensors [31].

Student Utterances Correctly Classified with the Help of Multimodal Features
ACK utterances misclassified as Q with multimodal features
ok so I just ask please give me your name
ok I get it now
I understand now
think I got it
I know I was trying to figure out what line it comes from
oh i see
S utterances misclassified as Q with multimodal features
just pops up in blue
closest experience I have to java is playing runexcape
sorry to take too long time, I am usually not creative at all so it usually takes long time to think about something
but I think I got it
totally guessed what I needed to do
I am not understanding
well that didn’t turn out right
and the name of the variable is the only other thing I could think of

Table 4: Sample utterances that were incorrectly classified when multimodal features were used but were correctly classified by the model that did not use posture and gesture features.

8. CONCLUSION AND FUTURE WORK

Understanding and modeling students in learning environments is a crucial step to better support learning. To this end, learning analytics approaches that mine student interactions within learning environments hold great promise. Textual analytics, a branch of learning analytics, has been well studied in the literature; however, multimodal features are only just beginning to be explored for developing rich understanding of students within tutorial dialogue. This paper has focused on investigating the extent to which multimodal features of posture and gesture during computer-mediated tutoring improve unsupervised classification of student dialogue acts. The experiments showed that incorporating multimodal features regarding posture and gesture improved the accuracy of dialogue act models significantly and that it is possible to predict the dialogue act of an upcoming utterance better than majority baseline chance before the utterance is observed. Furthermore, detailed inspection of clusters revealed information about which dialogue acts benefit more from the multimodal features. We found that some patterns in features such as higher lower and mid torso distance indicating students moving farther from the computer can confuse the

Sample Student Utterances From Clusters
ACK utterances misclassified as S without multimodal features
ok i am getting it
that makes sense
awesome thank you
ok got it
got it! I just thought it is an example
Q utterances misclassified as S without multimodal features
why is not it prompting me to enter my name?
are we going to learn how the player enters their name next? there is no box to type it in
do I have to?
just means to assign the name write?
so what happens if I do not put what the java is expecting
just comments are just a way to write notes to others to help them understand right?
how do you run it? the run button is not available to press
so the computer reads it from right to left?
name a variable first and then store the value for what I want to call it?
so the contents that come after string are what exactly? declaring a variable?
so anything that someone types in the comments box that will be used with the scanner?
would you still put the prompt after those lines of codes or should you move it up to prompt the user right after you print the game name
S utterances misclassified as ACK without multimodal features
not exactly sure how to go about this one
beautiful!
S utterances misclassified as Q without multimodal features
I guess I am not sure what the codes are currently displayed under the java code section
does not have a problem quitting
Q utterances misclassified as ACK without multimodal features
so you want me to redo number but change the adreamgame name to something else?
could I type in string the adventure quest? or would I need to put in quotes or something?
so I could have put the system.out.println command on line number and the input statement on line number and it would still work?
should I do another player input code?
spaces or no spaces?
oh so I need to insert it before the scanner player input line

Table 3: Sample utterances that were correctly classified with the help of multimodal features and their incorrect classifications by the model that did not utilize posture and gesture.

dialogue act classifier to classify questions although manual tags indicate statements. In addition, experiments showed that multimodal features are especially helpful for distinguishing questions that are very similar to statements in structure. These findings are important for understanding student utterances without needing manual annotations.

As the field moves toward richer automatic understanding of student utterances, these models will find broad application in contexts such as MOOCs and ITSs. Because the unsupervised model does not require manual labeling and the multimodal features are automatically extracted, the approach presented in this paper can be used across massive student data to understand more about whether and how students learn.

In future work, it will be important to continue to build and enrich unsupervised dialogue act classification models in order to understand better how students interact in learning environments. As multimodal data streams from learning interactions become more common, it will be important to utilize as many information sources as possible, including multimodal features to better understand students, the dynamics of learning, and therefore to provide more effective learning environments.

9. ACKNOWLEDGMENTS

The authors wish to thank the members of the Center for Educational Informatics at North Carolina State University for their helpful input. This work is supported in part by the National Science Foundation through Grant DRL-1007962, IIS-1409639 and the STARS Alliance, CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

10. REFERENCES

[1] K. E. Arnold and M. D. Pistilli. Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 267–270, 2012.

[2] P. Blikstein. Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 110–116, 2011.

[3] P. Blikstein. Multimodal learning analytics. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 102–106, 2013.

[4] D. Bohus and E. Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 5, 2010.

[5] K. E. Boyer, J. Grafsgaard, E. Y. Ha, R. Phillips, and J. C. Lester. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the International Conference of the Association for Computational Linguistics*, pages 1190–1199, 2011.

[6] K. E. Boyer, E. Y. Ha, R. Phillips, M. D. Wallis, M. A. Vouk, and J. C. Lester. Dialogue act modeling in a complex task-oriented domain. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–305. Association for Computational Linguistics, 2010.

[7] L. Chen and B. D. Eugenio. Multimodality and dialogue act classification in the RoboHelper project. In *Proceedings of the Annual Meeting of Special Interest Group on Discourse and Dialogue*, pages 183–192, 2013.

[8] M. M. Chiu and B. Hall. Statistical discourse analysis of online discussions: Informal cognition, social metacognition and knowledge creation. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 217–225, 2014.

[9] C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy. Visualizing patterns of student engagement and performance in MOOCs. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 83–92, 2014.

[10] E. Coopey, R. B. Shapiro, and E. Danahy. Collaborative spatial classification. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 138–142, 2014.

[11] N. Crook, R. Granell, and S. Pulman. Unsupervised classification of dialogue acts using a Dirichlet process mixture model. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 341–348, 2009.

[12] S. D'Mello, R. Dale, and A. Graesser. Disequilibrium in the mind, disharmony in the body. *Cognition & Emotion*, 26(2):362–374, 2012.

[13] A. Ezen-Can and K. E. Boyer. Unsupervised classification of student dialogue acts with query-likelihood clustering. In *Proceedings of the International Conference on Educational Data Mining*, pages 20–27, 2013.

[14] A. Ezen-Can and K. E. Boyer. Combining task and dialogue streams in unsupervised dialogue act models. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 113–122, 2014.

[15] A. Ezen-Can and K. E. Boyer. Toward adaptive unsupervised dialogue act classification in tutoring by gender and self-efficacy. In *Extended Proceedings of the International Conference on Educational Data Mining (EDM)*, pages 94–100, 2014.

[16] R. Ferguson, Z. Wei, Y. He, and S. B. Shum. An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 85–93, 2013.

[17] A. C. Graesser, N. K. Person, and J. P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9(6):495–522, 1995.

[18] J. F. Grafsgaard, K. E. Boyer, R. Phillips, and J. C. Lester. Modeling confusion: facial expression, task, and discourse in task-oriented tutorial dialogue. In *Proceedings of the Conference on Artificial Intelligence in Education*, pages 98–105, 2011.

[19] J. F. Grafsgaard, R. M. Fulton, K. E. Boyer, E. N. Wiebe, and J. C. Lester. Multimodal analysis of the implicit affective channel in computer-mediated textual communication. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 145–152, 2012.

[20] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester. Embodied affect in tutorial dialogue: Student gesture and posture. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 1–10, 2013.

[21] E. Y. Ha, J. F. Grafsgaard, C. M. Mitchell, K. E. Boyer, and J. C. Lester. Combining verbal and nonverbal features to overcome the 'information gap' in task-oriented dialogue. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 247–256, 2012.

[22] D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343, 1975.

[23] D. B. Jayagopi and D. Gatica-Perez. Discovering group nonverbal conversational patterns with topics. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pages 3–6, 2009.

[24] I. Jraidi, M. Chaouachi, and C. Frasson. A dynamic multimodal approach for assessing learners' interaction experience. In *Proceedings of the International Conference on Multimodal Interaction*, pages 271–278, 2013.

[25] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of the ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120, 1998.

[26] S. P. Leeman-munk, E. N. Wiebe, and J. C. Lester. Assessing elementary students' science competency with text analytics. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*,

pages 143–147, 2014.

[27] D. J. Litman and S. Pan. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2-3):111–137, 2002.

[28] D. McNeill. *Gesture and thought*. University of Chicago Press, 2008.

[29] C. M. Mitchell, E. Y. Ha, K. E. Boyer, and J. C. Lester. Learner characteristics and dialogue: Recognizing effective and student-adaptive tutorial strategies. *International Journal of Learning Technology (IJLT)*, 8(4):382–403, 2013.

[30] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the International Conference on Very Large Data Bases*, pages 144–155, 1994.

[31] L. Paquette, R. Baker, M. Sao Pedro, J. Gobert, L. Rossi, A. Nakama, and Z. Kauffman-Rogoff. Sensor-free affect detection for a simulation-based science inquiry learning environment. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, volume 8474, pages 1–10, 2014.

[32] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422, 2009.

[33] M. Rodrigo and R. Baker. Comparing learners' affect while using an intelligent tutor and an educational game. *Research and Practice in Technology Enhanced Learning*, 6(1):43–66, 2011.

[34] V. Rus, C. Moldovan, N. Niraula, and A. C. Graesser. Automated discovery of speech act categories in educational games. In *Proceedings of the International Conference on Educational Data Mining*, pages 25–32, 2012.

[35] B. Samei, H. Li, F. Keshtkar, V. Rus, and A. C. Graesser. Context-based speech act classification in intelligent tutoring systems. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, pages 236–241, 2014.

[36] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the International Conference on Human-Robot Interaction*, pages 305–311, 2011.

[37] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):443–492, 1998.

[38] V. Southavilay, K. Yacef, P. Reimann, and R. A. Calvo. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 38–47, 2013.

[39] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[40] D. R. Traum. Speech acts for dialogue agents. In *Foundations of Rational Agency*, pages 169–201. Springer, 1999.

[41] A. K. Vail and K. E. Boyer. Adapting to personality over time: Examining the effectiveness of dialogue policy progressions in task-oriented interaction. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 41–50, 2014.

[42] A. K. Vail and K. E. Boyer. Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes. In *Proceedings of the International Conference on Intelligent Tutoring Systems (ITS)*, 2014.

[43] K. Van Turnhout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 175–182, 2005.

[44] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé. When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1):3–62, 2007.

[45] J. B. Wiggins, J. F. Grafsgaard, C. M. Mitchell, K. E. Boyer, E. N. Wiebe, and J. C. Lester. Exploring the relationship between self-efficacy and the effectiveness of tutorial interactions. In *Proceedings of the 2nd Workshop on AI-supported Education for Computer Science (AIEDCS)*, pages 31–40, 2014.

[46] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek. Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 145–149, 2013.

[47] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3):129–164, 2009.

[48] M. Worsley. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 353–356, 2012.

[49] M. Worsley and P. Blikstein. Towards the development of multimodal action based assessment. In *Proceedings of the International Conference on Learning Analytics Knowledge (LAK)*, pages 94–101, 2013.