Towards Emotion Prediction in Spoken Tutoring Dialogues

Diane Litman
Dept. of Computer Science
LRDC, Univ. of Pittsburgh
Pittsburgh PA, 15260, USA
litman@cs.pitt.edu

Kate Forbes
LRDC, Univ. of Pittsburgh
Pittsburgh PA, 15260, USA
forbesk@pitt.edu

Scott Silliman
LRDC, Univ. of Pittsburgh
Pittsburgh PA, 15260, USA
scotts@pitt.edu

Abstract

Human tutors detect and respond to student emotional states, but current machine tutors do not. Our preliminary machine learning experiments involving transcription, emotion annotation and automatic feature extraction from our human-human spoken tutoring corpus indicate that the spoken tutoring system we are developing can be enhanced to automatically predict and adapt to student emotional states.

1 Introduction

Connections between learning and emotion are well-documented (Coles, 1999), and studies have shown considerable benefits of spoken tutoring (Hausmann and Chi, 2002). Human tutors can respond to both the content of student speech and the manner with which it is spoken (e.g. 'confidently' or 'uncertainly'), but most intelligent tutoring dialogue systems are text-based and thus limited in their ability to recognize such learning states (Rose and Freedman, 2000; Rose and Aleven, 2002). Building spoken dialogue tutoring systems has great potential benefit, for speech is the most natural and easy to use form of natural language interaction, and it supplies a rich source of prosodic and acoustic information about the speaker's current mental state, which can be used to monitor the pedagogical effectiveness of student-computer interactions. The success of computer-based tutoring systems could increase if they predicted and adapted to student emotional states, e.g. reinforcing positive states, while rectifying negative states (Evens, 2002).

Although (Ang et al., 2002; Litman et al., 2001; Batliner et al., 2000) have hand-labeled naturally-occurring utterances in a variety of corpora for various emotions, then extracted acoustic, prosodic and lexical features and used machine-learning techniques to develop predictive models, little work to date has addressed emotion detection in computer-based educational settings. In this paper we describe preliminary annotation of positive, negative, and neutral emotions in a human-human tutoring corpus and discuss the results of pilot machine learning experiments whose goal is to develop computational models of specific emotional states (Section 3) for use in a spoken dialogue system (Section 2).

2 The ITSPOKE System and Corpus

We are developing a spoken dialogue system, called ITSPOKE (Intelligent Tutoring SPOKEn dialogue system), which uses as its "back-end" the *text-based* Why2-Atlas dialogue tutoring system (VanLehn et al., 2002). In Why2-Atlas, a student types an essay answering a qualitative physics problem and a computer tutor then engages him/her in dialogue to provide feedback, correct misconceptions, and elicit more complete explanations, after which the student revises his/her essay, thereby ending the tutoring or causing another round of tutoring/essay revision. To date we have interfaced the Sphinx2 speech recognizer with stochastic language models trained from example user utterances, and the Festival speech synthesizer for text-to-speech, to the Why2-Atlas back-end, and are adapting the knowledge sources needed by the spoken language components; e.g. we have developed a set of dialogue-dependent language models using 4551 student utterances from the Why2-Atlas 2002 human-computer typed corpus and will enhance them using student utterances from our human-human spoken corpus.

Our human-human spoken corpus contains spoken dialogues collected via a web interface supplemented with a high quality audio link, where a human tutor performs the same task as ITSPOKE and Why2-Atlas. Our subjects are U. Pittsburgh students who have taken no college level physics and are native speakers of (Amer.) English. Our experimental procedure, taking roughly 7 hours/student over 1-2 sessions, is as follows: students 1) take a pretest measuring their physics knowledge, 2) read a small document of background material, 3) use the web and voice interface to work through up to 10 training problems with the human tutor, and 4) take a post-test similar to the pretest. We have to date collected 63 dialogues (1290 minutes of speech from 4 females and 4 males) and transcribed 20 of them. A corpus example is shown in Figure 1, containing the problem, the student's essay, and an annotated excerpt from the subsequent dialogue.

PROBLEM: If a car is able to accelerate at 2 m/s2, what acceleration can it attain if it is towing another car of equal mass?

ORIGINAL ESSAY: If the car is towing another car of equal mass, the maximum acceleration would be the same because the car would be towed behind and the friction caused would only be by the front of the first car.

. . . *dialogue excerpt at 6.5 minutes into session* . . .

TUTOR: Now this law that force is equal to mass times acceleration, what's this law called? This is uh since this it is a very important basic uh fact uh it is it is a law of physics. Um you have you have read it in the background material. Can you recall it?

STUDENT: Um no it was one of Newton's laws but I don't remember which one. (laugh) *(EMOTION=NEGATIVE)*

TUTOR: Right, right, that is Newton's second law of motion.

STUDENT: Ok, because I remember one, two, and three, but I didn't know if there was a different name *(EMOTION=POSITIVE)*

TUTOR: Yeah that's right. You know Newton was a genius and uh he looked at a large number of experiments and experimental data that was available and from that he could come to this general law...

STUDENT: mm-hm *(EMOTION=NEUTRAL)*

Figure 1: Human-Human Spoken Corpus Example

3 Predicting Emotional Speech

For this pilot study, we annotated 14 transcribed dialogues from 7 students, 2 dialogues per student. First, turn boundaries were manually annotated (based on consensus labelings from two coders). Each turn was then manually annotated for *speaker affect* (by a single coder) using three general categorizations: *negative* (e.g.'uncertain', 'frustration'), *positive* (e.g. 'confident', 'certain'), or *neutral/indeterminate*, as shown in Figure 1. Table 1 shows the distribution of our labeled turns.

neutral	*positive*	*negative*	total
248	167	141	553

Table 1: Labeled Turn Counts: ITSPOKE Pilot Corpus

We next conducted experiments using the RIPPER (Cohen, 1996) rule induction machine learning program, which takes as input the classes to be learned (e.g. our emotion annotations), the names and possible values in a feature set (discussed below), and training examples, each specifying its class and feature values (e.g. the labeled student turns in our pilot corpus), then outputs a classification model for classifying future examples, expressed as an ordered set of *if-then* rules. RIPPER's "set-valued" features allow us to represent the speech recognizer's best hypothesis and/or the turn transcription as a set of words, and its rule output is an intuitive way to gain insight into our data.

For our first pilot machine learning experiment, our feature set consisted of SUBJECT ID and PROBLEM ID, both representing system state, TURN START-TIME (relative to start of dialogue) and TURN DURATION, both representing timing information, TEXT IN TURN (transcription), and NUMBER OF WORDS IN TURN. Figure 2 presents the ruleset that was learned for this classification task. For example, the first learned rule states that if the duration of the turn is greater than 0.65 seconds and the transcribed text of the turn contains the lexical item "I", then the turn is predicted to be labeled *EMOTION=NEGATIVE*. The estimated mean error and standard error of this ruleset is 33.03% +/- 2.45%, based on 25-fold cross-validation.

if (duration $\geq$ 0.65) $\wedge$ (text has "I") **then** *neg*
else if (duration $\geq$ 2.98) **then** *neg*
else if (duration $\geq$ 0.93) $\wedge$ (startTime $\geq$ 297.62) **then** *pos*
else if (text has "right") **then** *pos*
else *neutral*

Figure 2: All-Features Ruleset for Emotion Prediction

For comparison, our feature set in our second pilot machine learning experiment consisted of just TEXT IN TURN. The ruleset learned for this classification task contained 21 rules; Figure 3 presents an (ordered) excerpt[1]. Estimated mean error and standard error of this ruleset is 39.03% +/- 2.40%, based on 25-fold cross-validation.

if (text has "I") $\wedge$ (text has "don't") **then** *neg*
else if (text has "um") $\wedge$ (text has "<hn>") **then** *neg*
else if (text has "the") $\wedge$ (text has "<fs>") **then** *neg*
else if (text has "right") **then** *pos*
else if (text has "so") **then** *pos*
else if (text has "(laugh)") $\wedge$ (text has "that's") **then** *pos*
else *neutral*

Figure 3: Text-Feature Ruleset for Emotion Prediction

Although both these error rates are still fairly high, they are a significant improvement over a majority class

[1] <hn> = human noise (e.g. sighs and coughs), and <fs> = false start (e.g. "I th- think")

baseline that always predicts the majority class in our corpus (*neutral/indeterminate*) - which has an error rate of 55.69%. Moreover, many of the learned rules contain features that are intuitively associated with the predicted emotion; for example, disfluencies such as false starts are often associated with negative emotions such as 'uncertainty', as are lexical items such as "um" used in combination with human noises such as sighs.

4 Future Directions

Even using a small corpus classified by one coder and predicted using only a handful of features, our results suggest that there are indeed features that can automatically distinguish emotions in tutoring dialogues. We will next explore the utility of a wider variety of features representing many knowledge sources (including acoustic, prosodic, lexical, syntactic, semantic, discourse, and local and global contextual dialogue features), using ablation studies. We will perform our learning using and comparing large corpora of both human-human and human-computer data for training and testing, and will evaluate our results using a variety of metrics (e.g. recall, precision, and F-measure). We will also investigate a variety of emotion annotations with the goal of producing a reliable annotation scheme for the emotions associated with our tutoring domain. Previous studies have shown low inter-annotator reliability (around 70%, Kappa values around 0.47 (Narayanan, 2002)), which originates partly in vague descriptions of the emotions to be labeled.

Finally, we hope to use this work to demonstrate that enhancing a spoken dialogue tutoring system to automatically predict and then dynamically respond to student emotional states will measurably improve system performance. Our enhancements will be motivated by tutoring literature (Evens, 2002; Aist et al., 2002) that addresses how a tutor might make use of such information if it could be inferred, as well as by looking at how the human tutor actually responded to emotionally labeled turns. Our methodology will build on previous adaptive (non-tutoring) dialogue systems (see (Litman and Pan, 2002)); however, our system will predict and adapt to both problematic and positive dialogue situations in tutoring.

Acknowledgments

This research is supported by the NSF under Grant No. 9720359 to the Center for Interdisciplinary Research on Constructive Learning Environments (CIRCLE) at the University of Pittsburgh and Carnegie-Mellon University.

References

G. Aist, B. Kort, Rob R.lly, J. Mostow, and R. Picard. 2002. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proc. of ITS*.

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

A. Batliner, R. Huber, H. R. Niemann, E. Nöth, J. Spilker, and K. Fischer. 2000. The recognition of emotion. In *Proc. of the ISCA Workshop on Speech and Emotion*.

William Cohen. 1996. Learning trees and rules with set-valued features. In *Proc. of AAAI*.

G. Coles. 1999. Literacy, emotions, and the brain. Reading Online, March 1999.

M. Evens. 2002. New questions for Circsim-Tutor. Presentation at the 2002 Symposium on Natural Language Tutoring, University of Pittsburgh.

Robert Hausmann and Michelene Chi. 2002. Can a computer interface support self-explaining? *The International Journal of Cognitive Technology*, 7(1).

Diane J. Litman and Shimei Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12.

D. Litman, J. Hirschberg, and M. Swerts. 2001. Predicting user reactions to system error. In *Proc.of ACL*.

S. Narayanan. 2002. Towards modeling user behavior in human-machine interaction: Effect of errors and emotions. In *Proc. of ISLE*.

C. P. Rose and V. Aleven. 2002. Proc. of the ITS 2002 workshop on empirical methods for tutorial dialogue systems. Technical report, San Sebastian, Spain, June.

C. P. Rose and R. Freedman. 2000. Building dialogue systems for tutorial applications. Technical Report FS-00-01 (Working Notes of the Fall Symposium), AAAI.

K. VanLehn, P. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. of ITS*.