

Affect Detection from Human-Computer Dialogue with an Intelligent Tutoring System

Sidney D'Mello¹ and Art Graesser²

¹ Department of Computer Science, The University of Memphis
Memphis, TN, 38152, USA
sdmello@memphis.edu

² Department of Psychology, The University of Memphis
Memphis, TN, 38152, USA
a-graesser@memphis.edu

Abstract. We investigated the possibility of detecting affect from natural language dialogue in an attempt to endow an intelligent tutoring system, AutoTutor, with the ability to incorporate the learner's affect into its pedagogical strategies. Training and validation data were collected in a study in which college students completed a learning session with AutoTutor and subsequently affective states of the learner were identified by the learner, a peer, and two trained judges. We analyzed each of these 4 data sets with the judges' affect decisions, along with several dialogue features that were mined from AutoTutor's log files. Multiple regression analyses confirmed that dialogue features could significantly predict particular affective states (boredom, confusion, flow, and frustration). A variety of standard classifiers were applied to the dialogue features in order to assess the accuracy of discriminating between the individual affective states compared with the baseline state of neutral.

1 Introduction

An emerging trend in the development of intelligent virtual agents (IVAs) has involved the modeling of the user's affective states, with the long-term goal of delivering a more engaging, adaptive, naturalistic experience [1, 2]. Over the last few years, a particular class of IVAs, namely intelligent tutoring systems (ITSs) with animated pedagogical agents [3-5], have been designed to assist learners in the active construction of knowledge, particularly at deeper levels of comprehension. Most of these systems provide one-on-one tutoring, which is known to be a powerful method of promoting knowledge construction [6], whereas others assist individual learners with a cast of animated agents that perform different functions [4, 7].

While ITSs have typically focused on the learner's cognitive states they can be endowed with the ability to recognize, assess, and react to a learner's affective state [8-11]. There is some evidence that an affect sensitive ITS would have a positive impact on learning. For example, Kim [12] conducted a study that demonstrated that the interest and self-efficacy of a learner significantly increased when the learner was

accompanied by a pedagogical agent acting as a virtual learning companion sensitive to the learner's affect. Linnerenbrink and Pintrich [13] reported that the posttest scores of physics understanding decreased as a function of negative affect during learning. Craig et al. [14] reported that increased levels of boredom were negatively correlated with learning of computer literacy, whereas levels of confusion and the state of flow (being absorbed in the learning process, [15]) were positively correlated with learning in an AutoTutor learning environment. AutoTutor is an intelligent tutoring system that helps learners construct explanations by interacting with them in natural language and helping them use simulation environments [9, 16]. The focus of this paper is on the transformation of AutoTutor into an affect-sensitive intelligent tutoring system [17, 18].

Much of the work in affect detection involves the use of bodily sensors that monitor facial expressions [19], gross body language [20], acoustic-prosodic vocal features [21-24], and physiological measures such as heart rate monitors, electromyography, skin conductance, etc. [25, 26]. This paper investigates a less frequently explored channel, namely human-computer natural language dialogue. There have been some investigations of emotions in human-human dialogues [21, 27] and human-computer dialogues [22], but the literature on automated affect detection is sparse. The use of dialogue to detect affect in learning environments is a reasonable information source to explore, as opposed to bodily sensors, because dialogue information is abundant in virtually all conversations and is inexpensive to collect.

Perhaps the most relevant work investigating dialogue and emotions has been conducted on the program ITSPOKE [23]. ITSPOKE integrates a spoken language component into the Why2-Atlas tutoring system [28]. The spoken student dialogue turns were analyzed on the basis of lexical and acoustic features, with codings of negative, neutral or positive affect. The algorithms were able to reach high levels of accuracy in detecting affect [22]. Another interesting use of natural language dialogue for affect detection is provided by Carberry, Lambert, and Schroeder [29] who developed an algorithm to recognize doubt by examining linguistic and contextual features in conjunction with world knowledge. The major difference between these research efforts and our approach is that we are concerned with a larger set of affective states (boredom, confusion, delight, flow, frustration, neutral, and surprise) as well as a novel set of dialogue features as will be elaborated below.

We begin this paper by describing the various information channels that are tracked during interactions with AutoTutor and are stored in its text log files. Next we describe a study used to systematically gather affect judgments (from four raters) and dialogue patterns while participants interacted with AutoTutor. The data collected in this study served as training and testing data for the machine learning algorithms, with the affect judgments of each judge representing the ground truth from his or her perspective. Statistical analyses assessed which of the affective states could be predicted from the dialogue features. A variety of machine learning algorithms were then applied to the features selected by the statistical methods in an attempt to assess the reliability in automatically detecting the learner's affect from AutoTutor's dialogue. We conclude by addressing limitations of this research and presenting options to alleviate some of the known problems.

2 AutoTutor's Mixed-Initiative Dialogue

The Tutoring Research Group (TRG) at the University of Memphis developed AutoTutor, a fully automated computer tutor that simulates human tutors and holds conversations with students in natural language [9, 30]. AutoTutor attempts to comprehend the students' natural language contributions and then responds to the students' verbal input with adaptive dialogue moves similar to human tutors. AutoTutor helps students learn by presenting challenging problems (or questions) from a curriculum script and engaging in a mixed-initiative dialogue while the learner constructs an answer.

AutoTutor provides *feedback* on what the student types in (positive, neutral, or negative feedback), *pumps* the student for more information ("What else?"), *prompts* the student to fill in missing words, gives *hints*, fills in missing information with *assertions*, identifies and corrects *misconceptions* and erroneous ideas, *answers* the student's questions, and *summarizes* topics. A full answer to a question is eventually constructed during this dialogue, which normally takes between 30 and 100 turns between the student and tutor for one particular problem or main question.

AutoTutor's knowledge about the topic being tutored (computer literacy in this study) is represented by Latent Semantic Analysis (LSA) [31] and a curriculum script on the material. LSA is a statistical technique that measures the conceptual similarity of two text sources. LSA computes a geometric cosine (ranging from -1 to 1) that represents the conceptual similarity between the two text sources.

A session with AutoTutor is comprised of a set of subtopics (main questions) that cover specific areas of the main topics (hardware, internet, and operating systems). Each subtopic has an associated set of expectations, potential dialogue moves to elicit expectations (e.g., hints, prompts, assertions), misconceptions, corrections of misconceptions, and other slots in the curriculum script that need not be addressed here. The expectations are ideally covered by a series of turns in AutoTutor's conversation with the student in an attempt to construct an answer to the current main question (subtopic). When an acceptable answer with the appropriate details is gleaned from the student's responses, AutoTutor moves on to the next subtopic. At the end of each student turn, AutoTutor maintains a log file that captures the student's response, a variety of assessments of the response, the feedback provided, and the tutor's next move.

As an example, the exchange below is one of the shorter tutorial dialogues between a college student and AutoTutor. This student was an above average student in the computer literacy class.

1. AUTOTUTOR: When you turn on the computer, how is the operating system first activated and loaded into RAM? [this is the *main question*]
2. STUDENT: When the computer is turned on a file is automatically booted up.
3. AUTOTUTOR: Anything else? [this is a *pump* for more information]
4. STUDENT: This file contains the operating system.
5. AUTOTUTOR: What about storage? [this is a *hint*]
6. STUDENT: The operating system is then read into RAM and activated.

7. AUTOTUTOR: Right. The CPU executes instructions in ROM, which helps the operating system to start. [after giving *positive feedback*, there is an *assertion*]
8. STUDENT: After being read from ROM the OS is activated.
9. AUTOTUTOR: Right. Let's review what we've just gone over. The operating system must first be stored on the hard disk. When you turn on the computer, the CPU executes instructions stored in ROM. These instructions help boot the operating system. The operating system is then loaded from the hard disk into RAM. [after *positive feedback* comes a *summary*]

We now briefly describe the relevant information channels that are available in AutoTutor's log files of the interaction history. The information can broadly be divided into five categories: temporal information, response verbosity, answer quality assessments, tutor directness, and the tutor's feedback.

2.1 Temporal Information

The temporal information can be viewed as a combination of global and local temporal markers that span the period of interaction. The *subtopic number* indicates the number of main questions answered. It provides a global measure of sequential position within the entire tutorial session. For example, for a one-hour session covering three subtopics, the third subtopic would indicate that the student is approximately in the 40-60 minute time span. The *turn* on the other hand, provides a local measure of the number of student contributions to the current question (subtopic). Finally, the student response time is the elapsed time (in milliseconds rounded to seconds) between the verbal presentation of the question by AutoTutor and the student submitting an answer.

2.2 Response Verbosity

The verbosity of the student's dialogue contributions is measured by the *number of characters* in the student's response. A qualitative classification of the student's contributions is provided by AutoTutor's Speech Act Classification system [32]. While the system classifies a response into a number of categories, those of interest to this research involved *frozen expressions* (e.g., I don't know, What did you say?) (coded as -1) and topic related *contributions* (scored as a 1).

2.3 Answer Quality

AutoTutor relies on LSA as its primary computation of the quality of student contributions in student turns. The primary measure of answer quality for a given turn is the *local good score*, which measures the student's contribution for that turn on the basis of its similarity to good answers (expectations). Therefore, a high local good

score reflects progress in answering the main question. A secondary measure of answer quality is the *global good score* which involves the same assessments as the local parameters, with the exception that the text used for the LSA match is an aggregation of all of the student's turns (1 through N) for a given subtopic. With this scheme, a student's past contributions to a subtopic (main question) are considered in AutoTutor's assessment of the student's current state. Additionally, a *delta local good score* and a *delta global good score*, measures changes in the local good and global good scores respectively. These measure the changes in student answer quality.

2.4 Tutor Directness

At the end of each student turn, AutoTutor incorporates the various LSA assessments when choosing its next pedagogically appropriate dialogue move. When AutoTutor tries to get a single expectation (*E*) covered (e.g., The hard disc is a storage medium), this goal is posted and is achieved by AutoTutor presenting a series of different dialogue moves across turns until the expectation *E* is expressed. It first gives a *pump* (What else?), then a *hint* (What about the hard disk?), then a *prompt* for specific information (i.e., an important word, The hard disk is a medium of what?), and then simply *asserts* the information (The hard disc is a medium for storage). After all of the expectations for the problem are covered a *summary* is provided by AutoTutor. Given this mechanism of encouraging the student to cover the expectations, the dialogue moves chosen can be ordered on a *directness* scale (ranging from -1 to 1) on the basis of the amount of information AutoTutor supplies to the learner. The ordering is *pump* < *hint* < *prompt* < *assertion* < *summary*. A pump conveys the minimum amount of information (on the part of AutoTutor) whereas a summary conveys the most amount of explicit information.

2.5 Tutor Feedback

AutoTutor's short *feedback* (positive, neutral positive, neutral, neutral negative, negative) is manifested in its verbal content, intonation, and a host of other non-verbal conversational cues. Examples of positive and negative feedback terms include "good job", "correct" and "wrong", "no" respectively. Similar to the directness scale constructed above, AutoTutor's feedback was mapped onto a scale ranging from -1 (negative feedback) to 1 (positive feedback).

3 Empirical Data Collection

The training and testing of the emotion classifier needs a gold standard for comparison. The appropriate gold standard is undoubtedly debatable, but there needs to be some plausible foundation for establishing ground truth, even though any gold standard proposed is open to challenge. One preliminary step in this process is to

examine how reliable humans are at classification of emotions. We investigated three potential measures of ground truth for emotion detection: the participants, novice judges, and trained judges.

We conducted a study which consisted of 28 participants interacting with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, internet, or operating systems. During the interaction process a video of the participant's face and a video of the screen were recorded. The judging process was initiated by synchronizing the video streams from the screen and the face and displaying to the judge. Judges were instructed to make judgments on what affective states were present in 20-second intervals at which the video automatically paused (freeze-framed). They were also instructed to indicate any affective states that were present in between the 20-second stops.

Four sets of emotion judgments were made for the observed affective states of each participant's AutoTutor session. For the self judgments, the participant watched his or her own session with AutoTutor immediately after having interacted with AutoTutor. Second, for the peer judgments, participants returned approximately a week later to watch and judge another participant's session on the same topic in computer literacy. Finally, two additional judges (called trained judges), who had been trained on how to detect facial action units according to Paul Ekman's Facial Action Coding System (FACS) [33], judged all of the sessions separately. The trained judges also had considerable interaction experience with AutoTutor. Hence, their emotion judgments were based on contextual dialog information as well as the FACS system.

A list of the affective states and definitions was provided for all judges. The states were boredom, confusion, flow, frustration, delight, neutral and surprise. The selection of emotions was based on previous studies of AutoTutor [14, 34] that collected observational data (i.e., trained judges observing learners) and *emote aloud* protocols while college students learned with AutoTutor.

Interjudge reliability was computed using Cohen's kappa for all possible pairs of judges: self, peer, trained judge1, and trained judge2. Cohen's kappa measures the proportion of agreements between two judges with correction for baserate levels and random guessing. There were 6 possible pairs altogether. The kappa's were reported in Graesser et al. [18]: self-peer (.08), self-judge1 (.14), self-judge2 (.16), peer-judge1 (.14), peer-judge2 (.18), and judge1-judge2 (.36). These kappa scores revealed that the trained judges had the highest agreement, the self-peer pair had lowest agreement, and the other pairs of judges were in between. It should be noted, however, that the kappa scores increase substantially when we focus on observations in which the learner declares they have an emotion, as opposed to points when they are essentially neutral. The kappa scores are on par with data reported by other researchers who have assessed identification of emotions by humans [22, 24]. More details on the collection of data in this study and follow up analyses are reported in Graesser et al. [18].

4 Results and Discussion

It is essential to have real-time automatic affect detection in order to achieve the larger goal of extending AutoTutor into an affect-sensitive ITS. Therefore, we applied several standard classification techniques in an attempt to detect the learner's affect from the various conversation features manifested through an interaction with AutoTutor.

The AutoTutor log files were mined to obtain information from the various dialogue channels described above. Four data sets, corresponding to each of the four judge's emotion judgments, were obtained by extracting the set of emotion judgments for each participant (according to the judge in question) and the set of dialogue features for each turn that were associated with the emotion. More specifically, the emotion judgment that immediately followed a dialogue move (within a 15 second interval) was bound to that dialogue move. This allowed us to obtain four sets of labeled dialogue data, each containing 1300 records, aggregated across the 28 participants.

4.1 Statistical Analyses

Multiple regression analyses were conducted to determine the extent to which the seven affective states of interest could be predicted from the various dialogue features. For each of the four data sets (self, peer, trained judge1, trained judge2), seven multiple regression analyses were performed, one for each of the affective states, yielding 28 models in all. The dependent variable for each multiple regression analysis was an affective state and the independent variables were the set of dialogue features.

The multiple regression analyses for the emotion data obtained from the trained judge ratings yielded statistically significant models for the affective states of boredom ($F_{\text{self}} = 5.90, R^2_{\text{adj}} = .184; F_{\text{peer}} = 8.12, R^2_{\text{adj}} = .211; F_{\text{judge1}} = 7.92, R^2_{\text{adj}} = .132; F_{\text{judge2}} = 12.30, R^2_{\text{adj}} = .140$), confusion ($F_{\text{self}} = 7.21, R^2_{\text{adj}} = .175; F_{\text{peer}} = 1.88, R^2_{\text{adj}} = .108; F_{\text{judge1}} = 2.37, R^2_{\text{adj}} = .075; F_{\text{judge2}} = 13.73, R^2_{\text{adj}} = .125$), flow ($F_{\text{judge1}} = 14.01, R^2_{\text{adj}} = .201; F_{\text{judge2}} = 12.21, R^2_{\text{adj}} = .139$), frustration ($F_{\text{self}} = 5.75, R^2_{\text{adj}} = .188; F_{\text{peer}} = 4.53, R^2_{\text{adj}} = .106; F_{\text{judge1}} = 10.09, R^2_{\text{adj}} = .094; F_{\text{judge2}} = 6.44, R^2_{\text{adj}} = .094$) and neutral ($F_{\text{self}} = 3.03, R^2_{\text{adj}} = .335; F_{\text{peer}} = 2.93, R^2_{\text{adj}} = .291; F_{\text{judge1}} = 2.19, R^2_{\text{adj}} = .026; F_{\text{judge2}} = 4.20, R^2_{\text{adj}} = .090$); all models were significant at the $p < .05$ level and $df_1 = 11, df_2 = 1261$. For the novice judges (self and peer), statistically significant models were discovered for boredom, confusion, frustration, and neutral, but not for flow. The multiple regression analyses failed to converge on significant models for the affective states of delight and surprise, indicating that these affective states cannot be predicted from the dialogue features. The signs (+, -) of the statistically significant standardized coefficients in the multiple regression analyses are presented in Table 1.

Table 1. Significant predictors for the multiple regression models for emotions in each data set

Dialogue Features	Boredom				Confusion				Flow				Frustration				Neutral				
	SF	PR	J1	J2	SF	PR	J1	J2	SF	PR	J1	J2	SF	PR	J1	J2	SF	PR	J1	J2	
Subtopic Number	+	+	+	+	-	-	-	-	-	-			+								
Turn Number	+	+	+	+	-												-				
Response Time					+			+													
No. Characters					-		-	-		+	+										
Global Good									-					-	-						
Delta Global Good																					
Local Good				-						+				+	+						
Delta Local Good																	-				
Speech Act	-				-		-	-									+			+	
Directness				+	+			-	-												
Feedback							-			+	+	+	-	-	-		+	+	+	+	

SF: Self Judgements, PR: Peer Judgements, J1: Trained Judge1, J2: Trained Judge2
 + or - indicates that the feature is a positive or negative predictor in the multiple regression model, with a significance level of $p < .05$.

A number of generalizations can be gleaned from Table 1 regarding the relationship between dialogue and affective states. If one considers the significant predictors in which the data from at least two judges agreed, a number of relationships surface. In particular, boredom occurs later in the session (high subtopic number), after multiple attempts to answer the main question (high turn number), and when there are more direct dialogue moves (high directness). Alternatively, confusion occurs earlier in the session (low subtopic number), with slower responses (long response time), shorter responses (less characters), with frozen expressions (negatively coded speech acts), and when the tutor is less direct in providing information. The analyses indicated that flow occurs earlier on in the session (low subtopic numbers), involves longer responses (more characters), and is accompanied by positive feedback from the tutor. Frustration was prevalent with good answers towards the immediate question (high local good score), but poor answers towards the broader topic (low global good score), and negative tutor feedback.

4.2 Machine Learning Experiments

The machine learning experiments focused on these significant predictors of the affective states, thereby reducing the number of features used to train and test the classifiers. In addition to potentially increasing classification accuracy by eliminating unrelated features, this feature selection procedure also offers significant computational advantages in terms of execution time, a crucial requirement for real time computation.

The Waikato Environment for Knowledge Analysis [35] was used to comparatively evaluate the performance of various standard classification techniques

in an attempt to detect affect from dialogue. The classification algorithms tested were a Naïve Bayesian classifier, a multilayer perceptron (neural network using back propagation for training), a nearest neighbor classifier, C4.5 decision trees, an additive logistic boosting classifier with a decision stump as the base learner, and support vector machines.

The classification process proceeded in two phases. In the first stage we grouped the four affective states of interest (boredom, confusion, flow, frustration) together and assessed the reliabilities of the various classification algorithms to discriminate among each affective state. In the second phase of the classification analyses, we were interested in the accuracies of detecting each of the four affective states from the base state of neutral.

4.2.1 Discriminating between Boredom, Confusion, Flow, and Frustration

The first set of classification experiments involved evaluating the classifiers on the four data sets (one for each judge’s ratings) in discriminating between boredom, confusion, flow, and frustration. To establish a uniform baseline (a chance value of 25%), we randomly sampled an equal number of observations from each affective state category. This process was repeated for 100 iterations and the reported reliability statistics were averaged across these 100 iterations. Each randomly sampled data set was evaluated on the 6 classification algorithms using k-fold cross-validation (k = 10). The dialogue features that were significant predictors of the multiple regression models listed in Table 1 were used for classification. The classification accuracies are presented in Table 2.

Table 2. Comparison of various classification techniques to detect learner’s affect

Classification Algorithm	Self		Peer		Judge 1		Judge 2	
	Acc	Kap	Acc	Kap	Acc	Kap	Acc	Kap
Additive Logistic Regression	35.0	.134	36.1	.147	47.0	.293	47.1	.295
Multilayer Perceptron	34.8	.130	34.7	.130	45.9	.278	45.2	.269
Naïve Bayes	35.3	.137	35.7	.142	45.9	.279	46.2	.282
Nearest Neighbor	28.7	.050	31.7	.089	40.7	.209	40.8	.210
C4.5 Decision Tree	31.1	.081	33.9	.119	42.0	.226	40.6	.208
Support Vector Machines	35.8	.144	36.9	.159	49.1	.321	48.4	.312

Acc: Classification accuracy (%), Kap: Cohen’s Kappa. Baseline rate (chance) is 25%.

The various classification algorithms were moderately successful in detecting affect, with the highest performance being 49.1%, a 96.4% improvement over the baseline. This was obtained from the affective judgments of trained judge1, which had a kappa score of .321 and was comparable to inter-judge reliability scores achieved by actual human coders. For example, Litman and Forbes-Riley [22] report kappa scores of around .4 in detecting positive, negative, and neutral affect. Shafran, Riley, and Mohri [24] report kappa scores ranging from .32 to .42 in coding affect. Additionally, this kappa value is on par with the kappa scores reported earlier ([18]) for the trained judges (kappa = .36).

The classification accuracies on the data based on the two trained judge’s ratings were on par and quantitatively higher than the accuracies in detecting affect based on the emotion ratings of the self and the peer. This trend is similar to that observed in the human judgments of emotions reported in Graesser et al. [18].

In order to assess class level accuracies, Table 3 lists the precision, recall, and F-measure scores obtained for the four affective states. The precision for class C is the proportion of samples that truly belong to class C among all the samples that were classified as class C. The recall score (sensitivity or true positive rate) provides a measure of the accuracy of the learning scheme in detecting a particular class. The F-measure provides a single metric of performance by combining the precision and recall. Since support vector machines constituted the most successful classifier the precision, recall, and F-measure scores presented in Table 3 are restricted to those obtained with this classifier.

Table 3. Detailed accuracies of the support vector machine classifier for each data set

Affective States	Self			Peer			Judge 1			Judge 2		
	PR	RC	FM	PR	RC	FM	PR	RC	FM	PR	RC	FM
Boredom	.40	.25	.31	.43	.30	.35	.51	.34	.40	.42	.24	.30
Confusion	.39	.33	.36	.33	.12	.17	.45	.29	.35	.52	.37	.42
Flow	.36	.27	.31	.37	.36	.36	.58	.59	.59	.56	.58	.56
Frustration	.33	.58	.42	.36	.70	.47	.45	.74	.56	.46	.76	.57

PR: Precision, RC: Recall, FM: F-Measure

When the same evaluation procedures were conducted on the affect data of the novice judges (self and peer), the F-measure indicated that the classifier was more successful in detecting frustration than the other three affective states. If one considers data from the self reports alone, classification accuracies for boredom and flow are identical and lower than that of confusion. The same trend is observed on the data from the peer judgments, with the exception that the F-measure for confusion was relatively low (.17). For the data collected from the trained judges’ identification of emotions, classification accuracies for frustration and flow were similar and quantitatively higher than those for boredom and confusion. On the basis of these results, we conclude that support vector machines offer reasonable accuracies in automatically discriminating between frustration, boredom, confusion and flow.

4.2.2 Discriminating between the Affective States and Neutral

Another important requirement for an emotion classifier is the ability to detect individual affective states from a baseline state of neutral. Therefore, additional analyses were conducted that assessed the reliability in detecting each of the four affective states (boredom, confusion, flow, and frustration) when compared to the neutral state. These analyses were conducted on each of the four data sets (self, peer, trained judge1, trained judge2). The classification procedures were similar to the random selection procedure described above. Table 4 presents overall classification

accuracies and Kappa scores for the classifier that yielded the best performance in detecting each of the four affective states from neutral.

Table 4. Classification accuracies in individually detecting boredom, confusion, flow, and frustration from neutral.

Affective States	Self		Peer		Judge 1		Judge 2	
	Acc	Kap	Acc	Kap	Acc	Kap	Acc	Kap
Boredom	61.3	.226	60.3	.206	62.5	.251	60.8	.216
Confusion	59.3	.187	58.1	.162	59.4	.188	61.0	.221
Flow	50.2	.003	53.5	.070	65.9	.319	63.8	.277
Frustration	62.1	.241	64.6	.292	71.8	.435	73.3	.466

Acc: Classification accuracy (%), Kap: Cohen’s Kappa. Baseline rate (chance) is 50%.

The reliabilities of the various classification algorithms in discriminating each of the four affective states from neutral followed a similar trend when the four affective states were considered together (Table 3), with the highest accuracies achieved in detecting frustration from neutral. Classification accuracies for the detection of boredom and confusion were moderate and comparable across the data sets provided by each of the four judges. The classifiers, when operating on the data set consisting of the trained judges’ emotion ratings, were quite successful in discriminating between flow and the baseline state of neutral. However, classifiers trained on emotion judgments of the novice judges failed to detect flow from neutral, with classification accuracies hovering around the chance rate.

5 Conclusion

Emotion measurement is a field resonating with murky, noisy, and incomplete data compounded with individual differences in experiencing and expressing emotions. On the basis of the natural language dialogue features alone, our results indicate that the standard classifiers were moderately successful in discriminating the affective states of boredom, confusion, flow, and frustration from each other, as well as from the base line state of neutral. A comparison of the accuracies obtained from the four human judges (self, peer, and 2 trained judges) revealed that classification models constructed on the basis of the trained judges’ emotion judgments consistently outperformed those of the novice judges. This trend is consistent with the inter-judge reliability results reported by Graesser et al. [18], thus offering convergent validity for the phenomenon that trained judges are better than untrained peers in detecting emotions. However, it is still not firmly established whether the trained judges or the self judgments are closer to the ground truth.

The reliability of the standard classifiers in detecting affect from dialogues validates any future efforts in pursuing more sophisticated classification techniques. For example, biologically motivated classifiers, based on the dynamic behaviors of neural populations involved in the olfaction processes of rabbits, have been

experimentally validated as powerful pattern classifiers for difficult, non-linearly separable, classification problems [36]. Other options to boost the accuracy of AutoTutor in modeling learner affect involve the use of bodily sensors that track facial features, posture patterns, and speech contours [17].

One of the known limitations of the data analyses presented in this paper is that each emotion judgment was analyzed only in the context of the immediately preceding turns of the student and tutor. Perhaps classification accuracies could be boosted by incorporating a broader scope of contextual information, including patterns of conversation that evolve over a series of turns leading up to an emotional experience. Future efforts will be directed towards the analysis of conversation features across a larger temporal resolution and number of turns.

The dialogue channels were unable to detect the affective states of delight and surprise. Perhaps these affective states are simply not manifested through AutoTutor's conversation features and their detection would require more sophisticated sensors. Delight and surprise are affective states that are generally expressed through animated facial features, so it may be possible to detect these states by means of the Facial Action Coding System; particular facial actions are known to be correlated with happiness (similar to delight) and surprise [33].

We conclude by speculating on the generalizability of the discovered relationships between the conversational cues and affective states. Although the features of dialogue we analyzed were specific to AutoTutor, a similar set of features would presumably be expected in any intelligent tutoring system, particularly in those that advocate deeper learning. The lower level features specific to AutoTutor (local good score, global good score, directness, etc.) can be generalized to generic categories of dialogue features, such as temporal assessments, response verbosity, student ability, tutor directness, and tutor feedback. We predict that these broad categories will replicate across most intelligent tutoring systems.

Acknowledgements

We would like to acknowledge our colleagues from the Emotive Computing Group <<http://emotion.autotutor.org>> at the University of Memphis including Patrick Chipman, Scotty Craig, Stan Franklin, Barry Gholson, Brandon King, Bethany McDaniel, Jeremiah Sullins, Kristy Tapp, and Amy Witherspoon for their valuable contributions to this research. We also thank our partners at the Affective Computing Research Group at MIT <<http://affect.media.mit.edu>>.

This research was supported by the National Science Foundation (REC 0106965 and ITR 0325428) and the DoD Multidisciplinary University Research Initiative administered by ONR under grant N00014-00-1-0600. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, DoD, or ONR.

References

1. Morgado, L. and Gaspar, G.: Emotion in Intelligent Virtual Agents: The Flow Model of Emotion. *Lecture Notes in Computer Science*, Vol. 2792. (2003) 31-38
2. Picard, R.W.: *Affective Computing*. MIT Press, Cambridge (1997)
3. Graesser, A., VanLehn, K., Rosé, C., Jordan, P., Harter, D.: Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine*, Vol. 22. (2001) 39-51
4. Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., Badler, N.: Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, Vol. 17. (2002) 54-63
5. Johnson, L.: Pedagogical Agent Research at CARTE. *AI Magazine*, Vol. 22. (2001) 85-94
6. Cohen, P.A., Kulik, J.A., Kulik, C.C.: Educational Outcomes of Tutoring: A Metaanalysis of Findings. *American Educational Research Journal*, Vol. 19. (1982) 237-248
7. McNamara, D.S., Levinstein, I.B., Boonthum, C.: iSTART: Interactive Strategy Trainer for Active Reading and Thinking. *Behavioral Research Methods, Instruments, and Computers*, Vol. 36. (2004) 222-233
8. Guhe, M., Gray, W.D., Schoelles, M.J., Ji, Q.: Towards an Affective Cognitive Architecture. In: Poster Session Presented at the Cognitive Science Conference. (2004)
9. Graesser, A.C., Chipman, P., Haynes, B., Olney, A.: AutoTutor: An Intelligent Tutoring System with Mixed-initiative Dialogue. *IEEE Transactions in Education*, Vol. 48. (2005) 612-618
10. Lepper, M.R. and Chabay, R.W.: Socializing the Intelligent Tutor: Bringing Empathy to Computer Tutors. In: *In Learning Issues for Intelligent Tutoring Systems*. (1988) 242-257
11. Lepper, M.R. and Woolverton, M.: The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. In: *Improving Academic Achievement: Impact of Psychological Factors on Education*. (2002) 135-158
12. Kim, Y.: Empathetic Virtual Peers Enhanced Learner Interest and Self-efficacy. Workshop on Motivation and Affect in Educational Software. In: 12th International Conference on Artificial Intelligence in Education. (2005)
13. Linnenbrink, E.A. and Pintrich, P.R.: The Role of Motivational Beliefs in Conceptual Change. In: *Reconsidering Conceptual Change: Issues in Theory and Practice*. (2002) 115-135.
14. Craig, S.D., Graesser, A.C., Sullins, J., Gholson, B.: Affect and Learning: An Exploratory Look into the Role of Affect in Learning. *Journal of Educational Media*, Vol. 29. (2004) 241-250
15. Csikszentmihalyi, M.: *Flow: The Psychology of Optimal Experience*. Harper-Row, New York (1990)
16. Graesser, A.C., Person, N., Harter, D., Tutoring Research Group.: Teaching Tactics and Dialogue in AutoTutor. *International Journal of Artificial Intelligence in Education*, Vol. 12. (2001) 257-279
17. D'Mello, S.K., Craig, S.D., Gholson, B., Franklin, S., Picard, R., Graesser, A.C.: Integrating Affect Sensors in an Intelligent Tutoring System. in *Affective Interactions: The Computer in the Affective Loop* In: Workshop at 2005 International Conference on Intelligent User Interfaces. (2005) 7-13
18. Graesser A.C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., Gholson, B.: Detection of Emotions During Learning with AutoTutor. In: 28th Annual Conference of the Cognitive Science Society, CogSci2006. In Press
19. Cohn, J.F. and Kanade, T.: Use of Automated Facial Image Analysis for Measurement of Emotion Expression. In: Coan, J.A., and Allen, J.B. (eds.): *The Handbook of Emotion Elicitation and Assessment*. Oxford University Press Series in Affective Science, New York Oxford. In Press

20. Mota, S. and Picard, R.W.: Automated Posture Analysis for Detecting Learner's Interest Level. In: Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI. (2003)
21. Forbes-Riley, K. and Litman, D.: Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. In: Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics, HLT/NAACL. (2004)
22. Litman, D.J. and Forbes-Riley, K.: Predicting Student Emotions in Computer-Human Tutoring Dialogues. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. (2004) 352-359
23. Litman, D.J. and Silliman, S.: ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In: Proceedings of the Human Language Technology Conference: 3rd Meeting of the North American Chapter of the Association of Computational Linguistics. (2004) 52-54
24. Shafran, I., Riley, M., Mohri, M.: Voice Signatures. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. (2003)
25. Nakasone, A., Prendinger, H., Ishizuka, M.: Emotion Recognition from Electromyography and Skin Conductance. In: Fifth International Workshop on Biosignal Interpretation. (2005) 219-222
26. Rani, P., Sarkar, N., Smith, C.A.: An affect-sensitive Human-Robot Cooperation: Theory and Experiments In: Proceedings of the IEEE Conference on Robotics and Automation. (2003) 2382-2387
27. Alm, C.O. and Sproat, R.: Perceptions of Emotions in Expressive Storytelling. In: InterSpeech (2005) 533-536
28. VanLehn, K., Jordan, P., Rosé, C.P., Bhembe, D., Bottner, M., Gaydos, A., et al.: The Architecture of Why2-atlas: A Coach for Qualitative Physics Essay Writing. In: Proceedings of the Sixth International Conference on Intelligent Tutoring (2002) 403-449
29. Carberry, S., Schroeder, L., Lambert, L.: Toward Recognizing and Conveying an Attitude of Doubt via Natural Language. Applied Artificial Intelligence, Vol. 16. (2002) 495-517
30. Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., Tutoring Research Group.: AutoTutor: A Simulation of a Human Tutor. Journal of Cognitive Systems Research, Vol. 1 (1999) 35-51
31. Landauer, T.K. and Dumais, S.T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, Vol. 104. (1997) 211-240
32. Olney, A., Louwse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., Graesser, A.: Utterance Classification in AutoTutor. In: Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing. (2003) 1-8.
33. Ekman, P. and Friesen, W.V.: The Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
34. D'Mello, S.K., Craig, S.D., Sullins, J., Graesser, A.C.: Predicting Affective States through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. International Journal of Artificial Intelligence in Education, Vol. 16. (2006) 3-28
35. Witten, I.H. and Frank E.: Data Mining: Practical Machine Learning Tools and Techniques. 3rd edn. Morgan Kaufmann, San Francisco (2005)
36. Kozma, R. and Freeman, W.J.: Chaotic Resonance: Methods and Applications for Robust Classification of Noisy and Variable Patterns. International Journal of Bifurcation and Chaos, Vol. 11 (2001) 1607-1629