Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs

Hua Ai, Diane J. Litman*, Kate Forbes-Riley', Mihai Rotaru", Joel Tetreault', Amruta Purandare**

Intelligent Systems Program*, Learning Research and Development Center', Dept. of Computer Science"
University of Pittsburgh, 210 S. Bouquet, Pittsburgh, PA, 15260, USA
{hua3,dlitman, forbesk, mir25, tetreaul, adp22}@pitt.edu

Abstract

In this study, we incorporate automatically obtained system/user performance features into machine learning experiments to detect student emotion in computer tutoring dialogs. Our results show a relative improvement of 2.7% on classification accuracy and 8.08% on Kappa over using standard lexical, prosodic, sequential, and identification features. This level of improvement is comparable to the performance improvement shown in previous studies by applying dialog acts or lexical-/prosodic-/discourse- level contextual features.
Index Terms: emotional speech, emotion detection, spoken dialog systems

1. Introduction

Emotion detection has been gaining increasing attention in spoken dialog systems. Information providing dialog systems use emotion detection to discover the problematic points in a conversation automatically so that the conversation can be passed onto a human operator at the appropriate time [1]. Equally, emotion detection is also important in intelligent spoken tutoring systems, as being able to detect and adapt to student emotions is considered to be an important strategy for closing the performance gap between human and computer tutors [2].

Previous research in emotion detection uses lexical and prosodic features as a basis, but has also shown the utility of incorporating other features. Identification features [3], dialog acts [4][5] and different levels of contextual features[1][6] are explored in different studies. Most of these features derive useful information from the dialog itself. For instance, dialog acts highlight the function an utterance plays within the context of a dialog, whereas contextual features model the phenomena in the larger structure that the current user turn is embedded in. However, in an application-oriented spoken dialog system where the user and the system complete some specific task together, we believe that user emotions are not only impacted by the factors that come directly from the dialog, but also by the progress of the task, which can be measured by metrics representing system and user performance. [7] and [8] both show that in spoken tutoring dialog systems, user emotions are strongly correlated with the system's performance. Other studies [9] in educational psychology point out that student emotions can impact their performance and learning, suggesting that student performance features can also be used as indicators of student emotion. In this paper, we augment the standard features used in most emotion prediction tasks in spoken dialog systems with system/user performance features to help inform student emotion classification in our tutoring dialogs.

2. Related Work

One basic question that has to be answered by research on emotion detection is which emotions to detect. Some earlier research [10] in automatic emotion detection attempts to detect full-blown emotions in non-naturally occurring speech, which is also devoid of context. However, research in the field of emotion detection in spoken dialog systems deals with the naturally occurring emotions of actual system users. Because it is harder to detect emotion in a more realistic setting [1], many studies collapse emotions into simpler classifications, such as two/three-way distinctions, to improve annotation and classification accuracies. In our study, we look at a three-way distinction of student emotion (certainty) which is of interest in recent tutorial dialog research [2].

Previous studies show that predictive models of emotion distinctions can be developed using features that are generally available in spoken dialog systems, typically lexical and prosodic features [3]. In order to fully utilize the contextual nature that the dialog structure provides, dialog act and other contextual information have also been used for emotion detection. [4][5][6] all report relative improvements in emotion detection accuracy (1-4%) when incorporating the dialog act of a user turn as a feature. [6] apply contextual information in lexical, prosodic, and discourse level and increase emotion classification accuracy by relatively 2.6%. As these previous studies mostly concentrate on mining information within the dialog itself, in this study we try to explore whether using system/user performance features can also improve emotion detection accuracy.

3. Corpus Description and Annotation

Our data consists of 100 dialogs (2252 student turns and 2854 tutor turns) between 20 students and ITSPOKE [11], a spoken dialog tutor built on top of the Why2-Atlas conceptual physics text-based tutoring system [12]. In ITSPOKE, a student first types an essay answering a qualitative physics question. A tutoring dialog is initiated by ITSPOKE after analyzing the essay to correct misconceptions and to elicit further explanations. The tutoring dialog consists of several sub-topic discussions which are manually authored by a human. After the dialog, the student revises the essay, thereby ending the tutoring

or causing another round of tutoring/essay revision. Therefore, in our system, essay revision is viewed as a sign of progress in the tutoring session. An excerpt of a tutoring dialog is shown in Figure 1.

$Tutor_1$: <Essay>	Read the question and enter your response in the essay window.
$Student_1$:	(***Types in and submits the essay***)
$Tutor_2$: <Velocity>	… So let's begin by describing the velocity of the pumpkin while the man is carrying it. How does its velocity compare to his?
$Student_2$:	They're the same (Correct)
$Tutor_3$: <Force>	Right. … What are the forces exerted on the pumpkin during the toss?
$Student_3$:	Gravity, vertically down. (Partially Correct)
$Tutor_4$: <Net force>	That's almost right. … So what is the direction of the NET force?
$Student_4$:	Zero. (Incorrect) (***… Dialog goes on…***) (***after 10 minute 42 second …***)
$Tutor_{15}$: <Sum_up>	Yep. To summarize: … Please give a try at writing the essay now.
$Student_{15}$:	(***Submits another essay***)
$Tutor_{16}$:	There are more points to include in your essay. What is the relationship between … (***Another tutoring dialog goes on …***)

Figure 1: Sample dialog from ITSPOKE

Prior to the present study, each turn in our corpus was manually annotated for certainness, using a scheme which has been applied to a comparable human-human tutoring corpus [13][14]. The annotators are instructed to tag a student emotion based on their human intuition. For example, if a student seems to be certain about his/her answer, the student turn is tagged as "certain". Four tags are defined in the coding manual: uncertain, certain, mixed (mixture of certain and uncertain), and neutral. One annotator from our group annotated the whole corpus using the four way distinction, while another colleague at Columbia University annotated it using a binary distinction of uncertain and not-uncertain. We combine our tags by collapsing "mixed" and "uncertain" into "uncertain" and "certain" and "neutral" into "not-uncertain" to compare with their annotation. A Kappa of 0.68 is obtained based on this binary distinction.

4. Emotion Classification

Our experiments apply a machine learning algorithm in WEKA [15] to all the student turns in our corpus to detect the emotion conveyed in each student turn. In our previous study [3], the AdaBoost J48 Decision tree gave the best emotion classification accuracy. Thus, we here use the boosted decision tree in all our classification experiments. All the experiments reported here are the results using 10-fold cross validation as provided by WEKA.

We perform a three-way certainty classification in this study. As our long-term goal is to trigger the intelligent tutoring spoken dialog system to automatically predict and adapt to student certainty, we collapse "mixed" into "uncertain" so that only the hypothesized useful triggering mechanisms certain-uncertain-neutral is kept. Some previous works [1] [3] use this same strategy as well.

In our classification experiments, each student turn is characterized by a set of features described in the following subsections. All the features we use can be obtained automatically in the running system. We first describe the features which have been used in our previous study on a smaller set of data (451 student turns) from our system, and then we introduce the new system/user performance features which we use in this study.

4.1. Previously used features

In this study, the only lexical feature (**LEX**) used is the Automatic Speech Recognizer (ASR) recognized student utterance (treated as bag of words) which is available from the system logs. We use the same set of automatically extracted/computed prosodic features (**PROS**) as our previous study [3] to represent the knowledge of pitch, energy, duration, tempo and pausing. The set of features consists of 12 raw prosodic features, 12 additional features that are gained by normalizing those raw features to the first dialog turn, and another 24 running totals and averages of first two sets of features.

We also take into account the gender, subject ID and problem ID as identification features (**ID**). Prior studies have shown that "subject" and "gender" features can play an important role in the emotion recognition. "subject ID" and "problem ID" are uniquely important in our domain since the student will use the system repeatedly and the problem will be discussed repeatedly across students. In our prior study, these 3 identification features were found to be useful.

In addition, we include turn sequence number (e.g. in Figure 1, the sequence number of the turn $Student_3$ is 3), turn beginning and ending time (in seconds) as sequential features (**SQ**), which are used in [5] as well as in our prior study [3].

4.2. System/user performance features

In this study, we also use features that characterize system/user performance (**PER**) to help inform the emotion detection of the present user turn.

The system performance is measured by ASR performance features, which are found to be significantly correlated with student emotional states in our system [8]. In this study, we only use the ASR performance features that can be automatically obtained from the system logs: the number of user turns that encounter ASR rejection errors and the percentage of such turns across the dialog so far.

11 features are used to measure the student performance. As we described in section 3, the discussion of each physics problem consists of several sub-topics (Shown in "<>" in Figure 1, e.g., sub-topic for the 2nd tutor/student turn is velocity). The sub-topics that the student discusses are used to measure the student performance because the tutor chooses to cover different sub-topics based on the student performance. We also count the times that a sub-topic is revisited in the dialog. These revisit counts serve as an indicator of the student performance since a sub-topic is revisited only if the student previously answers the problem incorrectly [16].

We further analyze the progression of our tutoring dialogs by investigating the transitions between sub-topics. In our system, sub-topics are embedded in a nested hierarchical structure, as in the Grosz and Sidner theory of discourse structure [17]. In order to complete a higher-level sub-topic, several lower-level sub-topics will be nested into the discussion. We automatically compute the "depth" of the sub-topics within the nesting structure and use it as a feature to indicate the current depth of the tutoring discussion (e.g., in Figure 1, sub-topic "net force" is nested in "force"; the depth for "force" is 1, and the depth for "net force" is 2). The average nesting depth so far is also computed as a feature.

We use the number of essay revisions to measure the progress of tutoring dialogs. As we described in section 3, in our system, the success of each sub-topic discussion is marked by an essay revision. More essay revisions indicate that the tutoring dialog covers a larger number of sub-topics in the discussion, thus is nearer to the end of the dialog. This feature is found to be useful in modeling user satisfaction and learning using a PARADISE framework [18].

The quality of each student's answer is measured by the correctness (correct/incorrect/partially correct) of the current student answer (e.g., in Figure 1, the 2nd student turn is "correct"), the percentage of correct student answers so far, and the times that the physics keywords appear in a student utterance (e.g., in Figure 1, the 3rd student turn contains one physics keyword, "Gravity"). Currently we automatically extract the keywords using an online physics dictionary from "Eric Weisstein's World of Physics" (http://scienceworld.wolfram.com/physics). Our ongoing work shows that the keyword counts are correlated with student learning, which is an important feature in evaluating system/user performance for tutoring systems.

Another two features are used to represent student prior domain knowledge: the pretest score on physics problems before interacting with the system and the quality of the student's first answer. We use a binary judgment (high/low) on the quality of the student's first essay based on the system's choice of first sub-topic. This information can be extracted automatically from the system log. The student's prior knowledge levels can possibly influence the student performances in tutoring [18].

5. Results

Table 1 summarizes our results using different combinations of features described in section 4 to automatically classify student certainty. The first column shows the features that are added in each run of our experiments. The 2nd and the 4th columns show the classification accuracy and Kappa, while the 3rd and the 5th columns show the relative improvements on the accuracy and the Kappa over the prior features. We report Kappa in addition to the overall classification accuracy because Kappa also considers inter-class agreement, which provides us a more comprehensive evaluation on the performance of the classifier.

The baseline performance in the first row represents classification using the majority class. Without using any features, we get an accuracy of 42.14% by always guessing "certain" for each student turn. Then we start adding more features. First, we apply the used features from our previous study to this larger corpus. The second row shows that by using lexical features only, the classification accuracy is 53.02%, and the Kappa is 0.24. By also adding prosodic features we obtain an accuracy of 56.08% and a Kappa of 0.31 in the third row. Similar to previous studies, we see a further increase on both the accuracy and the Kappa by adding the identification and sequential features. Finally, when adding system/user performance features, the classification accuracy increases to 59.41% and the Kappa increases to 0.36. A relative 2.7% improvement in the accuracy and 8% improvement in the Kappa are observed over the standard set of lexical, prosodic, sequential and identification features.

Features	Accuracy	Δ*Accuracy*	Kappa	Δ*Kappa*
Baseline	42.14%		0	
+LEX	53.02%	25.82%	0.24	
+PROS	56.08%	5.78%	0.31	25.80%
+ID, SQ	57.86%	3.17%	0.33	8.96%
+PER	59.41%	2.69%	0.36	8.08%

Table 1: Classification accuracy of student certainness given different feature sets

Class	Precision	Recall	F-measure
certain	0.62	0.673	0.645
uncertain	0.681	0.534	0.598
neutral	0.527	0.538	0.533

Table 2: Detailed accuracy by class when using all the features

In order to gain a sense of how useful the performance features are among all the features, we investigate the decision tree given by WEKA. All the performance features, except the raw count of ASR rejection turns, are used as decision nodes in the tree. Number of revisited sub-topics, correctness of student answers, depth of sub-topics, pretest score, and percentage of ASR rejections all appear in the upper ½ levels in the tree, which are usually considered to be more informative features in a decision tree comparing to those nodes in the lower levels. The tree nodes consist of mainly PROS, PER and ID features, with some LEX and other features in the lower-level nodes.

Table 2 shows the detailed accuracy by class when using all the features. We observe that the machine learning algorithm detects about 67% of the cases in which the students are certain about their answers, but only half of the cases in which they show uncertainty. Nevertheless, the precision of detection on uncertainty is slightly better than on certainty.

Whereas the overall performance of our features still shows a large room for improvement, our results demonstrate that the new performance features help to increase both the accuracy and the Kappa by a similar scope comparing to the previous studies, in which dialog acts [4][5] or lexical-/prosodic-/discourse level contextual features [6] are applied to improve emotion classification accuracy over using the standard lexical, prosodic, sequential, and identification features. In addition, while some of those previously used features are manually annotated, all the performance features we use here can be automatically extracted from a running system.

6. Discussion and Future Directions

In this study we investigate the impact of system/user performance features on student emotion detection in computer tutoring dialogs. We observe that by adding new system/user performance features to the standard feature set of prosodic,

lexical, sequential and identification features, we increase the classification accuracy by relatively 2.7% and Kappa by relatively 8.08%, which is comparable to the performance gain by adding dialog acts or lexical-/prosodic-/discourse level contextual features in previous studies.

Although our experiments are on tutoring dialogs, the system/user performance features can be generalized to information providing dialog systems. For example, we use number of essay submissions to indicate the progress in tutoring dialogs. In flight booking dialog systems, if we consider departure city, departure time and destination city as 3 slots that have to be filled in for the system to perform any kind of queries, then the progress of the dialog can be measured by the number of "slots" that have been filled in, assuming that more information the user has provided, the more likely the dialog is approaching the end. Similarly, user past experience with dialog systems may substitute for student prior knowledge on physics which is considered as a potential impacting factor of the user performance. We use the correctness of student answers to assess the student performance directly. Although in information providing systems, user utterances may not be measured as correct or incorrect, it is possible to judge whether the user is good at talking to the dialog system by looking into the style of user language. For example, if the user uses simple sentences which are basically keywords, it might be inferred that the user knows how to talk to a dialog system efficiently or the user does not have a high expectation on the system's understanding ability. In this case, the user is less likely to experience frustration or get angry.

In the future, we intend to incorporate our new system/user performance features in experiments studying student emotion detection in human-human tutoring dialogs, which is shown by the previous work [13] to be an easier task than the emotion detection on human-computer tutoring dialogs. As the best triggering mechanism for allowing the computer tutor to adapt to student certainty is still an open question, we are also interested in exploring other emotion classifications such as uncertain/not-uncertain and certain/not-certain.

7. Acknowledgements

This work is supported by NSF Awards #0328431 and #0325054.

8. References

[1] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E., "How to find Trouble in Communication," Speech Communication, vol. 40 (1-2), 2003.

[2] Bhatt, K., Evens, M., and Argamon, S, "Hedged responses and expressions of affect in human/human and human/computer tutorial interactions", In Proc. Cognitive Science, 2004.

[3] Litman, D. and Forbes-Riley, K., "Predicting Student Emotions in Computer-Human Tutoring Dialogues", in Proc. ACL, 2004.

[4] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A., "Prosody-based automatic detection of annoyance and frustration in human-computer dialog", in Proc. ICSLP, 2002.

[5] Lee, C. M. and Narayanan, S., "Towards detecting emotions in spoken dialogs", IEEE Transactions on Speech and Audio Processing 13(2), 2005.

[6] Liscombe, J., Riccardi, G., and Hakkani-Tür, D., "Using Context to Improve Emotion Detection in Spoken Dialogue Systems", in Proc. Interspeech, 2005.

[7] D'Mello, S. K., Craig, S. D., Sullins, J., and Graesser, A. C., "Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue", IJAIED, in press, 2005.

[8] Rotaru, M. and Litman, J., "Dependecies between student state and speech recognition problems in spoken tutoring dialogs", Coling/ACL, 2006.

[9] Pekrun, R., Goetz, T., Titz, W., and Perry., R. P., "Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research", Educational Psychologist, 37(2):91-105, 2002.

[10] Oudeyer, P., "Novel useful features and algorithms for the recognition of emotions in human speech", in Proc. Speech Prosody, 2002.

[11] Litman, D. and Silliman, S., "ITSPOKE: An Intelligent Tutoring Spoken Dialogue System", in Proc. of the HLT/NAACL, 2004.

[12] VanLehn, K., Jordan, P. W., Ros´e, C. P., Bhembe, D., B¨ottner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., Srivastava, R., and Wilson, R., "The architecture of Why2-Atlas: A coach for qualitative physics essay writing", in Proc. Intelligent Tutoring Systems Conference, 2002.

[13] Liscombe, J., Hirschberg, J., and Venditti, J., "Detecting Certianness in Spoken Tutorial Dialogues", in Proc. Interspeech, 2005.

[14] Forbes-Riley, K. and Litman, D., "Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus", in Proc. 6th SIGdial Workshop on Discourse and Dialogue, 2005.

[15] Witten, I. H. and Frank, E., "Data Mining:Practical Machine Learning Tools and Techniques with Java implementations", 1999.

[16] Tetreault, J. and Litman, J., "Using Reinforcement Learning to Build a Better Model Dialogue State", in Proc. EACL, 2006

[17] Grosz, B. and Sidner, C. L., "Attention, intentions, and the structure of discourse", Computational Linguistics, 12(3), 1986

[18] Forbes-Riley, K. and Litman, D., "Modelling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters", in Proc. HLT-NAACL, 2006.